Building tools to enable an automatic analysis of Delphi processes' results in health settings

Sarah dos Santos Magalhães

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisors: Prof. Mónica Duarte de Oliveira
Prof. Ana Catarina Lopes Vieira Godinho de Mato

Examination Committee

Chairperson: Prof. João Miguel Raposo Sanches
Supervisor: Prof. Ana Catarina Lopes Vieira Godinho de Mato
Members of the Committee: Prof. Klára Dimitrovová

October 2019

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Preface

The work presented in this thesis was performed at the company Decision Eyes (Lisbon, Portugal), during the period June-October 2019, under the supervision of Eng. João Bana. The thesis was co-supervised at Instituto Superior Técnico by Prof. Mónica Oliveira and Prof. Ana Vieira.

Abstract

Context: Health services have become basic in everyday consumer life. For that reason, it is extremely important to include decision-makers' and stakeholders' perspectives in health decision. Participative processes as the Delphi technique, have been increasingly used to include their perspectives in these decisions. However, the analysis of these processes requires a lot of manual work from the Delphi user regarding the treatment of the participants' answers and Delphi outputs. Therefore, it is extremely important to develop automatic tools to digest information and present outputs generated by Delphi processes in health settings.

Objective: This thesis proposes the development of a novel and innovative Decision Support System (DSS) to treat Delphi participants' responses and to provide the main features that describe Delphi processes through statistical outputs and addressing the current challenges identified in the field. The DSS will incorporate the analysis of three types of Delphi processes commonly used in healthcare – Delphi for selection of indicators, Delphi for weighting judgments and Delphi for shaping value functions.

Methods: A review of the literature focusing on Delphi processes and the main techniques that have been used to perform a complete analysis of its features was performed. From the information gathered it was possible to understand that many researchers defend the use of the statistical measures to analyse a Delphi since they provide the best information about these processes. A framework of a DSS was developed following a design and then the implementation of the *DelphiAnalysis* DSS was performed using Microsoft Excel. The tool was tested with data from a real healthcare project and compared with the available results obtained in the same project. Additionally, a webpage guide was developed to help the Delphi users who want to use the tool, along with a questionnaire to collect the opinion of Delphi experts about the tool and the webpage guide.

Results: The results obtained when testing the tool and when comparing them with the published values proved the efficiency of the DSS as it can provide all the planned statistical measures accurately and without errors. Also, all the results are provided in table formats that grant user-friendly outcomes. Delphi experts provided positive feedback regarding the DSS and the webpage created and provide some suggestions to improve the *DelphiAnalysis* DSS in future work.

Conclusions: The implemented tool proved to be useful to Delphi analysts that work with the three specific types of Delphi implemented as it provides a complete analysis in a short time. In the future, more tests should be done using different data. Also, some improvements can be made to make the DSS faster, with fewer limitations and with a better graphical interface.

Keywords: Health decision; Delphi processes; Statistical outputs; Decision Support System; Microsoft Excel

Resumo

Contexto: Os serviços de saúde têm-se tornado uma necessidade na vida quotidiana do consumidor. Processos participativos como processos Delphi têm sido cada vez mais usados para incluir as perspetivas dos stakeholders e decision-makers em decisões de saúde. No entanto, a análise destes processos ainda requer muito trabalho manual por parte do utilizador tanto no tratamento das respostas dos participantes como nos resultados do processo. Desta forma, é de extrema importância o desenvolvimento de ferramentas que façam uma digestão automática de informação e apresentem resultados gerados por processos Delphi utilizados em saúde.

Objetivo: Esta tese propõe o desenvolvimento de um sistema de apoio à decisão (DSS) novo e inovador para o tratamento das respostas dos participantes e que descreva as principais características de um processo Delphi através de outputs estatísticos e que vá de encontro com os desafios identificados na área. O DSS irá incorporar a análise de três tipos de Delphi amplamente utilizados na área da saúde: Delphi para seleção de indicadores, Delphi de pesos e Delphi de funções de valor.

Métodos: Foi feita uma revisão da literatura focada em processos Delphi e nas técnicas utilizadas para a realização de uma análise completa dos mesmos. A partir da informação recolhida concluiu-se que muitos investigadores defendem a utilização das medidas estatísticas mais comuns uma vez que são as que fornecem mais informação útil. Foi estruturado um modelo de desenho para o DSS a partir de um design existente, procedendo-se depois à sua implementação no Microsoft Excel. A ferramenta foi testada usando dados de um projeto real de saúde e os valores obtidos foram comparados com os valores publicados do projeto. Adicionalmente, uma página web foi desenvolvida para guiar os analistas de processos Delphi no uso da ferramenta. Um questionário foi também elaborado para obter opiniões de especialistas em Delphi relativamente à ferramenta e à página web desenvolvida.

Resultados: Ao comparar-se os resultados obtidos na ferramenta com os valores do projeto, comprovou-se a eficiência do DSS visto ser capaz de fornecer todas as medidas estatísticas de uma forma eficiente e sem erros. Além disso, os resultados são apresentados em formato de tabela o que garante resultados intuitivos. Os especialistas em processos Delphi deram feedback positivo tanto relativamente à ferramenta como à página web e ainda forneceram algumas sugestões do que pode ser melhorado em trabalhos futuros relativamente ao DSS *DelphiAnalysis*.

Conclusões: A ferramenta demonstrou ser útil para analisadores de processos Delphi que trabalham com os três tipos de Delphi implementados uma vez que oferece resultados úteis num curto período de tempo. Futuramente, mais testes devem ser realizados utilizando diferentes dados e melhorias podem ser desenvolvidas de modo a tornar a ferramenta mais rápida, com menos limitações e com uma melhor interface gráfica.

Acknowledgments

First, I would like to share with you how happy, proud and grateful I am for getting here. I will be eternally grateful for all my experiences, all the people I have met, and all the years spent in IST.

A special thanks to the professors Mónica Oliveira and Ana Vieira for their motivational support, understanding and knowledge sharing that made this work rewarding!

I truly acknowledge to João Bana e Costa, Teresa Rodrigues, Fábio Martins, and Mónica Nóbrega from Decision Eyes for providing all the help I ask them and certainly, for making my days happier during this process.

To those with whom I shared the best and worst moments during the 5 years of faculty and especially those who have been always by my side during the last period: Carolina Barata, Katrin Munzenrieder, Luís Gonçalves, Liliana Lameiras, María Miño, and Sofia Fonseca.

To my father and mother for making all this possible, for the greatest effort they have made but, distinctively, for their love and support. To my brothers, who are always there, no matter what. A special acknowledge to my uncles, Teresa and Orlando, and to my cousin Pedro, for being my second family and to my grandparents who always want to help me.

To the ones who still live in me: Dani, who taught me what it was to have a brother, who taught me how to spread love and smiles, and my grandmother and aunt Florinda, who taught me how to be good. I wish I could still have you with me.

Table of Contents

ABSTRACT ..... **IV**

RESUMO ..... **V**

ACKNOWLEDGMENTS ..... **VI**

LIST OF FIGURES ..... **X**

LIST OF TABLES ..... **XIV**

LIST OF ACRONYMS ..... **XV**

1. INTRODUCTION **14**

1.1. THESIS OBJECTIVE 15
1.2. THESIS STRUCTURE 15

2. CONTEXT **16**

2.1. DECISION-SUPPORT TOOLS IN HEALTHCARE 16
2.2. SCOPE OF THE DELPHI METHOD 17
2.3. USING THE DELPHI TECHNIQUE IN HEALTHCARE 17

3. LITERATURE REVIEW **20**

3.1. DELPHI TECHNIQUE 20
3.1.1. Types of Delphi 20
3.1.2. Characteristics of the Delphi technique 22
3.1.3. Strengths and Limitations of Delphi processes 23
3.1.4. How to perform a Delphi 24
3.1.5. Rigour of Delphi processes 26
3.2. DELPHI DESIGN 27

4. *DELPHIANALYSIS* DSS **30**

4.1. OVERVIEW OF THE STEPS TO TAKE BEFORE THE DEVELOPMENT OF THE DSS 30
4.2. IDENTIFICATION OF THE TYPES OF ANALYSIS AND DATA 30
4.3. METHODS USED IN LITERATURE 31
4.3.1. Types of scale and measurement scales 33
4.3.2. Parametric vs Non-parametric techniques 34
4.4. TYPES OF DELPHI TO BE ANALYSED WITHIN *DELPHIANALYSIS* 42
4.4.1. EURO-HEALTHY Project 44
4.4.2. Differences between one panel and multiple panels 46
4.5. SELECTION OF THE BEST STATISTICAL MEASURES TO IMPLEMENT 47
4.5.1. How to implement the selected statistical measures 54
4.6. DESIGN: ADAPTED FRAMEWORK OF THE DSS 54
4.6.1. Design and Architecture of a DSS 55

1. Outline the Decision Problem ... 57

a) Problem importance ... 57

4. Identify Design Context ... 59

5. Measure the outputs ... 59

4.7. IMPLEMENTATION OF THE *DELPHIANALYSIS* DSS ... 60

4.7.1. Global organization of the DSS ... 60

4.7.2. Input Data ... 61

4.7.3. Conversion of the Input Data ... 62

4.7.4. Statistical measures performed by the DelphiAnalysis DSS ... 63

4.7.5. Summary Page ... 69

4.7.6. MANOVA ... 70

4.7.7. "EXTRA" Tab ... 71

4.8. WEBPAGE ... 71

4.9. QUESTIONNAIRE TO DELPHI EXPERTS ... 72

5. RESULTS ... 74

5.1. RESULTS OBTAINED FOR DELPHI FOR SELECTION OF INDICATORS ... 74

5.1.2. Additional results for each round ... 76

5.1.3. Results of the "between rounds" section ... 76

5.1.4. "Summary page" results ... 77

5.1.5. "MANOVA" results ... 78

5.2. RESULTS OBTAINED FOR THE WEIGHTING JUDGMENT DELPHI ... 78

5.2.1. Frequency and percentage of votes on each category ... 79

5.2.2. Additional Results for each round ... 79

The Kendall's Coefficient of Concordance and respective p-value, the Fleiss' Kappa and the Spearman Rank Correlation Coefficient were calculated as well (Figure 36). ... 79

5.2.3. Results of the "Between rounds" section ... 80

5.2.4. "Summary Page" Results ... 81

5.2.5. "MANOVA" Results ... 82

5.3. RESULTS OBTAINED FOR THE VALUE FUNCTION DELPHI ... 82

5.3.1. First round Results ... 82

5.3.2. "Summary Page" Results ... 84

5.4. RESULTS OBTAINED IN THE QUESTIONNAIRE ... 85

6. DISCUSSION ... 88

6.1. ADVANTAGES OF THE DELPHIANALYSIS TOOL ... 88

6.2. DISADVANTAGES OF THE *DELPHIANALYSIS* TOOL ... 89

7. CONCLUSION ... 90

7.1. FUTURE WORK ... 90

7.2. FINAL REMARKS ... 91

REFERENCES **92**

APPENDIX A **98**

APPENDIX B **106**

APPENDIX C **108**

List of Figures

Figure 1 *- Delphi technique flow chart [40].* _______ *26*

Figure 2 - *Block diagram representing the steps made until the implementation of the tool.* _______ *30*

Figure 3 - *Flow diagram (based on [58]).* _______ *32*

Figure 4 - *Types of data and respective scales of measurement (based on [49] and [50]).* _______ *33*

Figure 5 - *Types of value function curves.* _______ *44*

Figure 6 *- Panel's names and number of members of the Web-Delphi weighting.* _______ *45*

Figure 7 - *Features to analyse according to the objectives of the Delphi process at the process level and final results (Delphi for selection of indicators using a Likert-scale).* _______ *49*

Figure 8 - *Scheme with the non-parametric techniques and correspondent feature that measures (Delphi for selection of indicators using a Likert-scale).* _______ *50*

Figure 9 - *Final techniques chosen to be implemented in the DSS (Delphi for selection of indicators using a Likert-scale).* _______ *51*

Figure 10 - *Objectives, features to analyse and final tools chosen to be implemented in the DSS (Delphi for qualitative weighting judgments using MACBETH).* _______ *52*

Figure 11 - *Objectives, features to analyse and final tools chosen to be implemented in the DSS (Delphi for shaping the value function using MACBETH).* _______ *53*

Figure 12 - *IQR (red zone).* _______ *54*

Figure 13 - *Functional components of a DSS (based on [91]).* _______ *55*

Figure 14 *- Workflow for DSS design (based on [48]).* _______ *56*

Figure 15 - *Framework design of a DSS (based on [57]).* _______ *56*

Figure 16 - *Descriptive Table (Delphi for selection of indicators). The right side of the table is where the user can introduce the categories used in the Delphi process. The categories shown in the left side of the table are the ones recognized by the DSS.* _______ *62*

Figure 17 - *Scheme of the relationship between the votes of the panelists and the shape of the respective value functions.* _______ *68*

Figure 18 – *Part of the table with the percentages of votes on each value function shape for each indicator.* _______ *68*

Figure 19 - *Part of the "Summary page" of the selection of indicators Delphi (about the "Agree" category).* _______ *70*

Figure 20 - *How to apply the Wilks Lambda test and its correspondent p-value. The range needed in the formula is selected in the figure.* _______ *71*

Figure 21 - *DelphiAnalysis DSS webpage guide (https://healthdelphianalysis.wixsite.com/delphianalysis).* _______ *71*

Figure 22 - *Table with the resultant percentages of the first 5 indicators, in round 1, using the data from the EURO-HEALTHY project.* _______ *75*

Figure 23 - *Table with the percentages of the first 5 indicators, in round 1, obtained in the EURO-HEALTHY project.* _______ *75*

Figure 24 - *Resultant mode, median and IQR for the first 5 indicators (Round 1) obtained in the tool.* *75*

Figure 25 - *Resultant mode, median and IQR for the first 5 indicators (Round 1) obtained in the EURO-HEALTHY project.* 75

Figure 26 - *Resultant W, K and r for the Round 1 using the data from the EURO-HEALTHY Project.* 76

Figure 27 - *How much participants changed their opinion between rounds 1 and 2 (first 5 indicators). The indicators that show the best stability (less changes of opinion among panelists) are highlighted with yellow.* 76

Figure 28 - *Resultant Spearman Rank Correlation Coefficient for the first 5 indicators between rounds 1-2.* 76

Figure 29 - *Part of the table that shows how many times each participant changed his opinion and table with the three that changed it most times (highlighted with purple on the left side Table).* 77

Figure 30 - *How much the votes of each participant changed between rounds 1 and 2 (for the first 5 indicators). Since it is a 5-level scale, jumps of two categories is considered drastic (highlighted with green).* 77

Figure 31 - *Part of the results present on the "summary page" between rounds 1 and 2.* 77

Figure 32 - *Results about how inter-reliability and stability evolved during the process for each indicator.* 78

Figure 33 - *Results for the first round MANOVA: Wilk's Lambda and respective p-value obtained for the responses given in the EURO-HEALTHY project, according to the tool.* 78

Figure 34 - *Results of the frequency and percentages of votes provided by the panelists in round 1 (first 5 indicators).* 79

Figure 35 - *Mode, Median and IQR (Round 1, first 5 indicators).* 79

Figure 36 - *Results of the analysis of the level of agreement and the overall stability for the first round of this Delphi process.* 80

Figure 37 - *Frequency and percentage of how many participants changed their vote between rounds 1 and 2 (first 5 indicators).* 80

Figure 38 - *Spearman's Correlation Coefficient between round 1 and round 2, for the first 5 indicators.* 80

Figure 39 - *Table that shows how many times each participant changed his vote and table with the 3 panelists that changed opinion most times.* 81

Figure 40 - *How much the percentages of the two higher categories changed between rounds 1 and 2.* 81

Figure 41 - *W and K obtained in rounds 1 and 2 concluding about how it changed.* 81

Figure 42 - *Analysis provided by the DSS to conclude about inter-reliability and stability through the Delphi process.* 81

Figure 43 - *Wilk's Lambda and correspondent p-value obtained for the evaluation of MANOVA between experts and stakeholders.* 82

Figure 44 - *"Gaps" row filled with the jumps of each range performance defined in the EURO-HEALTHY project. The indicators where replaced by their description.* 82

Figure 45 - *Frequency of votes on each category for the three gaps of the first two indicators (Round 1).* 82

***Figure 46 -** Percentage of votes on each category for the three gaps of the first two indicators (Round 1) and the percentage of some useful combinations of votes. 83*

***Figure 47 -** Mode and IQR of each of the three gaps of the first two indicators (Round 1). 83*

***Figure 48 -** Table with the value function shapes obtained from the conversion of the votes of the first 10 panelists (Round 1, first 5 indicators). 83*

***Figure 49 -** Table with the value function shapes obtained from the conversion of the votes of the first 10 panelists in the EURO-HEALTHY project (Round 1, first 5 indicators). 83*

***Figure 50 -** Percentages of each shape of the value function for the first 5 indicators (Round 1). 84*

***Figure 51 -** Mode and bi-mode, for the first 5 indicators and which curve achieved more than 50% of the votes, if existent (Round 1). The “X” represents inexistence. 84*

***Figure 52 -** How much decreased or increased the percentage of votes on linear and concave shapes, between round 1 and 2, for the first 5 indicators. 85*

***Figure 53 -** How much the percentage of participants who changed their votes varied throughout the Delphi process, concluding about the overall inter-reliability. 85*

***Figure 54 -** Part of the first input table for the selection of indicators Delphi with a random example. 98*

***Figure 55 -** Percentages of votes for the first 6 indicators obtained in round 2. The indicators not filled were the ones that obtained majority and were discarded for the next round evaluation. 99*

***Figure 56 -** Mode, Median and IQR for the first 6 indicators obtained in round 2. 99*

***Figure 57 -** Kendall’s Correlation Coefficient obtained in round 2. 99*

***Figure 58 -** Fleiss’ Kappa obtained in round 2. 100*

***Figure 59 -** Percentages of votes for the first 6 indicators obtained in round 3. The indicators not filled were the ones that obtained majority and were discarded for the next round evaluation. 100*

***Figure 60 -** Kendall’s Correlation Coefficient obtained in round 3. 100*

***Figure 61 -** Fleiss’ Kappa obtained in round 3. 100*

***Figure 62 -** Frequency and percentages of how participants changed their opinions between rounds 2 and 3 (for the first 6 indicators). 100*

***Figure 63 -** Spearman’s Correlation Coefficient obtained between round 2 and 3 for the first 6 indicators. 100*

***Figure 64 -** Table with the three participants that changed their opinion most times between rounds 2 and 3. For example, the “7” means “Respondent 7”. 100*

***Figure 65 -** Percentages of votes for the first 5 indicators obtained in round 2. 101*

***Figure 66 -** Kendall’s Correlation Coefficient obtained in round 2. 101*

***Figure 67 -** Fleiss’ Kappa obtained in round 2. 101*

***Figure 68 -** Percentages of votes for the first 5 indicators obtained in round 3. 102*

***Figure 69 -** Kendall’s Correlation Coefficient obtained in round 3. 102*

***Figure 70 -** Fleiss’ Kappa obtained in round 3. 102*

***Figure 71 -** Percentage of how many people changed their votes between rounds 2 and 3. 102*

***Figure 72 -** Spearman’s Correlation Coefficient obtained between rounds 2 and 3. 102*

***Figure 73 -** Table with the three participants that changed their opinion most times. 102*

Figure 74 - *Results obtained for the second round for the first 4 indicators: percentages of each type of shape, mode and bi-mode and curves that obtained more than 50% of the votes.* _______________ *103*

Figure 75 - *Results for the first 4 indicators obtained in round 3: percentages of each type of shape, mode and bi-mode and curves that obtained more than 50% of the votes.* _____________________ *103*

Figure 76 - *Frequency and percentage of how many participants changed their votes between rounds 2 and 3, for the first 5 indicators.* ____ *105*

Figure 77 - *The three participants that changed their opinion the most between rounds 2 and 3.* __ *105*

Figure 78 - *Where to download the three Excel files of the tool.* ______________________________ *106*

Figure 79 - *First part of the "How to use the TOOL" section.* _________________________________ *107*

Figure 80 - *"MANOVA" part in the "how to use" section.* ____________________________________ *107*

List of Tables

Table 1 - *Modifications of Delphi: definition and main advantages and disadvantages [18].* 22

Table 2 - *Types of Delphi designs and its principal focus, aim and panelists (combination of information from [15] and [55]).* 31

Table 3 - *Summary of the characteristics applicable to each measurement scale (based on [48]).* 34

Table 4 - *Statistical measures used to test the validity of Delphi surveys and types of data that is analysed by them.* 35

Table 5 - *Type of information that each procedure uses and give and if they are relative to the iteration process or the final results of a Delphi study.* 36

Table 6 - *Coefficient of variation and consensus (from [62]).* 38

Table 7 - *How to apply the selected methods in the tool.* 54

Table 8 - *Number of votes on each category of the Likert scale for each statement provided.* 85

Table 9 - *Number of "yes" or "no" to the question "Would you use the DelphiAnalysis DSS to help you?".* 86

List of Acronyms

ANOVA	Analysis of Variance
APMO	Average Percent of Majority Opinions
CV	Coefficient of Variance
DM	Decision Makers
DSS	Decision Support System
EU	Europe
HTA	Health Technology Assessment
ICC	Intraclass Correlation Coefficient
ID	Identification
IQR	Interquartile Range
MACBETH	Measuring Attractiveness by a Categorical Based Evaluation Technology
MANOVA	Multivariate Analysis of Variance
MCDA	Multicriteria Decision Analysis
PHI	Population Health Index
SD	Standard Deviation
TCC	Tetrachronic Correlation Coefficient

1. Introduction

Healthcare is a complex, multidimensional field that requires research to provide services of quality to the target population [1]. Higher life expectancy and higher patient expectations came with the technological development and medical research improvements, increasing the demand for health services. Therefore, resources need to be allocated efficiently and fairly to attend all the necessities of the society [2]. However, the allocation of resources that gives legitimacy to decisions is not easy due to the complexity of decisions related with healthcare, taking into account the variety of aspects that must be taken into consideration. For example, the involvement of stakeholders, the fact that evaluators in health settings are the decision-makers (DM) and their individual opinions that can bring conflicts [2]. Making decisions based on unstructured processes can lead to a lack of predictability and to inconsistencies that in turn can compromise the credibility of the evaluation [3]. To improve health quality it is necessary to identify the problem and make changes to improve health services, implementing qualitative and quantitative methods as structured approaches involving multiple criteria and improving the quality of decisions [4]. According to Belton and Stewart [5], Multicriteria Decision Analysis (MCDA) is "an umbrella term to describe a collection of formal approaches, which seek to take explicit account of multiple criteria in helping individuals or groups explore decisions that matter" and is intended to help DM to reach a decision based on the most appropriate evidence [3]. Regarding the social component of developing MCDA methods to improve health services and that consider stakeholders' and DM' opinions, interviews, observation and analysis of documents stand out as collecting data techniques [4]. However, these methods carry problems regarding the pressure and fear felt by the participants [6]. To solve these inconsistencies, alternative techniques, as the consensus ones, are required. The most common consensus procedures are the nominal group technique, consensus conference and Delphi processes [7].

The Delphi technique is a communication method that allows participants to express their opinions, anonymously [8]. The method is based on series of surveys to be filled by the participants during the rounds of the process. Feedback is provided to the panelists between rounds so they can adapt their opinion in the next survey. Therefore, this method considers perspectives provided by experts dealing with anonymity, controlled feedback, iteration and statistical group response, as explained in Chapter 3 [8], [9]. These processes have stood out in the healthcare for their ability to avoid conflicts, pressure and bias of the stakeholders and DM [10]. More recently, online platforms have emerged to help with the implementation of Delphi processes. However, these platforms only facilitate the application of the process, but they do not analyse the results. Therefore, manual work from the researcher is still needed, respecting the treatment of participants' answers and the presentation of outcomes, taking valuable extra time.

In this thesis, attention is given to Delphi processes used in healthcare as Delphi processes for selection of indicators and Delphi processes used in multicriteria as Delphi processes for weighting judgments and Delphi processes for shaping value functions. The goal is to offer statistical outputs to analyse Delphi processes' results by developing a prototype tool to help decision analysts and health DM to access Delphi results and make analyses in a more expedite way.

1.1. Thesis Objective

This dissertation was developed under the Master of Biomedical Engineering and the EIT Health MSc Technological Innovation in Health involving the collaboration of the Decision Eyes.

There are much literature about Delphi processes, its relationship with healthcare, which measures have been applied to analyse them and the existence of web platforms that allow performing a Delphi survey easily. However, analyses are still mostly manually done by decision analysts, taking some extra time. The aim of this thesis is to develop a novel Decision Support System (DSS) to help digesting the information given by the participants and to present outcomes accomplished by Delphi processes in health. Tools should assist performing a complete analysis of the Delphi responses, providing information about the principal features that describe the method, through statistical outputs. A webpage to guide the user in how to use the DSS and a questionnaire to validate the guide and DSS will be developed to complement the work.

A literature review on Delphi processes will be the starting point to understand which features and techniques should be analysed by the tool. A framework for DSS design will be used to construct a new DSS with all the requirements needed. The final aim of the thesis is the implementation of a novel and innovative decision tool, to be an example of what can be done in the future to improve the analysis of other types of Delphi that can be performed in several health contexts. The implementation will be further performed in Microsoft Excel, with mathematical programming code being developed and some already existing functions being used. A guide to help the users of the DSS was also prepared within a webpage format. Then, the DSS will be tested using data from a real healthcare project and Delphi experts will be questioned regarding the usefulness of the DSS and guide.

1.2. Thesis Structure

In Chapter 2, the context is provided. This section works as a starting point for the development of a DSS to enable Delphi analyses in MCDA and health contexts.

In Chapter 3, a literature review is presented. All the information about Delphi processes and how these processes should be analysed is given. Works performed in the past, their advantages and disadvantages and the types of analysis that can be performed will be studied. The proposed methodology adopted in this thesis is presented in Chapter 4. The explanation of the novel DSS in terms of its design, architecture and the implementation is provided, as well as the webpage guide and questionnaire to obtain experts' feedback are described.

In Chapter 5 the results from applying the tool and their comparison with the ones obtained in the real project are provided along with a discussion, present in Chapter 6, about the results, advantages and limitation of the DSS.

Chapter 7 provides final remarks about the work developed in this thesis and reflects upon future work and methodology improvements.

2. Context

In this chapter key concepts about consensus methods and its applications on health settings are introduced. Specifically, a brief description of participatory tools in healthcare starts the Chapter. Also, the Delphi technique and its relationship in healthcare, as a methodology to support decision-making are introduced.

2.1. Decision-support tools in healthcare

Healthcare is a multidimensional and complex field of study. According to the Institute of Medicine, healthcare quality is "the degree to which health services for individuals and populations increase the likelihood of desired health outcomes and care consistent with current professional knowledge" [1]. Therefore, health quality can only be assured if good qualitative and quantitative measures existed since they are required to assess good information [4]. Qualitative methods involve the collection, analysis, and organization of important data and include a variety of methods such as interviews, analysis of documents and careful observations. Interviews, for example, are one of the most used in healthcare concerns [4], [11]. Individual face-to-face interviews are useful to flexible topics with open-ended questions to explore attitudes or experiences and to obtain details about a specific issue or experience [4]. Focus group interviews are use the interaction between research participants to generate important data [6]. On the other hand, quantitative methods (e.g. meta-analysis) and other techniques (e.g. consensus methods) are used to solve inconsistencies in the results of published studies [6]. Consensus methods are a way to deal with conflicting scientific evidence which aims to overcome disadvantages in decision-making of groups or committees. These methods attempt to assess the extent of agreement about a subject, solving disagreements at the same time [7]. The most known consensus methods are the nominal group technique, the consensus conference and, the Delphi technique [7].

The nominal group technique uses a highly structured group to gather relevant information about an issue. It consists of two rounds, where panelists rate, discuss and, re-rate the important items about a question posed to the group. This process is repeated to achieve a higher level of consensus [8]. In the context of healthcare, the method has been used mostly to evaluate clinical interventions and for the identification of measures for clinical trials [7]. This method encourages the contribution of everyone providing equal participation among the group and has been seen as a useful way for idea generation in assembly discussions [8].

Consensus conferences is another method to achieve agreement on a matter of concern. A sample of individuals is invited to a conference where the importance of the subject is debated [8]. The group presents the pros and cons of the issue and, in the final step, delegates can vote to show their opinions on the topics. This method can be expensive, the group can dictate the direction of the discussion and it is a face-to-face discussion, which can lead to discomfort to the group members to present their own decisions [8].

The Delphi method provides an opportunity for experts to communicate their opinions about an issue, anonymously, avoiding domination of the consensus process by one expert, which sets it apart from the other methods [1]. This approach allows the panelists to change their opinions according to the information provided in the feedback about the responses of the rest of the panel [1]. Another advantage of the Delphi technique is that a large number of individuals across different locations and areas of expertise can be part of the method, eradicating geographical barriers and saving time, money and inconvenience [1], [12]. Delphi studies have then, the potential to provide valuable information in the health field and facilitate wider group participation [1], [12]. From this perspective, it is clear that the Delphi technique is a powerful and promising research tool to apply in many areas of healthcare, as explained in the following section.

2.2. Scope of the Delphi method

The Delphi method is a data collection method used to identify research first concerns in different fields [8]. Mathematical skills are not required for design, implementation or analysis of what makes the method interesting and useful in many areas. For example, the Delphi technique can be used in varied ways in the social sciences, as a committee evaluation or a decision-making tool or for forecasting [13]. Consequently, the potential and dynamics of the method are being recognized in a diversity of study fields, including economic and financial settings, civic planning and healthcare [8]. The Delphi technique and its characteristics, advantages and disadvantages are discussed in detail in Chapter 3.

2.3. Using the Delphi technique in healthcare

As the base of health is to be improved, the Delphi technique seems to be a useful too as it uses a group of experts that provide opinions and judgments allowing to guide to best practice [14].

Many clinical issues do not yield to stepwise quantitative data analysis. Instead, professionals use their experience to assist decision making in practical choices [15]. The decision-making process is the basis of every clinical practice since every action must be evaluated. However, decision-making is a complex task, particularly, medical decision-making, as various factors that influence decisions and patients must be part of the process, involving their preferences and values, what may lead to disagreements about the best course of action [16].

The Delphi technique has been used in many different fields, being health research one of the most common [8]. Anthony R. Romano (2010) argued that the method is suitable for questions that cannot be dealt analytically or when the face-to-face meetings are expensive. Additionally, he defended that the method is convenient when the only information existent is the judgments of experts who are geographically distant from each other [17]. Different characteristics are required, and different periods of time are available for each specific application of the method, so it was necessary to make some adaptations to improve the performance of the method [8]. Taking into account that there are no universal guidelines or rules on the use of the method, it suffered a lot of modifications and changes over the years, what reflects its wide flexibility and led to the appearance of many types of Delphi techniques [8], [18]. In this light, it is easy to understand that the policy Delphi, whose main goal is to generate a wide range of opinions about a certain topic, is extensively used in health questions to aggregate all the important judgements about a subject [15]. For example, the policy Delphi was applied to achieve national level policy making on child health indicators in Hong Kong. Moreover, the classical method has been used to achieve consensus in health issues and to discover factors influencing dental decision-making [15].

Other Delphi applications have included forecasting disease patterns and health funding requirements [8]. As a forecasting tool, the Delphi method has been used to predict developments of many healthcare areas, including child and maternal health. In 1971, this method was used to predict how improvements in nutrition, family income, and prenatal care would impact on birth weight and subsequent intellectual development [8].

Thus, it is possible to conclude that the Delphi technique is commonly used to identify issues and their solutions, selecting topics, planning, forecasting and defining research questions, what remarks its flexibility to address different fields in healthcare and its versatility as it can be implemented in various points of the research [15]. However, despite being a widely used method, it still presents major challenges that need to be overcome in order to improve healthcare and the approaches where Delphi processes can be useful.

3. Literature Review

The main objective of this Chapter is to provide a revision on what a Delphi process is, which are the existent modifications of the method, which are its main characteristics, how to apply the methods and its advantages and disadvantages. The main types of data and scales used in statistics is also described in order to understand further analysis performed in the thesis. The Chapter ends with the description of the main steps to construct a Decision Support System.

3.1. Delphi technique

As briefly explained before, the Delphi technique is a structured communication method whose first goal was to obtain consensus about an important topic. Based on the ideas that individual statistical predictions are stronger than unstructured [8] and that group opinion is more trustworthy than an individual belief, the method relies on a panel of experts that are questioned about their opinion, anonymously, through series of surveys with controlled opinion feedback [8], [9]. A multi-staged survey is presented to the participants in order to collect their opinions about an important topic. Each round is followed by feedback reporting the opinions of all the participants and it can be accompanied with qualitative notes or quantitative statistical measures that will help the participants to adapt their opinion regarding the knowledge shared by all the panelists [8].

According to Hasson [8], Keeney and McKenna, Lynn *et. al* (1998) defined the Delphi technique as an iterative process designed to combine expert opinion into group consensus. However, with the evolution of the technique, it is no longer seen as a method for solely reaching consensus but rather for collecting different opinions and points of view of people with high knowledge on a particular subject [10]. Monica R. Geist (2010) [10] defended that one of the goals of the method was to avoid negative face-to-face interactions between groups, stimulating the sharing of individual beliefs without fear or pressure [19], and overcoming geographical barriers that may exist [10].

Nowadays, there are a lot of modifications and variations of the Delphi method that have been emerged to facilitate specific issues, such as the ‘modified Delphi’, the ‘policy Delphi’, the ‘real-time Delphi’ [20].

3.1.1. Types of Delphi

Classical Delphi

The classical Delphi consists of a systematic technique that uses several survey rounds to obtain the expert panel opinions about a certain topic or issue. It is a paper-and-pencil version and is traditionally sent by post [20]. The first round is defined as being qualitative once it is an open-ended questionnaire that allows the experts to report meaningful statements [8]. The researcher has the responsibility of analyzing and condense the panel responses to provide feedback and create a new questionnaire for the next round [14]. Traditionally, the number of rounds made is the needed to achieve consensus [8].

Modified Delphi

The modified Delphi is one of the most common variations of the technique. The key difference between this type of Delphi and the original relies on the first round, that is no more an open-ended questionnaire obtained from the experts' opinion but it is formed from summarized reviews of the literature and interviews with experts [21]. Although the procedure and the purpose of the method are the same as those of the original technique, it brings some big advantages as a solid background in previous researches, decrease of personal bias and it allows the existence of quantitative data [22]. Some modified Delphi processes replace the first round by one-to-one interviews or even group conferences, which can be seen as an advantage too, since this approach orientates panelists and ensures that everyone starts from a common base [23], [24]. S.Keeney *et al.* [23] stated that McKenna (1994) argued that using face-to-face interviews in the first round increases the return rates of postal questionnaires.

Policy Delphi

The main interest of the policy Delphi technique is not getting consensus but generate a wide range of different opinions on a certain topic, using a panel of experts [25]. The plan is to generate divergent opinions, alternatives and ideas resorting to debate, which allows addressing different health issues [20], [26]. This Delphi variation permits the identification of agreement or disagreement points, qualitative and quantitative data and future issues, always based on evidence [20]. Despite all the advantages, there are some weaknesses regarding the vast diversity of views, the lack of concern about disagreement and the inability to provide an evaluation in depth on the solutions [25].

Real-Time Delphi

As new internet technologies emerged, a new approach of the Delphi technique was developed, which switched the classic paper questionnaires by a web-based survey [27]. Real-Time Delphi is a "round-less" method seeing that typical rounds don't exist; the web page automatically updates giving direct feedback when a respondent is assessing [10]. Each member of the panel can change his answers as many times as he wants until the end of a pre-defined amount of time, shortening the time frame required to perform the same study using the classical Delphi method [27],[28]. The major strengths of the method are its efficiency and some features such as the ease of use and the lofty response rates. However, a big disadvantage is the fact that it is not possible to track the progress of the riposte [28], [29].

Other types of Delphi techniques

As mentioned previously, many different designs of the Delphi process appeared to address special situations with specific characteristics, aims, advantages and disadvantages [20]. Despite the most common designs described above, others should be present as they fit to perform important analysis in the real world. A summarized table with more modifications that can be used in some investigations is shown below.

Table 1 - *Modifications of Delphi: definition and main advantages and disadvantages [18].*

Delphi type	Definition/Aim	Advantages	Disadvantages
e-Delphi	Internet-based platform designed to facilitate communication between the researcher and the panelists; it works to establish consensus [30].	- Suitable for organizing the gathered data; - Time and cost savings; - Convenience for the administrator and participants [30].	- Internet-use difficulties and challenges; - Place the entire data into the computer and control its accuracy; -Internet access remains expensive [30].
Technological	Use of technology to calculate statistical measures to provide instant feedback and record the responses provided through hand-held keypads [8];	- Time-saving; - Good to predict future events; - More quantitative Delphi approach [31].	- Impossible to track the evolution of responses; - More difficult to ask and explore open-ended questions [31].
Online	Questionnaires are answered online [8]; offer promise for future research, model building or theory validation [32]; the aim is to maximize the range of expert opinions and not consensus [33].	- Easy to use: chat room or forums are adequate [20]; - Explores the barrier factors to the adoption of mobile data services [33].	- Experts can choose to adopt more consensual answers and to express fewer clear opinions; - does not use the standard statistical tests [33].
Argument	The aim is to develop relevant arguments and reveal the reasons for the different opinions [20] and critiques to the other arguments until a consensus is achieved [34].	- Carry out factual important judgments [8]; - A wide range of opinions; - Can be collected at any time; - Allows any contributor to add new arguments [34];	- Panelist must be able to approach the topic from different perspectives, which is difficult; - Traditional argument aggregation is difficult [34].
Disaggregative policy	The objective is to build scenarios about the future [8]; the aim of consensus is not adopted, responses are grouped to several clusters [35].	- Experts are asked about their future, preferences and probabilities [20];	- Can't be taken as a granted as scenarios can be created according to subjective views instead of real quality material [35].

3.1.2. Characteristics of the Delphi technique

Identically to any other method, the Delphi technique has some problems that must be overcome. Four main features have been implemented to help the method to triumph over these weaknesses: anonymity, iteration, controlled feedback and statistical group response [10]. These characteristics demark the Delphi method relatively to classic techniques such as face-to-face meetings and/or interviews as they provide less pressure once responses are anonymous and the possibility of rethinking their answers based on what the other members said [36].

Anonymity

Questionnaires are used to address the different panel opinions without matching them with their identification, which allows anonymity [26], [36]. Therefore, this characteristic grants the respondents not to feel physiologically pressured by other participants and admits that any given response has equal weightiness for the closing analysis [8], [10]. However, complete anonymity cannot be guaranteed for two main reasons: first, the researcher always knows the members of the panel and their answers and second, the panel members may know each other and share opinions between them. These two reasons lead to the term "quasi-anonymity" to better describe this important feature of the Delphi process [8].

Iteration

Iteration is given through successive rounds allowing participants to adjust their judgments in consecutive rounds [8]. On the first round, panelists are presented with the subject of the study and they need to generate their statements about what they think about it [10]. Afterward, the researcher summarizes and organizes the responses and give them back allowing the members to modify their responses on the second round [10]. The process is repeated as many rounds as necessary until reaching consensus or during a pre-determined number of rounds [8].

Controlled feedback

As mentioned above, between successive rounds, the researcher takes care of the members' responses, providing quantitative and/or qualitative data (e.g. comments, notes), and presents it to the group members as controlled feedback. This information compiles the group opinion and their justifications, which allows the experts to change their answers between iterations reducing discord among the panel [10], [36].

Statistical group response

The statistical group response expresses quantitative measures (e.g. mean, median, standard deviations) of the judgments given by the entire panel group. The overall opinion of the final round is defined as an average or the different ideas are rated numerically based in some statistic measure and then, used as quantitative feedback [10], [26].

3.1.3. Strengths and Limitations of Delphi processes

The Delphi design provides advantages when compared with other approaches that are not suitable for a specific study. For example, *Yang et al.* (2012) argued that this method is appropriate for researches that present subjective inputs, unpredictable judgments and long time frames [26], [37]. Withal, the principal wealth of this technique comes in achieving consensus in areas of uncertainty [38]. Other strengths referred to as important are the flexibility and the simplicity of the method that support the adaptation to specific studies, cost-effectiveness, anonymity, presence of controlled feedback that helps participants to continue motivated as it permits knowledge sharing [37], [38] and it is ideal for situations where exists geographical barriers [26].

Although all the advantages of the method, the Delphi approach also presents some weaknesses. However, the majority of the drawbacks comes from the research or the panel group [37]. As examples of this type of flaws, there are the researcher's and the experts' bias, tendency to eliminate extreme positions to achieve central consensus, the requirement of written skills, time-consuming problems when dealing with a complex issue and others [26]. Regarding the method itself, S. Thangaratinam and C. W. Redman stood out some problems like lack of empirical rules or guidelines, definition of level of consensus and size of expert panel [39].

Ironically, some of the advantages of the method are disadvantages at the same time. Regarding anonymity, some aspects can be seen as problems like the respondents' not assuming responsibility for their opinions and isolate themselves, complicating the connection of ideas among the experts. Plus, the researcher may know some participants or some experts may know each other, being impossible to guarantee total anonymity [8], [39], as explained previously. About iteration, if a study accomplishes a vast number of questions or if the questions are complex, it can lead to fatigue which increases dropout rates [10].

3.1.4. How to perform a Delphi

Although the existence of many modifications, the classical Delphi survey follows several steps that should begin with the analysis of the suitability of the method, availability of resources and the definition and establishment of the necessary level of consensus [38]. First, it is important to identify the nature and extension of the issue being studied and understand if the Delphi approach is adequate to deal with the problem [24]. Time and cost must be taken into account as well as the choice of a good researcher and panelist members. Questionnaires must be elaborated with accuracy to avoid ambiguous interpretations and structure, type of answers and way to measure consensus should be thought and organized before the beginning of the study. Moreover, during the research, it is really important to keep the panel motivated and send reminders to each participant to enhance the response rates [24].

First round

The classical approach starts with a qualitative, open-response round. Many ways to collect data can be adopted like asking the experts to provide a word, a phrase or a note or even to provide as many good ideas as they have in their minds. Usually, the chosen way to collect data depends on how complex the study is, allowing different scopes of information. Anyhow, ambiguous questions should be avoided to obtain accurate data. Data is analysed and organized to use as an input of the subsequent questionnaire [38], [24].

Subsequent rounds

First of all, the questionnaires provided are based on the information collected in the previous round. Commonly, rating or raking techniques are adopted in these subsequent rounds to provide more specific ideas. As previously stated, feedback about the given ideas is presented with the new questionnaires so that consensus can be reached [38].Nowadays, it has been noted that three rounds should be the maximum to avoid withdrawals but also to counterbalance time and costs [38].

Expert panel

An expert is someone who has ample knowledge about a specialized topic. Therefore, an expert panel is a group of experts that are specialists in one specific area or field [8].

The selection of the experts is not an easy task and it also brings methodological concerns considering that just because they have knowledge in an issue, does not mean for sure that they are experts. Adler and Ziglio (1996) [39] have identified four requirements that a person should have to be considered an expert: knowledge and experience, enthusiasm, enough time to cooperate and communication skills. Another key aspect is the size of the panel [8]. There is no universal agreement on the number of experts required to constitute the panel [39]. However, it is known that the sample size depends on the purpose of the study and it must contain the necessary people to cover an entire span of opinions [38], [39]. Many studies include selection criteria (e.g. number of publications in the area, specific competences or even the years working in that specific field) to facilitate a good choice of the panel members [8].

Consensus

Consensus was the first main goal of the classical Delphi and it continues to be one of the important aims of some of the types of the Delphi techniques. Thus, it is really important to understand the definition of consensus or, at least, the acceptable level of consensus that is required in specific researches [39]. According to the Cambridge dictionary, consensus is “a generally accepted opinion or decision among a group of people” which is almost impossible to reach in a Delphi survey [19]. However, there are no universal rules that dictate when consensus is attained. As a matter of fact, it can be achieved in a variety of ways: defining a pre-determined percentage level of consensus, measuring the stability of responses between rounds or even through the aggregation of judgments [39], what brings discussion about not being consensus what they get, but agreement [19]. However, the real concern is to describe how it is going to be measured in each case of study. Whichever the selected manner to measure consensus, it is crucial to keep on mind that accomplish consensus about an issue does not mean that the correct answer has been found [8].

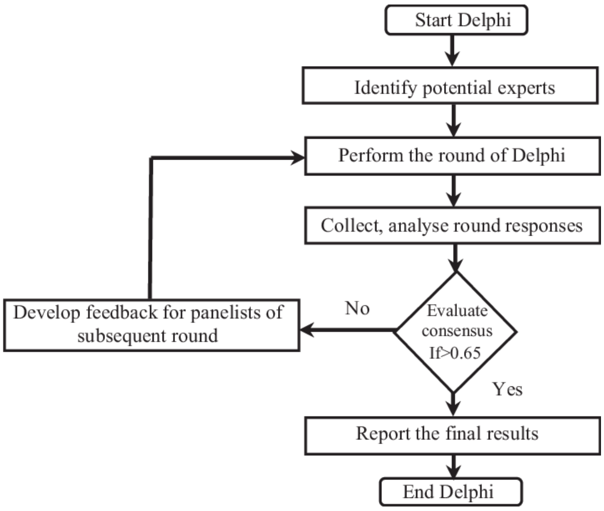

Start Delphi
Identify potential experts
Perform the round of Delphi
Collect, analyse round responses
Evaluate consensus If>0.65
No
Yes
Develop feedback for panelists of subsequent round
Report the final results
End Delphi

***Figure 1** - Delphi technique flow chart [40].*

3.1.5. Rigour of Delphi processes

Alike in any research method, on Delphi studies, establishing rigour in both qualitative and quantitative manners is essential to assure dependable results [20]. Yet, regarding the evolution and changes of the Delphi technique, this process can become challenging. However, quantitative research usually relies on the evaluation of reliability and validity to assess rigour [40].

Reliability refers to the stability of measurement under equivalent conditions. This is, if a specific study with some information is conducted by different groups of panelists, they should obtain the same conclusions to be considered reliable [41]. Two mechanisms enhance reliability in the Delphi process: the decision-making process without face-to-face meetings that avoid personal bias and having a big panel size [40]. Contrary to the quantitative approach, qualitative rigour is measured through trustworthiness elements: credibility, dependability, confirmability, and transferability. Credibility measures how much data can be believed. Dependability measures the consistency or stability of the collected information. Confirmability refers to the degree of objectivity in quantitative data and transferability relies on the applicability of the findings [20].

Regarding validity, Felicity Hasson and Sinead Keeney [20], suggested that "validity it is divided into external, which measures the generalizability of the findings and internal, which refers to the confidence we place in the cause and effect relationship" and argued that there are several methods to measure validity as content and criterion. Content validity measures if a specific instrument covers all the different aspects under investigation. Regarding the Delphi method, the use of an open-ended first round that allows enthusiastic experts to provide an ample range of important items and aspects in the field may help to increase the content validity [40], [41]. Criterion-related validity is found when a test is effective predicting indicators of a construct. Within this type of validity, two important sub-types differ in the timing: concurrent validity, which is demonstrated when a test and a measurement previously validated are correlated and predictive validity, regularly established as accuracy, which aims to predict a measure [40].

Thereby, the existence of many rounds of questionnaires in the Delphi technique enhances the concurrent validity [41].

3.2. Delphi design

Any research method needs effective planning to be successful. Although the wide use of the Delphi technique, this method does not provide universal guidelines or rules to follow. For that reason, for each specific study, the researcher needs to perform a protocol with detailed steps to pursue [24]. The Delphi technique is based on the judgments of a panel member, which cannot be equated to measurements since they can introduce situation or personal bias. To understand if the results of the Delphi technique are feasible, it is important to understand the validity of the results [42]. Technical performance is truly important as it provides a basis for constructive feedback, assess competency, monitoring the rate of technical skills and, consequently, identify how it is possible to obtain more realistic results [43]. However, planning a Delphi study is a particularly complex task once the choice of a specific characteristics can bring advantages and disadvantages at the same time. Thence, it is crucial to analyse all the hypothesis for each choice and select the best one in order to obtain the most accurate results, saving time and money as much as possible [43].

Over the past years, there are been identified some aspects to consider when designing a Delphi study that requires attention. First of all, the definition of consensus. One of the major aims of this methodology is to achieve consensus or agreement about a certain topic. However, the definition of consensus depends on the topic itself and the implication of the research [44]. There are different manners to describe when or how consensus can be achieved and it can evaluate if an agreement exists or defined as a stopping guideline [44]. If the first way is selected, the most common measures used are the statistical approach and percentage levels, but it can be also determined through the aggregate of judgment using measures of central tendency as mean, median and mode, and by confirming the stability of responses. However, the stability of responses is targeted by different opinions since some people defend that is a measure of internal reliability and not consensus [24]. Nowadays, it is defended that reliability, consensus, and agreement are three different things that must be understood and strictly defined before the realization of the study. Reliability measures the "proportional consistency of variance among raters", consensus measures if the participants agree with each other and agreement measures if participants agree with the statement under consideration [44]. Other ideas were shared about the best policy to measure consensus. For example, Keeney *et al.* (2006) argued that the use of confidence intervals could help to determine cut-off points according to the goals of the study [45]. Another study suggested that using a combination of statistics would reduce subjectivity and ensure the validity of results being the variance in response (IQR) an objective and rigorous way of determining consensus [46]. In short, it is important to understand that the stricter the criteria chosen, the more difficult is to obtain consensus and regardless the picked metric, the definition and level of consensus should be explained before data collection [24]. The stability of responses refers to the level of agreement between rounds and it is really useful to ensure the reliability and stability of the results [44]. In 1979 it was proposed a chi-square test, χ^2, as one way to measure the stability of responses [47]. However, more recently, these tests have been rejected as they determine "the independence of the rounds from responses found in them" but not the stability itself. As stability refers to when responses don't change significantly between rounds, median and IQR through graphical presentations of means and standards deviations or with intraclass correlation coefficient, ICC, are other ways that have been used to measure if the answers are stable or not across rounds [44].

Other areas that need to be analysed before the study are the panel members, how to choose them, how many are necessary, how to keep them motivated and the survey itself, how many rounds should exist and how each round should be [44]. The panel group is supposed to be constituted by 'experts' that are specialists in a specific subject, people that have more knowledge and experience in the topic being studied than other people. Yet, these characteristics do not ensure expertise [44]. The goal here is to choose the best participants to provide useful information. Inclusion criteria and training programs can be used to ensure the knowledge level [24]. Another important question is whether is better to have a homogeneous or heterogeneous sample. It has been defended that heterogeneous groups can provide a bigger range of ideas and perspectives, assuring better performance [44]. However, some studies defend that the use of strict selection criteria, that establish a homogeneous group, can lead to high-quality responses [24]. Regarding the panel size, it is known that it depends on the type and complexity of the problem and the availability of resources [44]. However, it is really difficult to establish the ideal number of members. Some people defend that small panel size can provide a diversity of opinions, while others defend that a large panel is essential to generalize different points of view [24]. The ideal number of questionnaire rounds is also subjective. The more rounds there are, more participants' fatigue exist and, consequently, more dropout rates rise. Differently, the use of only two rounds does not allow testing the stability of responses. Taking into account these aspects, three rounds seem to be optimal [44]. The classical approach of the method emerged with an open-ended questionnaire in the first round, gathering qualitative information. After, a modified approach started with a first round made of data collected from literature and expert interviews. If the first alternative is chosen, a lot of data can be gathered, leading to time consuming and lengthy next rounds, which can lead to fatigue and withdrawal. However, this method allows that everyone starts from a common base, being easier to perform the analysis of the responses. On the other hand, the latter approach can limit the data collected and discard some useful information [24], [44]. Therefore, the design of the first round should be thinking carefully and chosen according to the type of study. Finally, attrition can be also a problem in the Delphi technique. There are various ways to minimize it like reminding participants that they are very important and that the study depends on them, make them feel interested and enthusiastic and also, send them reminders and thank you cards. The bigger recommendation is to have a short time frame to avoid participants' fatigue [44].

The intrinsic subjectivity associated with the Delphi design leads to a huge difficulty in analysing the method. Consensus, level of agreement, stability of responses and some of the features mentioned above are really useful to understand and validate the Delphi outcomes [24]. Many different tests have already been performed to evaluate these outcomes, whose choice depends on the type of Delphi and its specific characteristics (scales and types of data) used and type of analysis wanted. One objective of this thesis is to go further in the understanding of which analysis should be done for each specific case.

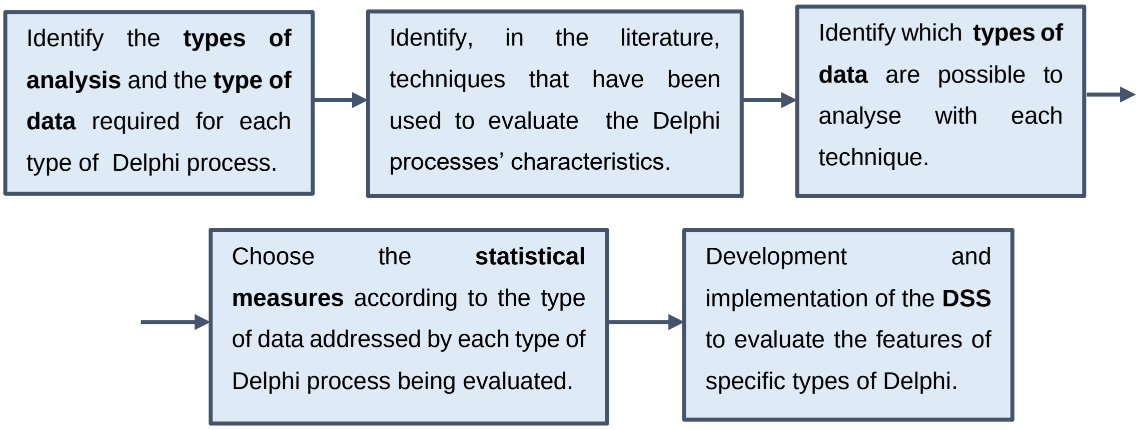
It is possible to understand that the wide flexibility of the method can be an advantage but also a big disadvantage as it brings an additional difficulty in the evaluation of this methodology, specifically because there are no concrete rules to follow as each case is different from the others. In this light, it is really important to spend time going further with researches in this area in order to specify and organize guidelines that can be used in the evaluation of Delphi processes.

4. *DelphiAnalysis* DSS

In this chapter, an overview with the main steps taken to create a novel DSS is presented – this DSS is named *DelphiAnalysis*. After, each step described in more detail to get all the information needed to choose the best options for the final implementation of the DSS. The design of the proposed DSS will be described in a user-friendly way to support the facilitator in guiding the process and the user in understanding the process of characterizing and analysing a Delphi process. This design requires to plan the requirements and features, with the inputs and outputs of each phase. The framework proposed by *Miah et. al* [48] was used to support the implementation of this tool.

4.1. Overview of the steps to take before the development of the DSS

To meet the objectives of this thesis - to develop and implement an automatic tool to digest and analyse information regarding the process and the outputs of three types of Delphi processes (Delphi for selection of indicators, Delphi for weighting judgments and Delphi for shaping value functions) -, some steps need to be followed, as shown in Figure 2. Different concepts need to be explained and correlated to be able to perform a correct evaluation of Delphi outcomes and their reliability. The tool needs to be convenient to evaluate the characteristics of specific types of Delphi processes, so it is important to understand which is the best approach to follow to assess relevant data for a specific type of Delphi process.



Identify the **types of analysis** and the **type of data** required for each type of Delphi process.

Identify, in the literature, techniques that have been used to evaluate the Delphi processes' characteristics.

Identify which **types of data** are possible to analyse with each technique.

Choose the **statistical measures** according to the type of data addressed by each type of Delphi process being evaluated.

Development and implementation of the **DSS** to evaluate the features of specific types of Delphi.

Figure 2 - *Block diagram representing the steps made until the implementation of the tool.*

4.2. Identification of the types of analysis and data

There are different types of Delphi designs, each one with a specific aim, more or less appropriated for a specific issue, as previously explained [8]. The inherent subjectivity of the method does not allow one rule or specific guidelines to be an exceptional manner to evaluate the characteristics of different types of Delphi processes. Instead, each case is a different case and the best way to perform the analysis should be thinking and chosen to fit that specific instance [8]. Then, it is important to organize useful information that will allow choosing the best procedures to implement in order to obtain an advantageous tool to evaluate a specific Delphi that can be employed within a specific situation, with many particular target participants and distinct administration requirements [20]. It is crucial to understand the purpose of each type of Delphi since the metrics used to evaluate them can be different according to its main objective. A summarized table with ten main categories that were identified by Hasson and Keeney [20] is presented below.

Table 2 - *Types of Delphi designs and its principal focus, aim and panelists (combination of information from [15] and [55]).*

Design type	As a forum for…	Aim	Types of participants
Classical	Facts	To collect opinion and gain consensus;	Many; Unbiased experts
Policy	Ideas	To generate opposing views on policy and potential resolutions;	Consider all relevant groups; Lobbyists
Modified	Facts and/or decisions	Aim varies (from predicting future events to achieving consensus);	Unbiased experts or decision makers
Ranking-type	Rankings	To reach group agreement about the importance of a set of issues; identify and rank key issues	Not a large number of experts;
Decision	Decisions that influence future directions	To prepare and support decisions; to create future in reality	Cover a high percentage of the relevant decision makers
Real-time	Facts	To collect opinion and gain consensus;	Many; Unbiased experts
e-Delphi	Facts or ideas	Aim varies according to the nature of the research;	Many; Unbiased experts
Technological	Facts and/or decisions	Aim varies (from predicting future events to achieving consensus);	Many; Unbiased experts or decision makers
Online	Facts and/or decisions	Aim varies (from predicting future events to achieving consensus);	Many; Unbiased experts or decision makers
Argument	Facts and/or decisions	To develop relevant arguments and expose the reasons for different opinions on a specific single issue;	Many; Unbiased experts or decision makers
Disaggregative policy	Decisions that influence future directions	Constructs future scenarios: panelists are asked about their probable and the preferable future;	Cover a high percentage of the relevant decision makers

4.3. Methods used in literature

The next step was to recognize which techniques have been used and the situations which they are applicable and useful. During May 2019, databases as PubMed, ScienceDirect, Web of Science and Google Scholar were used to find the best articles with the different forms to evaluate, qualitatively and quantitatively, Delphi processes and its outputs. Keywords like “Delphi” in conjunction with “evaluation” or “measurement”, and with each of the features incorporated in the method that was supposed to evaluate, such as “consensus”, “stability of responses”, “agreement”, and “feedback”, were used to find the articles with all the quantitative metrics applied. On the other hand, the combination of “Delphi” and “qualitative analysis” or “qualitative methods" were used to find articles related to the qualitative methods used in Delphi studies. No date limits were chosen but it was easy to realize that is a recent matter of study.

Retrieved articles were assessed one by one, titles and abstracts were read to identify the relevant studies and the rest were excluded, as shown in Figure 3. Citations present in these articles were also assessed to identify whether there were useful or not. The remaining articles were then fully analysed and the ones that summarized information already achieved in previous ones were discarded as well as the ones in which Delphi studies were not the principal data collection method used or when the approaches used to evaluate Delphi surveys were not metrics but other types of evaluators. After that, a set of 16 articles relative to quantitative metrics ([44], [46], [49], [53] and [57]-[68] and a set of 6 articles relative to qualitative metrics ([41], [55],[56] and [69]-[71]) were selected to be the basis of the tool that needs to be developed to elaborate a complete evaluation of the technique outcomes.

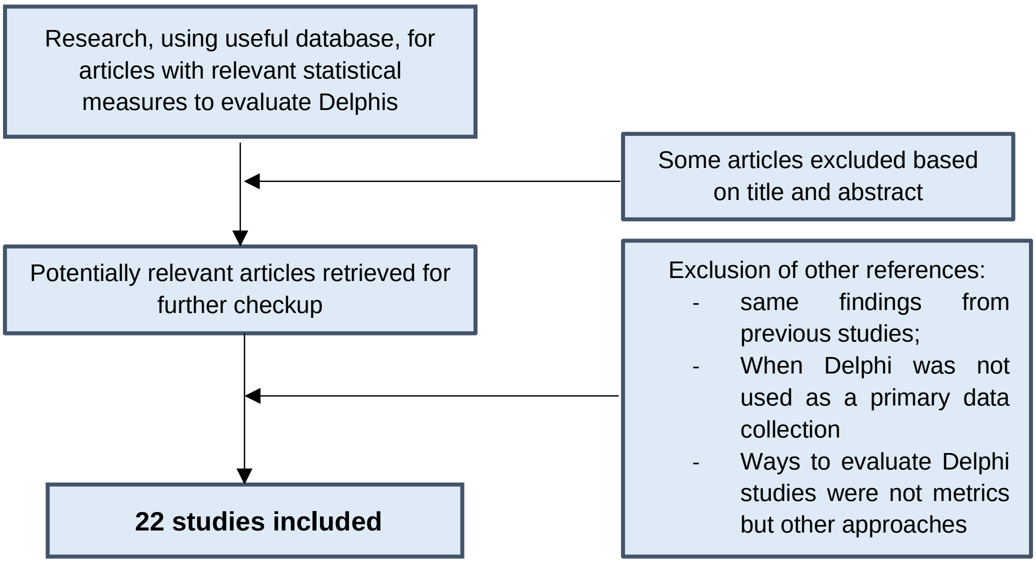

Research, using useful database, for articles with relevant statistical measures to evaluate Delphis
Some articles excluded based on title and abstract
Potentially relevant articles retrieved for further checkup
Exclusion of other references:
- same findings from previous studies;
- When Delphi was not used as a primary data collection
- Ways to evaluate Delphi studies were not metrics but other approaches
22 studies included

Figure 3 - *Flow diagram (based on [58]).*

During the research, it was possible to understand that measures of central tendency (mode, mean and median), measures of dispersion (variance, standard deviation and IQR) and frequency distribution (histograms and frequency polygons) should be incorporated into the Delphi assessment as they are the most basic ways to have useful information about statistical group responses and as they allow to provide quantitative feedback to the panel members [49]. These measures are not enough for a full analysis, but they can complement it and make it easier [49].

During the research, some important information was gathered for example the difference between parametric and non-parametric techniques and the existence of different types of scales and correspondent data that require different statistical measures to be analysed.

4.3.1. Types of scale and measurement scales

To evaluate the quality of Delphi processes according to the measures presented before, it is always necessary to recognize that measurement exists in many different forms and that scales of measurements fall into certain definite classes [50]. These classes are determined by the operations of "measuring" and by the mathematical properties of the scales. The statistical manipulations that can be legitimately applied to empirical data depend upon the type of scale against which the data are ordered [50]. In statistics, there are four data measurement scales: nominal, ordinal, interval and ratio, which are used when dealing with different types of data [51], as shown in Figure 4.

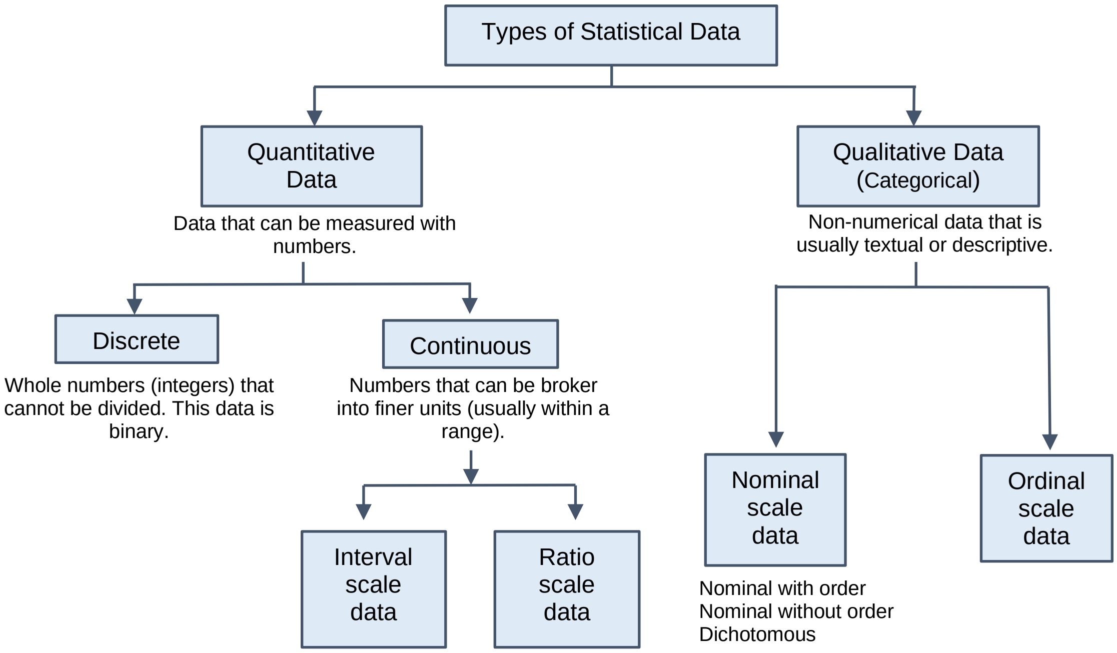

Types of Statistical Data
Quantitative Data
Data that can be measured with numbers.
Qualitative Data (Categorical)
Non-numerical data that is usually textual or descriptive.
Discrete
Whole numbers (integers) that cannot be divided. This data is binary.
Continuous
Numbers that can be broker into finer units (usually within a range).
Interval scale data
Ratio scale data
Nominal scale data
Nominal with order
Nominal without order
Dichotomous
Ordinal scale data

Figure 4 - *Types of data and respective scales of measurement (based on [50] and [52]).*

The different types of scales are described below.

Nominal Scales

Nominal scales contain rules for deciding if two objects are equivalent or not, i.e., for categorizing [52]. Equivalence means that two objects have a critical property in common [52]. The result of nominal scale is a series of classes which may be given a numeric designation. Both labels and categories can be nominal scales, but it is useful to distinguish them. Labels, numeric or not, are used to identify individual objects. In contrast, categories are grouping of objects (e.g. race, gender) [52]. Nominal scales can be transformed in any manner that does not assign the same number to different categories [52].

Ordinal Scales

Ordinal scale arises from the operation of rank-ordering, this is, it involves rules for deciding if an object is greater or less than another one, concerning a specific attribute [52]. In ordinal scales, the order of the values is what's important and significant. Jum C. Nunnally and Ira H. Bernstein [52] defend that he transformations permissible for this type of scale is more limited than it is for nominal scales as it must preserve the rank-order properties of the data.

Interval Scales

Interval scales define the unit of measurement as higher, equal or less [52]. These scales are quantitative as it is known the order and the exact differences between the values, this is, the distances among objects on the attribute. The only features that are not known are the absolute magnitudes of the attributes [52].

Ratio Scales

Ratio scales are those possible only when there exist operations for determining all four relations: equality, rank-order, equality of intervals and equality of ratios [50]. It is an interval scale with a rational. The presence of this zero makes ratios of any two measures meaningful [52]. An absolute zero is always implied, even though the zero value on some scales may never be produced [50].

Table 3 *- Summary of the characteristics applicable to each measurement scale (based on [50]).*

	Nominal	Ordinal	Interval	Ratio
The "order" is known		√	√	√
Frequency of Distribution	√	√	√	√
Mode	√	√	√	√
Median		√	√	√
Mean			√	√
Quantifies the difference between each value			√	√
Can multiple and divide values				√
Has a 'true' zero				√

4.3.2. Parametric vs Non-parametric techniques

Parametric techniques can be used when there are at least 30 expert members [53] or when the data is interval or ratio-scaled and they are those that make assumptions about the parameters of the sample distribution [49]. The latter ones are used when there are less than 30 experts, with categorical or ordinal data or even when the distribution of responses for each of the items is non-normal [53].

In the literature it was shown that most of the techniques used to evaluate the characteristics of Delphi processes are quantitative and many of them are to measure the reliability of these studies. However, qualitative analysis can be used to inform the development of Delphi studies [54]. It allows the scope of outcomes to be defined in a way which holds most relevance to stakeholders, it helps to identify the appropriate language for use in a Delphi survey and it can be compared with other stakeholder data or alternative sources of outcome data [54]. Also, these measures are performed to generate or develop analytical categories or theoretical explanations [55] and it is really useful when exploring the meanings of various phenomena [56]. For all these reasons, it was useful to look for qualitative measures as they complement the analysis of a Delphi study.

The most common quantitative and qualitative methods used to evaluate Delphi characteristics are following described.

Quantitative techniques

Two tables are displayed to facilitate the understanding of all the quantitative procedures. The first table is useful to better understand when and with which type of data we can use these tests. The second one presents the main objectives of each procedure, if they are relative to the iteration process or final results, which type of information they use and which type of information they give. Next, a small description of the methods is presented in order to understand their main purpose.

Table 4 - *Statistical measures used to test the validity of Delphi surveys and types of data that is analysed by them (information collected from [44], [46], [49], [53] and [57]-[68]).*

Name of the technique	**Parametric vs Non-Parametric**	**Number of participants**	**Type of data/distribution**
Mode	Parametric or non-parametric	Doesn't matter	Ordinal or Interval/Ratio; Not useful with large scales
Arithmetic mean	Parametric	At least 30 experts	Interval/ratio data that are not skewed;
Median	Non-Parametric	Less than 30 experts	Ordinal and interval/ratio data; Not useful with few values;
Interquartile range	Non-Parametric	Less than 30 experts	Ordinal, Interval or Ratio data
Standard Deviation (SD)	Parametric	At least 30 experts	Interval/Ratio data
Kendall's coefficient of concordance	Non-parametric	Less than 30 experts	Ordinal or interval data; Non-normal distribution;
APMO Cut off Rate	Parametric or non-parametric	Doesn't matter	Can be used with all types of data
Chi-square (χ^2) test	Non-parametric	Less than 30 experts	Nominal or ordinal data; Non-normal distribution;
Coefficient of variation (CV)	Parametric	At least 30 experts	Interval or ratio-scaled data; Normal distribution;
F-test	Parametric	At least 30 experts	Interval or ratio-scaled data; Normal distribution;
Pearson Correlation Coefficient	Parametric	At least 30 experts	Interval or ratio-scaled data; Normal distribution;
Paired t-test	Parametric	Parametric	Interval or ratio-scaled data; Normal distribution;
McNemar Change test	Non-parametric	Less than 30 experts	Nominal or ordinal data; Non-normal distribution;
Spearman's Rank Correlation Coefficient	Non-parametric	Less than 30 experts	Nominal or ordinal data; Non-normal distribution;
Wilcoxon Sign test	Non-parametric	Less than 30 experts	Nominal or ordinal data; Non-normal distribution;
ICC	Parametric or non-parametric	Depends on the case	Depends on the case

Kappa statistics	Non-parametric	Less than 30 experts	Nominal/ Ordinal data;
Kruskal- Wallis test	Non-parametric	Less than 30 experts	Ordinal or rank data
Wilks Lambda test	Non-parametric	Less than 30 experts	Nominal/ Ordinal data;
Scott's Pi Statistics	Non-parametric	Less than 30 experts	Nominal/ ordinal data

Table 5 - *Type of information that each procedure uses and give and if they are relative to the iteration process or the final results of a Delphi study (information collected from [44], [46], [49], [53] and [57]-[68]).*

Name of the procedure	Information that needs	Information that gives	Relative to
Mode	Histogram of rankings of the different experts of each item	The most popular rate of a specific item chosen by the experts	Iteration process and final results
Arithmetic Mean	Rankings of different experts of each item	Average of the rankings of each item	Iteration process and final results
Median	Histogram of rankings of the different experts of each item	Value that divides the higher half from the lower half of the rankings	Iteration process and final results
Interquartile Range (IQR)	Percentages of responses that falls into a specific score	Corresponds to half of the responses	Iteration process and final results
Standard deviation (SD)	Rankings/scores and its average	Quantifies the variation of each score relatively to the average	Iteration process and final results
Kendall's coefficient of concordance	Scale with levels of concordance	Strength of agreement between experts	Iteration process and final results
APMO Cut off Rate	"agree", "disagree" and "unable to comment" responses and their percentages	Gives the level or percentage of consensus	Iteration process
Chi-square test	Rankings/scores of each item	To see if there is a relationship between two variables.	Iteration process
Coefficient of variation	SD and mean of the rankings/scores of each item	Degree of consensus	iteration process
F-test	Variances of item scores of each round	Ratio of variances of item scores among experts	Iteration process and final results
Pearson Correlation Coefficient	Metric variables as responses	Level of association between two variables	Iteration process
Paired t-test	Mean of differences in responses	To see the change of opinion of the experts between rounds	Iteration process
McNemar Change test	Dichotomous responses	To analyse changes in the responses or compare distributions	Iteration process
Spearman's Rank Correlation Coefficient	Ranking-scale responses	Association between the ranks of the responses	Iteration process
Wilcoxon Sign test	Sum of the positive and negative ranks to determine the p-value	To analyse the change of responses	Iteration process
ICC	Scores of each item	See the similarity of the responses within a group	Iteration process
Kappa statistics	The probability that there is a chance agreement	Measure agreement between experts; Proportion of agreement beyond that expected by chance	Iteration process and final results

Kruskal-Wallis test	Information about the independent samples	If the samples were originated from the same distribution	Iteration process
Wilks Lambda test	Which are the groups, answers of the groups	If there are significant differences between group means	Iteration process and final results
Scott's Pi Statistic	Number of each option answers	Level of agreement and the agreement expected by chance	Iteration process and final results

- **Mode**

It is a measure of central tendency that refers to the proportion of experts who chose the most popular rate of a specific item/statement [57].

- **Arithmetic mean**

The mean is another measure of central tendency. It is the average of a set of numerical values and it is calculated by adding the values together and dividing by the total number of values. The fact that mean is only used with numerical values makes the use of it with ordinal scales a wrong procedure [58].

- **Median**

The median is a measure of central tendency and represents the value that divides the higher half from the lower half of the sample. Some authors argue that this measure should be used instead of the mean as it can be used with ranked data (ordinal, interval and ratio) [49].

- **Interquartile range**

The interquartile range one of the four measures of distribution. Is the measure of dispersion for the median and consists in half of the observations. An IQR less than 1 means that more than 50% of the opinions fall within one point on the scale [57].

- **Standard Deviation (SD)**

SD is a measure of dispersion or distribution that tries to capture the average distance each score is from the average. The combination of SD with mean is commonly used for consensus evaluation since it indicates the aggregate judgments and agreement [49].

- **Kendall's coefficient of concordance (W)**

This is a non-parametric test used to assess the level or strength of consensus between participants [46], [59]. It ranges from 0 to 1, in which a value of 0.1 indicates a very weak agreement while 0.7 means a strong agreement. Regarding this, it is supposed that the coefficient increases over the rounds. Moreover, it is also used to determine the inter-judge reliability [60]. This method is only available for Delphi surveys that use levels of concordance.

- **Average Percent of Majority Opinions (APMO) Cut-off rate**

When this metric is used, it is necessary to express the participants' comments "agree", "disagree" and "unable to comment" in percentages per statement. After that, the number of majority agreements and disagreements can be calculated taking into account that the majority means a percentage above 50%. Majority of agreements and disagreements are sum up and then divided by the total number of responses that gives a percentage. If that percentage is higher than a predetermined one, consensus is considered reached [49].

- **Chi-square test (χ^2)**

Chi-square is a non-parametric statistical hypothesis test to assess whether there is a relationship between two variables. It has been proposed to check for stability of responses and later discarded as it determines the independence of the Delphi rounds from responses obtained in them and not the stability of responses between rounds [49].

- **Coefficient of Variation**

The coefficient of variation (CV) is a parametric test described as the ratio of the standard deviation of a specific item rating score to its corresponding mean among the panel members [61]. Therefore, for each item only one CV exists. A value above 1 indicates that the responses of the panelists are scattered compared to the mean of responses. Contrary, a small value indicates that the data scattered compared to the mean is small [53]. English and Kernan (1976) used the coefficient of variation to determine the stopping rule. If the value for an item was found high (>0.8), the corresponding item/statement was needed to be modified and an additional round of data collection about that item was necessary [62].

Table 6 *- Coefficient of variation and consensus (from [62]).*

Coefficient of variation	Decision Rule
$0 < CV < 0.5$	Good degree of consensus. No need for additional round.
$0.5 < CV < 0.8$	Less than satisfactory degree of consensus. Possible need for additional round.
$CV > 0.8$	Poor degree of consensus. Definite need for additional round.

On the other hand, Dajani (1979) argued that to measure the stability of the responses for an item, an absolute CV different can be measured by subtracting the CVs obtained in two consecutive rounds, for that specific item [63]. Stability is reached when the absolute value of the difference is small and close to zero [53].

- **F-test**

F-test is a parametric method for interval/ratio-scaled data that uses a F-value obtained by calculating the ratio of variances of an item scores among experts [53]. This type of test can be used when researchers want to examine independent samples. The F-test can also be used to examine the mean differences among more than two groups [46]. When the F-ratio is equal to

1, the variances of both rounds are the same and perfect stability of consensus is reached. If there is a big deviation from 1, there is not a good stability and another round of Delphi is necessary to achieve stability of consensus [64].

- **Pearson Correlation Coefficient**

The level of agreement between two round ratings among panelists on each item is an alternative way to measure stability of responses and the *Pearson Correlation Coefficient* was used for this end [61]. It is a parametric measure that quantifies the level of association between two variables, this is, in Delphi surveys it is the relationship between the responses of the experts for each item between two consecutive rounds. A high positive coefficient (close to one) indicates that the responses about an item among the group are similar which mean great stability and consensus for that specific item. Contrary, a coefficient correlation close to zero indicates no relationship between the experts' opinions, consensus not reached and the item needs to be included in the next round to be evaluated again [53].

- **Paired t-test**

This parametric method is used to evaluate whether or not the mean of the difference in responses between two successive rounds about an item is close to zero. If so, this demonstrates that there is no change of opinion between the rounds. This method is the same as the F-test but this one is used when the same people are tested twice and not different groups. It is applied to analyse if there are "significant differences between the means for Delphi theses of successive rounds" [46]. This value is given by the p-value associated with the t-test statistic. A small p-value indicates that there is little change in the ratings so the item can be removed from the next rounds [53].

- **McNemar Change Test**

This non-parametric test is used to analyse changes in the responses. It is a test to compare two dependent samples in terms of their distributions across nominal-scaled data and on Delphi processes it can be used to quantify the level of change between rounds, in a positive or negative direction [61]. However, it is only applicable when there are dichotomous responses [61].

- **Spearman's Rank Correlation Coefficient**

Spearman's rank correlation is a non-parametric test that represents the association between the ranks of the ratings of the panel regarding one item. This means that this approach can only be used with Delphi surveys that use ranking scales [53]. This coefficient is a value between -1 and $+1$ with $+1$ indicating a perfect positive correlation between responses on an item from two successive rounds. The closer r_s is to zero, the less correlation between rankings and the closer r_s is to -1, the greater the correlation between the responses of the item but in an opposite direction indicating disagreement among the panelists. It is also possible to compare the calculated result with the critical value (obtained from a table of critical values of the Spearman's rank correlation coefficient, at $\alpha = 0.5$) [53]. If the obtained value is smaller than the critical one, this means that the relationship between the ratings of the panel members on that item is not significantly strong and that item should be included in the next round [53].

- **Wilcoxon Sign test**

The Wilcoxon Sign test is a non-parametric method used to compare two related samples or repeated measurements on a single sample using the ranks of the pairs of scores formed by the matched pairs in the sample, this is, it assesses whether or not the ranks of the difference in responses to a specific item from two consecutive rounds is zero. So, it provides the sum of each of the positive and negative ranks of the differences between consecutive rounds with a Z statistic and its asymptotic p-value. If the Wilcoxon coefficient is not significantly different from zero it indicates a little change in responses and therefore great consensus and stability of the item under analysis [49], [53].

- **Intraclass Correlation Coefficient (ICC)**

ICC is one way to assess reliability of Delphi studies accurately and effectively, being useful with numerical ratings and normal distributed data [65]. ICCs are useful to assess consistency or conformity between two or more quantitative measurements. Different types of ICCs may fit better to specific data and they can be either parametric or non-parametric and it has been used to assess the consistency of responses and to measure agreement among the panelists [49].

- **Kappa Statistic**

There are ways to measure agreement among experts in which the calculations do not take into account the agreement expected purely by chance. For example, if two experts agree purely by chance, they are not really agreeing. Kappa statistics are measures of "true agreement" indicating the level of agreement beyond that expected by chance [66]. The difference between Kappa techniques is specially in the value of the probability that there is a chance agreement because of the different assumptions made by the experts regarding the ratings [57]. However, Kappa statistics are measures only for nominal scale agreement and assumes that rating has no natural ordering [44]. The two most common Kappa statistic methods used in Delphi processes are the *Cohen's Kappa* and the *Fleiss' Kappa*. The former works for two raters, whereas the latter applies to any fixed number of raters. Both measures are suitable for nominal scale agreement and assume that the ratings have no natural ordering and they take into account the possibility of the agreement or disagreement occurring by chance [49].

- **Kruskal-Wallis test**

It is a non-parametric statistical test alternative to the one-way ANOVA [59]. The test determines whether the median of two or more groups are different using hypotheses. The Kruskal-Wallis tests if there is a significant difference between groups. However, it won't tell which are the groups that are different. To understand which ones are the different it is necessary a post hoc test [71]. The *Kruskal-Wallis* test compares the ranks [67]. If the test value is less than the critical value at a specific significance level, this indicates that consensus among different subgroups is achieved [59].

- **Wilks Lambda test**

This test is used with the same purpose as the *Kruskal-Wallis* test, but it is reported in results from MANOVA. More specifically, it tests how well each level of independent variable contributes to the model [68]. The closer the *Wilks Lambda* value is to zero, the more well separated the groups are according to the independent variable and the more contradicts the null hypothesis which assumes the equality for the panelists. However, this test should be complemented with the analysis of the correspondent p-value. On the case of Delphi processes, it is an useful method to check if the field of expertise influenced the responses [68].

- **Scott's Pi Statistic**

This coefficient is used to measure the inter-rater reliability [68]. More specifically, it measures the observed level of agreement between the panelists. *Scott's Pi Statistic* is a suitable for nominal and ordinal data with three or more coders instead of two as in the *Cohen's Kappa Coefficient* [68]. This statistic ranges from -1 to 1, with 1 representing perfect agreement, 0 indicating completely random agreement and -1 indicating perfect disagreement [68].

Qualitative techniques

Qualitative data can be non-numeric, textual (fieldnotes or transcripts), oral or visual (illustration of the quantitative data). The more usual way is textual and it can be text in open questions in a questionnaire, written arguments for the quantitative statements or interview talk [56]. Every Delphi study that has an open-ended questionnaire in the first round or the ones that use arguments supporting the quantitative statements needs qualitative evaluation [69]. The main types of Delphi that need qualitative analysis are the ones with an open-ended first round as the classical Delphi, Policy Delphi, Modified Delphi and the Argument Delphi.

- **Content analysis**

This type of analysis is a powerful technique used to identify reference modes and to estimate parameters from textual data. Moreover, it is a deductive coding technique since the researcher starts by defining a set of codes to be used in the process [70]. Content analysis may involve the use of qualitative software, described below.

- **Coding analysis**

At the first level of coding, distinct concepts and categories in the data will be the basic units of the analysis. Basically, it's breaking down the data into first level concepts or master headings and second level categories or subheadings.

- **Grounded Theory**

Consists of a set of techniques to identify themes or concepts across texts and link these concepts to create meaningful theories [70]. More generally, it is a research approach which denotes the discovery of theoretical ideas and forcing a certain theory to emerge. Grounded theory is supposed to emerge without the researcher's interference [71].

- **Softwares**

Qualitative softwares such as QSR NUD*IST or Ethnograph [41] has been used to handle qualitative data as they enable complex organization and retrieval of data [55]. Data collected from this initial stage are analysed by grouping similar items together. When there are various terms used to one issue, the researcher groups try to provide a universal description for them. These grouping systems and descriptions need to be verified to ensure that the data is fairly represented [41].

4.4. Types of Delphi to be analysed within *DelphiAnalysis*

After a detailed analysis of all the qualitative and quantitative concepts and tools, it was necessary to choose the right ones to implement within the *DelphiAnalysis* DSS, so as to enable analyses of the process and final results of a Delphi process. As explained previously, some techniques are better than others according to the type of data under analysis, according to the type of Delphi process. Thus, it is not possible to create a great tool to evaluate the features of every types of Delphi studies. Then, the idea was to develop a useful DSS to be applied in real healthcare context, to analyse three specific types of Delphi that have been used in many different health contexts:

1. Delphi for indicators selection using a Likert scale
2. Delphi for qualitative weighting of indicators using the MACBETH model
3. Delphi for shaping the value function for each indicator

and test it using the value results obtained in a project already developed that used these designs of Delphi processes, the EURO-HEALTHY project [68]. This project is one example of the many applications in the healthcare area that used these types of Delphi, being an inspiration for the creation of this tool and an ally data from the project can be used to test the DSS.

1. Delphi for indicators selection (using a Likert scale)

Indicators are essential instruments for monitoring and evaluation population health. In this light, the selection of indicators should reflect scientific evidence on health as well as the views of health experts and stakeholders [68]. For this reason, the Delphi technique has been widely used for quality-indicator development in healthcare since it allows gathering opinions and knowledge of an ample range of individuals with diverse backgrounds and located in various regions ensuring anonymity [68]. The method has been already used in a vast health contexts and to identify many different types of indicators as prescribing indicators, indicators reflecting patient and general practitioner perspectives of chronic illness, performance indicators for emergency medicine and indicators for cardiovascular disease [1], being essential in the healthcare area. As described, there are diverse applications of this type of Delphi that can include selection of indicators using a specific scale. Likert scale is one of the most common and it is widely used in this field [72].

The Likert scale is a psychometric scale used to measure attitude or conviction when it is equated with strength and intensity that is commonly used in research that employs questionnaires that cover a range of opinions on a topic [72]. The participants of the survey need to vote according to their opinions, perceptions and behaviors to the topics under analysis. The usual form of a Likert scale consists of statements to which the respondent needs to indicate the degree of his agreement or disagreement using the following options: *Strongly disagree*, *Disagree*, *Neither agree nor disagree*, *Agree* and *Strongly Agree*. Normally, the scale presents a neutral point even if it is a scale with 7 levels instead of 5 [72]. The direction of the measure is indicated with the agreement or disagreement and the intensity with being strong or not [72]. It is important to note that Likert scaling assumes that the distances between each answer choice are equal [72].

2. Delphi for qualitative weighting of indicators (using the MACBETH model)

Weights are used in indicator aggregation allowing developers or users to assign different weights to the indicators [73]. Weighting technology can be classified in two categories: statistical-based or participatory-based methods. In the first case, weights are assigned based on the analysis of the data of the indicators. In the second case, weights are given based on experts' opinions as it happens using a Delphi methodology [73]. The main goal of this type of Delphi is to collect qualitative weighting judgments on the indicators in a way to indicate how important is to close a gap in its performance range [81].

This type of Delphi is mostly used in multicriteria modelling processes since weighting judgements is a modelling activity [74]. Nowadays, multicriteria decision conferencing processes have been adopted for multicriteria modelling, when there is a small group in decision conferences. To extend this framework to collaborative contexts in which it is important to capture a diversity of points of view from different experts and stakeholders, multicriteria decision conferencing can be combined with a Delphi process (p.e. Delphi for qualitative weighting judgements or Delphi with the construction of a value function) using specific multicriteria methods, as the MACBETH model [74].

MACBETH is an interactive, constructive approach for decision aid. It is a multi-criteria approach motivated by multi-attribute value theory used for the quantification of value judgments [75]. Technically, MACBETH uses a chain of linear programs for assigning numbers to the elements of a set based upon qualitative judgments on the difference of attractiveness between two action at a time, expressed by a decision maker [75]. In Delphi processes as the ones treated in this thesis, the judgments will not be quantified in numerical values by linear programming since the paper of the participants is, in this case, to transmit their opinions about the topics under analysis, according the MACBETH semantic scale of attractiveness (*Extreme, Very Strong, Strong, Moderate, Weak, Very Weak* and *Not Important*) [75].

The MACBETH method assists the evaluator obtaining value functions and weights for each concern, which are the main functions used in this dissertation [76].

3. Delphi for shaping the value function of each indicator

Value functions are functions of states (or state-action) that estimates how good it is for the agent to be in a given state or how good it is to perform a given action in a given state [74]. The objective of this type of Delphi is to determine the shape of the value function that characterizes each indicator (to obtain consensus about the type of curve), which indicates what is the added value to population health of improving performance along the indicator range [74]. When using the MACBETH model as an assistant, value functions are obtained by asking the evaluator to judge the difference of attractiveness between different levels of performance of each indicator using the correspondent scale. This is, in the first round of a Delphi process, the performance range of each indicator is divided in three sub ranges and each participant provides its opinion by voting according to the MACBETH semantic scale. The votes on the three sub ranges is organized in order to attribute a value function curve shape that represents the global opinion about the performance range of the indicator. Figure 5 shows the types of curves that can be obtained: linear, concave, convex or S-shaped (S-sigmoid or S-seat).

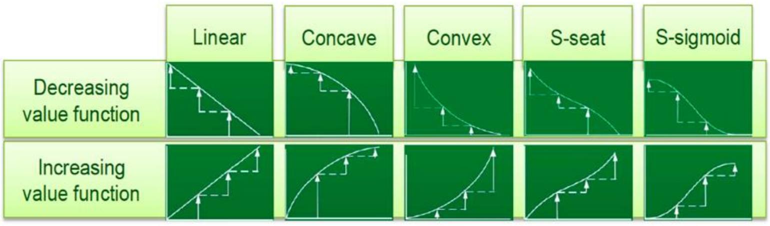


Figure 5 - *Types of value function curves.*

The subsequent rounds of the Delphi process are slightly different since the participants vote directly on the type of shape that they think is the best to describe the performance range instead of voting on the three sub ranges of performance according to the MACBETH semantic scale.

The value function of each indicator is defined on its range from minimum to maximum performance. It is useful as it serves to convert performance of the indicator into a value for population health, allowing to understand what is the value added to population health of improving performance along the indicator range [77].

4.4.1. EURO-HEALTHY Project

The EURO-HEALTHY project was a three-year research project (2015-2017) aiming to advance knowledge about which policies that have the highest potential to enhance health and health equity across European regions. Participatory processes involving multidisciplinary experts and key stakeholders at different geographical locations were adopted to build sound methods and tools to evaluate and monitor the overall population health in Europe [77]. The project recognized as critical to develop and test methodologies that could inform about the best policies to improve health and health equity while accounting for cost, doability and power issues and in light of the EURO-HEALTHY scenarios [77]. We herein briefly describe three web-Delphi processes carried out within the process of developing a population health index to evaluate population health across European regions.

Web-based Delphi indicators selection

The 130 indicators included in this Delphi study as potential indicators of population health determinants and health outcomes were selected from the literature [68]. The role of the panel was to review these indicators and state the level of agreement about how relevant each indicator would be in evaluating Europe's population health. A three-round web-based Delphi in which the participants were asked to indicate their level of agreement or disagreement with indicator's relevance for evaluating population health in Europe was then conducted [68]. An "identity card" for each indicator for on-line search of information and scientific evidence was always available for the panelists [68]. In the first round, the participants were asked to show their agreement or disagreement with the following statement: "*This indicator is relevant to the evaluation of the Europe's population health*" through a 5-level Likert scale [68]. In the second and third rounds the participants were presented with feedback about the previous results and which indicators had been approved or rejected. The indicators that did not reach agreement were included for re-evaluation with panelists taking the option to change or maintain their original answers [68].

Web-based Delphi for weighting

Four Web-based Delphi weighting panels were formed by participants and stakeholders that were selected based on their area of expertise, ability to provide unbiased judgments and availability to participate (Figure 6) [68]. A common Delphi design was implemented simultaneously to all of them, with members of each panel answering to Delphi questions (related with their areas of expertise) in the same time period.

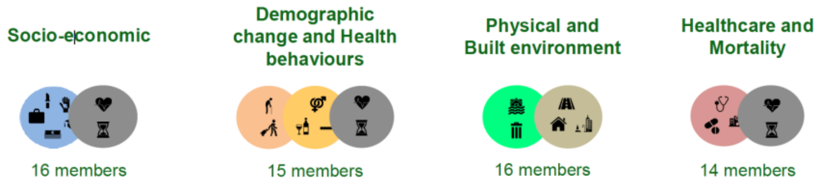


Figure 6 *- Panel's names and number of members of the Web-Delphi weighting.*

As in the previous case, an "identity card" was available for the participants [68]. Also, for each indicator, it was available the range between the worst and the best performance across the European regions [68]. In the first round, participants were presented with a list of indicators and their respective gaps. Each participant was asked to answer the question "*To reduce inequalities in Europe, how important is to close this gap?*" considering how big is the gap and how important is to develop policies to close this gap to reducing health inequalities in Europe [68]. The answers were provided with a MACBETH qualitative judgment scale from not important to extremely important [68]. The weighting judgments given for each indicator in each round were analysed to detect the existence of group consensus on a majority judgment.

Web-based Delphi for defining the shapes of value functions

The Web-based Delphi value function process was constituted again by the four panels above mentioned with a common design being implemented for all. Each panel was answering to questions related with their area of expertise but with a distinct number of individuals answering.

The range of performance of each indicator was divided in three equal pieces representing three changes. In the first round, each participant evaluated the contribution of changes in performance within each indicator according to the contribution of each indicator to the population health [68]. Answers were provided with the MACBETH qualitative judgment scale (from very weak contribution to extreme contribution) [68]. The sequence of these three judgements has implicit a value function shape. In the second round, participants had access to a table where they could see the implicit value functions and could change them having feedback of other's answers to help. In the third round, participants were presented with the distribution of the panel answers given in the second round.

4.4.2. Differences between one panel and multiple panels

The choice of panel members is crucial in the design of a DSS. Panel members should be representative of their profession or area of expertise to have the power to implement their findings [78]. Although there is no agreement about how much members should be part of the panel, it is known that the members need to have availability to participate, interest in the area and enough background on the issue being studied [24]. The size of the panel is not the only feature that can vary from Delphi to Delphi. The members can be grouped forming one global panel or they can be divided into different groups forming more than one panel [24]. Usually, when there is more than one panel, this means that there are different groups that answer exactly the same questions but not simultaneously. It is a way to simplify the process, make it easier to collect data and analyse it. However, since all the members are answering the same questions, the analysis can be done together or can be compared with each other [24]. For example, in the EURO-HEALTHY project, the Delphi process for selection indicators was constituted by only one panel that answered to all questions while in the other two cases, 4 panels completely independent did not answer the same questions, only the ones correlated with their area of expertise. In this case, the responses cannot be analysed together or even compared. Therefore, the analysis must be independent.

In this sense, it is extremely important to understand the type of panel that we are dealing before performing statistical analysis of the responses provided in a Delphi process. For example, in this thesis, when testing the *DelphiAnalysis* DSS with the EURO-HEALTHY data, it is very important to understand that all participants can be taking into account for the case of the Delphi for selection of indicators but in the other two Delphi processes, the participants to have in consideration at each time need to be carefully selected based on their areas of expertise.

4.5. Selection of the best statistical measures to implement

Belton *et. al* (2019) argued that there are many design features and methods, described in literature, to analyse Delphi processes [79]. However, there is not a full explanation or a guide with the main steps to understand how to implement an appropriate Delphi procedure leading to several mistakes in both evaluation and interpretation of a Delphi [79].

The authors defend that the feedback provided to the participants is a key design issue as it has implications for how panelists respond to subsequent rounds and for the overall effectiveness of the process in gaining group consensus response or stability of responses [79]. For ordinal scales as Likert-scale and MACBETH, feedback reporting central tendencies and dispersion it's very beneficial since it allows the participants to see how their answers compares to the group opinion as a whole [79]. They defend that central tendency measures can efficiently and effectively depict an aggregated response for several panelists [79], being for this reason, crucial to approach these measures in Delphi analyses, as they are simple measures rich in useful information both for the evaluator and/or participants. In this sense, measures of central tendency and dispersion will be the first ones to be implemented in the tool, but they will be complemented by additional ones. As mentioned previously, mean and SD are not appropriate when dealing with ordinal scale although they are misused many times. The writers recommend providing the median and inter-quartile range for responses made to individual ordinally-measured question items [79]. They also defend that can be helpful to display the type of responses and provide the feedback in a tabular fashion, graphically or both and using visual summaries such as bar charts and/or boxplots [79].

According to the authors, Von der Gratcht (2012) [80], argued that a panel's opinions must first be relatively stable before consensus can be meaningfully assessed so they defend measuring both consensus and stability on a round-by-round basis [79]. As a consensus criteria can be used the same or similar being reported by a pre-determined percentage of panelists (e.g. 80%) or particular levels of statistical dispersion as measured by inter-quartile ranges [79]. On the other hand, qualitative feedback creates challenges around how to aggregate the responses once the facilitator aggregation may introduce researcher bias [79].

Regarding consensus or agreement among panelists, Meijering *et. al* (2013) [19] concluded that many Delphi studies failed in offering a good interpretation of these aspects. They declared that with the same data, different indices suggest different levels of agreement and agreement between rounds [19]. Through a simulation, they showed that it is impossible to understand which measure is the most suitable for measuring consensus or the level of agreement in Delphi studies so they advise researchers to transparently describe the indexes they will use to evaluate these attributes [19].

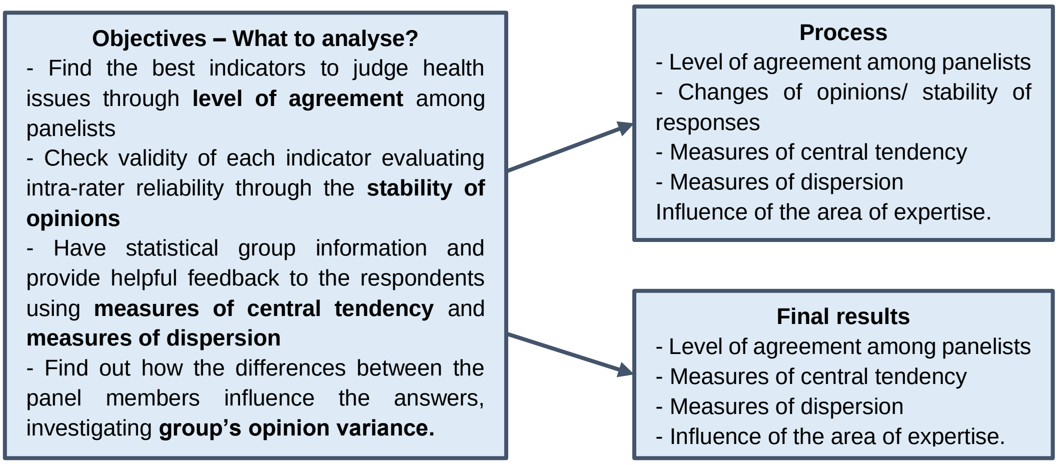
Taking this information into consideration, in this thesis it was decided to implement statistical methods on the DSS, as they are defended to be the ones that offer more information about important features [79]. Measures of central tendency and dispersion are useful to give feedback to the participants and also to provide information about consensus, which is the information used in all Delphi studies to make evaluations [79]. For that reason, these types of statistical methods were considered crucial to implement in the tool instead of complex ones, hard to interpret that can lead to misinterpretations. Consensus and level of agreement will be measured with some statistical approaches that will be chosen and described later since no methods were found to be better than others [19]. Qualitative feedback is not going to be implemented inasmuch it is not as important as numerical data, especially for the types of Delphi under analysis, and because it may introduce bias [79].

The DSS tool could potentially have all the statistical tools described in the literature review but not all statistical tools are adequate for all Delphi processes. Furthermore, as the *DelphiAnalysis* DSS is expected to be one first prototype to help evaluators performing their job, one aimed at considering the most useful statistical tools for the three types of Delphi processes that are considered. According to Belton *et. al* (2019), these methods are the measures of central tendency and dispersion and visual summaries of information [79]. We now discuss which statistical tools are useful and will be programmed for each of the three Delphi processes above described.

1. Delphi for selection of indicators using a Likert-Scale

A Delphi process that aims to select indicators using a Likert-scale is a ranking-type Delphi. As shown in Table 4, the main goal of this type of Delphi is to reach agreement between the panelists about the importance of a set of key issues, ranking them, this is, finding the best indicators to characterize or analyse a specific topic [55]. Additionally, the level of importance is, in this case, measured through the Likert-scale that is an ordinal scale. Ordinal data is treated with **non-parametric methods** [53], what needs to be taken into account in the choice of the techniques to implement in the DSS. Usually the indicators to evaluate are pre-selected, so no qualitative measures are required here.

There are different attributes important to understand in this type of Delphi process, concerning the problem process and concerning final results. Moreover, there are information that is useful not only to the facilitators/ decision analysts but also to the health user [79]. As explained before, the aim is to find the best indicators to judge health issues, which are the ones indicated through consensus and/or agreement among the panelists. Then, the first feature that really matters to the facilitators is the level of agreement between the opinions of the members [79], which is important to analyse during the process in order to know which indicators are seen as very important in the first round and which of them need a re-evaluation and in the final results to see the final level of agreement concerning each indicator. It is also necessary to provide feedback to the decision-makers about how much consensus exist for each indicator [79]. Since the indicators will be used in real healthcare context, the facilitator must evaluate whether an indicator has appropriate characteristics for the concept being assessed, so validity needs to be used as a selection criteria [1]. The validity is related with the consensus criteria: stricter criteria will give the results greater validity but will measure consensus harder [79]. In this light, researchers need to ensure that the chosen approach provides a level of confidence in the outcome that is suited to the needs of the research topic [79]. The DSS should report information about the level of agreement or consensus and validity but the consensus criteria depend on the choice of the evaluator. It is important to analyse the intra-rater reliability, i.e., the consistency or changes in ratings given by the same person across multiple rounds [79] and it is crucial to analyse this feature during the process since it allows to see the direction of change of the answers. Measures of central tendency and measures of dispersion are useful at both the process and final levels, as they give advantageous information to the facilitator about statistical group responses and they allow to provide helpful feedback to the respondents in the following rounds of the Delphi process [79]. Furthermore, they allow to infer about the variety of responses and the level of consensus between the answers [79]. Additionally, at the end of the process, can be practical to the facilitator to explore whether the different areas of expertise or different panel groups somehow influenced the responses, since information given from different geographical areas can influence the responses [68]. More specifically, if the Delphi process is constituted by only one panel, it can be important to evaluate whether the area of expertise, geographical locations or other characteristics influenced the answers provided.



Objectives – What to analyse?
- Find the best indicators to judge health issues through **level of agreement** among panelists
- Check validity of each indicator evaluating intra-rater reliability through the **stability of opinions**
- Have statistical group information and provide helpful feedback to the respondents using **measures of central tendency** and **measures of dispersion**
- Find out how the differences between the panel members influence the answers, investigating **group's opinion variance.**

Process
- Level of agreement among panelists
- Changes of opinions/ stability of responses
- Measures of central tendency
- Measures of dispersion
Influence of the area of expertise.

Final results
- Level of agreement among panelists
- Measures of central tendency
- Measures of dispersion
- Influence of the area of expertise.

Figure 7 - *Features to analyse according to the objectives of the Delphi process at the process level and final results (Delphi for selection of indicators using a Likert-scale).*

Considering Table 4, it is easy to identify which are the non-parametric methods. After the evaluation of each non-parametric measure provided in the section 4.1.3., it was possible to identify the purpose of each approach, to measure the level of agreement, the stability of responses, the central tendency and the dispersion.

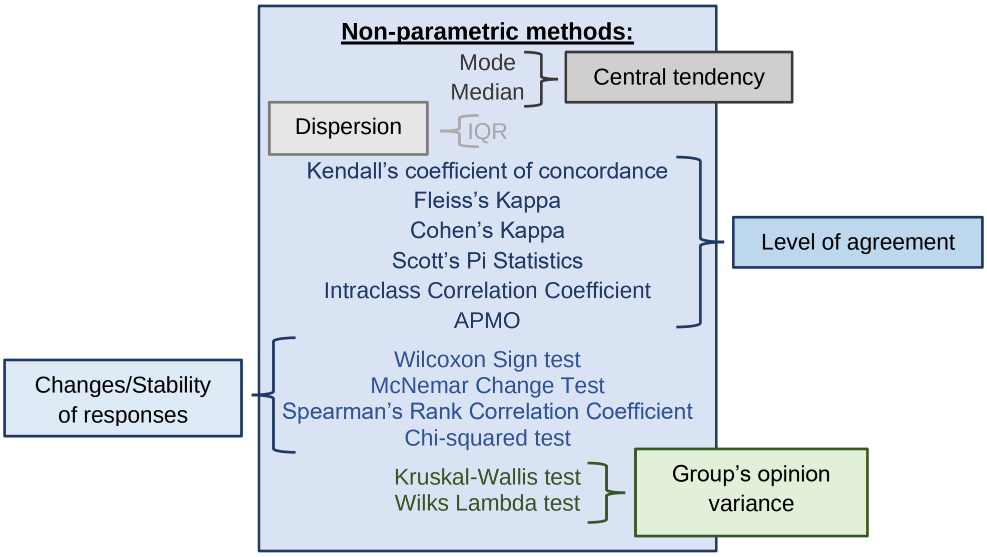


Figure 8 - *Scheme with the non-parametric techniques and correspondent feature that measures (Delphi for selection of indicators using a Likert-scale).*

Mode, median and mean are the most common measures of central tendency. The mean is a measure that only works with interval or ratio data, which is not the case of a Likert-scale (ordinal data) [52]. According to Nunnally and Bernstein (1994) [52], central tendency may be described in terms of median or mode as they will change predictably with permissible transformations. Median is useful with ranked data and the mode is useful as it is often used as a measure of consensus so, it was decided to implement both. Measures of dispersion are convenient to describe data, this is, to show the extent of variability, how much distribution is stretched or squeezed [81]. Interquartile Range, dispersion, variance and standard deviation are some examples of measures of dispersion. Usually, each measure of central tendency relates with one of dispersion. The measure of dispersion that relates with the mean is the SD so, as the mean was discarded, so the SD. The median combines with the interquartile range, which was then selected to be implemented [19]. Moreover, the IQR can be used as a particular level of statistical dispersion to evaluate consensus [79], making it even more valuable.

To measure the level of agreement or disagreement between the raters (inter-rater reliability), two methods were selected to be implemented: the *Kendall's Coefficient of Concordance* since it is really useful using levels of concordance (e.g. Likert scale) and it assesses the strength of consensus between raters, and the *Fleiss' Kappa* because it takes into account the agreement obtained by chance and it can be used with 3 or more raters [19], [79]. *Scott's Pi* and *Cohen's Kappa* also take into account pure chances. The first one was not selected just because it has the same purpose as *Fleiss' Kappa* and the latter because it only uses two raters.

Relatively to the stability of responses between consecutive rounds (change of opinions), the first basic measure that will be implemented is the difference between the rakings given by each participant to each indicator between consecutive rounds, to see if they vary their opinion drastically or moderately. It is also important to evaluate if the differences of opinions are provided always by the same participants or by different ones. Regarding other statistical methods usually used in Delphi processes: the *Chi-Squared* test was rejected based on the disapproval that was found in recent literature; the *Wilcoxon Sign Test* is for continuous data and not efficient if there are a lot of tied ranks, which may be the case as we are dealing with a 5-categorie scale and the *McNemar Change* test cannot be used since it has a dichotomous trait, with matched pairs of subjects [61]. In this light, it will be implemented the *Spearman's Rank Correlation Coefficient* to complement the analysis of stability, which is very appropriate to use with ranked data and to inspect the changes of opinion between respondents [59]. This method will be used to compare the answers of all the participants for each indicator, between two consecutive rounds to understand if there is a high stability between responses or not.

To conclude, the group's opinion variance needs to be evaluated through a MANOVA to examine whether the responses given by the panel were statistically different across groups (e.g. experts vs stakeholder or fields of expertise) [68]. More specifically, the *Wilks Lambda* test will be implemented on the tool. Figure 9 shows the tools selected to be implemented in the *DelphiAnalysis* DSS file correspondent to the Delphi for selecting indicators.

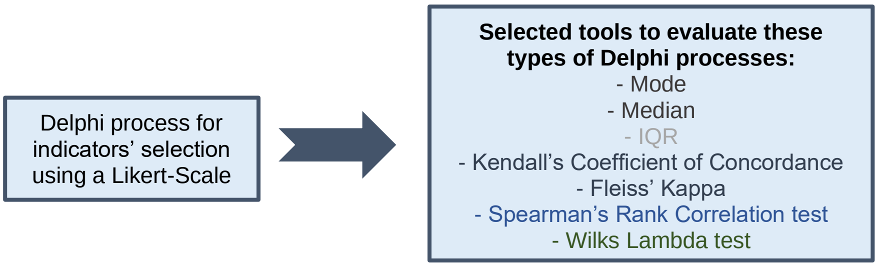

Delphi process for indicators' selection using a Likert-Scale
Selected tools to evaluate these types of Delphi processes:
- Mode
- Median
- IQR
- Kendall's Coefficient of Concordance
- Fleiss' Kappa
- Spearman's Rank Correlation test
- Wilks Lambda test

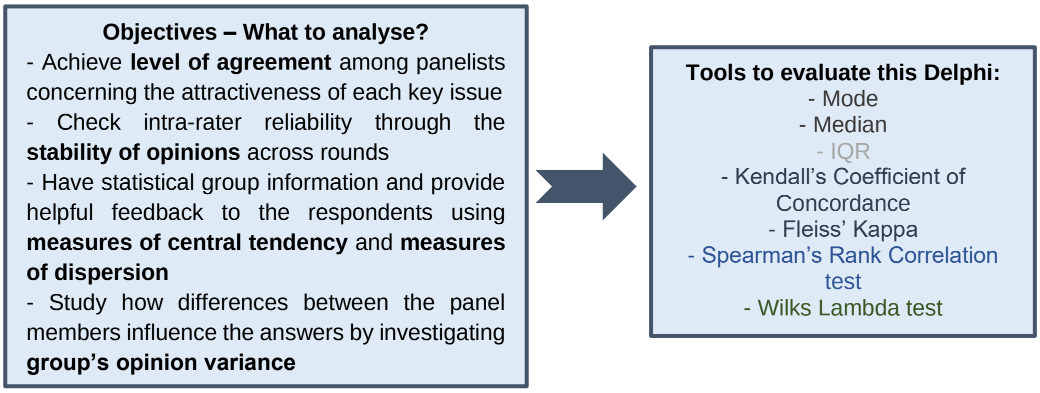
Figure 9 - *Final techniques chosen to be implemented in the DSS (Delphi for selection of indicators using a Likert-scale).*

2. Delphi for qualitative weighting judgments using the MACBETH method

A Delphi process that aims to collect qualitative weighting judgements on the indicators based on the scale of the MACBETH model is also a ranking-type Delphi. The main goal of these types of processes is to understand the difference of attractiveness between two actions at a time in a qualitative scale [75]. The raters need to give a qualitative judgement indicating their opinion about the level of attractiveness about each key issue, usually according to a seven-level MACBETH scale [75], which is an ordinal scale. As in the previous case, the evaluation needs to be done with **non-parametric techniques**. Also, the key issues are usually provided, not collected through an open ended first round so, no qualitative analysis is required.

Performing the same analysis as in the previous case, it was defined that the features to analyse relatively to the process and the final results are the same according to the objectives of this type of Delphi. The main goal of this Delphi is to identify the weight of importance about each health indicator, so, once again, it is important to understand the level of agreement between the participants [79] and to see the validity of results according to the chosen consensus criteria [1], both during the process and final results and use this information to provide helpful feedback to the panelists [79]. It is also important to evaluate the intra-rater reliability across multiple rounds through the difference of rankings given by each participant throughout the process. Measures of central tendency and measures of dispersion will be implemented to analyse the process and final results giving information about statistical group responses and providing feedback to the respondents in the following rounds of the Delphi process [79]. Finally, at the end of the process, can be practical to the facilitator to explore whether the differences between panel groups somehow influenced the responses [68].

In this case, the choice of the statistical measures to implement in the DSS is natural since the objectives and characteristics of this type of Delphi are the same as in the previous one. Therefore, the tools to be implemented in the DSS file correspondent to the Delphi for weighting judgments is shown in the next figure.



Objectives – What to analyse?
- Achieve **level of agreement** among panelists concerning the attractiveness of each key issue
- Check intra-rater reliability through the **stability of opinions** across rounds
- Have statistical group information and provide helpful feedback to the respondents using **measures of central tendency** and **measures of dispersion**
- Study how differences between the panel members influence the answers by investigating **group's opinion variance**

Tools to evaluate this Delphi:
- Mode
- Median
- IQR
- Kendall's Coefficient of Concordance
- Fleiss' Kappa
- Spearman's Rank Correlation test
- Wilks Lambda test

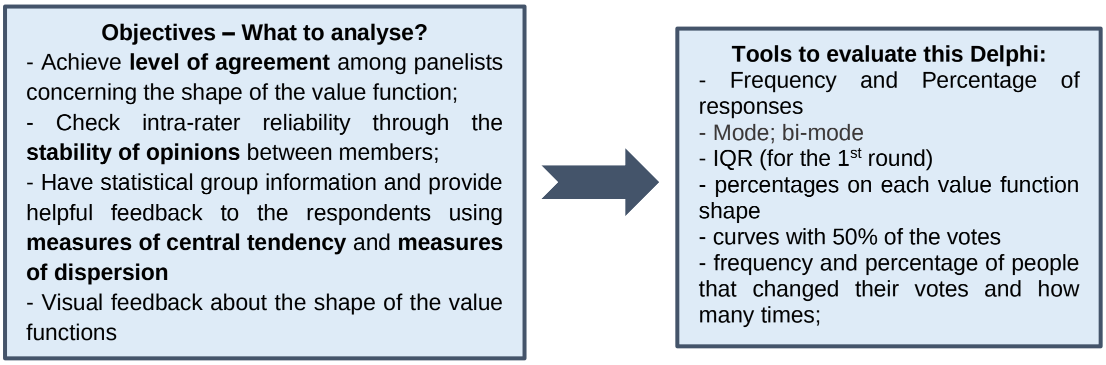
Figure 10 - *Objectives, features to analyse and final tools chosen to be implemented in the DSS (Delphi for qualitative weighting judgments using MACBETH).*

3. Delphi for shaping the value functions for each indicator/key issue using MACBETH

The main goal of this type of Delphi is to determine the shape of the value function that fits for each key issue, this is, the shape that characterizes the opinions of the participants [77]. To do so, it is necessary to know the range of performance of each key issue. The range of performance need to be divided in 'jumps' (changes of performance) and every respondent must vote the level of attractiveness of each one of the 'jumps', using the correspondent scale. The answers will give a value-function, which describes the importance of different changes within the range of performance [77]. The objective of the first round is to determine the shape of the value function according to the votes of the panelists. For the remaining rounds, the objective is to understand if the members agree with the previous round shape in the light of the results obtained by the all group or if they want to change it [77]. The percentage of each shape of value functions needs to be given as a feedback element to the decision makers. Once again, the scale used is an ordinal one so we will implement **non-parametric methods**.

According to the objectives of this type of Delphi, only simple statistical measures can be applied since the objective is to analyse shape of value functions and not numerical values. The number of votes on each categorie will be counted and converted into percentages to be easier to see the agreement between the answers of the participants. Like in the other cases, the non-parametric methods as the mode and interquartile will be used to infer about the central tendency, dispersion and consensus both at the process level and final results. In the first round, a shape will be attributed to each participant's answers and then, the percentages of each shape of value function for each indicator will be calculated. This results will be organized in a tabular fashion so that they can be given as feedback to the second round. It can be important to calculate which are the most voted types of shape and if some of them have more than 50% of votes. On the remaining rounds, the participants will not vote on the "jumps" according to the MACBETH scale but they will vote on the type of shape of value function (linear, concave, convex, s-seat or s-sigmoid) that best fits according to the importance of the range of each indicator. Once again, the only suitable methods here is the mode, the interquartile range and, across the rounds, it is also important to understand which members changed their votes, if are always the same participants changing their opinion and if the changes are drastic or moderate. This kind of information is always crucial in every Delphi since they give many important information to the evaluator and the DM as feedback being the only suitable methods when evaluating functions. This kind of Delphi can't be treated with numerical data since the participants will vote on a shape and not on categories of an ordinal scale. In this sense, the analysis of this Delphi will be slightly different from the other two.

The IQR can only be applied in the first round, in which the participants are voting on the importance of the ranges, according to the MACBETH scale but the mode and bi-mode will be applied for all rounds. From the first round to the second one, the answers of the participants will be translated in the respective shapes that they represent. Between the rounds, analysis of intra-reliability will be added as in the previous cases. Figure 12 summarizes the tools that will be implemented in the DSS file correspondent to the Delphi for obtaining value functions.



Objectives – What to analyse?

- Achieve **level of agreement** among panelists concerning the shape of the value function;
- Check intra-rater reliability through the **stability of opinions** between members;
- Have statistical group information and provide helpful feedback to the respondents using **measures of central tendency** and **measures of dispersion**
- Visual feedback about the shape of the value functions

Tools to evaluate this Delphi:

- Frequency and Percentage of responses
- Mode; bi-mode
- IQR (for the 1st round)
- percentages on each value function shape
- curves with 50% of the votes
- frequency and percentage of people that changed their votes and how many times;

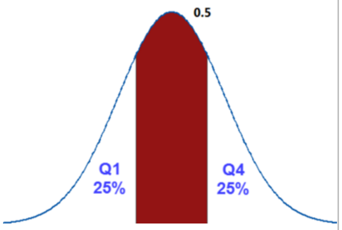
Figure 11 - *Objectives, features to analyse and final tools chosen to be implemented in the DSS (Delphi for shaping the value function using MACBETH).*

For all the three Delphis presented, a page with a summarization of the information obtained in the results will be provided to help the researcher to understand how the percentages, level of agreement, stability of responses evolved during throughout the rounds of the process.

4.5.1. How to implement the selected statistical measures

After the selection of the methods to implement in the DSS it is necessary to understand how they work to be able to implement them correctly in the tool. Despite the brief explanation of all the statistical measures, presented in section 4.1.3, it is necessary to describe explain how to implement the chosen approaches, which formulas will be necessary to carry out the tool.

Table 7 - *How to apply the selected methods in the tool.*

IQR	The IQR is given by: $\boldsymbol{IQR} = \boldsymbol{Q_3} - \boldsymbol{Q_1}$ and it can be obtained following the steps above: 1. Find the median; 2. Form two groups: parameters above and below the median (if there is an even set of data, include each of these values in the each of the groups, according the order); 3. Q_1 is a median in the lower half; Q_3 is the median of the upper half; 4. Apply the formula indicated above [82].  ***Figure 12 -*** *IQR (red zone).*
Kendall's Coefficient of Concordance	If *i* represents each object and *j* each rater, then S_{ij} is the rating given by the judge *j* about the object *i*. For each indicator, $\boldsymbol{S_i} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{m}} \boldsymbol{S_{ij}}$ is the sum of every votes given by the raters. If $\bar{S}$ is the mean of S_i, then the squared deviation is $\boldsymbol{S} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{S_i} - \bar{\boldsymbol{S}})^{\boldsymbol{2}}$ [83]. The Coefficient is given by $\boldsymbol{W} = \frac{\boldsymbol{12S}}{\boldsymbol{m^2(n^3-n)-mT}}$, where T is a correction factor ($\boldsymbol{T} = \sum_{\boldsymbol{k=1}}^{\boldsymbol{g}} (\boldsymbol{t_k}^{\boldsymbol{3}} - \boldsymbol{t_k})$) used in the cases that the ranks are tied [84].
Fleiss' Kappa	This values is given by: $\boldsymbol{K} = \frac{\bar{\boldsymbol{P}} - \overline{\boldsymbol{P_e}}}{\boldsymbol{1-}\overline{\boldsymbol{P_e}}}$, where $(1 - \bar{P}_e)$ gives the degree of agreement that is attainable above chance and $(\bar{P} - \bar{P}_e)$ gives the degree of agreement achieved above chance [85]. Let N represent the total number of subjects, n the number of ratings per subject and k the number of categories. If $i = 1, \dots, N$ represents the subjects and $j = 1, \dots, k$, represents the categories of the scale, then n_{ij} represents the number of raters who assigned the i^{th} subject to the j^{th} category, and so $\boldsymbol{p_j} = \frac{\boldsymbol{1}}{\boldsymbol{Nn}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{n_{ij}}$ is the proportion of all assignments that were to the j^{th} category [85]. The number of rater-rater pairs that are in agreement is given by $\boldsymbol{P_i} = \frac{\boldsymbol{1}}{\boldsymbol{n(n-1)}} [(\sum_{\boldsymbol{j=1}}^{\boldsymbol{k}} \boldsymbol{n_{ij}}^{\boldsymbol{2}}) - \boldsymbol{n}]$ and its mean by $\bar{\boldsymbol{P}} = \frac{\boldsymbol{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{P_i}$. Additionally, $\bar{P}_e$ is given by $\overline{\boldsymbol{P_e}} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{k}} \boldsymbol{p_j^2}$.
Spearman's Correlation Test	This coefficient is given by: $\boldsymbol{r_s} = \boldsymbol{1} - \frac{\boldsymbol{6 \sum d_i^2}}{\boldsymbol{n(n^2-1)}}$ where d_i is the difference between the ranks of the responses on the i^{th} item of the Delphi and n is the number of experts [86]. However, it is also possible to compute the average S*pearman Correlation Coefficient* of all possible pairs of judges through the formula: $\bar{\boldsymbol{r}}_{\boldsymbol{s}} = \frac{\boldsymbol{mW-1}}{\boldsymbol{m-1}}$ where m is the total number of raters and *W* is the Kendall's Correlation Coefficient[87].
Wilks Lambda	There is *a* ***RealStatistics Resource*** pack available for the Excel that incorporates the *Wilks Lambda* function and a function to calculate its associated p-value.

4.6. Design: Adapted Framework of the DSS

In this chapter, the design and architecture of a DSS will be presented in order to understand how to correctly implement one. The steps and checkpoints for each stage of the DSS tool will be described following a design proposed by Miah *et al.* [48].

4.6.1. Design and Architecture of a DSS

A DSS is a computer-based system that combines data from various sources and formats and supports choice by assisting the decision maker in the organization of information and modeling outcomes according to a specific model. Besides, it presents the user with advantageous information and display graphical interfaces [88], [89]. Computerized DSSs presents a lot of advantages: performing large number of computations in a short time, complex information and/or relationships may be searched, processed and transmitted quickly, allowing to evaluate more alternatives that humans could do by themselves and also avoiding human errors [90]. The DSS architecture typically contains user interface, knowledge acquisition interface, knowledge base and inference engine, and the framework is composed by three main subsystems shown in Figure 14 - the dialogue, the input management and the knowledge management subsystems [91]. The dialogue subsystem serves to integrate other subsystems and assures user-friendly communications between the decision-maker and the DSS [91]. The input management subsystem organizes and manages all the inputs whichever the type and the quantity of data inputs. The knowledge management subsystem retains all the multi-criteria analysis methods available in the DSS [91].

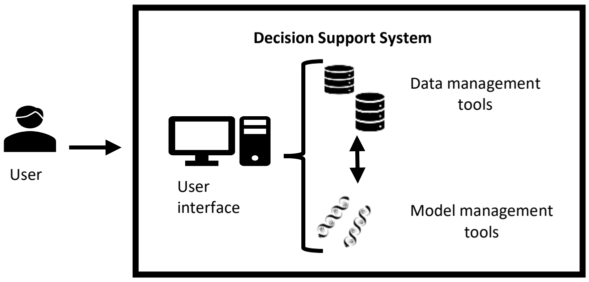


Figure 13 - *Functional components of a DSS (based on [91]).*

A proper DSS design requires the identification of the nature of the target decision problem and a well-known strategy of the best way to support the decision process [48]. To illustrate how it's possible to evaluate DSS development within a socio-technical context, is presented a workflow based in the one proposed by Miah *et al*. [48] that consists in six stages containing checkpoints that must be taken into consideration when developing the DSS, as shown in Figure 14.

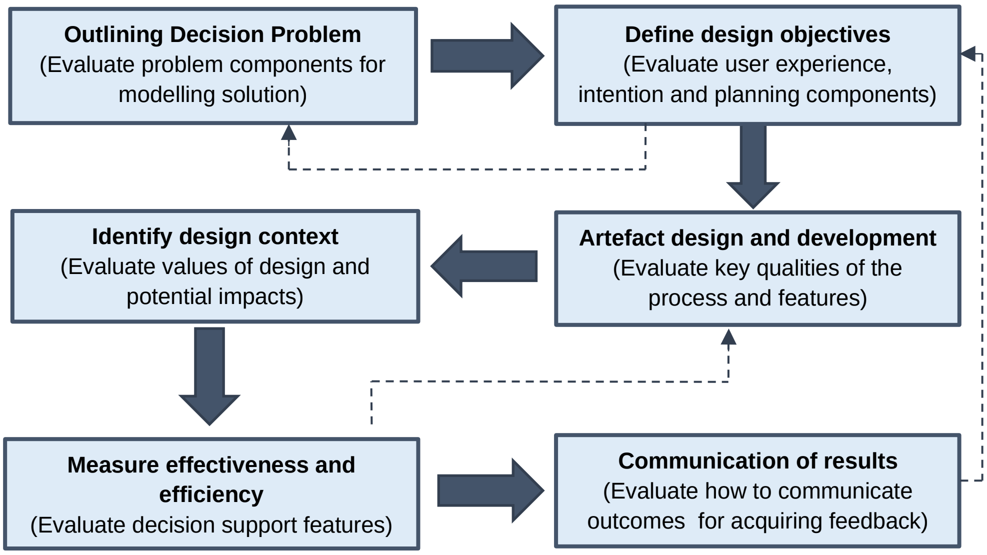

Outlining Decision Problem
(Evaluate problem components for modelling solution)
Define design objectives
(Evaluate user experience, intention and planning components)
Identify design context
(Evaluate values of design and potential impacts)
Artefact design and development
(Evaluate key qualities of the process and features)
Measure effectiveness and efficiency
(Evaluate decision support features)
Communication of results
(Evaluate how to communicate outcomes for acquiring feedback)

Figure 14 *- Workflow for DSS design (based on [48]).*

The first stage dwell in the definition of the decision problem, where it is necessary to explicit the decisions to be made, the agents involved, the importance of the problem and its complexity [89]. The information provided in the first phase will allow designing the objectives of the tool - determine the inputs and outputs of each stage, the resources needed and which measures to use. The third stage is to identify the main problems and to define the approaches that will be used to solve those problems. The next stage consists in determining the context in which the tool will be used and tested. In the fifth stage, it is defined the appropriate ways to evaluate and analyse the outputs, concluding about the performance of the DSS. Finally, it is crucial to determine which inputs will be asked, how the outputs will be presented and to determine whether the outputs match discipline knowledge, as shown in Figure 15.

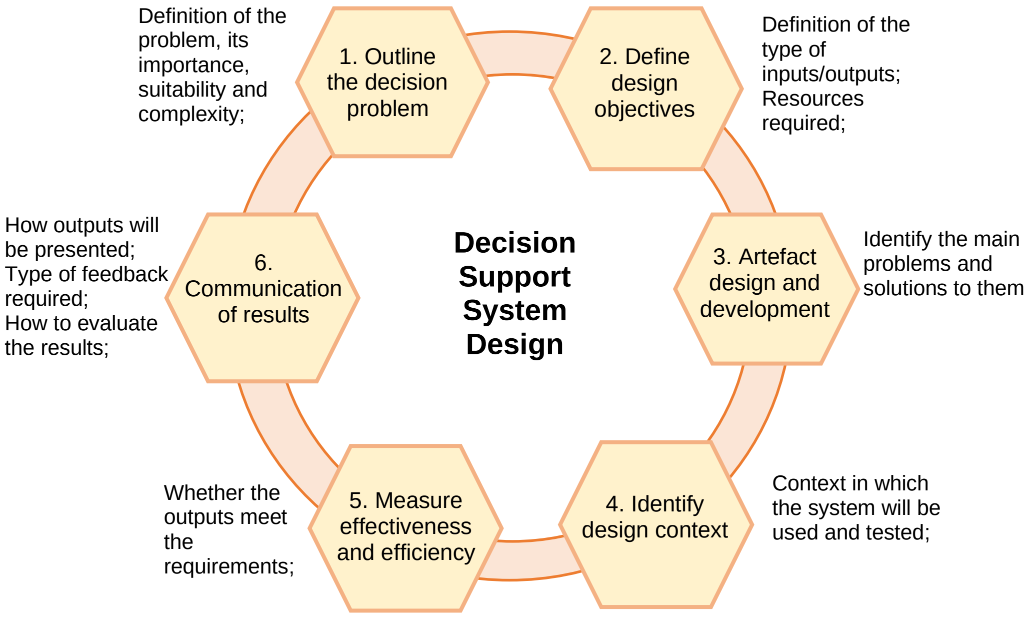

Definition of the problem, its importance, suitability and complexity;
1. Outline the decision problem
2. Define design objectives
Definition of the type of inputs/outputs; Resources required;
How outputs will be presented; Type of feedback required; How to evaluate the results;
6. Communication of results
Decision Support System Design
3. Artefact design and development
Identify the main problems and solutions to them
Whether the outputs meet the requirements;
5. Measure effectiveness and efficiency
4. Identify design context
Context in which the system will be used and tested;

Figure 15 - *Framework design of a DSS (based on [57]).*

1. Outline the Decision Problem

a) Problem importance

Delphi processes have been widely used in the healthcare field to help decision-making processes. It is based in the assumption that opinions coming from a group of experts are more valuable than a single opinion and that it can count of opinions from all over the world since no physical meetings are required [8], [9]. These processes need to be evaluated deeply when concerning to healthcare considering that can influence doctors and/or patients' life [16]. In this light, it is really useful to have a tool that can perform automatically the statistical measures that allow to see the reliability of the matter being studied, increasing the accuracy of the results and decreasing the time dispended in this evaluation.

b) Problem suitability for decision-makers

The Delphi process is a structured communication technique which relies on the opinions of a panel of experts to gather important information that will lead to useful forecasts or decisions [14]. However, several studies have pointed out that, despite its usefulness, it is a time-consuming method as it is constituted by questionnaires that need to be analysed at each round of the process [26]. An intuitive, user-friendly DSS to operationalize automatically all the calculations needed to evaluate both the process itself and the final results would be a huge step to rescue some time to the analyst. The *DelphiAnalysis* DSS aims to help the decision analysts and/or health users that have some knowledge in Delphi processes to analyse the main features of these processes quickly and without mistakes. The tool should be easy enough so any evaluator can understand how to use it and the key conditions behind the model. To assure that the tool is correctly used and interpreted, a webpage to guide the user must exist. Usually Delphi processes are not large so the DSS will permit a maximum of 200 indicators and 350 participants to be part of the study.

c) Problem complexity/simplicity

The model to be developed only intends to facilitate the calculations that need to be made by the health user to be able to evaluate the results of Delphi processes, to make them flawless and reducing the time dispended to perform them by his own. Since there are many types of Delphi processes that require different analysis according to their objectives, it is extremely difficult to create a tool that permits all possible evaluations. The main idea is then, the elaboration of a tool able to help the evaluator in three types of Delphi commonly used in healthcare. Therefore, it is not expected that this approach will be used to help any type of problems but rather to be helpful for usual real and specific Delphi studies performed in the health field. Statistical measures will be implemented in Excel, as it is intended to be a user-friendly prototype ensuring ease of use and some flexibility. The tool will provide the results of the statistical calculations and may incorporate some notes about the outcomes, but the objective of the tool is not teaching how to interpret the results.

2. Define Design Objectives

a) Whether quantitative or qualitative inputs/outputs are to be used

The inputs must be qualitative judgements provided by the participants as Likert or MACBETH semantic categories while the outputs will be mostly quantitative as the categories are transformed in numerals to use statistical procedures. As previously explained, the evaluator must have knowledge on Delphi processes and in how to interpret the results obtained in the DSS.

b) Appropriateness of objectives

The proposed DSS aims to enable automatic calculations on the above described statistical operators and tests only with the answers provided by the Delphi respondents. These calculations can be done for each round or all of them at the same time. If all the rounds are filled, the tool should also offer an analysis about what happens between the rounds. The results can be used to perform a quick evaluation of the responses during the process or about the final issues and it can be easily used to extract the necessary information to provide a helpful feedback to the panel members between successive rounds. Although the analysis of the results needs to be performed by the DSS user, it is no longer his job making intermediate calculations by hand or worrying about possible miscalculations. Besides, the time spent to analyse an overall Delphi process will be greatly reduced.

c) Resources required by design objectives

In a preliminary phase, the user needs to define the type of Delphi he is going to evaluate (between the three incorporated in the DSS). The DSS should receive input data presented in an Excel file format. Otherwise, the analyst needs to organize the information and put it on the right places in the input tables presented in the tool. However, the DSS can adapt the semantic words to describe the categories of the scale to the ones it can automatically treat. Additionally, a web page needs to be created to give some guidance to the user. To do so, the Wix platform that allows to build professional webpages will be used. The platform chosen can incorporate text, images, videos, and different sections necessary to be intuitive.

3. Artefact design and development

a) Design and development approach used

The tool will be developed in Microsoft Excel, which is used to create spreadsheets and includes an intuitive interface, powerful graphing and calculations tools. [92]. In this light, it seems to be an adequate platform to implement the calculations needed in the DSS, to organize all the information necessary to provide feedback to the participants and to generate more accessible results, facilitating their interpretation. For a prototype, it can be considered a good DSS because of its ease of use, without the necessity of learning additional concepts. Some mathematical programming was developed within the Microsoft Excel, making use of the existent functions and of the *Data Analysis Toolpack* which had to be downloaded. This pack incorporates predefined operations that can be used to perform advanced statistical methods [92]. Finally, the web page will be created using the online Wix platform, which will provide a helpful guide to the user. The Wix platform enables the edition of web pages to create a new one. The webpage will incorporate several sections describing Delphi processes, my dissertation work and how *DelphiAnalysis* emerged. Additionally, it will explain how to use the *DelphiAnalysis* DSS - how to transform the data to be suitable to the tool and how to calculate the *Wilks Lambda* test and the correspondent p-value.

b) Innovative aspects of the DSS

Delphi processes are widely used in health issues and some platforms have been developed to help Delphi achievement. However, although there are many techniques described in literature that have been used to analyse Delphi processes' characteristics, there are no platforms or DSS tools, to the best of knowledge, that calculate and summarize them automatically without the need of manual work by the Delphi user. This DSS will be useful within the healthcare field because it will incorporate an automatic analysis for the three types of Delphi above described that are widely used in that field. Although its limitations, it is a great asset to Delphi users since it allows to obtain the most common analysis of Delphi features without requiring manual work by the researchers.

4. Identify Design Context

The system must be enough to be used in every context that encompasses a selection of indicators using a Likert scale, to weight judgments using a MACBETH scale or to shape a value function according to an indicator performance range. It can be used by any kind of Delphi evaluator, either for professional or for academic purposes, as long as the user understands the process, what it requires and how to analyse it. The DSS will be tested using health indicators used in a real healthcare project and some of the results will be compared with the ones obtained in this project. Additionally, a questionnaire will be developed and presented to the Delphi experts of the Decision Eyes company to understand their opinion about the tool to validate it and to find what could be improved in future works.

5. Measure the outputs

It is expected that the DSS returns the results of all the calculations of the statistical procedures implemented, without errors. Each formula must be coded in a specific way to able to evaluate the exact data format received by the tool and provide the analysis results for each round and between consecutive rounds. All the calculations must be automatically performed as the user uploads the answers of each round in specific tables ready to receive that information.

6. Communication of Results

a) Determine how the outcomes are to be presented

Once the tool will be implemented in Excel, the outcomes will be given in an Excel spreadsheet. For each type of Delphi, it will exist a worksheet to provide the answers of the participants (input data), which will be converted into numerical data to be possible to apply statistical measures. A new worksheet will present all the outcomes for each round and between the rounds. Additionally, a “summary page” spreadsheet will present the results obtained easily to understand what happen during all the Delphi process. An extra sheet will exist to make some supplementary calculations required to implement the methods that doesn’t provide any useful information for the users. The spreadsheets that will contain the results will be very easy to interpret since the results will be presented in tables, with auxiliary titles and some small notes.

b) **Determine whether the communication structure is appropriate for the target audience**

The communication structure allows the evaluator to see the results presented in tables or right after the name of the procedure being applied and it presents additional informative notes to help and guide the user. The target audience will be the evaluator itself that needs to understand the results to make an appropriate analysis of the Delphi process. Also, the results can be used to provide feedback to the participants, so they need to understand the results as well. However, the user does not need to evaluate the implemented code or access the extra calculations made. The results are shown in an isolated spreadsheet in a straightforward manner, which facilitates their understanding.

4.7. Implementation of the *DelphiAnalysis* DSS

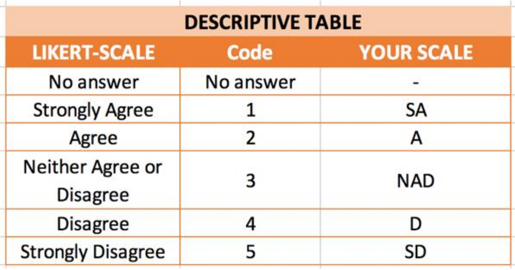
Having already the background about how to create a DSS and which the best techniques to implement in the *DelphiAnalysis* DSS , it is possible to start the development of the model. The development of this DSS consists in the implementation of statistical measures which allow to automatically analyse the features of the three above described Delphi processes only using the answers provided by the Delphi participants. As previously mentioned, the main purpose of this DSS is to facilitate the Delphi user by presenting him effective results of useful measures that allow to analyse the most common features of a Delphi process, saving time. It is assumed that who will use the tool knows what a Delphi process is and has enough knowledge to understand the results provided by the tool. However, some notes are provided in the DSS. Afterward, it will be possible to test if the DSS works correctly by testing it with the results of a real context project and validate it through a questionnaire answered by Delphi processes’ experts, to understand what is their opinion about the *DelphiAnalysis* DSS, its usefulness and what could be improved in the future.

4.7.1. Global organization of the DSS

For each type of Delphi process, a different Excel file will be created, and each file will have sections (Excel tabs) to execute different functions. The first tab is where the user can do the upload of the input data in an organized way and where he defines how the categories of the ordinal scale are mentioned in his work. The second tab is where the tool automatically converts the input data to a numerical format, that the tool can read and treat. The results are presented in the third tab. It starts presenting the results for rounds 1, 2 and 3 and then the results between rounds 1 and 2 and rounds 2 and 3. The fourth separator presents a “summary page” which organizes some of the information obtained in the results in a more intuitive way to the user and to present some conclusions about what is happening with the Delphi process’ features during the process. Another tab called “MANOVA” exists in the *DelphiAnalysis* files for the Delphi for selecting indicators and the Delphi for weighting judgments and is intended to display the results of a *Wilks Lambda* test and its respective p-value. However, the results of the MANOVA cannot be obtained automatically since it depends on the range of input data. Therefore, some guidance is provided to the Delphi user to help him to perform these tests. This tab does not exist for the Delphi for shaping value functions because the answers provided by the participants are not from an ordinal scale, being impossible to convert them in numerical data. The last tab is an extra page that was created to perform additional calculations required to implement the statistical measures. In this section and in Appendix A, the mathematical programming and the visual part of the DSS will be displayed presented.

4.7.2. Input Data

The first step was to define how the data would be uploaded in the tool. As the input data of these types of Delphi are the panelists’ votes according to ordinal scales or value function shapes, the easier way to organize them is using tables. Therefore, the first constraint applied to the DSS was a limitation on the number of participants and indicators under analysis - 350 and 200 respectively. Three tables (for rounds 1, 2 and 3) were then created with 200 columns correspondent and 350 lines to upload the responses of the Delphi participants. Once the tables were created it was necessary to make the tool able to treat the data provided by the user, whatever the type of data of the responses - numbers, words or acronyms -, so that it would not be necessary to change it manually. For that reason, a descriptive table (see Figure 16) was created in the top of the first Excel tab to convert the received data to the one that is read by the DSS. For the value function Delphi type, it was necessary to create two descriptive tables since the answers provided in the first round were based on a MACBETH semantic scale and in the following rounds, the answers were on the type of shape. Another characteristic implemented in was that if the user decides to substitute the indicators line by a description or identify the respondents with another ID different from the ones provided by default, he can do it on the first input table and the rest of the tables will automatically adapt to the same labels. It is important to notice that the ID of the participant must be the same during all the process and that the indicators must be distributed in the same order in all the three input tables, otherwise, the DSS will not be able to compare the correct responses. From round to round, it is necessary to have all the participants and indicators on the list and they need to be distributed in the same order in the three rounds. If a panelist did not answer anything in the first round, instead of being discarded from the list, it continues with “no answer” in all the following rounds. The indicators that may be discarded or accepted by the majority can be left with empty cells, but they need to occupy their position on the list.



DESCRIPTIVE TABLE		
LIKERT-SCALE	Code	YOUR SCALE
No answer	No answer	-
Strongly Agree	1	SA
Agree	2	A
Neither Agree or Disagree	3	NAD
Disagree	4	D
Strongly Disagree	5	SD

Figure 16 - *Descriptive Table (Delphi for selection of indicators). The right side of the table is where the user can introduce the categories used in the Delphi process. The categories shown in the left side of the table are the ones recognized by the DSS.*

In the value function Delphi, for the first round all works the same way - the participants need to vote in the three performance sub-ranges of each indicator according to a semantic ordinal scale. However, the user needs to fill an extra table with the sub-ranges/gaps of the performance range. Moreover, this sub-ranges need to be filled in a very specific structure - the first number followed by a space, then a hyphen, space, and the second number.

For the Delphi for selection of indicators and weighting judgments, it may be useful to perform a MANOVA to study the variance of opinion between the participants. In this light, in the first tab of these types of Delphi, there are another three tables below the ones described above, to insert the information about the participants that the user wants to compare using a MANOVA. First, he needs to have the respondents divided by the groups he wants to compare and distinguish them with numbers on the second column of the table (e.g. geographers – GROUP 1; economists – GROUP 2). After, he only needs to insert the answers on the table.

The first tab is designed to the user as he needs to fill everything in the right way and after, all the results will appear automatically on the third and fourth tab of the Excel files.

4.7.3. Conversion of the Input Data

All the three Excel files have the input tab where it is performed the conversion of the votes provided by the Delphi participants to the data read by the tool. This second tab presents three tables (one for each round), similar to the ones to insert the input data, which will present numerical values, necessary to perform statistical measures. In the case of the value function, for rounds 2 and 3, it only converts the names of the categories used in the Delphi process to the names recognized by the DSS. The mathematical programming needed to convert the input data to numbers was the creation of a succession of "*IF*" functions One example of the programming of a descriptive table is presented below:

$$\boxed{\begin{gathered} = IF(Input1!\,C23 = "";"----"; IF(Input1!\,C23 = Input1!\,\$D\$11;"";\\ IF(Input1!\,C23 = Input1!\,\$D\$12; Input1!\,\$C\$12;\\ IF(Input1!\,C23 = Input1!\,\$D\$13; Input1!\,\$C\$13;\\ IF(Input1!\,C23 = Input1!\,\$D\$14; Input1!\,\$C\$14;\\ IF(Input1!\,C23 = Input1!\,\$D\$15; Input1!\,\$C\$15;\\ IF(Input1!\,C23 = Input1!\,\$D\$16; Input1!\,\$C\$16))))))). \end{gathered}} \qquad (1)$$

$C23$ is the Excel cell where the first answer is presented (respondent 1, indicator 1). If the cell is empty, the sign “----" is returned, if the answer is the one presented in the first line of the right side of the descriptive table (answer provided by the respondent), it will return the first line of the left side of the descriptive table (word used to the describe the response provided and which is recognized by the tool) and so on. This process is repeated for every cell of the table to convert all the responses.

The user does not need to access these data since the analysis is performed automatically in another tab of the Excel, which is an important part to the researcher. Nevertheless, this spreadsheet will remain available to the user so he can use the numerical data in case he wants to perform additional tests to those provided by the tool.

4.7.4. Statistical measures performed by the *DelphiAnalysis* DSS

The statistical measures provided by the tool are the same for both selection of indicators and weighting judgment Delphis, but slightly different for the value function Delphi. Regarding the first two cases, there is a question, that the user needs to answer, at the beginning of the results tab before the analysis of each round - “*How many participants answered?*”. The participants that did not answer to any indicator in each one of the rounds need to be discarded of this count. For example, if there is a total of 50 participants and 2 of them didn’t answer to any indicator, the answer will be 48 participants, which doesn’t mean that for one indicator you don’t have 3 or more empty answers since the participants can decide not to answer to some of them. After this question, the results are presented. For each round, it is shown the frequency of responses, this is, how many times each category of the scale was voted by the respondents, including how many of them did not answer and the correspondent percentages. Some examples of the mathematical programming applicated in the selection of indicator Delphi are shown below.

1. For the frequency of votes “*Strongly Agree*” for indicator 1
   - The answers provided about the indicator 1 from are displayed in column C between the lines 6 and 355 in the table presented in the second tab.
   - $C6$ is the first answer provided.
   - $C12$ represents the “*Strongly Agree*” in the descriptive table.

$$\boxed{= IF(N.Input1!\,C6 = "----";"----";COUNTIF(N.Input1!\,\$C\$6{:}\,\$C\$355;Input1!\,\$C\$12))} \qquad (2)$$

This formula says that if the first answer cell is filled with “----", then the answer cell is empty, and the indicator was not evaluated by the respondent and so the frequency is filled with the same symbol. Otherwise, it counts how many times the “*Strongly Agree*”, $C12$, appears between the cells $C6$ and $C355$. The $ symbol is used to fix the cell, for example, the position of the “*Strongly Agree*” it’s always the same in the descriptive table so it is necessary to fix the cell, $\$C\$12$, in the column (represented by a letter) and line (represented by a number). This was performed for all the indicators and each of the 5-level Likert scale categories.

2. Frequency of people who did not answer for indicator 1

$$\boxed{= IF(N.Input1!\,C6 = "----";"----";COUNTBLANK(N.Input1!\,\$C\$6{:}\,\$C\$355))} \tag{3}$$

This formula is similar to the previous one but instead of counting the times that a category appears, it counts how many empty cells exist between the cells $C6$ and $C355$, which represent the participants who did not answer.

3. Frequency of people who answered for indicator 1
   - The formula deals with the first tab where the tables have the direct answers of the participants. The empty cells are the participants and indicators that do not exist;
   - The answers are between cells $C23$ and $C372$;
   - $D11$ corresponds to the "*Not Answered*" category in the descriptive table.

$$\boxed{\begin{array}{c} = IF(D7 = "----";"----"; \\ 350 - COUNTBLANK(Input1!\,C23{:}\,C372) - COUNTIF(Input1!\,C23{:}\,C372;Input1!\,\$D\$11)) \end{array}} \tag{4}$$

The same line of thought was used in this case, but now from the 350 panelists that may exist, we subtract the empty cells and the ones who did not answer.

4. Percentage of votes of "*Strongly Agree*" for indicator 1
   - $D14$ is the total number of people who answered, this is, the result obtained in the previous formula.

$$\boxed{= IF(D\$7 = "----";"----";D7/(D\$14))} \tag{5}$$

The percentage of each category is given by the number of participants who voted in that category, obtained with the formula presented in 1. divided by the total number of panelists who answered, obtained using formula 3. The tool also presents the percentage of some combinations of categories but to obtain them it was only necessary to sum up the respective percentages.

The next statistical measures implemented are the mode and median, as measures of central tendency and the interquartile range. Some examples for the selection of indicators Delphi are presented below.

1. Mode for indicator 1
   - The fixed values $\$C\$12$ to $\$C\$16$ correspond to each category of the Likert scale;

$$\boxed{\begin{array}{c} = IF(N.Input1!\,C716 = "----";"----"; \\ IF(MAX(IndicSelecD.RESULTS!\,D157{:}\,D161) = IndicSelecD.RESULTS!\,D157;Input1!\,\$C\$12; \end{array}}$$

$$\begin{gathered}IF(MAX(IndicSelecD.RESULTS!D157:D161)=IndicSelecD.RESULTS!D158;Input1!\$C\$13;\\ IF(MAX(IndicSelecD.RESULTS!D157:D161)=IndicSelecD.RESULTS!D159;Input1!\$C\$14;\\ IF(MAX(IndicSelecD.RESULTS!D157:D161)=IndicSelecD.RESULTS!D160;Input1!\$C\$15;\\ IF(MAX(IndicSelecD.RESULTS!D157:D161)=IndicSelecD.RESULTS!D161;Input1!\$C\$16))))))\end{gathered} \tag{6}$$

The objective of this formula is to find the category that has a higher percentage. It checks the percentages calculated before and it picks the maximum value. If this value is in the first cell, it concludes that the higher percentage is for the "*Strongly Agree*". If the maximum it is not in the first cell, it will check the other ones until it finds the higher value.

2. Median for indicator 1

$$=IF(N.Input1!C6="----";"----";MEDIAN(N.Input1!\$C\$6:\$C\$355)) \tag{7}$$

The Excel has incorporated a "*MEDIAN*" function that automatically calculates the median of an interval of values.

3. Interquartile range (IQR) for indicator 1

To calculate the IQR, it is necessary to calculate the first and third quartiles and then perform a subtraction between them.

$$=IF(N.Input1!C6="----";"----";QUARTILE(N.Inpu1!\$C\$6:\$C\$355;1)) \tag{8}$$

The formula "*QUARTILE*" was applied to calculate these quartiles. The number "1" as input of this functions represents the first quartile. To calculate the third one, the "1" is substituted by a "3".

To study the level of agreement, the *Kendall's Coefficient of Concordance* (*W*) and the *Fleiss' Kappa* (*K*) were implemented. Some of the intermediate values were calculated in the extra tab. The overall *Spearman Rank Correlation Coefficient* (*r*) was also determined to have an idea of the overall stability of each round.

1. *Kendall's Coefficient of Concordance*

The first steps are the identification of how many indicators were under evaluation and how many raters were participating. The number of raters comes from the answer "*How many participants answered?*" answered by the user. The number of indicators comes from how many values were present in the first line of the frequency table displayed in the results, using the existent function "*COUNT*". After, all the ranks obtained for each indicator were summed as shown below.

$$=IF(SUM(N.Input1!\$C\$6:\$C\$355)=0;"----";SUM(N.Input1!\$C\$6:\$C\$355)) \tag{9}$$

The square of this sum for each indicator was calculated as well as the total sum of the square values. As explained in Table 7, the *Kendall's Coefficient of Concordance* is given by the following formula when ties exist.

$$W = \frac{12S}{m^2(n^3-n)-mT} \tag{10}$$

The next steps were calculating the numerator and denominator. The numerator was calculated directly using the function "*DESVQ*" of the Excel, which returns the sum of squares of deviation and the denominator was calculated having in consideration the *T* factor described in Table 7. The example for respondent 1 for the "*Strongly Agree*" is also presented by formula 12.

$$\boxed{= IF(D44 = "----"; "----"; 12 * DESVQ(D44{:}GU44)).} \tag{11}$$

$$\boxed{= COUNTIF(N.Input1!\$C6{:}\$GT6; Input1!\$C\$12)} \tag{12}$$

After, the sum of the subtraction of each of these values cubed minus the value itself was performed to obtain *T.* These calculations were performed for each one of the participants and the last thing to do was to sum the *T* value obtained for each one of them. Afterward, we were able to calculate the denominator.

$$\boxed{= IF(E42 = 0; "----"; (E43\char`^2) * ((E42\char`^3) - E42) - (E43 * EXTRA1!I10))} \tag{13}$$

Note that $E42$ and $E43$ represent the number of indicators and number of raters, respectively. The *Kendall's Correlation Coefficient* was obtained dividing the numerator by the denominator.

2. *Fleiss' Kappa*

Once again it was necessary to define how many indicators, raters and categories were under evaluation. The first two values were set up the same way as in the previous case. The number of categories it's always 5 in the case of the Delphi for selection of indicators and 7 in the case of the weighting judgments Delphi. Subsequently, the number of cells were calculated multiplying the number of indicators by the number of raters. After, the number of votes on each category of the scale for each indicator was calculated, using the same sightline. We added the total votes in each category for all the indicators and then we calculated the P_j (described in Table 7), for each category. P_j for the "*Strongly Agree*" of the first Delphi is given by:

$$\boxed{= IF(HD12 = 0; "----"; (1/IndicSelecD.RESULTS!\$D\$60) * HD12)} \tag{14}$$

where $HD12$ is the sum of the votes on the "*Strongly Agree*" category for all the indicators and $IndicSelecD.RESULTS!\$D\$60$ is the number of cells. After having P_j we were able to calculate $\bar{\bar{P_e}}$, which is no more than the sum of the square of P_js.

In parallel, we had to calculate P_i. An intermediate calculation was made to obtain the square of the sum of votes on all categories for each indicator and then the number of raters was subtracted to these values. The P_i was then obtained for each indicator as shown in formula 15.

$$\boxed{\begin{array}{c} = IF(L12 = \text{"----"}; \text{"----"}; \\ ((1/(IndicSelecD.RESULTS!\,\$E\$57 * (IndicSelecD.RESULTS!\,\$E\$57 - 1)) * EXTRA1!\,L21))). \end{array}} \quad (15)$$

Summing these values for each indicator, the global P_i was achieved, which allowed to perform the next step, the computation of $\bar{\bar{P}}$. $EXTRA1!\,L18$ represents the total P_i.

$$\boxed{= IF(E56 = 0; \text{"----"}; (1/E56) * EXTRA1!\,L18)} \quad (16)$$

To obtain the *Fleiss' Kappa* value, it was necessary to applicate the formula described in Table 7,using the $\bar{\bar{P}}$ and $\bar{\bar{P_e}}$ described atop.

3. Global *Spearman Rank Correlation Coefficient*

The global *Spearman Rank Correlation Coefficient* depends on *Kendall's Correlation Coefficient* and on the number of raters. Since we had already these two values defined, it was only to directly applicate the formula demonstrated in Table 7.

In the case of the value function Delphi, the analysis for each round did not present many of these measures since this Delphi process does not deal with numerical data. It requires a simpler analysis and so the *Kendall's Correlation Coefficient* and MANOVA could not be applied. For round 1, the statistical measures are the same and had a similar implementation. Additionally, a table that converts the judgments of the participants on the correspondent value function shapes is presented. The formula used to achieve it was:

$$\boxed{\begin{array}{c} = IF(N.Input3!\,D8 = \text{"----"}; \text{"----"}; IF(N.Input3!\,D8 = \text{""}; \text{""}; \\ IF(AND(N.Input3!\,D8 = N.Input3!\,E8; N.Input3!\,E8 = N.Input3!\,F8); \text{"}Linear\text{"}; \\ IF(AND(N.Input3!\,D8 < N.Input3!\,E8; N.Input3!\,E8 < N.Input3!\,F8); \\ \text{"}Concave\text{"}; IF(AND(N.Input3!\,D8 = N.Input3!\,E8; N.Input3!\,E8 < N.Input3!\,F8); \text{"}Concave\text{"}; \\ IF(AND(N.Input3!\,D8 < N.Input3!\,E8; N.Input3!\,E8 = N.Input3!\,F8); \text{"}Concave\text{"}; \\ IF(AND(N.Input3!\,D8 > N.Input3!\,E8; N.Input3!\,E8 > N.Input3!\,F8); \text{"}Cnvex\text{"}; \\ IF(AND(N.Input3!\,D8 = N.Input3!\,E8; N.Input3!\,E8 > N.Input3!\,F8); \text{"}Convex\text{"}; \\ IF(AND(N.Input3!\,D8 > N.Input3!\,E8; N.Input3!\,E8 = N.Input3!\,F8); \text{"}Convex\text{"}; \\ IF(AND(N.Input3!\,D8 > N.Input3!\,E8; N.Input3!\,E8 < N.Input3!\,F8); \text{"}S - Seat\text{"}; \\ IF(AND(N.Input3!\,D8 < N.Input3!\,E8; N.Input3!\,E8 > N.Input3!\,F8); \text{"}S - Sigmoide\text{"})))))))))) \end{array}} \quad (17)$$

which presents the implementation of conditions required to be considered a certain type of shape. For example, for a decreasing concave function the condition is that the vote on the first gap in the performance range is bigger than the second one and that this is larger than the latter gap vote ($vote\ (gap1)\ >\ vote\ (gap\ 2)\ >\ vote\ (gap\ 3)$). However, it can be considered concave too if the first and second votes are equal and bigger than the last one - ($vote\ (gap1) =$

$vote\ (gap\ 2)\ >\ vote\ (gap\ 3))$ - or if the second and third votes are equal and smaller than the first one - $vote\ (gap1) >\ vote\ (gap\ 2) =\ vote\ (gap\ 3))$. These conditions come from the fact that the first vote given by the participant is the correspondent to the last part of the function since they start voting in the higher values of the range performance, described in the *xx* axis (gap 1 in Figure 17).

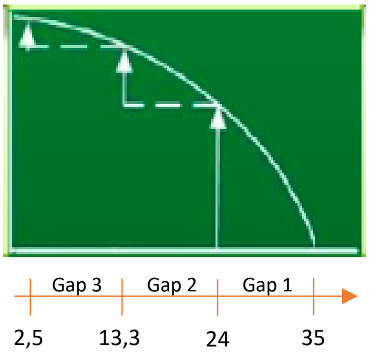
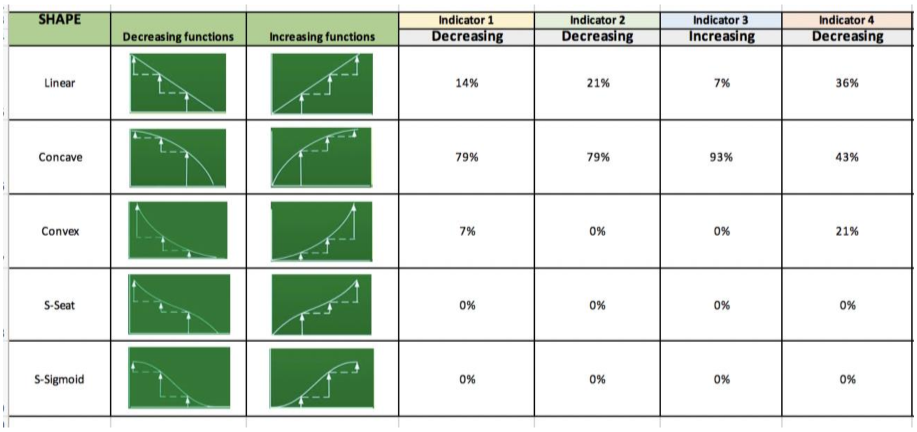


Figure 17 - *Scheme of the relationship between the votes of the panelists and the shape of the respective value functions.*

Afterward, a table with the percentages of votes on each shape is presented. The implementation was similar to the calculation of the other percentages. Plus, having the correct shapes defined, the mode and bi-mode were calculated as well as the existence of some shape with more than 50% of the votes, using the equations described in appendix A. For rounds 2 and 3, the evaluation was the same as the performed in the round 1, after having the shape of the functions.



SHAPE	Decreasing functions	Increasing functions	Indicator 1 Decreasing	Indicator 2 Decreasing	Indicator 3 Increasing	Indicator 4 Decreasing
Linear			14%	21%	7%	36%
Concave			79%	79%	93%	43%
Convex			7%	0%	0%	21%
S-Seat			0%	0%	0%	0%
S-Sigmoid			0%	0%	0%	0%

Figure 18 – *Part of the table with the percentages of votes on each value function shape for each indicator.*

The DSS also contains analysis between rounds 1 and 2 and rounds 2 and 3. For the first two types of Delphi, the *Spearman's Correlation Coefficient* was calculated for each indicator to understand how stable the responses were between rounds. To calculate it, we used the formula "*SCORREL*" provided by the "*Real Statistics Resource Pack*" available for Excel, which automatically calculates this coefficient between two sets of values, in this case, values obtain in consecutive rounds.

$$\boxed{\begin{gathered} = IF(D221 = 0; 1; \\ IF(N.Input1!C361 = "----"; "---"; \\ SCORREL(N.Input1!C\$6: N.Input1!C\$355; N.Input1!C\$361: N.Input1!C\$710))) \end{gathered}} \tag{18}$$

The inter-reliability was studied counting and calculating the percentage of how many participants changed their opinion between the rounds, and how many times each participant changed his vote. The equations are shown in the Appendix A. A table with the three participants that changed their opinion the most is also displayed. The objective is to understand if it is always the same person changing opinion or not. It was necessary to calculate the correct ranking for each participant and then applicate the "*MATCH*" function to obtain their positions in the list. If the answer provided in the table is 3 it means that is "Respondent 3".

$$\boxed{= MATCH(SMALL(EXTRA1!GX\$1078: GX\$1427; 1); EXTRA1!GX\$1078: GX\$1427; 0)} \tag{19}$$

This example has a "1" in the last argument of "*MATCH*" function because we wanted to know the one who changed opinion the most. To find the second one, it is necessary to substitute the "1" for "2" and so on. Additionally, to understand if the changes of opinion are drastic or not, a table that presents the height of the "jump" between categories was built. For example, in the selection of indicators Delphi, which uses a 5-level Likert scale, if a participant changes his opinion from "Strongly Agree", which is the highest category to "Moderately Agree", a "jump" of **3** categories. To implement this measure, it was necessary to create an extra table which shows if the response of each panelist between rounds were the same or not, as shown in Appendix A. The mathematical programming used to implement this table is presented in formula 20.

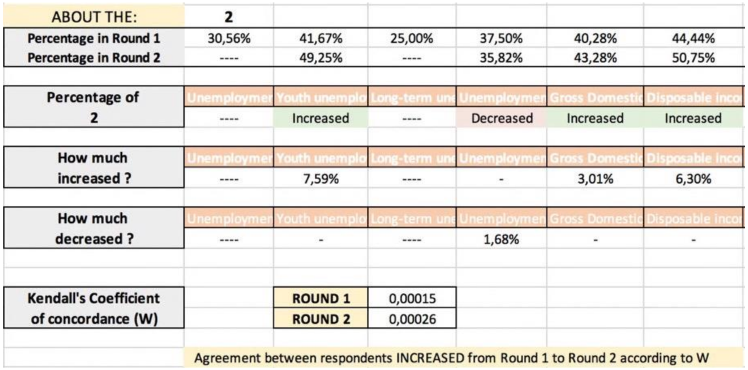
$$\boxed{\begin{gathered} = IF(N.Input1!C6 = "----"; "----"; \\ IF(EXTRA1!C1078 = "----"; "----"; \\ IF(OR(N.Input1!C6 = ""; N.Input1!C361 = ""); "-"; ABS(N.Input1!C361 - N.Input1!C6)))) \end{gathered}} \tag{20}$$

The formula "*ABS*" gives absolute values. Therefore, the jumps will be displayed with a positive number, no matter the direction of the changing votes. For the value function Delphi, the evaluation between rounds was similar but the *Spearman Correlation Coefficient* and this last table were not possible to calculate.

4.7.5. Summary Page

The tool also produces a "*Summary Page*" with the most important points of the results. In the Delphi for selection of indicators, this page analyses the two higher categories of the Likert scale closely. On the other hand, when using the 7-level MACBETH scale it is important to check more precisely what happens with the "*Extremely Important*", "*Very Strongly Important*" and

"*Strongly Important*" categories during the process. For some key issues, it is important to see a significant increase in these percentages; for others, it might be better to see the opposite. Therefore, this page shows the modifications that happened during the process. It also displays the values obtained for the level of agreement statistical measures. For the value function Delphi, the tool presents a similar evaluation for each value function shape, to see how the percentages changed during the process. This page also demonstrates how the opinion changing varied between rounds 1-2 and 2-3, i.e., how inter-reliability varied, and for the first two types of Delphi, how much the *Spearman's Correlation Coefficient* changed during the process, concluding about stability. The mathematical programming was based in "*IF*" conditions and subtraction equations between the percentages of two successive rounds (see Appendix A).

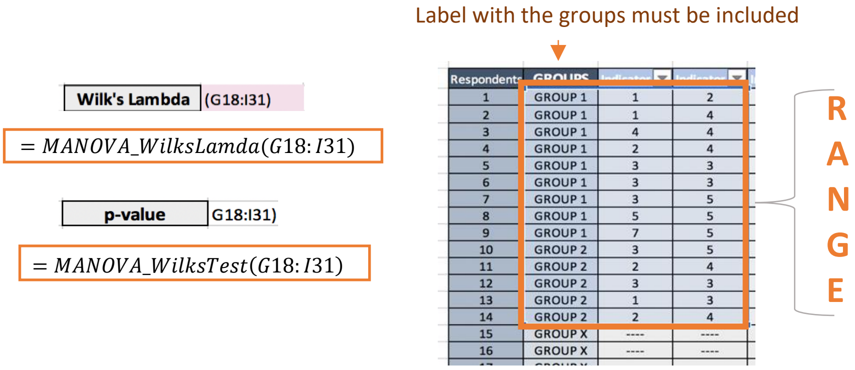


ABOUT THE:	2					
Percentage in Round 1	30,56%	41,67%	25,00%	37,50%	40,28%	44,44%
Percentage in Round 2	----	49,25%	----	35,82%	43,28%	50,75%
Percentage of	Unemploymen	Youth unemplo	Long-term une	Unemploymen	Gross Domestic	Disposable inco
2	----	Increased	----	Decreased	Increased	Increased
How much	Unemploymen	Youth unemplo	Long-term une	Unemploymen	Gross Domestic	Disposable inco
increased ?	----	7,59%	----	-	3,01%	6,30%
How much	Unemploymen	Youth unemplo	Long-term une	Unemploymen	Gross Domestic	Disposable inco
decreased ?	----	-	----	1,68%	-	-
Kendall's Coefficient		ROUND 1	0,00015			
of concordance (W)		ROUND 2	0,00026			
	Agreement between respondents INCREASED from Round 1 to Round 2 according to W					

Figure 19 - *Part of the "Summary page" of the selection of indicators Delphi (about the "Agree" category).*

4.7.6. MANOVA

The "*MANOVA*" tab only exist for the first two types of Delphi and it is destined to the users who want to study the changes in the opinions' variance. The *RealStatistics Resource* pack has built-in functions that allow performing different analyses only by applying a function, without the need for intermediate calculations. *MANOVA* is one type of analysis that has many functions available in this pack, for example functions to calculate the *Wilks' Lambda* test and its p-value. However, to apply directly this function it is necessary to select the range of the data, which varies with the data of each user. For that reason, these values will not appear automatically, the researcher needs to apply the formula on the range of his data. Indications are provided to facilitate the process, as shown below. It is important to note that when selecting the data, the label must be embedded. It is also shown how to calculate the respective p-value which is multiple times used to complement the evaluation of the Wilks test.



Wilk's Lambda	(G18:I31)

$= MANOVA\_WilksLamda(G18{:}\,I31)$

p-value	G18:I31)

$= MANOVA\_WilksTest(G18{:}\,I31)$

Respondents	GROUPS	Indicator	Indicator
1	GROUP 1	1	2
2	GROUP 1	1	4
3	GROUP 1	4	4
4	GROUP 1	2	4
5	GROUP 1	3	3
6	GROUP 1	3	3
7	GROUP 1	3	5
8	GROUP 1	5	5
9	GROUP 1	7	5
10	GROUP 2	3	5
11	GROUP 2	2	4
12	GROUP 2	3	3
13	GROUP 2	1	3
14	GROUP 2	2	4
15	GROUP X	----	----
16	GROUP X	----	----

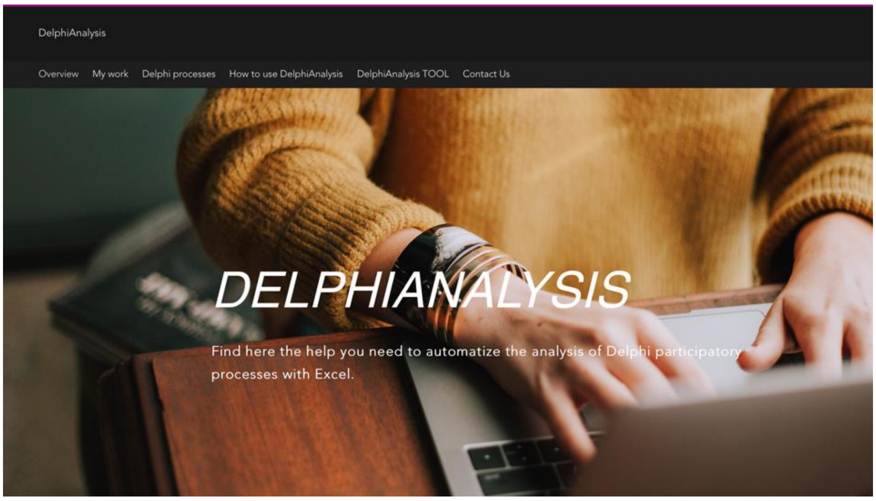
Figure 20 - *How to apply the Wilks Lambda test and its correspondent p-value. The range needed in the formula is selected in the figure.*

4.7.7. "*EXTRA*" Tab

The tool presents a page called "*EXTRA*" to perform intermediate calculations to implement the required statistical measures, explained along with the other calculations. This spreadsheet does not bring any useful information to the user so it will not available to change it.

4.8. Webpage

A webpage was created using the Wix platform, which is a cloud-based web development platform that allows to create professional websites and mobile sites through online drag and drop tools and with true creative freedom [93]. The main objective was to create a guide about how to correctly use the tool, showing the researchers how to change the descriptive tables in the right way and how to implement the *Wilks Lambda* and its p-value in the "MANOVA" tab. The web page will contain the three Excel files that make up the tool and can be downloaded. Additionally, it is shown where the researcher can get the *Real Statistics Resource* Pack for free. This webpage will also have a section about the work perform in this dissertation to explain how the *DelphiAnalysis* tool emerged. Details about the web site and how it looks like are provided in Appendix B.



Find here the help you need to automatize the analysis of Delphi participatory processes with Excel.

Figure 21 - *DelphiAnalysis DSS webpage guide (https://healthdelphianalysis.wixsite.com/delphianalysis).*

4.9. Questionnaire to Delphi experts

A short questionnaire was developed to understand the opinion of Delphi experts about the implemented tool. The main topics covered in the questions were the functionality of the appliance, whether they found it useful or not for those analysing the three specific types of Delphi in question, whether they would use it, some pros and cons and what could be improved in the future. The questions are mainly closed-ended but the respondents may suggest some improvements in the last part of the survey through open-ended answers. The questionnaire was filled by four Delphi experts from the Decision Eyes company. The planned survey is described in Appendix C and the answers provided are presented and discussed in the Results section.

The main objective of this part of the work was to obtain validation by experts in the field regarding the tool and the outcomes provided by it.

5. Results

In this chapter, the DSS will be tested using real values from the EURO-HEALTHY project to prove it works well and without errors. The results will be presented along with the comparison of the existing results obtained by the project itself. It is important to note that the results will be compared with the real ones, but they will not be commented since the objective of this tool is not to analyse the results but facilitate the work of the researchers to do that analysis. Also, to validate the DSS, an evaluation of the tool will be made by professionals who work with Delphi processes every day through a small questionnaire which aims to understand their opinion about the usefulness of the tool, as previously explained.

As the results present extremely big tables, only part of them will be shown. Also, since the results for the Delphi for selection of indicators and weighting judgments are the same, only the first will be deeply analysed as well as the results for the Delphi for shaping value functions.

5.1. Results obtained for Delphi for selection of indicators

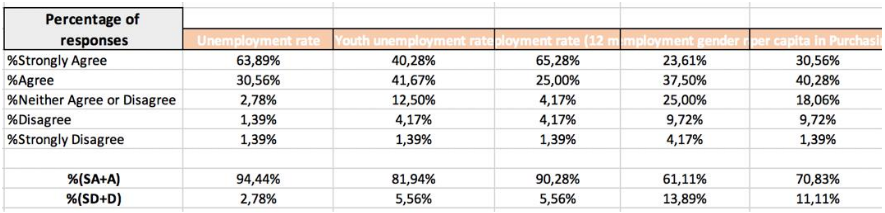
The EURO-HEALTHY project has developed a Delphi process to select important indicators for monitoring and evaluating population health with 51 experts and 30 stakeholders from different areas of knowledge and geographies participated. The participants formed only one panel that evaluated 130 indicators of health determinants and health outcomes. The panelists were required to indicate their level of agreement with the statement "*This indicator is relevant to the evaluation of Europe's population health*" using the 5-level Likert scale described above.

First of all, all the excel files with the responses had to be analysed to understand the semantic words used in the project to describe each category of the scale. The first thing noticed was that they had the answers provided in numbers and not words and also that the "5" was the "*Strongly Agree*" option, contrary to the order used by the tool. Then, it was necessary to change the right side of the descriptive table shown in Figure 16. Moreover, they had the indicators distributed by rows and the participants by columns and so the answers were paste in a transposed way to the table intended to receive the answers of each round. In the EURO-HEALTHY project, the panelists that did not answer were discarded as well as the indicators that obtained an absolute majority. However, as explained before, this tool needs to consider them all the time in the labels even if the answers stay empty.

Going to the third tab of the Excel file, the question "*How many participants answered?*" need to be answered by the user. From all the participants, the only ones that do not count are the ones that did not vote in any indicator. In this case, there were 81 participants assigned but 9 of them did not participate at all, so the answer is 81-9=**72** participants that answered. Below this question, the results for these data are displayed.

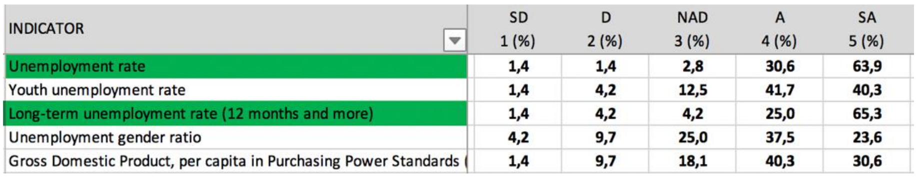
5.1.1. Frequency and percentage of votes on each category

The frequency and percentage of votes on each category of the Likert-scale were calculated as well as the percentage of the combination of some of them. The table with the resultant percentages for the first 5 indicators is presented below along with the percentages obtained in the EURO-HEALTHY project.



Percentage of responses	Unemployment rate	Youth unemployment rate	ployment rate (12 m	mployment gender r	per capita in Purchasi
%Strongly Agree	63,89%	40,28%	65,28%	23,61%	30,56%
%Agree	30,56%	41,67%	25,00%	37,50%	40,28%
%Neither Agree or Disagree	2,78%	12,50%	4,17%	25,00%	18,06%
%Disagree	1,39%	4,17%	4,17%	9,72%	9,72%
%Strongly Disagree	1,39%	1,39%	1,39%	4,17%	1,39%
%(SA+A)	94,44%	81,94%	90,28%	61,11%	70,83%
%(SD+D)	2,78%	5,56%	5,56%	13,89%	11,11%

Figure 22 - *Table with the resultant percentages of the first 5 indicators, in round 1, using the data from the EURO-HEALTHY project.*

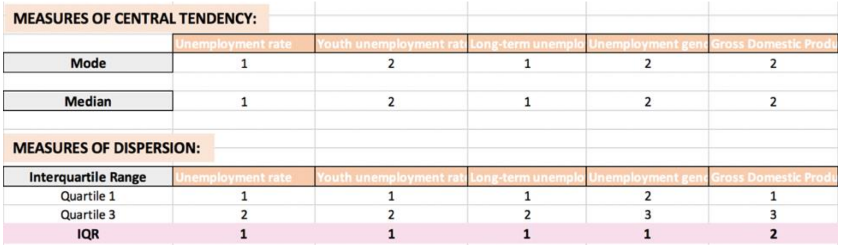


INDICATOR	SD 1 (%)	D 2 (%)	NAD 3 (%)	A 4 (%)	SA 5 (%)
Unemployment rate	**1,4**	**1,4**	**2,8**	**30,6**	**63,9**
Youth unemployment rate	**1,4**	**4,2**	**12,5**	**41,7**	**40,3**
Long-term unemployment rate (12 months and more)	**1,4**	**4,2**	**4,2**	**25,0**	**65,3**
Unemployment gender ratio	**4,2**	**9,7**	**25,0**	**37,5**	**23,6**
Gross Domestic Product, per capita in Purchasing Power Standards (	**1,4**	**9,7**	**18,1**	**40,3**	**30,6**

Figure 23 - *Table with the percentages of the first 5 indicators, in round 1, obtained in the EURO-HEALTHY project.*

Comparing the values obtained with the ones of the EURO-HEALTHY project, we can see that the values are more or less the same, differing only because of the rounding. It is possible to conclude that the DSS is calculating the values correctly.

The next results are the median, mode and the IQR. Comparing them with the ones obtained in the EURO-HEALTHY project and remembering that the order of the categories was the opposite (in the Tool, the “SA” category is represented by 1 and in the results of the project, it is represented by 5), we conclude that they are the same.



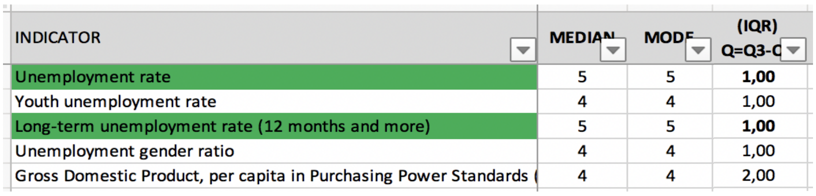
MEASURES OF CENTRAL TENDENCY:

	Unemployment rate	Youth unemployment rat	Long-term unemplo	Unemployment gend	Gross Domestic Produ
Mode	1	2	1	2	2
Median	1	2	1	2	2

MEASURES OF DISPERSION:

Interquartile Range	Unemployment rate	Youth unemployment rat	Long-term unemplo	Unemployment gend	Gross Domestic Produ
Quartile 1	1	1	1	2	1
Quartile 3	2	2	2	3	3
IQR	**1**	**1**	**1**	**1**	**2**

Figure 24 - *Resultant mode, median and IQR for the first 5 indicators (Round 1) obtained in the tool.*

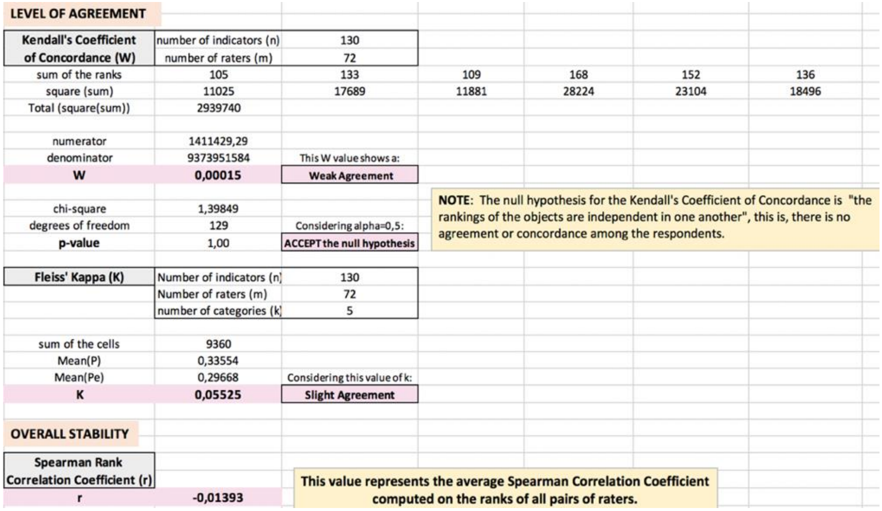


INDICATOR	MEDIAN	MODE	(IQR) Q=Q3-Q
Unemployment rate	5	5	**1,00**
Youth unemployment rate	4	4	1,00
Long-term unemployment rate (12 months and more)	5	5	**1,00**
Unemployment gender ratio	4	4	1,00
Gross Domestic Product, per capita in Purchasing Power Standards (	4	4	2,00

Figure 25 - *Resultant mode, median and IQR for the first 5 indicators (Round 1) obtained in the EURO-HEALTHY project.*

5.1.2. Additional results for each round

Additionally, in order to study the level of agreement between the raters, the *Kendall Coefficient of Concordance* (W) and the *Fleiss' Kappa Coefficient* (K) are calculated as well as the average *Spearman Rank Correlation Coefficient,* to have an idea of the stability of the data. For the first round, the results are presented below. The EURO-HEALTHY Project did not perform these kinds of analysis but the DSS seemed to work correctly.



LEVEL OF AGREEMENT						
Kendall's Coefficient of Concordance (W)	number of indicators (n)	130				
	number of raters (m)	72				
sum of the ranks	105	133	109	168	152	136
square (sum)	11025	17689	11881	28224	23104	18496
Total (square(sum))	2939740					
numerator	1411429,29					
denominator	9373951584	This W value shows a:				
W	0,00015	Weak Agreement				
chi-square	1,39849					
degrees of freedom	129	Considering alpha=0,5:				
p-value	1,00	ACCEPT the null hypothesis				
Fleiss' Kappa (K)	Number of indicators (n)	130				
	Number of raters (m)	72				
	number of categories (k)	5				
sum of the cells	9360					
Mean(P)	0,33554					
Mean(Pe)	0,29668	Considering this value of k:				
K	0,05525	Slight Agreement				

NOTE: The null hypothesis for the Kendall's Coefficient of Concordance is "the rankings of the objects are independent in one another", this is, there is no agreement or concordance among the respondents.

OVERALL STABILITY	
Spearman Rank Correlation Coefficient (r)	
r	-0,01393

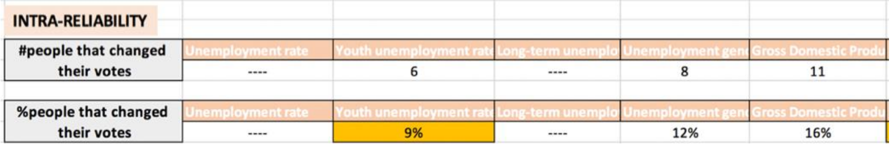
This value represents the average Spearman Correlation Coefficient computed on the ranks of all pairs of raters.

Figure 26 - *Resultant W, K and r for the Round 1 using the data from the EURO-HEALTHY Project.*

The results obtained for the next two rounds are present in the Appendix A.

5.1.3. Results of the "between rounds" section

In the project, the indicators that reached absolute majority were discarded and so the DSS can only analyse the ones present in two consecutive rounds. First, it is calculated how many participants changed their votes between rounds. These values allow understanding whether the participants had their ideas fixed and how many of them changed their mind regarding the feedback provided. After, to study the stability of responses, the Spearman's Correlation Coefficient was also calculated between successive rounds, as shown in Figure 28.

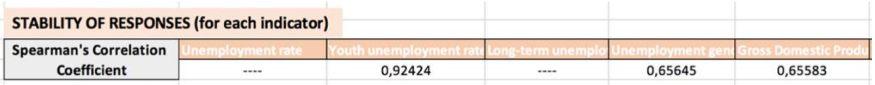


INTRA-RELIABILITY

#people that changed their votes	Unemployment rate	Youth unemployment rate	Long-term unemplo	Unemployment gen	Gross Domestic Produ
	----	6	----	8	11

%people that changed their votes	Unemployment rate	Youth unemployment rate	Long-term unemplo	Unemployment gen	Gross Domestic Produ
	----	9%	----	12%	16%

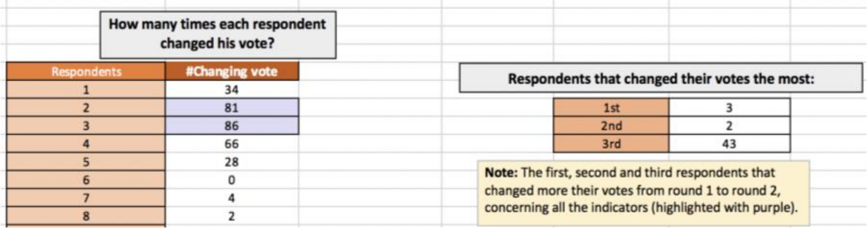
Figure 27 - *How much participants changed their opinion between rounds 1 and 2 (first 5 indicators). The indicators that show the best stability (less changes of opinion among panelists) are highlighted with yellow.*



STABILITY OF RESPONSES (for each indicator)

Spearman's Correlation Coefficient	Unemployment rate	Youth unemployment rate	Long-term unemplo	Unemployment gen	Gross Domestic Produ
	----	0,92424	----	0,65645	0,65583

Figure 28 - *Resultant Spearman Rank Correlation Coefficient for the first 5 indicators between rounds 1-2.*



How many times each respondent changed his vote?

Respondents	#Changing vote
1	34
2	81
3	86
4	66
5	28
6	0
7	4
8	2

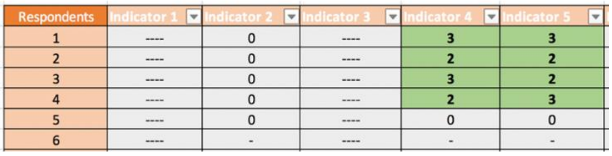
Respondents that changed their votes the most:

1st	3
2nd	2
3rd	43

Note: The first, second and third respondents that changed more their votes from round 1 to round 2, concerning all the indicators (highlighted with purple).

Figure 29 - *Part of the table that shows how many times each participant changed his opinion and table with the three that changed it most times (highlighted with purple on the left side Table).*

Additionally, a table with how much each opinion varied between the rounds were created. The variation is calculated through the absolute difference between the categories voted on the second round minus the category voted on the first round (e.g. $(|Strongly\ Agree\ (1) - Moderately\ Agree\ (3)| = 2)$, what allows understanding whether the change of opinion was drastic or soft.

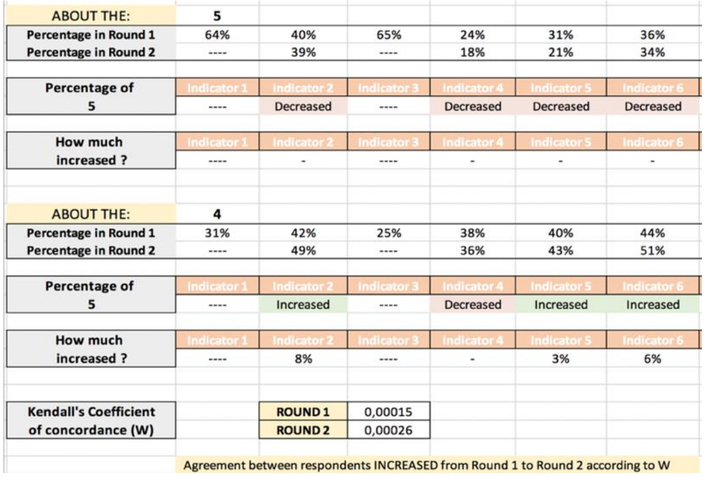


Respondents	Indicator 1	Indicator 2	Indicator 3	Indicator 4	Indicator 5
1	----	0	----	3	3
2	----	0	----	2	2
3	----	0	----	3	2
4	----	0	----	2	3
5	----	0	----	0	0
6	----	-	----	-	-

Figure 30 - *How much the votes of each participant changed between rounds 1 and 2 (for the first 5 indicators). Since it is a 5-level scale, jumps of two categories is considered drastic (highlighted with green).*

5.1.4. "Summary page" results

In the "Summary Page" tab, the DSS presents the information about the two highest categories of the Likert scale for two consecutive rounds and it shows if the percentages of votes on these categories increased or decreased and how much. The level of agreement obtained in two consecutive rounds are put side by side to conclude whether the level of agreement increased or not.



ABOUT THE:	5					
Percentage in Round 1	64%	40%	65%	24%	31%	36%
Percentage in Round 2	----	39%	----	18%	21%	34%

Percentage of 5	Indicator 1	Indicator 2	Indicator 3	Indicator 4	Indicator 5	Indicator 6
	----	Decreased	----	Decreased	Decreased	Decreased

How much increased ?	Indicator 1	Indicator 2	Indicator 3	Indicator 4	Indicator 5	Indicator 6
	----	-	----	-	-	-

ABOUT THE:	4					
Percentage in Round 1	31%	42%	25%	38%	40%	44%
Percentage in Round 2	----	49%	----	36%	43%	51%

Percentage of 5	Indicator 1	Indicator 2	Indicator 3	Indicator 4	Indicator 5	Indicator 6
	----	Increased	----	Decreased	Increased	Increased

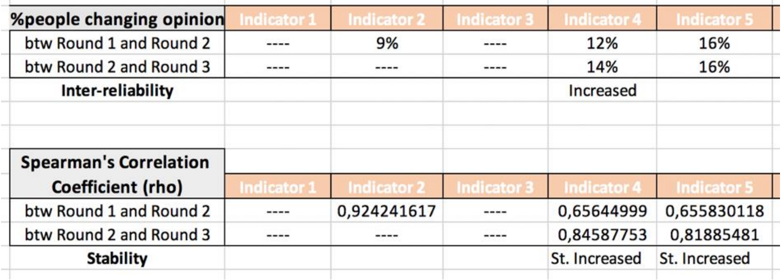
How much increased ?	Indicator 1	Indicator 2	Indicator 3	Indicator 4	Indicator 5	Indicator 6
	----	8%	----	-	3%	6%

Kendall's Coefficient of concordance (W)		
	ROUND 1	0,00015
	ROUND 2	0,00026

Agreement between respondents INCREASED from Round 1 to Round 2 according to W

Figure 31 - *Part of the results present on the "summary page" between rounds 1 and 2.*

The tool also compares the percentage of people that changed their opinion between rounds 1 and 2 with the values obtained between rounds 2 and 3 in order to understand if the inter-reliability increased during the process or not. It is shown how the stability evolves during the process, analysing the Spearman's Correlation Coefficient, as displayed in Figure 32.



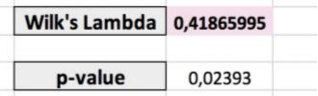
%people changing opinion	Indicator 1	Indicator 2	Indicator 3	Indicator 4	Indicator 5
btw Round 1 and Round 2	----	9%	----	12%	16%
btw Round 2 and Round 3	----	----	----	14%	16%
Inter-reliability				Increased	

Spearman's Correlation Coefficient (rho)	Indicator 1	Indicator 2	Indicator 3	Indicator 4	Indicator 5
btw Round 1 and Round 2	----	0,924241617	----	0,65644999	0,655830118
btw Round 2 and Round 3	----	----	----	0,84587753	0,81885481
Stability				St. Increased	St. Increased

Figure 32 - *Results about how inter-reliability and stability evolved during the process for each indicator.*

5.1.5. "MANOVA" results

In the "MANOVA" tab, the Delphi user can calculate the *Wilk's Lambda* test and its associated p-value. In the EURO-HEALTHY project, they evaluated if the area of expertise had an influence on the responses and also if stakeholders and experts had significant differences in their answers. However, the results could not be compared as it was not possible to have access to the full set of participants' information. To test this functionality of the tool, with the available information, some participants were divided according four fields of expertise: Economics and Social Environment; Environment Health, Ecological Systems and Sustainability; Epidemiology, Social Medicine and Public Health, and Health Geography, Demography and Sociology, which were denominated GROUP 1, 2, 3 and 4, respectively. Figure 33 displays the values obtained.



Wilk's Lambda	0,41865995
p-value	0,02393

Figure 33 - *Results for the first round MANOVA: Wilk's Lambda and respective p-value obtained for the responses given in the EURO-HEALTHY project, according to the tool.*

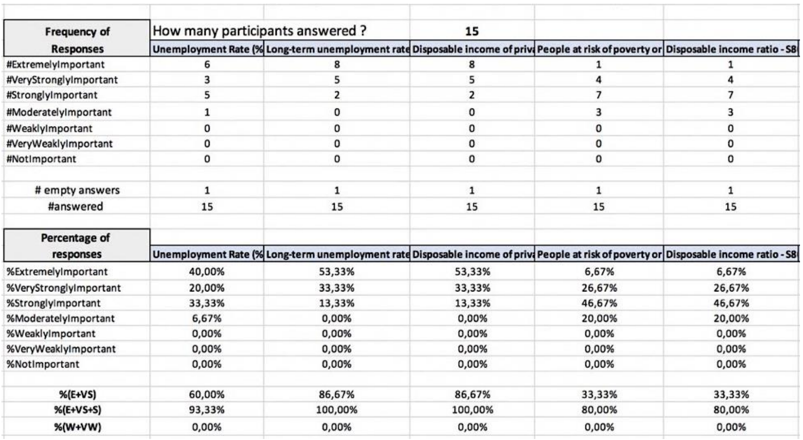
5.2. Results obtained for the weighting judgment Delphi

A weighting judgment Delphi process was also developed in the EURO-HEALTHY project. The main objective was to know how important each panelist considered to close the gaps of the performance range of each indicator. The panelists voted according to the 7-level MACBETH scale above described. Here, the panelists were divided into different panels, where each group only voted for the indicators of their area of expertise to increase the reliability of the results. Therefore, the answers provided by different groups could not be compared. The results present will be about the socioeconomic determinants that included 10 indicators and 16 panelists and will be briefly described as they are the same as in the previous case.

The descriptive table was already in the correct form since the categories used in the real project were the same as the provided by default in the DSS. The question "*How many participants answered?*", was answered with 15 for the three rounds $(16\, participants - 1\, \text{"}Not\ Answered\text{"} = \mathbf{15})$.

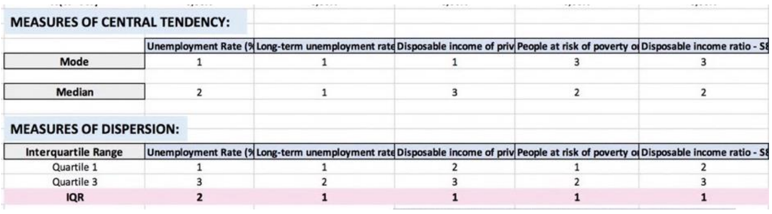
5.2.1. Frequency and percentage of votes on each category

The results of the frequency and percentages of votes are shown in Figure 34 and the mode, median and IQR in Figure 35.



Frequency of Responses	How many participants answered ?		15		
	Unemployment Rate (%	Long-term unemployment rate	Disposable income of priv	People at risk of poverty or	Disposable income ratio - S8
#ExtremelyImportant	6	8	8	1	1
#VeryStronglyImportant	3	5	5	4	4
#StronglyImportant	5	2	2	7	7
#ModeratelyImportant	1	0	0	3	3
#WeaklyImportant	0	0	0	0	0
#VeryWeaklyImportant	0	0	0	0	0
#NotImportant	0	0	0	0	0
# empty answers	1	1	1	1	1
#answered	15	15	15	15	15
Percentage of responses	Unemployment Rate (%	Long-term unemployment rate	Disposable income of priv	People at risk of poverty or	Disposable income ratio - S8
%ExtremelyImportant	40,00%	53,33%	53,33%	6,67%	6,67%
%VeryStronglyImportant	20,00%	33,33%	33,33%	26,67%	26,67%
%StronglyImportant	33,33%	13,33%	13,33%	46,67%	46,67%
%ModeratelyImportant	6,67%	0,00%	0,00%	20,00%	20,00%
%WeaklyImportant	0,00%	0,00%	0,00%	0,00%	0,00%
%VeryWeaklyImportant	0,00%	0,00%	0,00%	0,00%	0,00%
%NotImportant	0,00%	0,00%	0,00%	0,00%	0,00%
%(E+VS)	60,00%	86,67%	86,67%	33,33%	33,33%
%(E+VS+S)	93,33%	100,00%	100,00%	80,00%	80,00%
%(W+VW)	0,00%	0,00%	0,00%	0,00%	0,00%

Figure 34 - *Results of the frequency and percentages of votes provided by the panelists in round 1 (first 5 indicators).*



MEASURES OF CENTRAL TENDENCY:

	Unemployment Rate (%	Long-term unemployment rate	Disposable income of priv	People at risk of poverty o	Disposable income ratio - S8
Mode	1	1	1	3	3
Median	2	1	3	2	2

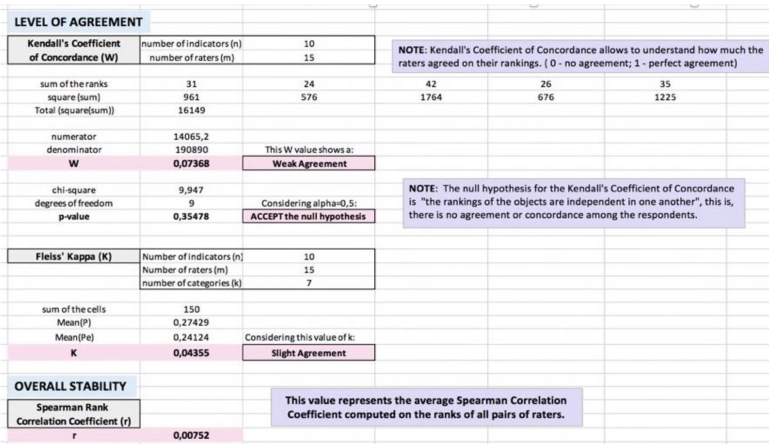
MEASURES OF DISPERSION:

Interquartile Range	Unemployment Rate (%	Long-term unemployment rate	Disposable income of priv	People at risk of poverty o	Disposable income ratio - S8
Quartile 1	1	1	2	1	2
Quartile 3	3	2	3	2	3
IQR	2	1	1	1	1

Figure 35 - *Mode, Median and IQR (Round 1, first 5 indicators).*

5.2.2. Additional Results for each round

The *Kendall's Coefficient of Concordance* and respective p-value, the *Fleiss' Kappa* and the *Spearman Rank Correlation Coefficient* were calculated as well (Figure 36).



LEVEL OF AGREEMENT

Kendall's Coefficient of Concordance (W)	number of indicators (n)	10			
	number of raters (m)	15			
sum of the ranks	31	24	42	26	35
square (sum)	961	576	1764	676	1225
Total (square(sum))	16149				
numerator	14065,2				
denominator	190890	This W value shows a:			
W	0,07368	Weak Agreement			
chi-square	9,947				
degrees of freedom	9	Considering alpha=0,5:			
p-value	0,35478	ACCEPT the null hypothesis			
Fleiss' Kappa (K)	Number of indicators (n)	10			
	Number of raters (m)	15			
	number of categories (k)	7			
sum of the cells	150				
Mean(P)	0,27429				
Mean(Pe)	0,24124	Considering this value of k:			
K	0,04355	Slight Agreement			

NOTE: Kendall's Coefficient of Concordance allows to understand how much the raters agreed on their rankings. (0 - no agreement; 1 - perfect agreement)

NOTE: The null hypothesis for the Kendall's Coefficient of Concordance is "the rankings of the objects are independent in one another", this is, there is no agreement or concordance among the respondents.

OVERALL STABILITY

Spearman Rank Correlation Coefficient (r)	
r	0,00752

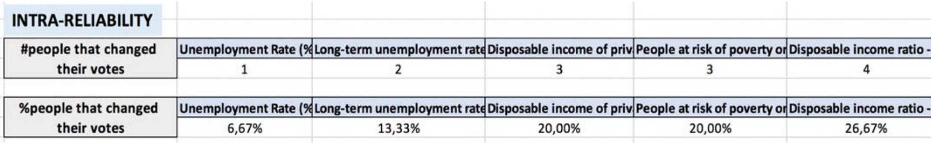
This value represents the average Spearman Correlation Coefficient computed on the ranks of all pairs of raters.

***Figure 36 -** Results of the analysis of the level of agreement and the overall stability for the first round of this Delphi process.*

The analysis obtained for the second and third rounds are present in the Appendix A.

5.2.3. Results of the "Between rounds" section

The tool also provides an analysis between the rounds 1 and 2 and rounds 2 and 3. To evaluate the intra-reliability of the process, the DSS calculates the frequency and percentage of participants that change their votes between two successive rounds. To conclude about the stability, the difference of the *Spearman's Correlation Coefficient* between rounds was calculated as well.

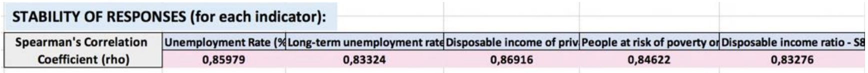


INTRA-RELIABILITY

#people that changed their votes	Unemployment Rate (%	Long-term unemployment rate	Disposable income of priv	People at risk of poverty or	Disposable income ratio -
	1	2	3	3	4

%people that changed their votes	Unemployment Rate (%	Long-term unemployment rate	Disposable income of priv	People at risk of poverty or	Disposable income ratio -
	6,67%	13,33%	20,00%	20,00%	26,67%

***Figure 37 -** Frequency and percentage of how many participants changed their vote between rounds 1 and 2 (first 5 indicators).*

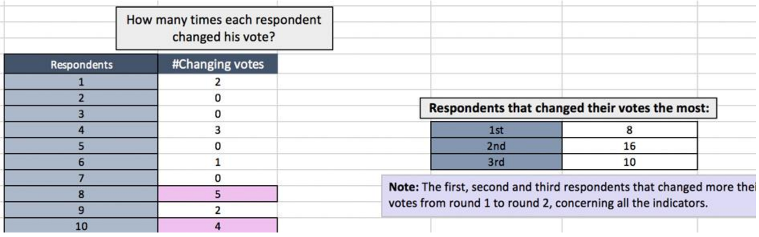


STABILITY OF RESPONSES (for each indicator):

Spearman's Correlation Coefficient (rho)	Unemployment Rate (%	Long-term unemployment rate	Disposable income of priv	People at risk of poverty or	Disposable income ratio - S8
	0,85979	0,83324	0,86916	0,84622	0,83276

***Figure 38 -** Spearman's Correlation Coefficient between round 1 and round 2, for the first 5 indicators.*

The DSS counts how many times each participant changed opinion and which ones changed it the most, as shown in Figure 39. The "jumps" between the votes of two successive rounds for each indicator are also calculated.



How many times each respondent changed his vote?

Respondents	#Changing votes
1	2
2	0
3	0
4	3
5	0
6	1
7	0
8	5
9	2
10	4

Respondents that changed their votes the most:

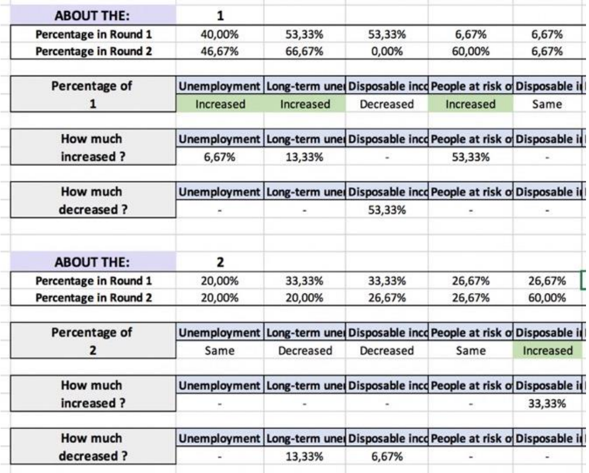
1st	8
2nd	16
3rd	10

Note: The first, second and third respondents that changed more thei votes from round 1 to round 2, concerning all the indicators.

Figure 39 - *Table that shows how many times each participant changed his vote and table with the 3 panelists that changed opinion most times.*

5.2.4. “Summary Page” Results

The “Summary Page” shows how much the percentage of votes increased or decreased between rounds for the highest three categories of the semantic MACBETH scale. The *Kendall’s Coefficient, Fleiss’ Kappa* and *Spearman Correlation Coefficient* were calculated between rounds.



ABOUT THE:	1				
Percentage in Round 1	40,00%	53,33%	53,33%	6,67%	6,67%
Percentage in Round 2	46,67%	66,67%	0,00%	60,00%	6,67%

Percentage of 1	Unemployment	Long-term une	Disposable inc	People at risk o	Disposable i
	Increased	Increased	Decreased	Increased	Same

How much increased ?	Unemployment	Long-term une	Disposable inc	People at risk o	Disposable i
	6,67%	13,33%	-	53,33%	-

How much decreased ?	Unemployment	Long-term une	Disposable inc	People at risk o	Disposable i
	-	-	53,33%	-	-

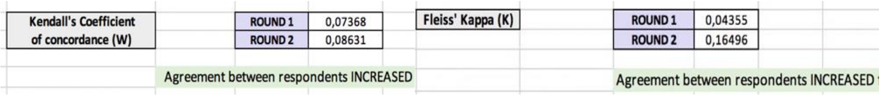
ABOUT THE:	2				
Percentage in Round 1	20,00%	33,33%	33,33%	26,67%	26,67%
Percentage in Round 2	20,00%	20,00%	26,67%	26,67%	60,00%

Percentage of 2	Unemployment	Long-term une	Disposable inc	People at risk o	Disposable i
	Same	Decreased	Decreased	Same	Increased

How much increased ?	Unemployment	Long-term une	Disposable inc	People at risk o	Disposable i
	-	-	-	-	33,33%

How much decreased ?	Unemployment	Long-term une	Disposable inc	People at risk o	Disposable i
	-	13,33%	6,67%	-	-

Figure 40 - *How much the percentages of the two higher categories changed between rounds 1 and 2.*



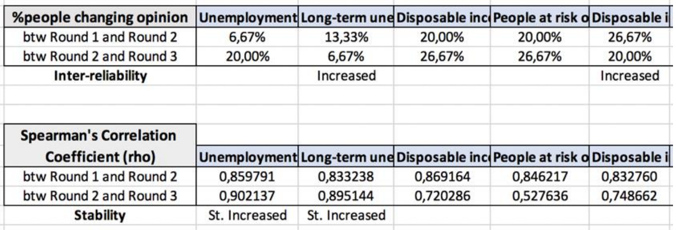
Kendall's Coefficient of concordance (W)		
	ROUND 1	0,07368
	ROUND 2	0,08631

Agreement between respondents INCREASED

Fleiss' Kappa (K)		
	ROUND 1	0,04355
	ROUND 2	0,16496

Agreement between respondents INCREASED

Figure 41 - *W and K obtained in rounds 1 and 2 concluding about how it changed.*



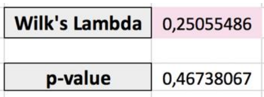
%people changing opinion	Unemployment	Long-term une	Disposable inc	People at risk o	Disposable i
btw Round 1 and Round 2	6,67%	13,33%	20,00%	20,00%	26,67%
btw Round 2 and Round 3	20,00%	6,67%	26,67%	26,67%	20,00%
Inter-reliability		Increased			Increased

Spearman's Correlation Coefficient (rho)	Unemployment	Long-term une	Disposable inc	People at risk o	Disposable i
btw Round 1 and Round 2	0,859791	0,833238	0,869164	0,846217	0,832760
btw Round 2 and Round 3	0,902137	0,895144	0,720286	0,527636	0,748662
Stability	St. Increased	St. Increased			

Figure 42 - *Analysis provided by the DSS to conclude about inter-reliability and stability through the Delphi process.*

5.2.5. “MANOVA” Results

Regarding the MANOVA, the field of expertise of the participants could not be studied as they were part of the same panel group. However, some participants were experts (GROUP 1) and some stakeholders (GROUP 2), so it was possible to analyse whether this distinction lead to differences on the results of the *Wilks Lambda* value and its respective p-value.

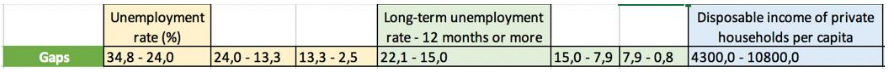


Wilk's Lambda	0,25055486
p-value	0,46738067

Figure 43 *- Wilk’s Lambda and correspondent p-value obtained for the evaluation of MANOVA between experts and stakeholders.*

5.3. Results obtained for the value function Delphi

A value function Delphi was also performed in the EURO-HEALTHY project to obtain the shapes of a value function that would describe the performance range of each indicator. This Delphi required a different analysis compared with the other two cases because it does not deal with numerical data, as previously explained. The first step was to adapt the descriptive table present by default in the DSS, as in the other cases. In the EURO-HEALTHY project, the performance range was divided into three equal parts, so it was necessary to fill the “*Gaps*” row with the “jumps” defined from the performance range, as shown in Figure 44.



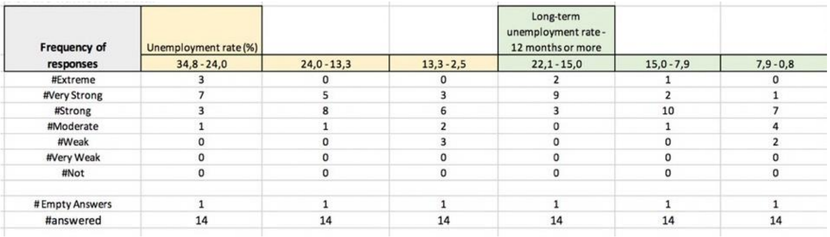
	Unemployment rate (%)			Long-term unemployment rate - 12 months or more			Disposable income of private households per capita
Gaps	34,8 - 24,0	24,0 - 13,3	13,3 - 2,5	22,1 - 15,0	15,0 - 7,9	7,9 - 0,8	4300,0 - 10800,0

Figure 44 *- “Gaps” row filled with the jumps of each range performance defined in the EURO-HEALTHY project. The indicators where replaced by their description.*

After, it is necessary to fill the input tables with the answers provided by the panelists. Going to the third tab, we have the results.

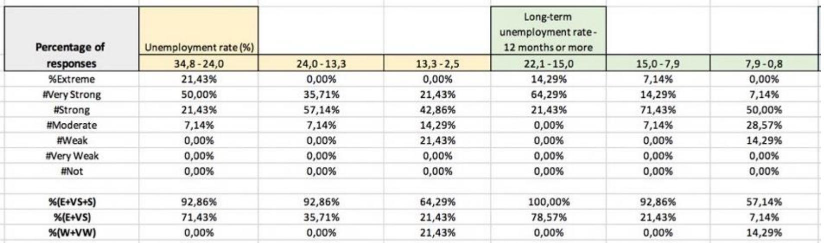
5.3.1. First round Results

For the first round, the DSS provides the frequency and percentages of votes on each category of the scale for the three gaps of each indicator. Additionally, it presents the sum of the percentages of some combinations of categories that may be useful to the researcher to analyse. The mode and interquartile range are also calculated, as shown below for some indicators.



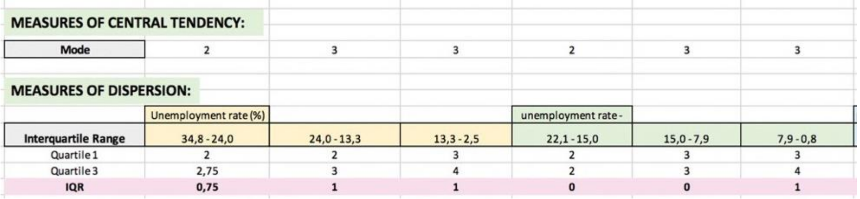
Frequency of responses	Unemployment rate (%)			Long-term unemployment rate - 12 months or more		
	34,8 - 24,0	24,0 - 13,3	13,3 - 2,5	22,1 - 15,0	15,0 - 7,9	7,9 - 0,8
#Extreme	3	0	0	2	1	0
#Very Strong	7	5	3	9	2	1
#Strong	3	8	6	3	10	7
#Moderate	1	1	2	0	1	4
#Weak	0	0	3	0	0	2
#Very Weak	0	0	0	0	0	0
#Not	0	0	0	0	0	0
# Empty Answers	1	1	1	1	1	1
#answered	14	14	14	14	14	14

Figure 45 *- Frequency of votes on each category for the three gaps of the first two indicators (Round 1).*



Percentage of responses	Unemployment rate (%) 34,8 - 24,0	24,0 - 13,3	13,3 - 2,5	Long-term unemployment rate - 12 months or more 22,1 - 15,0	15,0 - 7,9	7,9 - 0,8
%Extreme	21,43%	0,00%	0,00%	14,29%	7,14%	0,00%
#Very Strong	50,00%	35,71%	21,43%	64,29%	14,29%	7,14%
#Strong	21,43%	57,14%	42,86%	21,43%	71,43%	50,00%
#Moderate	7,14%	7,14%	14,29%	0,00%	7,14%	28,57%
#Weak	0,00%	0,00%	21,43%	0,00%	0,00%	14,29%
#Very Weak	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
#Not	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
%(E+VS+S)	92,86%	92,86%	64,29%	100,00%	92,86%	57,14%
%(E+VS)	71,43%	35,71%	21,43%	78,57%	21,43%	7,14%
%(W+VW)	0,00%	0,00%	21,43%	0,00%	0,00%	14,29%

Figure 46 - *Percentage of votes on each category for the three gaps of the first two indicators (Round 1) and the percentage of some useful combinations of votes.*



MEASURES OF CENTRAL TENDENCY:

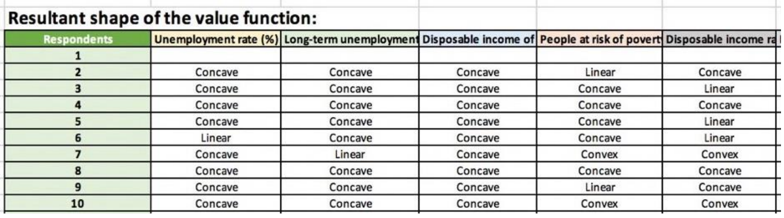
Mode	2	3	3	2	3	3

MEASURES OF DISPERSION:

Interquartile Range	Unemployment rate (%) 34,8 - 24,0	24,0 - 13,3	13,3 - 2,5	unemployment rate - 22,1 - 15,0	15,0 - 7,9	7,9 - 0,8
Quartile 1	2	2	3	2	3	3
Quartile 3	2,75	3	4	2	3	4
IQR	0,75	1	1	0	0	1

Figure 47 - *Mode and IQR of each of the three gaps of the first two indicators (Round 1).*

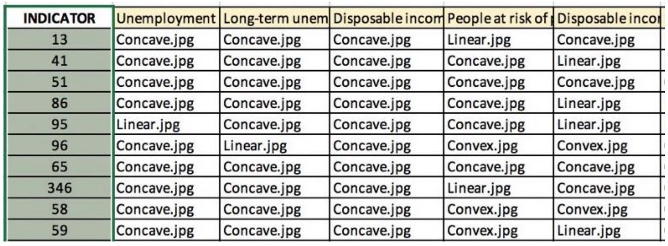
The value function allows to see the behavior of the performance range of each indicator and its shape depends on the votes given for the three "jumps" defined from the range. In this light, the tool presents a table, displayed in part in Figure 48, with the value function shape resultant from the votes on the three "jumps" of each indicator. The shapes presented in Figure 49 are the ones obtained in the EURO-HEALTHY project.



Resultant shape of the value function:

Respondents	Unemployment rate (%)	Long-term unemployment	Disposable income of	People at risk of povert	Disposable income ra
1					
2	Concave	Concave	Concave	Linear	Concave
3	Concave	Concave	Concave	Concave	Linear
4	Concave	Concave	Concave	Concave	Concave
5	Concave	Concave	Concave	Concave	Linear
6	Linear	Concave	Concave	Concave	Linear
7	Concave	Linear	Concave	Convex	Convex
8	Concave	Concave	Concave	Concave	Concave
9	Concave	Concave	Concave	Linear	Concave
10	Concave	Concave	Concave	Convex	Convex

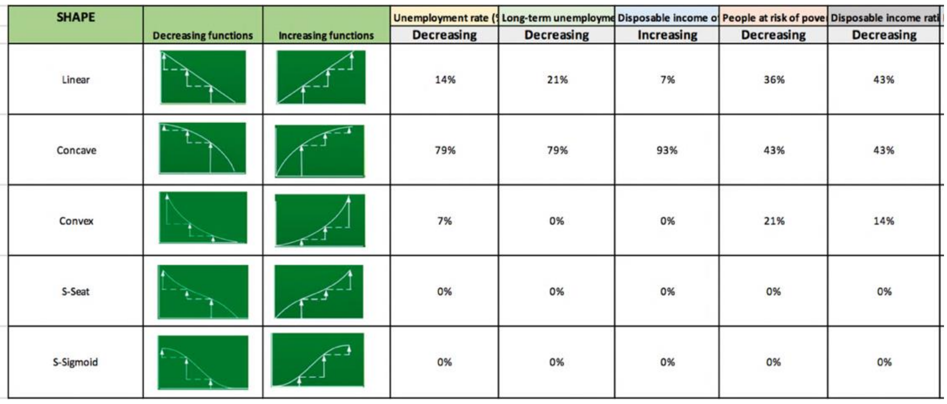
Figure 48 - *Table with the value function shapes obtained from the conversion of the votes of the first 10 panelists (Round 1, first 5 indicators).*



INDICATOR	Unemployment	Long-term unem	Disposable incom	People at risk of	Disposable inco
13	Concave.jpg	Concave.jpg	Concave.jpg	Linear.jpg	Concave.jpg
41	Concave.jpg	Concave.jpg	Concave.jpg	Concave.jpg	Linear.jpg
51	Concave.jpg	Concave.jpg	Concave.jpg	Concave.jpg	Concave.jpg
86	Concave.jpg	Concave.jpg	Concave.jpg	Concave.jpg	Linear.jpg
95	Linear.jpg	Concave.jpg	Concave.jpg	Concave.jpg	Linear.jpg
96	Concave.jpg	Linear.jpg	Concave.jpg	Convex.jpg	Convex.jpg
65	Concave.jpg	Concave.jpg	Concave.jpg	Concave.jpg	Concave.jpg
346	Concave.jpg	Concave.jpg	Concave.jpg	Linear.jpg	Concave.jpg
58	Concave.jpg	Concave.jpg	Concave.jpg	Convex.jpg	Convex.jpg
59	Concave.jpg	Concave.jpg	Concave.jpg	Convex.jpg	Linear.jpg

Figure 49 - *Table with the value function shapes obtained from the conversion of the votes of the first 10 panelists in the EURO-HEALTHY project (Round 1, first 5 indicators).*

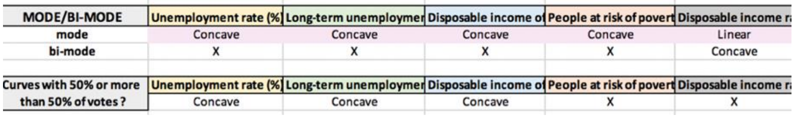
Comparing the shapes obtained by the DSS and in the project, it is possible to conclude that they are the same, which means the tool performing this conversion correctly. Another table with the percentages of each type of curve obtained through the votes is presented. This table summarizes if the range performance of each indicator is decreasing or increasing to understand which image represents the real shape.



SHAPE			Unemployment rate (%	Long-term unemployme	Disposable income o	People at risk of pove	Disposable income rati
	Decreasing functions	Increasing functions	Decreasing	Decreasing	Increasing	Decreasing	Decreasing
Linear			14%	21%	7%	36%	43%
Concave			79%	79%	93%	43%	43%
Convex			7%	0%	0%	21%	14%
S-Seat			0%	0%	0%	0%	0%
S-Sigmoid			0%	0%	0%	0%	0%

Figure 50 - *Percentages of each shape of the value function for the first 5 indicators (Round 1).*

The next part of the results provides the most common shape voted (mode), if there are two types of curves with the same percentage of votes (bi-mode) and the curves with more than 50% of votes, if existent.



MODE/BI-MODE	Unemployment rate (%)	Long-term unemploymen	Disposable income of	People at risk of povert	Disposable income ra
mode	Concave	Concave	Concave	Concave	Linear
bi-mode	X	X	X	X	Concave

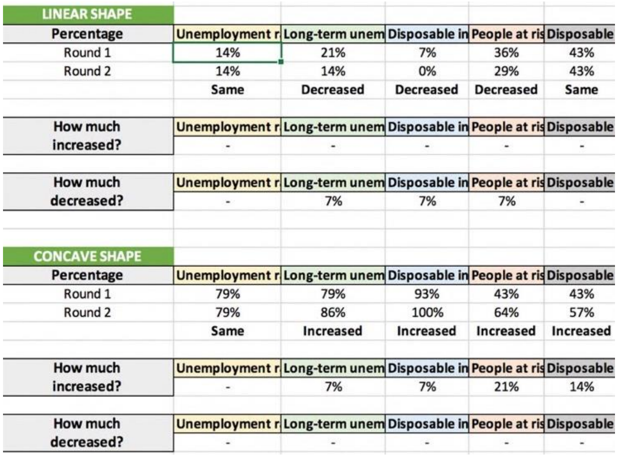
Curves with 50% or more than 50% of votes ?	Unemployment rate (%)	Long-term unemploymen	Disposable income of	People at risk of povert	Disposable income ra
	Concave	Concave	Concave	X	X

Figure 51 - *Mode and bi-mode, for the first 5 indicators and which curve achieved more than 50% of the votes, if existent (Round 1). The "X" represents inexistence.*

In the second and third rounds, the participants vote directly on the shape that they think is the best to characterize the performance range of each indicator, instead of voting in each "jump" according to an ordinal scale. On the first tab of the Excel, there is a second descriptive table, placed before the tables to be filled with the answers of rounds 2 and 3, that was used to adapt the data of the EURO-HEALTHY project to the one used by the tool, regarding the names used to characterize the shapes. Since the responses provided now are the shapes of the value functions, for round 2 and 3, the DSS only analyses the frequency and percentages of votes on each shape, the mode, bi-mode and if there were any shape with more than 50% of the votes, as in the second part of the results for the first round. The results for these two rounds are presented in the Appendix section.

5.3.2. "Summary Page" Results

The "Summary Page" for this Delphi presents a comparison between rounds 1 and 2 and rounds 2 and 3 about how much the percentages on each shape increased or decreased between two consecutive rounds. Results between rounds 1 and 2, for the linear and concave shapes of the first 5 indicators are presented below.



LINEAR SHAPE					
Percentage	Unemployment r	Long-term unem	Disposable in	People at ris	Disposable
Round 1	14%	21%	7%	36%	43%
Round 2	14%	14%	0%	29%	43%
	Same	Decreased	Decreased	Decreased	Same

How much increased?	Unemployment r	Long-term unem	Disposable in	People at ris	Disposable
	-	-	-	-	-

How much decreased?	Unemployment r	Long-term unem	Disposable in	People at ris	Disposable
	-	7%	7%	7%	-

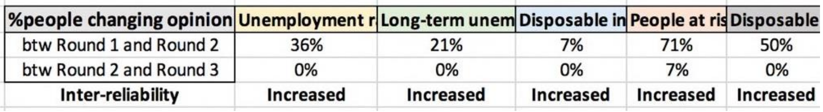
CONCAVE SHAPE					
Percentage	Unemployment r	Long-term unem	Disposable in	People at ris	Disposable
Round 1	79%	79%	93%	43%	43%
Round 2	79%	86%	100%	64%	57%
	Same	Increased	Increased	Increased	Increased

How much increased?	Unemployment r	Long-term unem	Disposable in	People at ris	Disposable
	-	7%	7%	21%	14%

How much decreased?	Unemployment r	Long-term unem	Disposable in	People at ris	Disposable
	-	-	-	-	-

Figure 52 - *How much decreased or increased the percentage of votes on linear and concave shapes, between round 1 and 2, for the first 5 indicators.*

Additionally, a table comparing the percentages of how many people changed their opinion throughout the Delphi, concluding whether the inter-reliability increased or decreased during the process is also displayed, as shown in Figure 53.



%people changing opinion	Unemployment r	Long-term unem	Disposable in	People at ris	Disposable
btw Round 1 and Round 2	36%	21%	7%	71%	50%
btw Round 2 and Round 3	0%	0%	0%	7%	0%
Inter-reliability	Increased	Increased	Increased	Increased	Increased

Figure 53 - *How much the percentage of participants who changed their votes varied throughout the Delphi process, concluding about the overall inter-reliability.*

5.4. Results obtained in the questionnaire

The questionnaire was given to four Delphi facilitators/decision analysts. However, only three answered. Their responses were then collected and analysed. In the most important questions of the questionnaire, it was intended to collect their opinions about some statements, according to a 5-level Likert scale, as following presented in tables 8 and 9.

Table 8 - *Number of votes on each category of the Likert scale for each statement provided.*

	How many votes on each category?				
Statements	**Strongly Agree**	**Agree**	**Neither Agree nor Disagree**	**Disagree**	**Strongly Disagree**
“The Webpage is useful”	2	1	-	-	-
“The tool is useful”	2	1	-	-	-
“The tool is user-friendly”	-	3	-	-	-

Table 9 - *Number of "yes" or "no" to the question "Would you use the DelphiAnalysis DSS to help you?".*

	Yes	**No**
"Would you use the *DelphiAnalysis* DSS to help you?"	3	-

As shown in the tables, it is possible to conclude that, globally, the feedback provided regarding the webpage and the DSS was positive. The website was considered useful for all the respondents and the tool was considered useful and user-friendly. However, the votes on the latter aspect were not on the highest category of the scale, what was justified by the fact that the tool was implemented in Excel and not using a better graphical interface or a website that could increase intuitiveness. Nonetheless, all agreed that the tool is advantageous and that they would use it to analyse these processes. This idea was reinforced by commentaries saying that the tool is very complete and gives all the answers that most experts need when evaluating Delphi processes. Other suggestions were provided as to incorporate more notes to help people that do not work with Delphi studies every day, to create an automatic way to import the input data and improve the interaction with the user. As positive aspects stand out the ease in getting and reading the results, how quickly the tool offers the results and the fact that the analysis provided is complete and diverse. As negative falls, the pointed aspects were the fact that the tool becomes slow sometimes and that it could be more intuitive and without some of its limitations if implemented in another language.

6. Discussion

The proposed methodology was developed to analyse three specific types of Delphi processes, where the best statistical measures were chosen according to the type of data and main objectives of each type. When testing the tool, all the statistical outputs are calculated immediately without any errors. By comparing the results with the available ones of the EURO-HEALTHY project, it became clear that the *DelphiAnalysis* DSS can lead to correct outcomes. Therefore, it is evident that the tool can be useful in automatically obtaining the most common techniques used to perform a complete analysis of Delphi processes, saving time and manual work from the user. However, the tool was implemented in Microsoft Excel, what brought some limitation on the number of participants, indicators and rounds that can be performed. Also, it did not allow to perform graphical analysis, which is often important when interpreting these kind of results. In general, it can be said that the DSS is useful since it automatically provides the most common statistical measures usually used to describe these Delphi processes. Therefore, despite all the limitations, the model can be seen as a novel and innovative DSS example to be followed and improved in the future. Some specific positive and negative aspects about of proposed DSS are presented below.

Regarding the webpage created, the feedback given by Delphi experts shown that the guide is necessary to help the users and that it provides useful information about how to correctly use the DSS. Also, Delphi experts defend that the webpage is user-friendly and intuitive.

The questionnaire was answered by three Delphi experts and allowed to understand their opinion about the *DelphiAnalysis* tool. In general, the information collected was that the DSS is intuitive and useful to provide a complete analysis required for most people working with these types of Delphi. Therefore, these surveys allowed to validate the DSS and to have some feedback of what can be improved in the DSS in future works.

6.1. Advantages of the DelphiAnalysis Tool

The developed DSS allows to fully analyse the three specific types of Delphi processes described, using the most common techniques - it includes measures of central tendency, dispersion, methods to conclude about the level of agreement, stability, inter and intra-reliability, which are the most important features to evaluate a Delphi process. Moreover, the tool can even be eventually used to analyse another type of Delphi if the user can adapt their data to be suitable to the tool. The descriptive tables allow the DSS to work independently of how each user defines the categories of the scale, which brings some flexibility to the tool. The results are shown in user-friendly tables, which make them convenient to be provided as feedback to the respondents. Beyond the results, it is still provided a summary page that organizes and complements the information accessed to be easily understood by the researcher. Therefore, the tool allows to quickly access to the most required statistical measures used by Delphi experts when analysing a process.

6.2. Disadvantages of the *DelphiAnalysis* Tool

During the development of this thesis, it became clear that analysing a Delphi process is not a linear process as the evaluation techniques should address the requirements of each situation. The major difficulty is to build a tool that can be able to evaluate every types of Delphi, which is not the state of this DSS as it only allows to analyse three specific standards. The next complication is that the proposed tool presents some limitations as the incorporation of only 200 indicators, 350 participants and 3 rounds. If the process is larger than this, it's not possible to evaluate it entirely at the same time using this tool. Also, the user needs to upload the answers manually which may take some extra time if he does not have the answers organized in Excel files. If during the process the researcher discard any participant or indicator, he needs to manually incorporate them in all the rounds in this DSS so it can make the correct comparisons between two successive rounds. In the value function Delphi, the gaps in the performance range also need to be filled manually in a really specific way, which may be a frustration if many indicators are under analysis. Another disadvantage is that the Excel files can work slow or block for a few seconds when the user is filling the tables with the participants' answers because of the quantity of content provided in the Excel files. However, since the results appear automatically, the overall time continues to be short.

These are the principal limitations of the tool along with the fact that it always provides the same evaluation, independently of the which is the study case. In this light, the prototype can certainly be improved in many ways, especially to make it even more automatic, introducing more or new important statistical measures or even in appearance by having a better graphical interface.

7. Conclusion

The main objective of this thesis was to develop a DSS to automatically analyse Delphi processes usually used in health settings as Delphi processes for selection of indicators, for weighting judgments and for shaping value functions that characterize the performance range of an indicator. Therefore, an independent Excel file to analyse each of these cases were developed along with a webpage guide that could give researchers some help about how to correctly use the *DelphiAnalysis* DSS. A questionnaire was also developed to collect some Delphi experts' opinions about the developed methodologies. With the work presented during the thesis, it is possible to see that the first objective was clearly fulfilled and that it was still possible to go further by creating adjacent tools that provide useful information. The prototype was developed and tested in the context of healthcare using data from a real case study: the EURO-HEALTHY project. The proposed tool addresses the pre-defined objectives as it calculates many statistical measures commonly used to analyse these types of processes. This tool designed following a proposed framework for DSS design, is the first appliance that automatically calculates statistical outputs to analyse three specific types of Delphi. In this light, all the resources and features required were planned before implementation. The DSS was developed in Excel as it was intended to be as user-friendly as possible and focused on the evaluator's understanding.

The resultant outcomes shown that the tool is working without any problems and, through the questionnaire answered by Delphis' experts, it was made clear that the tool can be useful for researchers working within the health area. However, being a novel prototype of this nature, it has several dimensions where it can be improved or even continued to be more complete or independent. Globally, it can be seen as an example for future developments, as the ones presented below.

7.1. Future Work

More tests should be performed in order to test and validate the *DelphiAnalysis* DSS, using different types of data and data from different contexts. More and new statistical measures that can complement the global evaluation of the features that describe these Delphi processes can be implemented or similar tools adapted to analyse different types of Delphi can be created. Another mathematical programming, more sophisticated and adapted to statistics can be used to implement a new version of the *DelphiAnalysis* DSS or new ones in order to build tools with fewer limitations, better interactions between the user and the DSS and a more intuitive and complete interface. Also, some of the recommendations given in the questionnaire could be implemented as provide more guide notes in the tool.

It would be great to have more Delphi experts' answering the questionnaire to validate the DSS or organize an interview with a group of professionals who could give good advices about improvements that can be done according to their necessities.

7.2. Final Remarks

The different perspectives of health stakeholders and DM are essential for decision-making in healthcare, mainly in clinical decisions, health technology assessment and strategic planning [14], [15]. To assess them, many participative approaches have been used recently, in hospitals, health care centers or private health care providers. These participative processes have been increasingly recognized as an added value in decision-making processes as they promote consensus among health stakeholders that may be at distinct locals and support evidence-based decisions [14]. However, it is complicated to treat participants' answers and organize Delphi outputs what explains the great importance of creating automatic tools to support the analysis of Delphi processes used in the field. Being health an area that impacts all the society, it is of great importance to continue developing new methodologies that can cope with time pressure and the discipline knowledge.

The work developed in this dissertation is an example to follow, to explore and develop new tools that can improve the treatment of Delphi participants' responses and statistical outputs.

References

[1] O. Sibony, R. Boulkedid, C. Alberti, M. Loustau, and H. Abdoul, “Using and Reporting the Delphi Method for Selecting Healthcare Quality Indicators: A Systematic Review,” *PLoS One*, 2011, vol. 6.

[2] G. Montibeller, “Resource Allocation and Priority Setting in Health Care : A Multi-criteria Decision Analysis Problem of Value ?,” vol. 8, pp. 76–83, 2017.

[3] P. Thokala *et al.*, “Multiple criteria decision analysis for health care decision making - An introduction: Report 1 of the ISPOR MCDA Emerging Good Practices Task Force,” *Value Heal.*, vol. 19, pp. 1–13, 2016.

[4] C. Pope, “Qualitative methods in research on healthcare quality,” *Qual. Saf. Heal. Care*, vol. 11, pp. 148–152, 2002.

[5] B. V. S. TJ, "*Multiple Criteria Decision Analysis: An Integrated Approach.*", Massachussetts: Kluwer Academic Publishers, 2012.

[6] J. Kitzinger, “Qualitative Research: Introducing focus groups,” *Bmj*, vol. 311, no. 7000, p. 299, 1995.

[7] J. Jones, D.Hunter, "Qualitative Research: Consensus methods for medical and health services research", BMJ publishing Group Ltd, 1995, vol. 311, pp. 376–380.

[8] S. Thangaratinam and C. W. Redman, “The Delphi technique,” *Obstet. Gynaecol.*, 2005 vol. 7, pp. 120–125.

[9] H. A. Linstone and M. Turoff, “Delphi: A brief look backward and forward,” *Technol. Forecast. Soc. Change*, 2011, vol. 78 pp. 1712–1719.

[10] M. R. Geist, “Using the Delphi method to engage stakeholders: A comparison of two studies,” *Eval. Program Plann.*, 2010, vol. 33, pp. 147–154.

[11] P. Gill, K. Stewart, E. Treasure, and B. Chadwick, “Methods of data collection in qualitative research: Interviews and focus groups,” *Br. Dent. J.*, vol. 204, no. 6, pp. 291–295, 2008.

[12] Cantrill J A, Sibbald B, and Buetow S, “The Delphi and nominal group techniques in health services research,” *Int. J. Pharm. Pract.*, vol. 4, pp. 67-74., 1996.

[13] H. Strauss and L. H. Zeigler, “The Delphi technique and Its Uses I n Social Science Research *,” vol. 9, no. 4, pp. 253–259.

[14] J. de Meyrick, “The Delphi method and health research,” *Health Educ.*, vol. 103, no. 1, pp. 7–16, 2003.

[15] R. Balasubramanian and D. Agarwal, “Delphi Technique- A Review,” *Int. J. Public Heal. Dent.*, vol. 3, no. 2, pp. 16–25, 2012.

[16] M. Glatzer, C. M. Panje, C. Sirén, N. Cihoric, and P. M. Putora, “Decision Making Criteria in Oncology,” *Oncol.*, 2018.

[17] A. R. Romano, “Malleable Delphi : Delphi Research Technique , its Evolution , and Business Applications,” *Int. Reviwe Bus. Res. Pap.*, vol. 6, no. 5, pp. 235–243, 2010.

[18] A. J. Fletcher and G. P. Marchildon, “Using the delphi method for qualitative, participatory action research in health leadership,” *Int. J. Qual. Methods*, vol. 13, no. 1, pp. 1–18, 2014.

[19] J. V. Meijering, J. K. Kampen, and H. Tobi, “Quantifying the development of agreement among experts in Delphi studies,” *Technol. Forecast. Soc. Change*, vol. 80, no. 8, pp. 1607–1614, 2013.

[20] F. Hasson and S. Keeney, “Enhancing rigour in the Delphi technique research,” *Technol. Forecast. Soc. Change*, vol. 78, no. 9, pp. 1695–1704, 2011.

[21] R. Morgan *et al.*, “Consensus in controversy: The modified Delphi method applied to Gynecologic Oncology practice,” *Gynecol. Oncol.*, vol. 138, no. 3, pp. 712–716, 2015.

[22] R. L. Custer and B. R. Stewart, “The Modified Delphi Technique - A Rotational Modification,” *J. Vocat. Tech. Educ.*, vol. 15, no. 2, 1999.

[23] S. Keeney, F. Hasson, and H. McKenna, “Conducting the Research Using the Delphi Technique,” *Delphi Tech. Nurs. Heal. Res.*, no. 1994, pp. 69–83, 2011.

[24] S. Keeney, F. Hasson, and H. McKenna, “How to Get Started with the Delphi Technique,” *Delphi Tech. Nurs. Heal. Res.*, no. 1975, pp. 43–68, 2011.

[25] R. C. de Loë, N. Melnychuk, D. Murray, and R. Plummer, “Advancing the State of Policy Delphi Practice: A Systematic Review Evaluating Methodological Evolution, Innovation, and Opportunities,” *Technol. Forecast. Soc. Change*, vol. 104, pp. 78–88, 2016.

[26] A. B. Costello and J. W. Osborne, “Practical Assessment, Research & Evaluation,” *Pract. assessment, Res. Eval.*, vol. 10, no. 7, p. pp 1-10, 2005.

[27] J. Huck, M. Muszynska, S. Aengenheyster, L. Gerhold, M. Heiskanen-Schüttler, and K. Cuhls, “Real-Time Delphi in practice — A comparative analysis of existing software-based tools,” *Technol. Forecast. Soc. Change*, vol. 118, pp. 15–27, 2017.

[28] T. Gnatzy, J. Warth, H. von der Gracht, and I. L. Darkow, “Validating an innovative real-time Delphi approach - A methodological comparison between real-time and conventional Delphi studies,” *Technol. Forecast. Soc. Change*, vol. 78, no. 9, pp. 1681–1694, 2011.

[29] T. Gordon and A. Pease, “RT Delphi: An efficient, ‘round-less’ almost real time Delphi method,” *Technol. Forecast. Soc. Change*, vol. 73, no. 4, pp. 321–333, 2006.

[30] H. Donohoe, M. Stellefson, and B. Tennant, “Advantages and Limitations of the e-Delphi Technique,” *Am. J. Heal. Educ.*, vol. 43, no. 1, pp. 38–46, 2013.

[31] P. L. Davidson, “The delphi technique in doctoral research: considerations and rationale,” *Rev. High. Educ. Self-learning*, vol. 6, no. 22, pp. 53–65, 2013.

[32] J. M. Culley, “Use of a computer-mediated delphi process to validate a mass casualty conceptual model,” *CIN - Comput. Informatics Nurs.*, vol. 29, no. 5, pp. 272–279, 2011.

[33] M. Steinert, “A dissensus based online Delphi approach: An explorative research tool,” *Technol. Forecast. Soc. Change*, vol. 76, no. 3, pp. 291–300, 2009.

[34] S. E. Seker, “computarized,” *IEEE Access*, vol. 3, pp. 368–380, 2015.

[35] P. Tapio, “Disaggregative policy Delphi,” *Technol. Forecast. Soc. Change*, vol. 70, no. 1, pp. 83–101, 2002.

[36] “The Delphi Technique: characteristics and sequence model,” *Couper, M. R. (1984). Delphi Tech. Adv. Nurs. Sci. 7(1), 72–77.*

[37] J. R. Avella, “Delphi panels: Research design, procedures, advantages, and challenges,”

Int. J. Dr. Stud., vol. 11, pp. 305–321, 2016.

[38] C. Powell, “The Delphi technique: Myths and realities,” *J. Adv. Nurs.*, vol. 41, no. 4, pp. 376–382, 2003.

[39] S. Keeney, F. Hasson, and H. McKenna, “Debates, Criticisms and Limitations of the Delphi,” *Delphi Tech. Nurs. Heal. Res.*, pp. 18–31, 2011.

[40] D. Bartram, “Reliability and Validity,” *Test. people Pract. Guid. to Psychom.*, pp. 57–86, 1990.

[41] F. Hasson, S. Keeney, and H. McKenna, “Research guidelines for the Delphi survey technique,” *J. Adv. Nurs.*, vol. 32, no. 4, pp. 1008–1015, 2010.

[42] W. Fred, “An evaluation of Delphi,” *Technol. Forecast. Soc. Change*, vol. 40, no. 2, pp. 131–150, 1991.

[43] V. N. Palter, H. M. MacRae, and T. P. Grantcharov, “Development of an objective evaluation tool to assess technical skill in laparoscopic colorectal surgery: A Delphi methodology,” *Am. J. Surg.*, vol. 201, no. 2, pp. 251–259, 2011.

[44] T. E.G. and R. N., “Delphi methodology in health research: How to do it?,” *Eur. J. Integr. Med.*, vol. 7, no. 4, pp. 423–428, 2015.

[45] Keeney Sinead, Hasson Felicity, and McKenna Hugh, “Consulting the oracle: ten lessons from using the Delphi technique in nursing research,” *J. Adv. Nurs.* , vol. 52, no. 2, pp. 205–12, 2006.

[46] H. A. von der Gracht, “Consensus measurement in Delphi studies,” *Technol. Forecast. Soc. Change*, vol. 79, no. 8, pp. 1525–1536, 2012.

[47] S. Keeney, F. Hasson, and H. McKenna, “Analysing Data from a Delphi and Reporting Results,” *Delphi Tech. Nurs. Heal. Res.*, pp. 84–95, 2011.

[48] S. Miah, J. Debuse, and D. Kerr, “A Development-Oriented DSS Evaluation Approach: A Case Demonstration for Conceptual Assessment,” *Australas. J. Inf. Syst.*, vol. 17, no. 2, pp. 43–55, 2012.

[49] H. A. von der Gracht, “Consensus measurement in Delphi studies,” *Technol. Forecast. Soc. Change*, vol. 79, no. 8, pp. 1525–1536, 2012.

[50] “Stevens S. On the theory of scales of measurement.pdf,” *Science*, vol. 03, no. 268. pp. 677-680., 1946.

[51] T. G.R., *Integrating Quantitative and Qualitative Methods in Research*, George R. New York: University Press of America, 2000.

[52] J. C. Nunnally and I. H. Bernstein, “NunnallyBernstein 1994 Chapter 1 in Psychometric Theory.pdf.” p. 35, 1994.

[53] S. A. Kalaian and R. M. Kasim, “Terminating Sequential Delphi Survey Data Collection - Practical Assessment, Research &amp; Evaluation,” *Pract. Assessment, Res. Eval.*, vol. 17, no. 5, pp. 1–10, 2012.

[54] T. Keeley *et al.*, “The use of qualitative methods to inform Delphi surveys in core outcome set development,” *Trials*, vol. 17, no. 1, pp. 1–9, 2016.

[55] G. B. Wetherill and A. E. Maxwell, “Analysing Qualitative Data.,” *J. R. Stat. Soc. Ser. A*,

vol. 125, no. 2, p. 289, 2006.

[56] J. F. Amber *et al.*, “The unholy marriage? Integrating qualitative and quantitative information in Delphi processes,” *Trials*, vol. 13, no. 1, pp. 1–18, 2016.

[57] S. Birko, E. S. Dove, V. Özdemir, and K. Dalal, “Evaluation of nine consensus indices in delphi foresight research and their dependency on delphi survey characteristics: A simulation study and debate on delphi design and interpretation,” *PLoS One*, vol. 10, no. 8, pp. 1–14, 2015.

[58] G. Rowe, “EXPERT OPINIONS IN FORECASTING : THE ROLE OF THE DELPHI TECHNIQUE.”

[59] E. E. Ameyaw, Y. Hu, M. Shan, A. P. C. Chan, and Y. Le, “Application of Delphi method in construction engineering and management research: A quantitative perspective,” *J. Civ. Eng. Manag.*, vol. 22, no. 8, pp. 991–1000, 2016.

[60] A. Habibi, A. Sarafrazi, and S. Izadyar, “Delphi technique theoretical framework in qualitative research,” *Int. J. Eng. Sci.*, vol. 3, no. 4, pp. 8–13, 2014.

[61] Y. N. Yang, “Testing the stability of experts’ opinions between sucesive rounds of Delphi studies,” *Pap. Prep. Anu. Meting Am. Educ. Res. Asoc.*, no. 1979, pp. 1–16, 2003.

[62] J. M. English and G. L. Kernan, “The prediction of air travel and aircraft technology to the year 2000 using the Delphi method,” *Transp. Res.*, vol. 10, no. 1, pp. 1–8, 1976.

[63] J. S. Dajani, M. Z. Sincoff, and W. K. Talley, “Stability and agreement criteria for the termination of Delphi studies,” *Technol. Forecast. Soc. Change*, vol. 13, no. 1, pp. 83–90, 1979.

[64] H. A. Shah and S. A. Kalaian, “Which Is the Best Parametric Statistical Method For Analyzing Delphi Data?,” *J. Mod. Appl. Stat. Methods*, vol. 8, no. 1, pp. 226–232, 2017.

[65] M. R. Kastein, M. Jacobs, R. H. van der Hell, K. Luttik, and F. W. M. M. Touw-Otten, “Delphi, the issue of reliability. A qualitative Delphi study in primary health care in the Netherlands,” *Technol. Forecast. Soc. Change*, vol. 44, no. 3, pp. 315–323, 1993.

[66] J. Sim and C. C. Wright, “The Kappa Statistic in Reliability Studies : Use , Interpretation , and,” vol. 85, no. 3, pp. 257–268.

[67] J. W. Gooch, “Kruskal-Wallis Test,” *Encycl. Dict. Polym.*, no. 1, pp. 984–985, 2011.

[68] R. Almendra, Â. Freitas, J. C. Bana e Costa, M. D. Oliveira, P. Santana, and C. A. Bana e Costa, “Indicators for evaluating European population health: a Delphi selection process,” *BMC Public Health*, vol. 18, no. 1, pp. 1–20, 2018.

[69] P. Tapio, R. Paloniemi, V. Varho, and M. Vinnari, “The unholy marriage? Integrating qualitative and quantitative information in Delphi processes,” *Technol. Forecast. Soc. Change*, vol. 78, no. 9, pp. 1616–1628, 2011.

[70] L. F. Luna-Reyes and D. L. Andersen, “Collecting and analyzing qualitative data for system dynamics: Methods and models,” *Syst. Dyn. Rev.*, vol. 19, no. 4, pp. 271–296, 2003.

[71] E. N. S. C. D. E. L. Éducation and I. P. Administratives, “G t d s ’,” pp. 1–28, 2016.

[72] G. Albaum, “The Likert Scale Revisited,” *Mark. Res. Soc. Journal.*, vol. 39, no. 2, pp. 1–21, 2018.

[73] I. Juwana, N. Muttil, and B. J. C. Perera, “Indicator-based water sustainability assessment - A review,” *Sci. Total Environ.*, vol. 438, pp. 357–371, 2012.

[74] A. C. L. Vieira, M. D. Oliveira, and C. A. Bana e Costa, “Enhancing knowledge construction processes within multicriteria decision analysis: The Collaborative Value Modelling framework,” *Omega (United Kingdom)*, Accepted Version, pp. 1–15, 2019.

[75] C. A. Bana e Costa and J. C. Vansnick, “Uma nova abordagem ao problema da construção de uma função de valor cardinal: Macbeth,” *Investig. Operacional*, vol. 15, no. Junho, pp. 15–35, 1995.

[76] P. Karande and S. Chakraborty, “A Facility Layout Selection Model using MACBETH Method,” *Proc. 2014 Int. Conf. Ind. Eng. Oper. Manag.*, no. Mcdm, pp. 17–26, 2014.

[77] P. (coord. . Santana, *Promoting population health and equity in Europe: from evidence to policy*, University. 2017.

[78] C. Duffield, “The Delphi technique: a comparison of results obtained using two expert panels,” *Int. J. Nurs. Stud.*, vol. 30, no. 3, pp. 227–237, 1993.

[79] I. Belton, A. MacDonald, G. Wright, and I. Hamlin, “Improving the practical application of the Delphi method in group-based judgment: A six-step prescription for a well-founded and defensible process,” *Technol. Forecast. Soc. Change*, vol. 147, no. July, pp. 72–82, 2019.

[80] H. A. von der Gracht, “Consensus measurement in Delphi studies. Review and implications for future quality assurance,” *Technol. Forecast. Soc. Change*, vol. 79, no. 8, pp. 1525–1536, 2012.

[81] S. Manikandan, “Measures of dispersion,” *Journal of Pharmacology and Pharmacotherapeutics, 2011.* vol. 2, p. 315.

[82] M. Riaz, “On enhanced interquartile range charting for process dispersion,” *Qual. Reliab. Eng. Int.*, vol. 31, no. 3, pp. 389–398, 2015.

[83] P. Legendre, “Species associations: The Kendall coefficient of concordance revisited,” *J. Agric. Biol. Environ. Stat.*, vol. 10 , pp. 226–245, 2005.

[84] H. Herrmann and H. Bucksch, “Coefficient of Concordance,” *Dict. Geotech. Eng. Geotech.*, pp. 248–248, 2015.

[85] J. L. Fleiss, J. C. Nee, and J. R. Landis, “Large sample variance of kappa in the case of different sets of raters,” *Psychol. Bull.*, vol. 86, pp. 974–977, 1979.

[86] T. D. Gauthier, “Detecting trends using Spearman’s rank correlation coefficient,” *Environ. Forensics*, vol. 2, pp. 359–362, 2001.

[87] D. C. H. (University of Vermont), *Statistical Methods for Psychology*, Seventh Ed. .

[88] Kelyanmoy Deb et. al, “Evolutionary multi-criterion optimization,” First International Conference, EMO 2001, Zurich, Switzerland, p. 2577.

[89] V. L.Sauter, *Decision Support Systems for Business Intelligence*. Missouri: A John wiley & SONS, INC.PUBLICATION, 2011.

[90] U. Baizyldayeva and O. Vlasov, “Multi-Criteria Decision Support Systems. Comparative Analysis,” *Middle-East J. Sci. Res.*, vol. 16, pp. 1725–1730, 2013.

[91] O. O. Ajayi, T. O. Ojeyinka, O. G. Isheyemi, and M. A. Lawal, “A Rule-Based Higher Institution of Learning Admission Decision Support System,” *J. Inf. Eng. Appl.*, vol. 4, no. 1, pp. 7–18, 2014.

[92] T. F. Cargill, M. L. Berenson, and D. M. Levine, *Basic Business Statistics: Concepts and Application.*, vol. 75, no. 372. 2006.

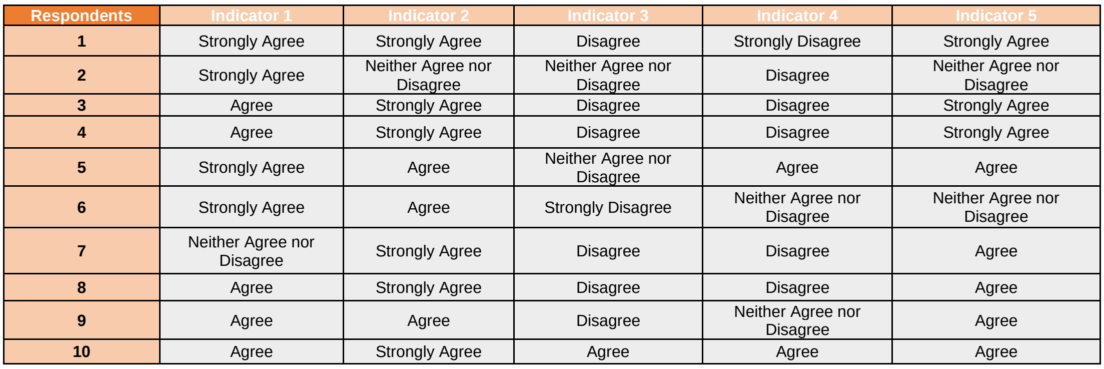
[93] “Wix Platform,” 2006. [Online]. Available: https://www.wix.com/about/us.

APPENDIX A

The visual part of the DSS, the results of the analysis performed with the data from the EURO-HEALTHY project and the mathematical programming that were not described in the section of the Results will be presented here.

1. Delphi for selection of indicators

The first tab of the tool is where the user can upload the answers obtained in the Delphi process. One example of the visual part of these tables is the following.



Respondents	Indicator 1	Indicator 2	Indicator 3	Indicator 4	Indicator 5
1	Strongly Agree	Strongly Agree	Disagree	Strongly Disagree	Strongly Agree
2	Strongly Agree	Neither Agree nor Disagree	Neither Agree nor Disagree	Disagree	Neither Agree nor Disagree
3	Agree	Strongly Agree	Disagree	Disagree	Strongly Agree
4	Agree	Strongly Agree	Disagree	Disagree	Strongly Agree
5	Strongly Agree	Agree	Neither Agree nor Disagree	Agree	Agree
6	Strongly Agree	Agree	Strongly Disagree	Neither Agree nor Disagree	Neither Agree nor Disagree
7	Neither Agree nor Disagree	Strongly Agree	Disagree	Disagree	Agree
8	Agree	Strongly Agree	Disagree	Disagree	Agree
9	Agree	Agree	Disagree	Neither Agree nor Disagree	Agree
10	Agree	Strongly Agree	Agree	Agree	Agree

Figure 54 - *Part of the first input table for the selection of indicators Delphi with a random example.*

After the three tables to upload the answers of each round, another three tables are presented to upload the answers of the participants that the user wants to analyse using a MANOVA. These tables are similar to the one presented in Figure 54. In the next spreadsheet, there are only equal tables to the ones showed above but with the data transformed in numbers. The third tab is where the results are provided. Regarding the analysis performed between rounds, to obtain the table that shows how many times each participant changed their votes it was necessary to apply the formula:

$$\boxed{= 200 - \left(CONTAR.SE(M234{:}HD234; 0) + COUNTIF(M234{:}HD234; \text{----}) + COUNTIF(M234{:}HD234; \text{-})\right).} \tag{21}$$

which subtracts the cells filled with “----", “-“ or zero of each row of the table that has the absolute differences between the votes given in two successive rounds. As previously mentioned, a table with the three participants that changed their votes the most is also displayed. To obtain them, the rankings of the “how many times each participant changed their votes” needed to be calculated using the “*RANK*” function, as demonstrated in formula 22.

$$\boxed{\begin{gathered} = IF(IndicSelecD.RESULTS!\,D235 = "----"; "----"; \\ IF(IndicSelecD.RESULTS!\,D235 = 0; "-"; \\ RANK.EQ(IndicSelecD.RESULTS!\,D235; IndicSelecD.RESULTS!\,D\$235{:}\,D\$584; 0))). \end{gathered}} \tag{22}$$

However, with this function the “jumps” with equal numerical value were ranked with the same number. To rank them with different numbers, a correction was necessary to be made as shown below.

$$\boxed{\begin{gathered} = IF(IndicSelecD.RESULTS!\,D235 = "----"; "----"; \\ IF(IndicSelecD.RESULTS!\,D235 = 0; "-"; \\ RANK.EQ(IndicSelecD.RESULTS!\,D235; IndicSelecD.RESULTS!\,D\$235{:}\,D\$584; 0) \\ +COUNTIF(IndicSelecD.RESULTS!\,\$D\$235{:}\,D235; IndicSelecD.RESULTS!\,D235) - 1)) \end{gathered}} \tag{23}$$

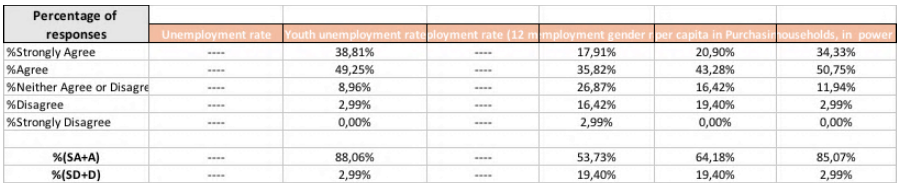
If the tool finds equal values (equal “jumps”), it ranks the second value with the first value minus one, the third value with the second value minus one and so on, which means that the early value is the one ranked with the higher number. Regarding the table of the absolute difference of votes between two consecutive rounds, to create it, a table which reported if an answer was the same between the rounds was built in the “*EXTRA*” tab, using the formula:

$$\boxed{\begin{gathered} = IF(N.Input1!\,C361 = "----"; "----"; \\ IF(OR(N.Input1!\,C6 = ""; N.Input1!\,C361 = ""); ""; \\ IF(N.Input1!\,C6 = N.Input1!\,C361; "Same"; "Different"))) \end{gathered}} \tag{24}$$

where the answers, from the same participant and respective to the same indicator were compared. If the answer was the same, the cell was filled with “*Same*”; otherwise the cell was filled with “*Different*”.

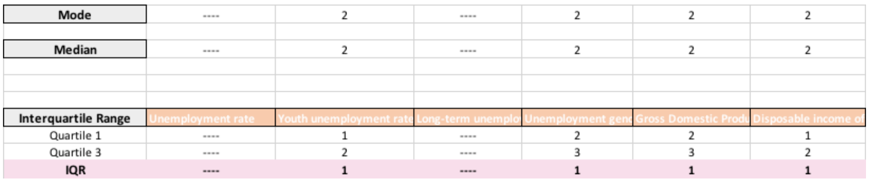
The most important results for rounds 2 and 3 are going to be presented above. It is important to note that the outcomes displayed are only for some of the indicators since it’s impossible to display a complete table with 130 indicators.

.



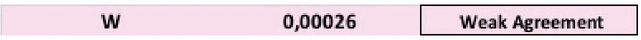
Percentage of responses	Unemployment rate	Youth unemployment rate	ployment rate (12 m	mployment gender r	per capita in Purchasin	ouseholds, in power
%Strongly Agree	----	38,81%	----	17,91%	20,90%	34,33%
%Agree	----	49,25%	----	35,82%	43,28%	50,75%
%Neither Agree or Disagre	----	8,96%	----	26,87%	16,42%	11,94%
%Disagree	----	2,99%	----	16,42%	19,40%	2,99%
%Strongly Disagree	----	0,00%	----	2,99%	0,00%	0,00%
%(SA+A)	----	88,06%	----	53,73%	64,18%	85,07%
%(SD+D)	----	2,99%	----	19,40%	19,40%	2,99%

Figure 55 - *Percentages of votes for the first 6 indicators obtained in round 2. The indicators not filled were the ones that obtained majority and were discarded for the next round evaluation.*



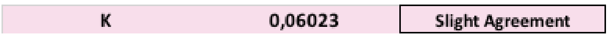
Mode	----	2	----	2	2	2
Median	----	2	----	2	2	2
Interquartile Range	Unemployment rate	Youth unemployment rate	Long-term unemplo	Unemployment genc	Gross Domestic Produ	Disposable income of
Quartile 1	----	1	----	2	2	1
Quartile 3	----	2	----	3	3	2
IQR	----	1	----	1	1	1

Figure 56 - *Mode, Median and IQR for the first 6 indicators obtained in round 2.*



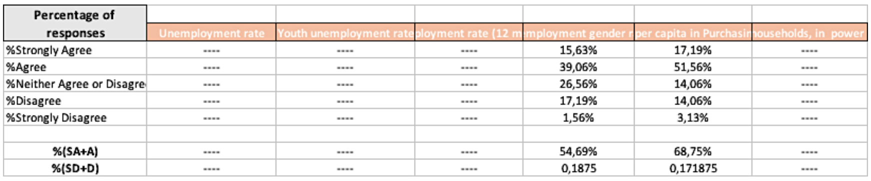
W	0,00026	Weak Agreement

Figure 57 - *Kendall’s Correlation Coefficient obtained in round 2.*



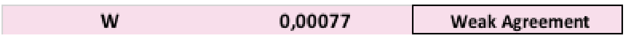
K	0,06023	Slight Agreement

***Figure 58 -** Fleiss' Kappa obtained in round 2.*



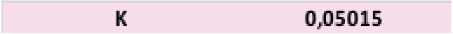
Percentage of responses	Unemployment rate	Youth unemployment rate	ployment rate (12 m	mployment gender r	per capita in Purchasin	ouseholds, in power
%Strongly Agree	----	----	----	15,63%	17,19%	----
%Agree	----	----	----	39,06%	51,56%	----
%Neither Agree or Disagre	----	----	----	26,56%	14,06%	----
%Disagree	----	----	----	17,19%	14,06%	----
%Strongly Disagree	----	----	----	1,56%	3,13%	----
%(SA+A)	----	----	----	54,69%	68,75%	----
%(SD+D)	----	----	----	0,1875	0,171875	----

***Figure 59 -** Percentages of votes for the first 6 indicators obtained in round 3. The indicators not filled were the ones that obtained majority and were discarded for the next round evaluation.*



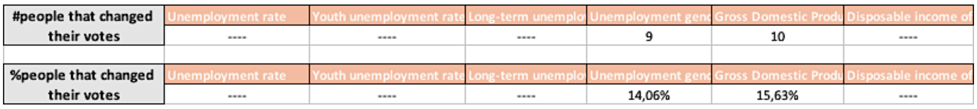
W	0,00077	Weak Agreement

***Figure 60 -** Kendall's Correlation Coefficient obtained in round 3.*



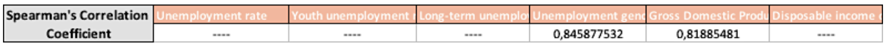
K	0,05015

***Figure 61 -** Fleiss' Kappa obtained in round 3.*



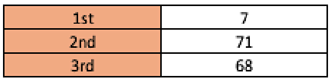
#people that changed their votes	Unemployment rate	Youth unemployment rate	Long-term unemplo	Unemployment gend	Gross Domestic Produ	Disposable income of
	----	----	----	9	10	----
%people that changed their votes	Unemployment rate	Youth unemployment rate	Long-term unemplo	Unemployment gend	Gross Domestic Produ	Disposable income of
	----	----	----	14,06%	15,63%	----

***Figure 62 -** Frequency and percentages of how participants changed their opinions between rounds 2 and 3 (for the first 6 indicators).*



Spearman's Correlation Coefficient	Unemployment rate	Youth unemployment r	Long-term unemplo	Unemployment gend	Gross Domestic Produ	Disposable income c
	----	----	----	0,845877532	0,81885481	----

***Figure 63 -** Spearman's Correlation Coefficient obtained between round 2 and 3 for the first 6 indicators.*



1st	7
2nd	71
3rd	68

***Figure 64 -** Table with the three participants that changed their opinion most times between rounds 2 and 3. For example, the "7" means "Respondent 7".*

The results provided in the summary page will not be described since they are similar to the ones described in the Results section. However, the mathematical programming not explained before is elucidated below. The analysis made in this summary page was about the higher two categories of the Likert scale. The following formula calculates if the percentage between two consecutive rounds increased, decreased or were the same. This example is for the Delphi's "*Strongly Agree*" category.

$$\boxed{\begin{array}{c} = IF(OR(C6 = "----"; C7 = "----"); "----"; \\ IF(C7 > C6; "Increased"; \\ IF(C7 = C6; "Same"; "Decreased"))) \end{array}} \qquad (25)$$

After, two formulas were created: if the previous answer was "*Increased*", the percentage of how much increased was calculated; on the other hand, if the answer was "*Decreased*", the percentage of how much decreased was calculated.

$$\boxed{= IF(C10 = \text{"}----\text{"}; \text{"}----\text{"}; IF(C10 = \text{"}Increased\text{"}; C7 - C6; \text{"}-\text{"}))} \tag{26}$$

$$\boxed{= IF(C10 = \text{"}----\text{"}; \text{"}----\text{"}; IF(C10 = \text{"}Decreased\text{"}; ABS(C7 - C6); \text{"}-\text{"})).} \tag{27}$$

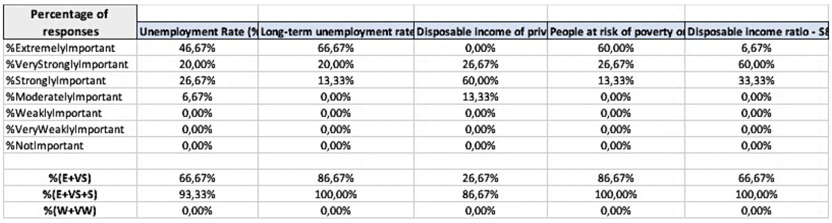
After, the *Kendall's Correlation Coefficient* and *Fleiss' Kappa* values from the two consecutive rounds were put side by side to simply compare them using a small sentence that describes if the values increased or decreased between the rounds.

$$\boxed{\begin{gathered} = IF(E34 > E33; \\ \text{" } Agreement\ between\ respondents\ INCREASED\ from\ Round\ 1\ to\ Round\ 2\ according\ to\ W\text{"}; \\ \text{" } Agreement\ between\ respondents\ DECREASED\ from\ Round\ 1\ to\ Round\ 2\ according\ to\ W\text{"}). \end{gathered}} \tag{28}$$

Additionally, the percentages of who changed votes between successive rounds were set side by side to understand how it varied during all the process and concluding if the inter-reliability increased or decreased. The same was done with the *Spearman's Correlation Coefficient* obtained between successive rounds concluding whether the stability increased or decreased throughout the Delphi process. The mathematical programming used was the same as explained in the results section.

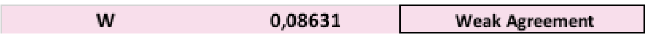
2. Delphi for weighting judgments

The most important results for rounds 2 and 3 and between these rounds are presented below.



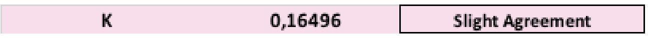
Percentage of responses	Unemployment Rate (%	Long-term unemployment rate	Disposable income of priv	People at risk of poverty o	Disposable income ratio - S8
%ExtremelyImportant	46,67%	66,67%	0,00%	60,00%	6,67%
%VeryStronglyImportant	20,00%	20,00%	26,67%	26,67%	60,00%
%StronglyImportant	26,67%	13,33%	60,00%	13,33%	33,33%
%ModeratelyImportant	6,67%	0,00%	13,33%	0,00%	0,00%
%WeaklyImportant	0,00%	0,00%	0,00%	0,00%	0,00%
%VeryWeaklyImportant	0,00%	0,00%	0,00%	0,00%	0,00%
%NotImportant	0,00%	0,00%	0,00%	0,00%	0,00%
%(E+VS)	66,67%	86,67%	26,67%	86,67%	66,67%
%(E+VS+S)	93,33%	100,00%	86,67%	100,00%	100,00%
%(W+VW)	0,00%	0,00%	0,00%	0,00%	0,00%

Figure 65 - *Percentages of votes for the first 5 indicators obtained in round 2.*



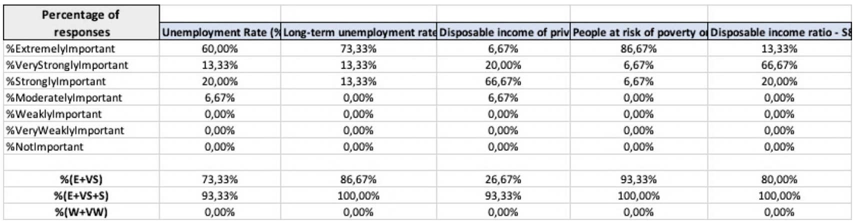
W	0,08631	Weak Agreement

Figure 66 - *Kendall's Correlation Coefficient obtained in round 2.*



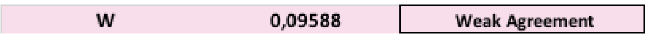
K	0,16496	Slight Agreement

Figure 67 - *Fleiss' Kappa obtained in round 2.*



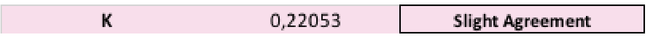
Percentage of responses	Unemployment Rate (%	Long-term unemployment rate	Disposable income of priv	People at risk of poverty o	Disposable income ratio - S8
%ExtremelyImportant	60,00%	73,33%	6,67%	86,67%	13,33%
%VeryStronglyImportant	13,33%	13,33%	20,00%	6,67%	66,67%
%StronglyImportant	20,00%	13,33%	66,67%	6,67%	20,00%
%ModeratelyImportant	6,67%	0,00%	6,67%	0,00%	0,00%
%WeaklyImportant	0,00%	0,00%	0,00%	0,00%	0,00%
%VeryWeaklyImportant	0,00%	0,00%	0,00%	0,00%	0,00%
%NotImportant	0,00%	0,00%	0,00%	0,00%	0,00%
%(E+VS)	73,33%	86,67%	26,67%	93,33%	80,00%
%(E+VS+S)	93,33%	100,00%	93,33%	100,00%	100,00%
%(W+VW)	0,00%	0,00%	0,00%	0,00%	0,00%

Figure 68 - *Percentages of votes for the first 5 indicators obtained in round 3.*



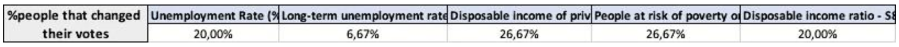
W	0,09588	Weak Agreement

Figure 69 - *Kendall's Correlation Coefficient obtained in round 3.*



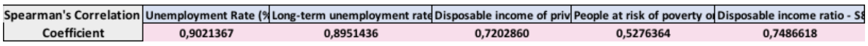
K	0,22053	Slight Agreement

Figure 70 - *Fleiss' Kappa obtained in round 3.*



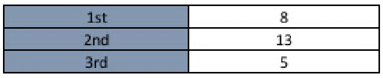
%people that changed their votes	Unemployment Rate (%	Long-term unemployment rate	Disposable income of priv	People at risk of poverty o	Disposable income ratio - S8
	20,00%	6,67%	26,67%	26,67%	20,00%

Figure 71 - *Percentage of how many people changed their votes between rounds 2 and 3.*



Spearman's Correlation Coefficient	Unemployment Rate (%	Long-term unemployment rate	Disposable income of priv	People at risk of poverty o	Disposable income ratio - S8
	0,9021367	0,8951436	0,7202860	0,5276364	0,7486618

Figure 72 - *Spearman's Correlation Coefficient obtained between rounds 2 and 3.*



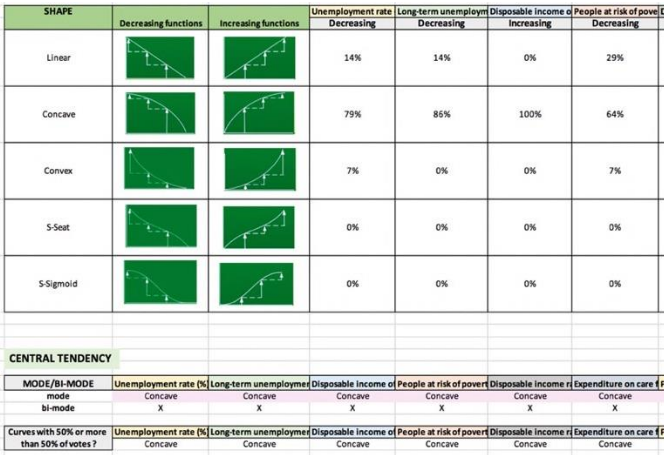
1st	8
2nd	13
3rd	5

Figure 73 - *Table with the three participants that changed their opinion most times.*

The mathematical programming of the summary page was the same as the one performed for the Delphi for selection of indicators but related with the three higher categories of the 7-level MACBETH scale.

3. Delphi to shape a value function

For the value function Delphi, the results of the first round were described in the results section. Here, we will provide the results obtained in rounds 2 and 3 and between them along with mathematical programming not described in the results section.



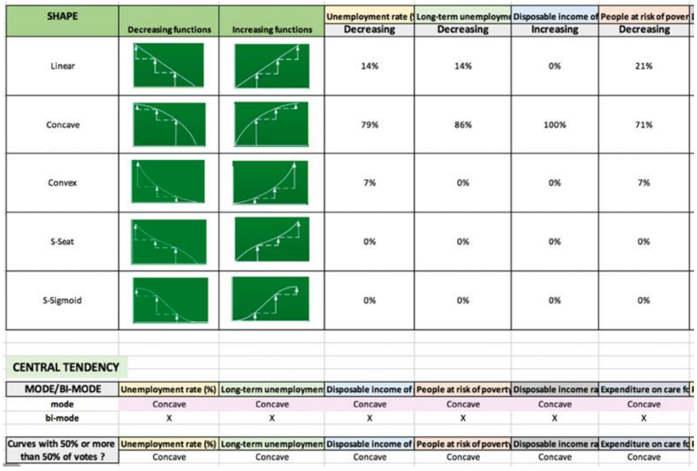
SHAPE	Decreasing functions	Increasing functions	Unemployment rate Decreasing	Long-term unemploym Decreasing	Disposable income o Increasing	People at risk of pove Decreasing
Linear			14%	14%	0%	29%
Concave			79%	86%	100%	64%
Convex			7%	0%	0%	7%
S-Seat			0%	0%	0%	0%
S-Sigmoid			0%	0%	0%	0%

CENTRAL TENDENCY

MODE/BI-MODE	Unemployment rate (%)	Long-term unemploymer	Disposable income of	People at risk of povert	Disposable income ra	Expenditure on care f
mode	Concave	Concave	Concave	Concave	Concave	Concave
bi-mode	X	X	X	X	X	X

Curves with 50% or more than 50% of votes ?	Unemployment rate (%)	Long-term unemploymer	Disposable income of	People at risk of povert	Disposable income ra	Expenditure on care f
	Concave	Concave	Concave	Concave	Concave	Concave

Figure 74 - *Results obtained for the second round for the first 4 indicators: percentages of each type of shape, mode and bi-mode and curves that obtained more than 50% of the votes.*



SHAPE	Decreasing functions	Increasing functions	Unemployment rate (Decreasing	Long-term unemployme Decreasing	Disposable income of Increasing	People at risk of pover Decreasing
Linear			14%	14%	0%	21%
Concave			79%	86%	100%	71%
Convex			7%	0%	0%	7%
S-Seat			0%	0%	0%	0%
S-Sigmoid			0%	0%	0%	0%

CENTRAL TENDENCY

MODE/BI-MODE	Unemployment rate (%)	Long-term unemploymen	Disposable income of	People at risk of poverty	Disposable income ra	Expenditure on care fo
mode	Concave	Concave	Concave	Concave	Concave	Concave
bi-mode	X	X	X	X	X	X

Curves with 50% or more than 50% of votes ?	Unemployment rate (%)	Long-term unemploymen	Disposable income of	People at risk of poverty	Disposable income ra	Expenditure on care fo
	Concave	Concave	Concave	Concave	Concave	Concave

Figure 75 - *Results for the first 4 indicators obtained in round 3: percentages of each type of shape, mode and bi-mode and curves that obtained more than 50% of the votes.*

The mathematical programming used to calculate these last results presented in images 74 and 75 are shown below:

- Mode:

$$\boxed{\begin{gathered} =\ IF(F\$475 = \text{"} - - - - \text{"}; \text{"} - - - - \text{"}; \\ IF(MAX(F\$475{:}\,F\$479) = F475; \text{"}Linear\text{"}; \\ IF(MAX(F475{:}\,F479) = F476; \text{"}Concave\text{"}; \\ IF(MAX(F475{:}\,F479) = F477; \text{"}Convex\text{"}; \\ IF(MAX(F475{:}\,F479) = F478; \text{"}S - Seat\text{"}; \\ IF(MAX(F475{:}\,F479) = F479; \text{"}S - Sigmoid\text{"}))))) \end{gathered}} \qquad (29)$$

where the cells from $F475$ to $F479$ are the percentages for each shape. The tool calculates the maximum value and compares each one of the percentages (for each shape) with this maximum; when it finds the one that matches, it picks the correspondent name of the shape.

- Bi-mode:

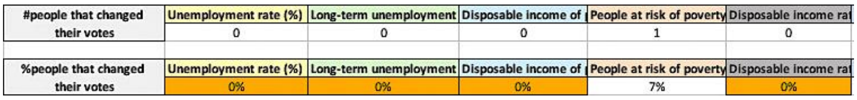
$$\boxed{\begin{gathered}
= IF(D485 = "----"; "----"; IF(D485 = \$C\$475; \\
IF(F476 = EXTRA3!\,C\$55; "Concave"; \\
IF(ValueFuncD.RESULTS!\,F477 = EXTRA3!\,C\$55; "Convex"; \\
IF(ValueFuncD.RESULTS!\,F478 = EXTRA3!\,C\$55; "S - Seat"; \\
IF(ValueFuncD.RESULTS!\,F479 = EXTRA3!\,C\$55; "S - Sigmoid"; \\
IF(D485 = "Linear"; "X"; \\
IF(ValueFuncD.RESULTS!\,D485 = ValueFuncD.RESULTS!\,\$C\$476; \\
IF(ValueFuncD.RESULTS!\,F475 = EXTRA3!\,C\$55; "Linear"; \\
IF(ValueFuncD.RESULTS!\,F477 = EXTRA3!\,C\$55; "Convex"; \\
IF(ValueFuncD.RESULTS!\,F478 = EXTRA3!\,C\$55; "S - Seat"; \\
IF(ValueFuncD.RESULTS!\,F479 = EXTRA3!\,C\$55; "S - Sigmoid"; \\
IF(ValueFuncD.RESULTS!\,D485 = \$C\$477; \\
IF(ValueFuncD.RESULTS!\,F475 = EXTRA3!\,C\$55; "Linear"; \\
IF(ValueFuncD.RESULTS!\,F476 = EXTRA3!\,C\$55; "Concave"; \\
IF(ValueFuncD.RESULTS!\,F478 = EXTRA3!\,C\$55; "S - Seat"; \\
IF(ValueFuncD.RESULTS!\,F479 = EXTRA3!\,C\$55; "S - Sigmoid"; \\
IF(ValueFuncD.RESULTS!\,D485 = \$C\$478; \\
IF(ValueFuncD.RESULTS!\,F475 = EXTRA3!\,C\$55; "Linear"; \\
IF(ValueFuncD.RESULTS!\,F476 = EXTRA3!\,C\$55; "Concave"; \\
IF(ValueFuncD.RESULTS!\,F477 = EXTRA3!\,C\$55; "Convex"; \\
IF(ValueFuncD.RESULTS!\,F479 = EXTRA3!\,C\$55; "S - Sigmoid"; \\
IF(ValueFuncD.RESULTS!\,D485 = \$C\$477; \\
IF(ValueFuncD.RESULTS!\,F475 = EXTRA3!\,C\$55; "Linear"; \\
IF(ValueFuncD.RESULTS!\,F476 = EXTRA3!\,C\$55; "Concave"; \\
IF(ValueFuncD.RESULTS!\,F477 = EXTRA3!\,C\$55; "Convex"; \\
IF(ValueFuncD.RESULTS!\,F478 \\
= EXTRA3!\,C\$55; "S - Seat"))))))))))))))))))))))))); "X"))
\end{gathered}} \tag{30}$$

The idea was to fix the maximum value (mode) and to check which from the others had the same value. If none of them is equal, then there is no bi-mode and a “X” will appear as an answer.

- Curves with more than 50% of the votes:

$$\boxed{\begin{gathered}
= IF(F475 = "----"; "----"; \\
IF(F475 \geq 0{,}5; \$C\$475; \\
IF(F476 \geq 0{,}5; \$C\$476; \\
IF(F477 \geq 0{,}5; \$C\$477; \\
IF(F478 \geq 0{,}5; \$C\$478; \\
IF(F479 >= 0{,}5; \$C\$479; "X")))))).
\end{gathered}} \tag{31}$$

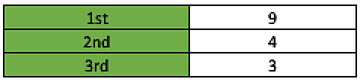
Here, all the percentages are compared with a value of 0,5, which corresponds to 50%. If there is none higher than 0,5, there are no shapes with more than 50% of the votes and a “X” is returned.



#people that changed their votes	Unemployment rate (%)	Long-term unemployment	Disposable income of	People at risk of poverty	Disposable income rat
	0	0	0	1	0

%people that changed their votes	Unemployment rate (%)	Long-term unemployment	Disposable income of	People at risk of poverty	Disposable income rat
	0%	0%	0%	7%	0%

Figure 76 - *Frequency and percentage of how many participants changed their votes between rounds 2 and 3, for the first 5 indicators.*



1st	9
2nd	4
3rd	3

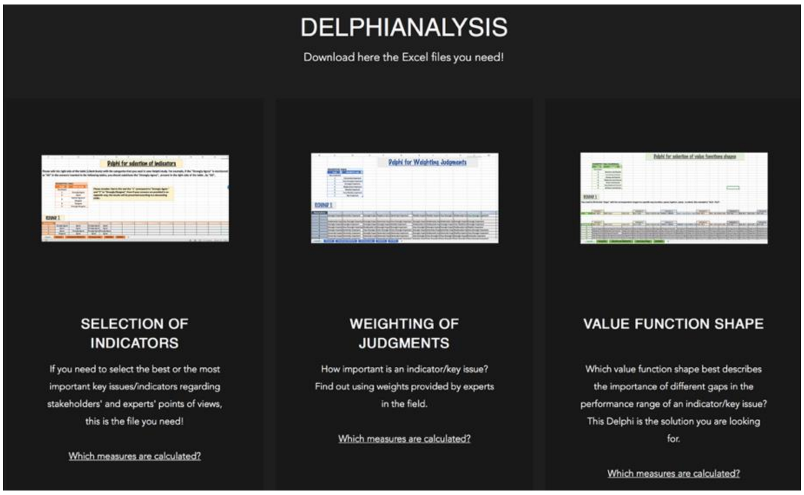
Figure 77 - *The three participants that changed their opinion the most between rounds 2 and 3.*

The information provided in the “summary page” was shown in part in the Results section. The same procedures were taken with all the categories. The intra-reliability and stability could not be studied in this type of Delphi since it does not deal with numerical data.

APPENDIX B

A webpage was created to help the Delphi analysts who aim to use the implemented DSS. It provides all the Excel files of the tool and some step guides about how to use the tool correctly. Moreover, it presents an introduction to Delphi processes and a page to contact us if they have some questions.

The webpage has a section that describes my thesis work and explains how the *DelphiAnalysis* DSS emerged. The section " *DelphiAnalysis* TOOL" is where the users can download the Excel files of the DSS and see which statistical measures are implemented in each case. Therefore, the user has the option to see if the statistical measures are interesting for the analysis he wants to perform, before downloading the files. Three different Excel files are presented, each relative to one of the three types of Delphi processes covered in this dissertation.



DELPHIANALYSIS

Download here the Excel files you need!

SELECTION OF INDICATORS

If you need to select the best or the most important key issues/indicators regarding stakeholders' and experts' points of views, this is the file you need!

Which measures are calculated?

WEIGHTING OF JUDGMENTS

How important is an indicator/key issue? Find out using weights provided by experts in the field.

Which measures are calculated?

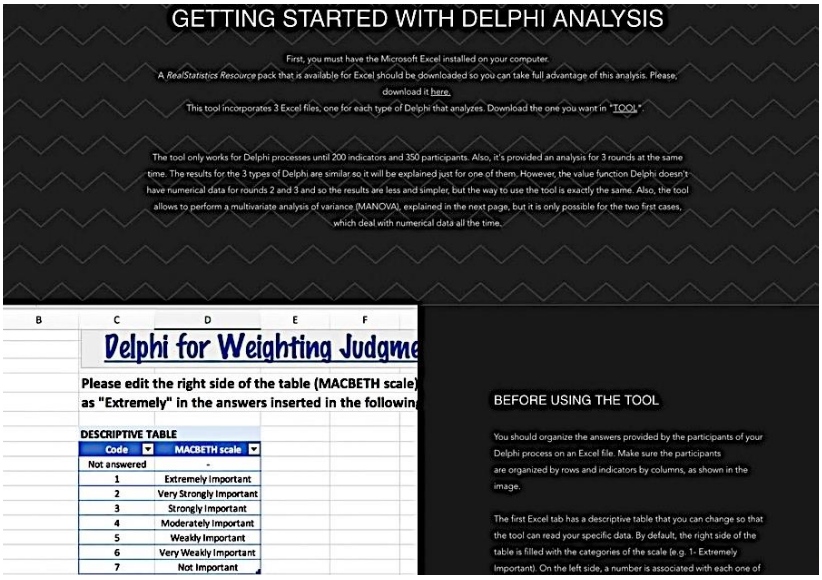
VALUE FUNCTION SHAPE

Which value function shape best describes the importance of different gaps in the performance range of an indicator/key issue? This Delphi is the solution you are looking for.

Which measures are calculated?

Figure 78 - *Where to download the three Excel files of the tool.*

The explanation of how to use the tool is presented in another section called "How to use" where the main and most important steps to use the DSS correctly are explained. Some videos are presented as well to elucidate the user more efficiently. There is a section that explains how to perform MANOVA tests. Basically, it explains how to apply the formulas for the *Wilks' Lambda* test and its p-value to a specific range of data. Some figures are provided below just to show how the webpage interface looks like.



GETTING STARTED WITH DELPHI ANALYSIS

First, you must have the Microsoft Excel installed on your computer.
A *RealStatistics Resource* pack that is available for Excel should be downloaded so you can take full advantage of this analysis. Please, download it here.
This tool incorporates 3 Excel files, one for each type of Delphi that analyzes. Download the one you want in "TOOL".

The tool only works for Delphi processes until 200 indicators and 350 participants. Also, it's provided an analysis for 3 rounds at the same time. The results for the 3 types of Delphi are similar so it will be explained just for one of them. However, the value function Delphi doesn't have numerical data for rounds 2 and 3 and so the results are less and simpler, but the way to use the tool is exactly the same. Also, the tool allows to perform a multivariate analysis of variance (MANOVA), explained in the next page, but it is only possible for the two first cases, which deal with numerical data all the time.

Delphi for Weighting Judgme

Please edit the right side of the table (MACBETH scale)
as "Extremely" in the answers inserted in the followin

DESCRIPTIVE TABLE

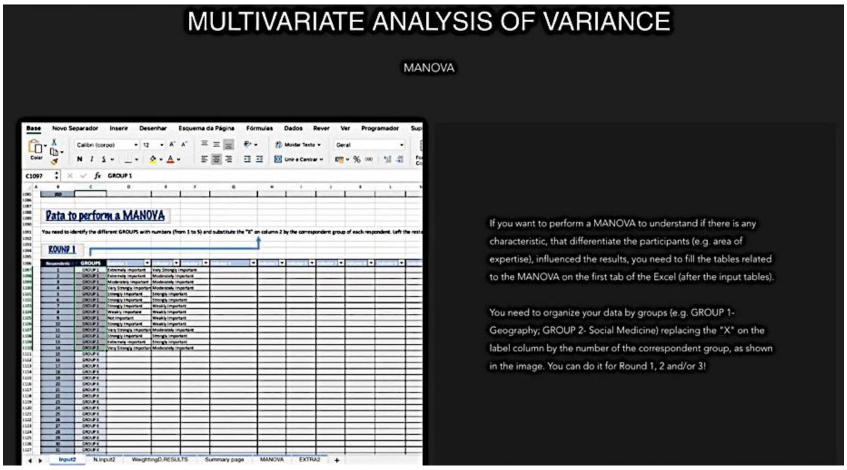
Code	MACBETH scale
Not answered	-
1	Extremely Important
2	Very Strongly Important
3	Strongly Important
4	Moderately Important
5	Weakly Important
6	Very Weakly Important
7	Not Important

BEFORE USING THE TOOL

You should organize the answers provided by the participants of your Delphi process on an Excel file. Make sure the participants are organized by rows and indicators by columns, as shown in the image.

The first Excel tab has a descriptive table that you can change so that the tool can read your specific data. By default, the right side of the table is filled with the categories of the scale (e.g. 1- Extremely Important). On the left side, a number is associated with each one of

***Figure 79 -** First part of the "How to use the TOOL" section.*



MULTIVARIATE ANALYSIS OF VARIANCE

MANOVA

If you want to perform a MANOVA to understand if there is any characteristic, that differentiate the participants (e.g. area of expertise), influenced the results, you need to fill the tables related to the MANOVA on the first tab of the Excel (after the input tables).

You need to organize your data by groups (e.g. GROUP 1- Geography; GROUP 2- Social Medicine) replacing the "X" on the label column by the number of the correspondent group, as shown in the image. You can do it for Round 1, 2 and/or 3!

***Figure 80 -** "MANOVA" part in the "how to use" section.*

Finally, there is a "Contact Us" section so the users can contact us if they have any additional questions about the tool.

APPENDIX C

As explained in the methodology, a questionnaire was developed about the tool and approaching the web page issue. This questionnaire is intended for professionals dealing with Delphi processes and who can provide their opinion to validate the implemented tool. The developed questionnaire is presented below.

QUESTIONNAIRE TO VALIDATE THE *DELPHIANALYSIS* TOOL

This questionnaire aims to collect the opinion of some Delphis experts about the implemented tool and its application, so as to understand the user experience and how to improve it. The tool calculates statistical measures that allow the evaluation of three types of Delphi widely used in healthcare: Delphis for selection of indicators, Delphis of weights and Delphis of value functions.

Please note that the analysis of responses will be anonymous.

Area of Expertise:

INSTRUCTIONS

1) FOR THE MULTIPLE CHOICE, COMPLETE THE QUESTIONNAIRE BY TICKING ONLY ONE OF THE OPTIONS PRESENTED WITH AN “X”.
2) FOUR OPEN QUESTIONS ARE PRESENT. PLEASE PROVIDE AN ANSWER WITH NO MORE THAN 5 LINES.

QUESTIONS

CLOSE-ENDED QUESTIONS

1. How much do you work with Delphi processes?

( ) Daily

( ) Almost every day

( ) Sometimes

( ) Rarely

( ) Never

2. How much do you work with the 3 specific types of Delphi described above?

( ) Daily

( ) Almost every day

( ) Sometimes

( ) Rarely

( ) Never

3. Did you understand how to use the tool by yourself or you needed the web page to guide you?

( ) Yes, I understood by myself

( ) No, I made some use of the web page to help me

( ) No, I needed to use intensively the web page to help me

4. Please provide your opinion regarding “The website is useful”.

( ) Strongly Agree

( ) Agree

( ) Neither Agree nor Disagree

( ) Disagree

( ) Strongly Disagree

5. Please provide your opinion regarding “The tool is user-friendly”.

( ) Strongly Agree

( ) Agree

( ) Neither Agree nor Disagree

( ) Disagree

( ) Strongly Disagree

6. Please provide your opinion regarding “The tool is useful”.

( ) Strongly Agree

( ) Agree

( ) Neither Agree nor Disagree

( ) Disagree

( ) Strongly Disagree

7. Would you use this tool to help you?

( ) Yes

( ) No

OPEN-ENDED QUESTIONS

Do you have specific comments or suggestions about the tool?

What do you think about the statistical measures implemented? Do you think there were best choices, or could some better ones have been chosen? If so, which ones?

Please enumerate some pros and some cons that you find in the tool?

What do you think it can be done to make it better in future work? Do you have any further suggestions?