Probing the Power of Apollo

Methodological Challenges and Opportunities of the Delphi Method for Developing Rankings

Jurian V. Meijering

Probing the Power of Apollo:

Methodological Challenges and Opportunities
of the Delphi Method for Developing Rankings

Jurian V. Meijering

Thesis committee

Promotor
Prof. Dr A. van den Brink
Professor of Landscape Architecture
Wageningen University

Co-promotor
Dr H. Tobi
Associate professor, Biometris
Wageningen University

Other members
Prof. Dr I.G. Klugkist, Utrecht University
Prof. Dr D. Borsboom, University of Amsterdam
Prof. Dr R. Leemans, Wageningen University
Prof. Dr K.J. Jørgensen, Norwegian University of Life Sciences, Ås, Norway

This research was conducted under the auspices of the Wageningen School of Social Sciences (WASS)

Probing the Power of Apollo:

Methodological Challenges and Opportunities of the Delphi Method for Developing Rankings

Jurian V. Meijering

Thesis

submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Tuesday 13 December 2016
at 4 p.m. in the Aula.

Jurian V. Meijering
Probing the Power of Apollo:
Methodological Challenges and Opportunities of the Delphi Method for Developing Rankings
184 pages.

PhD thesis, Wageningen University, Wageningen, NL (2016)
With references, with summary in English

DOI: 10.18174/389482
ISBN: 978-94-6257-918-7

Acknowledgements

Since the start of this PhD project I wondered how it would feel to write the acknowledgements of my thesis. Now, more than 5 years later it has finally come to this and I can tell you, it feels good. Although I thoroughly enjoyed the journey, it was not without any trouble. I'm therefore grateful that there were people who walked along with me and gave me support, advice, or just a listening ear. It is here that I would like to thank those people.

First of all, I would like to thank Hilde Tobi. When I started to work as a full-time lecturer within the Research Methodology Group of Wageningen University back in December 2009 you coached me, gave me the opportunity to obtain my University Teaching Qualification, and helped me to write a successful PhD research proposal. Once that proposal was accepted, you became my daily supervisor and co-promotor. In this capacity you proved to be invaluable. Almost every week we had a meeting in which you challenged my initial ideas, gave feedback on my work, and shared your vast knowledge of research methodology. In doing so you did not only improve the quality of the research reported in this thesis, but also shaped me into the research methodologist I am today.

Another important person, without whom this thesis would not have been possible, is Adri van den Brink. After my PhD proposal got accepted, I searched for a promotor capable and willing to supervise the project. Soon we came in contact with each other and after a fruitful meeting we decided to work together. Adri, throughout the project I got to know you as a dedicated promotor who closely monitored my progress, offered clear advice about the course of my research, and encouraged me to keep on thinking about the common thread running through the individual research papers. I'm also grateful for your enthusiasm about the Delphi method which resulted in a joint research paper as well as a chapter in your book about research methods in landscape architecture.

Aside from my promotor and co-promotor I would like to thank several other persons who contributed to the realization of this thesis. First, Kristine Kern who was involved in the project from the start as an expert on urban sustainability issues. I especially won't forget your company and support during my stay at the Leibniz Institute for Research on Society and Space in Germany. Second, I would like to thank the members of the reading committee: Professor Irene Klugkist, professor Denny Borsboom, professor Rick Leemans, and professor Karsten Jørgensen. Thank you so much for reviewing this thesis. Third, I would like to thank my colleagues of the Research Methodology Group, old and new, with whom I worked together during this PhD project: Jarl, Ruud, Peter, Vanessa, Matthijs, Giel, Willy, Fannie, Gerben, Johan, Jennifer, and Nicolette. In one way or another you all made a worthwhile contribution, even if it was just by providing for a pleasant work environment. As a group without a professor we had a hard time explaining our relevance for the education

and research at Wageningen University. I have good hopes that our move to Biometris will open up new opportunities for strengthening the role of research methodology within this university and our position as independent research methodologists.

Of course, I would never have been able to conduct this PhD project and write this thesis without the continuous support and love of my family and friends. Especially my mom, dad, and big brother Ramon. Thank you for a wonderful childhood and for always being there. I'm really blessed that to this day our family is still complete and that there is a happy home to which I can always return. It is an inspiration and an example of how to raise my own children. I'm also grateful for the second family I became part of when I met my wife. Dear Klaas and Margreet, thank you for making me feel welcome in your home and for your interest in my research.

Finally, I want to thank my wife Metske. At the start of this PhD project we moved to our new home in Ede. Here we got our two lovely daughters Milou and Feline. Although our life together has been great so far, we also both know that having a busy job and raising children can be stressful at times. Fortunately, we have each other when things go awry. I'm immensely grateful for your emotional support and understanding when I had to lock myself up in the attic or wasn't home in time for dinner. Without your warmth, love, and joy I would not have been where I am now.

Milou and Feline, although you are much too young to understand any of this, know that your existence has made my life so much more meaningful. Nothing puts things more into perspective than coming home and getting hugs, kisses, and smiles from you two. I would not trade those moments for anything in the world, not even a PhD!

Table of Contents

Chapter 1	**Introduction**		**11**
	1.1	Prologue	12
	1.2	What is the Delphi method?	12
	1.3	What are rankings?	16
	1.4	Using Delphi to develop rankings	17
	1.5	Overview of chapters	20
Chapter 2	**Quantifying the development of agreement among experts in Delphi studies**		**23**
	2.1	Introduction	25
	2.2	The indices under study	26
	2.3	Outline of the simulation study	30
	2.4	Results	32
	2.5	Discussion and conclusion	36
Chapter 3	**Exploring research priorities in landscape architecture: An international Delphi study**		**41**
	3.1	Introduction	43
	3.2	Methods	44
	3.3	Results: Most important domains	50
	3.4	Results: Most useful domains	55
	3.5	Discussion and conclusion	59
Chapter 4	**The effect of controlled opinion feedback on Delphi features: Mixed messages from a real-world Delphi experiment**		**65**
	4.1	Introduction	67
	4.2	Theory and hypotheses	68
	4.3	Methods	70
	4.4	Results	74
	4.5	Discussion and conclusion	77

Chapter 5 **Identifying the methodological characteristics of European green city rankings** **83**

5.1 Introduction 85
5.2 Literature review on city ranking methodology 86
5.3 Methods 91
5.4 Results 94
5.5 Discussion 103
5.6 Conclusion 105

Chapter 6 **Defining and measuring urban sustainability in Europe: A Delphi study on identifying its most relevant components** **109**

6.1 Introduction 111
6.2 Methods 112
6.3 Results 118
6.4 Discussion 123
6.5 Conclusion 126

Chapter 7 **Feeding back experts' own initial ratings in Delphi studies: Effects on opinion change and the level of agreement** **129**

7.1 Introduction 131
7.2 Materials and methods 133
7.3 Results 137
7.4 Discussion 140
7.5 Conclusion 142

Chapter 8 **Discussion and conclusions** **145**

8.1 Introduction 146
8.2 Overview of main findings 147
8.3 Limitations 150
8.4 Scientific contribution 151
8.5 Contribution to society 153
8.6 Future research 154

References 158
Summary 175
About the author 181
Completed training and supervision plan 182

Chapter 1

Introduction

1.1 Prologue

My first acquaintance with the Delphi method dates back to September 2003. At that time I studied Applied Communication Science at the University of Twente. During a lecture the Delphi method was explained to me as a data-collection procedure that allows experts to reach an agreement on some topic in several subsequent rounds. At first, I was impressed by the systematic nature of the method and its potential for dealing with complex societal problems. However, I was also told about the various methodological challenges surrounding the method. For instance, there were no guidelines on how to measure and report the level of agreement among experts. As a result, many Delphi studies simply reported that eventually consensus among experts was achieved. Something that I found difficult to accept.

Eight years later I became a PhD candidate within the Research Methodology Group of Wageningen University. Here my fascination for rankings was sparked. In our society rankings are everywhere and people generally believe that they give us a simple and decent overview of among others the best performing universities and most sustainable cities. Consequently, rankings shape people's decisions, for example regarding where to study or where to live, and are commonly used by organizations and governments to monitor the effectiveness of policies. It is therefore astonishing and also worrying that the methodological characteristics of rankings are rarely considered. By reading the literature I soon learned that within the ranking development process various methodological decisions need to be made that may severely influence ranking results (see for example Lun et al., 2006). In this regard some researchers mentioned that certain decisions may best be made and substantiated by consulting experts (see for example Morse & Fraser, 2005). At that moment I recalled the Delphi method and thought that it may provide opportunities for developing rankings. I searched the literature, but no research seemed to exist that explored these opportunities. Surprisingly, I did found out that little research was done into the methodological challenges of the Delphi method. I realized that to effectively use the Delphi method for developing rankings, these challenges also needed to be dealt with. Thus, together with my promotor and co-promotor, I decided to dedicate my PhD project to examining the methodology of the Delphi method and its potential for developing rankings.

1.2 What is the Delphi method?

The Delphi method is named after the Delphic Oracle, an ancient shrine whose ruins can still be admired near the Greek town of Delphi. As explained by Parke (1939), the ancient Greeks believed that the god Apollo took over the Oracle from the earth goddess Gé by slaying her visible manifestation in the form of the monstrous serpent Python. Ordinary individuals, kings, and ambassadors of communities would visit the Oracle to consult Apollo about

various important decisions, such as whom to marry or whether or not to wage a war. The consultation of Apollo was only possible on nine days of the year. On those days a long queue of enquirers waited outside Apollo's temple. An enquirer was allowed to enter the inner sanctuary of the temple after purifying himself with holy water, offering a sacred cake, and sacrificing a sheep or goat. Once inside, the enquirer did not directly speak to Apollo, but instead asked a question through a female priest called the Pythia (whose name is a reminiscent of the serpent Python). The Pythia, seated on a tripod and in a state of trance, then shouted Apollo's more or less coherent advice. According to Marchais-Roubelat and Roubelat (2011) the Delphi method is more than just a namesake of the Delphic Oracle. Like the Oracle, the Delphi method is used to help contemporary enquirers, such as academics, policy makers, and corporate managers, make and justify important decisions. However, instead of consulting the god Apollo, enquirers now consult a group of experts in the belief that the combined opinions of these experts will be superior to any single opinion.

The Delphi method as we know it today was developed in the 1950s at the RAND corporation by Olaf Helmer, Norman Dalkey, Ted Gordon, and other associates on behalf of the United States Air Force (Dalkey et al., 1969; Dalkey & Helmer, 1963; Linstone & Turoff, 2011). The original objective of the method was to "obtain the most reliable consensus of opinion of a group of experts" (Dalkey & Helmer, 1963, p. 458). In the first publicly reported Delphi study experts were asked to estimate, from the viewpoint of a Soviet strategic planner, the least number of atomic bombs necessary to basically destroy the United States munitions output (Dalkey & Helmer, 1963). Because of this, the Delphi method has been called a positive spin-off from the Cold War (Hargie & Tourish, 2000).

Since its public introduction in the 1960s the classical Delphi method has evolved into various types, each having a particular design and aim (Hasson & Keeney, 2011). Examples are the so-called modified, decision, and technological Delphi method. Although one could therefore state that *the* Delphi method does not actually exist, Linstone and Turoff (1975) provided a general definition that seems to underlie most types:

> "Delphi may be characterized as a method for structuring a group communication process so that the process is effective in allowing a group of individuals, as a whole, to deal with a complex problem" (p. 3).

They further explained that the use of the method is especially suited for problems for which accurate data are not available or too costly to obtain. Initially, the Delphi method was mainly used to deal with forecasting problems, for example to estimate the year in which certain technological developments would occur. Later, the method was also applied to deal with complex societal problems related to for example the environment and health (Linstone & Turoff, 1975). In this regard the Delphi method is often aimed at allowing experts to achieve agreement on various topics (Hasson & Keeney, 2011). Nonetheless, as

Linstone and Turoff (2011) stressed, the Delphi method should not aim to force an agreement among experts. Finding out on which aspects of a topic experts disagree may be just as worthwhile. Some Delphi types even explicitly aim to maximize the range of expert opinions (Steinert, 2009) or to elicit opposing views (Turoff, 1970).

Despite the existence of different types, the Delphi method has several defining characteristics (Dalkey et al., 1969; Diamond et al., 2014; Keeney et al., 2006; Landeta, 2006; Linstone & Turoff, 1975; Powell, 2003; Rowe et al., 1991). Any Delphi study consists of at least two subsequent rounds in which data are collected from a single sample of experts. In each round experts individually give their opinion on the topic of interest, usually by means of a standardized questionnaire. Experts remain anonymous throughout the study and are not supposed to directly communicate with each other. Instead, after each round the study moderator provides experts with controlled opinion feedback: a summary of the results from the previous round. Based on this feedback experts are allowed to change their opinion in the next round. This process may continue for several more rounds until experts' responses have stabilized, a certain level of agreement has been achieved, or a pre-specified number of rounds has been completed (usually not more than three). The final result of a Delphi study is a so called statistical group response in which the individual opinions of all experts who participated in the final round are combined.

The rationale of the Delphi method is based on the age-old adage: 'two heads are better than one' (Dalkey et al., 1969). Additionally, the idea is that through anonymity and controlled opinion feedback experts are able to share their knowledge and views without undue social pressures associated with ordinary (face-to-face) group discussions (Rowe et al., 1991). Nonetheless, within the general framework of the Delphi method many specific decisions still need to be made that may influence the functioning and thereby the results of a Delphi study. Several important decisions involve the sampling of experts, the design of the first round questionnaire, the provision of controlled opinion feedback, and the measurement of agreement.

In Delphi studies experts are usually selected on a non-random basis. To prevent selection bias, users of the Delphi method need to decide who is an expert by establishing selection criteria (Keeney et al., 2006). In this regard it has been recommended to select experts who have knowledge of and experience with the topic of interest and also have the capacity, willingness, and time to participate (Keeney et al., 2006; Powell, 2003; Skulmoski et al., 2007). Additionally, the sample as a whole should contain experts with varied backgrounds (e.g. academics and policy makers) to ensure the inclusion of a wide range of views (Rowe et al., 1991). A decision on the number of experts to include in a Delphi study is usually based on common sense, practical logistics and available resources like time and money (Keeney et al., 2006; Powell, 2003). According to Skulmoski et al. (2007) the number of experts should increase with the heterogeneity of the expert sample. A rather homogeneous sample may

consist of ten to fifteen experts, whereas a highly heterogeneous (international) panel may consist of several hundreds of experts.

Regarding the design of the first round questionnaire there are roughly two options (Hasson & Keeney, 2011; Hung et al., 2008; Keeney et al., 2006; Powell, 2003). The first option is to design a questionnaire consisting of open-ended questions that allow experts to freely elaborate on the topic of interest. Views and issues expressed are analysed by the research team with the aim of generating specific items that experts need to rate in subsequent rounds, for example by using a 5-point scale ranging from 'very unimportant' to 'very important'. The second option is to design a first round questionnaire that already includes a predefined list of items based on literature or previously conducted studies. As a predefined list of items may bias or limit experts' responses, it is recommended to allow experts to also express their own views (Keeney et al., 2006), for example by asking them to suggest additional items.

In the second and subsequent rounds expert receive controlled opinion feedback in the form of a summary of the results from the previous round. Users of the Delphi method need to decide what information to include in this feedback. Typically, a distinction is made between two types of feedback: summary statistics and rationales. Summary statistics are based on experts' ratings and show per questionnaire item a location and dispersion statistic (e.g. the median and interquartile range). Rationales consist of a summary of the explanations that experts gave for their ratings. Although some researchers recommended feeding back both types (Murphy et al., 1998; Rowe et al., 1991), others suggested to only feed back rationales as to prevent experts from simply changing their ratings in the direction of the majority opinion as indicated by the summary statistics (Bolger et al., 2011). A related issue concerns feeding back experts' own initial ratings from the previous round. Although feeding back this information has been recommended, it is by no means always provided (Boulkedid et al., 2011).

Due to controlled opinion feedback experts' ratings of questionnaire items tend to converge (Powell, 2003), although this is not always the case (Steinert, 2009). It is therefore important that users of the Delphi method decide on how to measure the level of agreement among experts in each round. Sadly, many Delphi studies fail to offer an interpretation of the meaning of agreement (Hasson & Keeney, 2011; Powell, 2003). As a result, agreement in Delphi studies is measured in different ways. Commonly, the level of agreement is measured per item as the percentage of experts that, for example, rated an item as 'very important' (Powell, 2003). More sophisticated indices of inter-rater agreement are available (Banerjee, 1999). Apart from Cronbach's alpha (see for example Graham et al., 2003) these indices hardly seem to be used in the context of Delphi studies.

1.3 What are rankings?

Although the exact origin of rankings is unknown, it is safe to assume that they have existed for centuries. Nonetheless, it was not until the 1980s that commercial magazines and media started to produce and publish rankings, thereby substantially increasing their accessibility and popularity (Espeland & Sauder, 2007; Lange, 2010). Today, rankings are everywhere. There are rankings of the most powerful people, the best companies to work for, the most prominent universities, the most sustainable cities, and the happiest countries to name only a few.

In general terms, a ranking consists of at least two objects (e.g. cities) that have been ordered based on their performance on a ranking attribute (e.g. urban sustainability). The objects in a ranking are distinct, non-overlapping, but more or less comparable entities that belong to a certain category (i.e. European capital cities). They are usually given ascending rank numbers, starting with 1 for the best performing object. As Jones (1971) explained, rankings are transitive in the sense that if object A has a higher rank than object B, and object B has a higher rank than object C, then object A has a higher rank than object C. Furthermore, although rank numbers provide information about the direction of the difference between any two objects (i.e. object A performs better than object B), they do not provide any information about the magnitude of the difference. Thus, the difference between the objects A and B may be equal to the difference between the objects B and C in terms of rank numbers, but wholly unequal in terms of their actual performance on the ranking attribute.

The usefulness of rankings is a topic of intense debate (Espeland & Sauder, 2007; Lange, 2010; Rauhvargers, 2011). Proponents point at the potential of rankings for reducing complexity and making information more accessible to end-users, such as the general public. Furthermore, rankings may provide insight into the strengths and weaknesses of objects which in turn may increase the performance of objects on the ranking attribute. Opponents argue that rankings oversimplify the performance of objects on complex attributes which may cause end-users to misinterpret and misuse ranking results. Furthermore, rankings may actually have negative impacts for low ranking objects in the form of among others a reduced budget and less customers. In turn this may decrease their performance on the ranking attribute even further. The most important argument against rankings involves the many methodological issues related to their development. These issues give rise to doubts about the extent to which rankings reflect the relative performance of objects on the ranking attribute (Jacobs et al., 2005; Lun et al., 2006; McManus, 2012).

Generally, there are two approaches to developing a ranking. The most straightforward approach is to ask experts to rate objects according to their performance on a ranking attribute, after which the individual expert ratings are aggregated into an overall rating (e.g.

by taking the mean) and corresponding rank number for each object (Giles & Garand, 2007; Liao et al., 2014; Williams & Van Dyke, 2008). A more sophisticated approach is to operationalize a ranking attribute into various indicators that each aim to measure the performance of the objects on a specific aspect of the ranking attribute (Lange, 2010). For each indicator, data on the objects are collected. These usually involve secondary data of a quantitative nature (e.g. official statistics from local city authorities on the amount of particulate matter in the air). Nonetheless, primary data may also be collected, for example by asking experts to rate objects on a specific indicator. Data collected on the various indicators are aggregated into composite index values on the basis of which the objects are ranked.

Both the fully expert-based approach and the indicator-based approach have certain methodological issues. With regard to the expert-based approach, experts may not be sufficiently familiar with all objects. As a result, they may not be able to make a proper estimation of the actual performance of each object on the ranking attribute (Liao et al., 2014; Williams & Van Dyke, 2008). Experts may also have very different interpretations of the ranking attribute, especially when it involves a complex multi-dimensional attribute such as urban sustainability. Although the indicator-based approach largely avoids the issues of the expert-based approach, it has some methodological issues of its own. Two major issues involve the definition and operationalization of the ranking attribute and the aggregation of indicators into a composite index value. How a ranking attribute such as urban sustainability is defined, determines the selection of indicators and thereby ranking results (McManus, 2012; Wilson et al., 2007). It is therefore important that developers of indicator-based rankings carefully define the ranking attribute and justify the specific selection of indicators. To aggregate the data collected on the different indicators into a composite index value, various normalization, weighting and aggregation techniques are available. For example, indicators may be assigned different weights that reflect their relative contribution to measuring the ranking attribute. Again, the selection of specific techniques ought to be carefully substantiated as it may severely influence ranking results (Floridi et al., 2011; Jacobs et al., 2005; Lun et al., 2006).

1.4 Using Delphi to develop rankings

The Delphi method seems to provide opportunities for developing rankings. With regard to a fully expert-based approach, the Delphi method enables experts to independently and iteratively rate objects according to their performance on a ranking attribute. Moreover, through the provision of controlled opinion feedback experts are able to share their views and knowledge. This may help them to make more informed judgements in each subsequent round. Additionally, the Delphi method may provide insight into the stability of objects' rank

positions across rounds as well as the extent to which experts were able to agree on the performance of the objects on the ranking attribute.

A so called ranking-type Delphi method has been developed in the field of information systems research to identify and rank management issues (Dickson et al., 1984; Schmidt, 1997). This specific type of Delphi method distinguishes three phases that each may consist of multiple rounds (Pare et al., 2013). In the first phase experts are stimulated to identify as many issues as possible. A consolidated list of identified issues is presented to the experts in phase two with the request to select the most important ones. Issues selected by a majority of experts are put on a short-list (consisting of about 20 issues). In the third phase experts are asked to actually rank the issues on this short-list, usually in at least two subsequent rounds. Although the ranking-type Delphi method may be useful for obtaining a ranking of objects, there are some problems. Most importantly, asking experts in the last phase to actually rank objects forces them to make a specific distinction between objects that they may not actually perceive. Besides, the possibilities for analysing ranking data are rather limited. A different approach to the use of the Delphi method for obtaining a ranking of objects may thus be desirable.

With regard to developing rankings based on an indicator system, the Delphi method seems to provide opportunities as well. Developers of rankings often aim to measure the performance of objects on a complex ranking attribute (e.g. urban sustainability). Usually, there is no clear definition and operationalization of such an attribute, which makes it hard to determine which indicators need to be selected. In this case the Delphi method may be used to allow experts to define and operationalize the ranking attribute, for example by identifying and rating its most relevant components. In this way the use of the Delphi method may result in a framework consisting of several components about which experts generally agreed that they are most relevant for defining and measuring the ranking attribute. Such a framework may then guide a subsequent selection of indicators. Additionally, based on experts' ratings, weights may be assigned to the components that reflect the relative contribution of the underlying indicators in measuring the ranking attribute.

The use of the Delphi method for the development of rankings also brings about various challenges. These mainly involve methodological issues inherent to the Delphi method itself. As explained in section 1.2, within the general framework of the Delphi method many specific decisions need to be made regarding the sampling of experts, the design of the first round questionnaire, the provision of controlled opinion feedback, and the measurement of agreement. These decisions are important because they may influence the functioning of the Delphi method and thereby its usefulness for developing rankings.

Quite some time ago Rowe et al. (1991) already stressed the need for more research into the Delphi method. Despite their appeal, only a limited number of empirical studies into the method have been conducted (see for example Bolger et al., 2011; Campbell et al., 1999; Hussler et al., 2011). As a result, evidence-based guidelines on how to design a Delphi study are in short supply (Hung et al., 2008; Keeney et al., 2006). Furthermore, almost no research seems to exist that explicitly explored the opportunities of the Delphi method for developing rankings. Therefore, the following overall research question underlies this PhD project:

> *What are the methodological challenges and opportunities of the Delphi method for developing rankings?*

With this overall research question as a starting point I conducted six empirical studies. Three of these studies focussed on the methodological issues of the Delphi method. Although the Delphi method has various methodological issues, I decided to focus on two issues that are at the core of any Delphi study: the measurement of agreement and the provision of controlled opinion feedback. Consequently, in this PhD project I conducted one study into the measurement of agreement and two studies into the provision of controlled opinion feedback.

As stated in section 1.2, multiple indices are available that may be used to measure the level of agreement among experts in Delphi studies. Research into how these indices behave within and across the rounds of a Delphi study did not seem to exist. Therefore, a simulation study was conducted in which the behaviour of nine different indices was examined within various Delphi scenarios.

With regard to the provision of controlled opinion feedback, a few experiments have been conducted (e.g. Best, 1974; Bolger et al., 2011; Rowe & Wright, 1996). Most of these examined the effect of different types of feedback on experts' degree of opinion change and forecast accuracy (i.e. the degree to which experts' judgements correspond to verifiable true values). Similar experiments including other relevant Delphi outcome measures, such as the degree to which experts conformed their ratings to the majority opinion and achieved agreement, could not be found (except for a rather peculiar study by Gowan & McNichols, 1993). Furthermore, most Delphi experiments had an artificial set-up in the sense that university students or staff were asked to give their opinion on rather trivial topics about which they had no expert knowledge. These so called laboratory Delphi experiments have been heavily criticized and dismissed as largely inappropriate (Rowe & Wright, 1999). In this PhD project it was therefore decided to conduct two Delphi experiments within real-world Delphi studies in which actual experts on the topic of interest participated. Both experiments examined the effect of different types of feedback on several relevant Delphi outcome measures, such as the degree to which experts changed their ratings towards the majority opinion and achieved agreement. In response to the debate about what information to

include in feedback (see section 1.2), experts in the first experiment received feedback of rationales either with or without summary statistics. In the second experiment experts received feedback that either included or excluded their own initial ratings.

Opportunities and challenges related to the actual application of the Delphi method for developing rankings were explored by means of two Delphi studies. The first Delphi study was conducted within the field of landscape architecture, a highly interdisciplinary and practice-oriented field that covers a great variety of research domains (van den Brink & Bruns, 2014). Drawing on an international sample of landscape architecture experts from academia and practice, the Delphi study aimed to rank research domains according to their importance for landscape architecture research and their usefulness for landscape architecture practice. A second Delphi study was conducted in response to a study in which I examined the methodological characteristics of several existing indicator-based urban sustainability rankings. It appeared that most of these rankings did not properly define the ranking attribute urban sustainability. Presumably, because many definitions of urban sustainability exist and there is no consensus on what the concept actually entails (Huang et al., 2015). Therefore, the second Delphi study aimed to identify which components are most relevant for defining and measuring the ranking attribute urban sustainability according to a European sample of urban sustainability experts.

Both Delphi studies were designed in a similar way. A heterogeneous sample of experts was assembled by means of selection criteria and search strategies. To limit the number of rounds and thereby drop-out of experts, the first round questionnaire included a predefined list of items (i.e. research domains in the first Delphi study and components of urban sustainability in the second Delphi study). Experts were given the opportunity to suggest additional items which were used to expand the list for subsequent rounds. In both the first and second round experts were asked not to rank, but to rate the items using an ordinal rating scale. This enabled experts to give items similar ratings. In the third and final round, however, experts had to select a limited number of items. In this way the most important and useful landscape architecture research domains as well as the most relevant components for defining and measuring urban sustainability could be clearly identified.

1.5 Overview of chapters

The remainder of this thesis is organized as follows. Chapter 2 reports on the simulation study that examined the behaviour of nine agreement indices within various Delphi scenarios. Based on the results of the simulation study several agreement indices were selected and applied in studies that are reported in subsequent chapters. Chapter 3 discusses the Delphi study that was conducted in the field of landscape architecture with the aim of developing two expert-based rankings. This study included the first experiment into

the provision of controlled opinion feedback, which is described in chapter 4. Chapter 5 reports on a study into the methodological characteristics of several indicator-based urban sustainability rankings. Following this study, chapter 6 describes a Delphi study about the definition and measurement of the ranking attribute urban sustainability. This study included the second experiment into the provision of controlled opinion feedback, which is reported in chapter 7. Finally, chapter 8 provides an overview of the main findings of the studies in light of the overall research question and discusses the contribution of this thesis to science and society.

Chapter 2

Quantifying the development of agreement among experts in Delphi studies

This chapter is published as:
Meijering, J. V., Kampen, J. K., & Tobi, H. (2013). Quantifying the development of agreement among experts in Delphi studies, *Technological Forecasting & Social Change*, 80, 1607-1614.

Abstract

Delphi studies are often conducted with the aim of achieving consensus or agreement among experts. However, many Delphi studies fail to offer a concise interpretation of the meaning of consensus or agreement. Whereas several statistical operationalizations of agreement exist, hardly any of these indices is used in Delphi studies. In this study, computer simulations were used to study different indices of agreement within different Delphi scenarios. A distinction was made between indices of consensus (DeMoivre index), agreement indices (e.g., Cohen's kappa and generalisations thereof), and association indices (e.g., Cronbach's alpha, intra-class correlation coefficient). Delphi scenarios were created by varying the number of objects, the number of experts, the distribution of object ratings, and the degree to which agreement increased between subsequent rounds. Each scenario consisted of three rounds and was replicated 1000 times. The simulation study showed that in the same data, different indices suggest different levels of agreement, and also, different levels of change of agreement between rounds. In applied Delphi studies, researchers should be more transparent regarding their choice of agreement index and report the value of the chosen index within every round as to provide insight into how the suggested agreement level has developed across rounds.

Keywords: Delphi, consensus, agreement, association, methodology, simulation

2.1 Introduction

The Delphi method is a data-collection method originally developed by Dalkey and Helmer of the RAND Corporation in the 1950s as part of a US military defence project (Dalkey & Helmer, 1963; Linstone & Turoff, 1975). Delphi studies are usually conducted to achieve a level of agreement among experts about a topic. Experts are questioned about their opinion by means of a standardized questionnaire in a number of successive rounds. Communication among experts is avoided in order to stimulate independent thought and prevent group pressure. Instead, communication is controlled by the researcher who provides feedback to the experts, usually in the form of a summary of the findings of the previous round. Based on this feedback, experts may change their opinion about the topic in the next round. This procedure continues until a certain level of agreement has been achieved (Dalkey & Helmer, 1963; Hung et al., 2008; Keeney et al., 2006; Keeney et al., 2001; Landeta, 2006; Linstone & Turoff, 1975; Powell, 2003; Rowe et al., 1991).

Since its public introduction in the 1960s, the Delphi method has been used in different domains (Gupta & Clarke, 1996) (e.g., health care, management, education) and in different formats (Keeney et al., 2006). It does have, however, some unresolved methodological issues (Gupta & Clarke, 1996; Hung et al., 2008; Keeney et al., 2006; Keeney et al., 2001; Landeta, 2006; Powell, 2003; Rowe et al., 1991). These issues concern, among others, the identification and selection of experts, the organization of feedback, the process of opinion change, and the reliability and validity of Delphi results (Bolger et al., 2011; Bolger & Wright, 2011; Ecken et al., 2011; Frewer et al., 2011; Gnatzy et al., 2011; Goluchowicz & Blind, 2011; Hasson & Keeney, 2011; Hussler et al., 2011; Landeta, et al., 2011; Parente & Anderson-Parente, 2011).

Another methodological issue concerns the definition and measurement of agreement (Hasson & Keeney, 2011; Hung et al., 2008; Keeney et al., 2006; Powell, 2003). Powell (2003) reviewed a selection of Delphi studies and concluded that agreement is defined and achieved in many different ways, although “setting a percentage level for inclusion of items appears to be a common interpretation” (p. 379). This means that within a Delphi round experts evaluate a number of items or objects (e.g., guidelines in a checklist, items in a questionnaire, indicators of a concept) using some sort of rating scale (e.g., a 5-point ordinal scale ranging from ‘very unimportant’ to ‘very important’), and agreement is expressed per object by calculating the percentage of experts that evaluates an object in a certain way (e.g., the percentage of experts that rates an object as ‘very important’). The standard deviation is also often used as a measure to express agreement among experts, although this measure should not be used when objects are rated on an ordinal scale.

Other ways to measure agreement among experts can be found in the literature. Different indices exist that quantify the level of agreement among experts in a single number. Several review studies compared indices by discussing their development, strengths, and weaknesses (Banerjee, 1999; Dijkstra & van Eijnatten, 2009; Hubert, 1977). Most indices are rarely or never used in Delphi studies. It is, therefore, unknown how these indices behave when used to measure the level of agreement within Delphi studies.

Because experts may change their opinion based on the feedback that they receive from the researcher, it is expected that the level of agreement among experts increases in every subsequent round. In other words: there may be a certain level of conformity that causes the level of agreement among experts to increase across rounds. It is unknown, however, how different indices cope with this level of conformity.

The objective of this study was, therefore, to find out how different agreement indices behave, not only within, but also across the rounds of a Delphi study by applying them within Delphi scenarios that have been created using computer simulations.

2.2 The indices under study

Delphi studies are often conducted with the aim of achieving consensus among experts (Bunting, 2010; Eberman & Cleary, 2011; Graham et al., 2003). The Cambridge dictionary defines consensus as: “a generally accepted opinion or decision among a group of people” (Cambridge Dictionaries Online, 2012). This implies that within a Delphi study consensus on an object occurs only if all experts attribute the same rating to that object. It is actually very difficult to achieve consensus in a Delphi study (Keeney et al., 2006). The concept of agreement is less strict than the concept of consensus and occurs if within a group of experts at least two experts attribute the same rating to an object. Consensus can, therefore, be considered a special case of agreement. The concept of association is less strict than the concept of agreement and requires only that “the category of one response can be predicted from the category of the other” (Dijkstra & van Eijnatten, 2009). This implies that association allows for a systematic difference between the ratings of two experts. When there is agreement between two experts there is also association between the ratings of both experts. The opposite is not necessarily true. The ratings of two experts can be perfectly associated even though they have not attributed the same rating to any of the objects. With these different operationalizations in mind, the existing indices were classified as consensus, agreement, or association indices. An overview of consensus, agreement, and association indices is found in table 2.1.

Table 2.1
Overview of consensus, agreement, and association indices.

Name of index (denotation)	Index type	Minimum value	Maximum value
DeMoivre index (DM_t)	Consensus	0	1
Strict agreement index (SA_t)	Agreement	0	1
Corrected strict agreement index (CSA_t)	Agreement	0	?
Light's kappa (KL_t)	Agreement	–1	1
Fleiss' kappa (KF_t)	Agreement	–1	1
Conger's kappa (KC_t)	Agreement	–1	1
Cronbach's alpha (CA_t)	Association	–∞	1
Intra-class correlation coefficient absolute (ICA_t)	Association	–1	1
Intra-class correlation coefficient consistency (ICC_t)	Association	–1	1
Kendall's measure of concordance (W_t)	Association	0	1

Consensus index

Hubert (1977) formulated DeMoivre's definition of agreement (after the French mathematician Abraham de Moivre): "an agreement occurs if and only if all raters agree on the categorization of an object" (p. 298). This definition matches the above mentioned definition of consensus. Based on DeMoivre's definition, a consensus index was developed. Consensus between two experts r and r' regarding rating x of object i in round t(t = 1, …, T) is defined as

$$c_{irr't} = \begin{cases} 1 \text{ if } x_{irt} = x_{ir't} \\ 0 \text{ otherwise} \end{cases}.$$

When there are M experts, consensus on an object is achieved when there is consensus within all possible expert pairs (M(M–1)/2). Consensus on object i in Delphi round t was thus defined as

$$c_{it} = \begin{cases} 1 \text{ if } \sum_{r'>r} c_{irr't} = M(M-1)/2 \\ 0 \text{ otherwise} \end{cases},$$

where c_{it} = 0 or 1 (0 denoting no consensus, 1 denoting consensus). For all N objects taken together, the DeMoivre index

$$DM_t = \frac{\sum_i c_{it}}{N},$$

denotes the number of objects that received the same rating from all experts as a proportion of the total number of objects within Delphi round t.

Agreement indices

In addition to the DeMoivre index, an agreement index was developed that matches the concept of agreement as mentioned above. We define the percentage of equal ratings of object i in round t over the M experts in Delphi round t as

$$a_{it} = \frac{\sum_{r'>r} c_{irr't}}{M(M-1)/2},$$

where $0 \leq a_{it} \leq 1$ (0 denoting total disagreement, 1 denoting consensus). For all N objects taken together, the index

$$SA_t = \frac{\sum_i a_{it}}{N},$$

denotes the mean proportion of equal ratings across experts and objects within Delphi round t.

The calculation of SA_t does not take into account that some agreement among experts occurs by chance. Particularly when there is a low number of possible rating categories, chance agreement might constitute a substantial part of the observed agreement among experts. In practice, the level of chance agreement is unknown and will be estimated from the data. With simulated data, however, it is possible to compute the percentage of chance agreement. For instance, when the first round in a simulated Delphi scenario is based on random data, no agreement beyond chance is expected and the level of agreement SA_1 found in the first round represents the chance agreement. The corrected strict agreement index

$$CSA_t = SA_t - SA_1,$$

corrects the observed level of total agreement for the level of chance agreement ($t > 1$).

Cohen's kappa (Cohen, 1960) quantifies the level of agreement between two experts who classify a number of objects into a number of mutually exclusive categories. It is based on the definition that agreement only occurs when two experts have attributed the same rating to an object. Cohen's kappa corrects for chance agreement by incorporating the proportion of expected agreement between two experts into its calculation. Cohen's kappa typically takes on values between −1 and 1. A value of 0 means that there is no agreement between the two experts beyond chance, while a value of 1 signifies perfect agreement between the two experts. A value of −1 indicates that the ratings are each other's mirror image. Because Cohen's kappa was designed for only two experts, different generalizations of Cohen's kappa

were developed that are suitable for cases with more than two experts. Light (1971) developed a kappa that takes the average of all pairwise kappas that can be calculated within the group of experts. Cohen's and Light's kappa take into account the fact that two experts might have different distributions of object ratings. Fleiss' kappa (Fleiss, 1971) deviates from this in that it assumes that the distribution of object ratings for the entire population of experts is known and is taken to be equal for all experts. As such, Fleiss' kappa can actually be regarded as a generalization of Scott's pi (Scott, 1955; Siegel & Castellan, 1988). Conger (1980) adjusted Fleiss' kappa in such a way that differences between experts in the distribution of object ratings are again taken into account. Light's, Fleiss' and Conger's kappa in round t will be denoted respectively as KL_t, KF_t, and KC_t in the remainder of this study.

Association indices

Cronbach's alpha is typically used to assess the reliability or internal consistency of a group of items that make up a scale (Cronbach, 1951, 2004). Cronbach's alpha has also been used in Delphi studies to measure the level of agreement among experts (Bederman et al., 2010; Graham et al., 2003). Because Cronbach's alpha is based on the covariances between all expert ratings, it is actually an association index instead of an agreement index as defined in this study. Cronbach's alpha typically takes on values between 0 and 1, although technically it can take on any value below 0. Values close to 0 mean that the ratings of the experts are completely unrelated to one another, while values close to 1 mean that the ratings are strongly associated. In the remainder of this study Cronbach's alpha in round t will be denoted as CA_t.

There are different operationalizations of the intra-class correlation coefficient (Field, 2005; Shrout & Fleiss, 1979). All are based on an analysis of variance model. According to Field (2005), a two-way random effects model is most suitable for measuring the level of agreement among experts. Within this model, a decision must be made about whether to measure the intra-class correlation coefficient according to an 'absolute' or 'consistency' definition (Field, 2005). The difference between the two is that the relative differences between the ratings of the experts are taken into account in the former and are ignored in the latter. Either way, the intra-class correlation coefficient is based on the concept of association. The intra-class correlation coefficient typically takes on values between 0 and 1 (although values between 0 and -1 are possible). The 'absolute' and 'consistency' intra-class correlation coefficients in round t will be denoted respectively ICA_t and ICC_t in the remainder of this study.

When experts rank a number of objects, the association between their rankings can be determined using Kendall's measure of concordance (Siegel & Castellan, 1988). As an

association index, Kendall's measure of concordance has been used in Delphi studies to measure the level of agreement among experts (Bunting, 2010; Nevo & Chan, 2007; Rushton & Moore, 2010; Schmidt, 1997). When a rating scale with a limited number of categories is used to rank objects, many tied rankings per expert might occur. In this case, a correction for ties needs to be incorporated in the calculation. Kendall's measure of concordance takes on values between 0 and 1, with values close to 0 indicating a very weak association and values close to 1 indicating a strong association. In the remainder of this study, Kendall's measure of concordance (corrected for ties) in round t will be denoted as W_t.

2.3 Outline of the simulation study

Computer simulations were used to study the behaviour of the indices within different Delphi scenarios. A script was developed in R (R Core Team, 2011) for simulating data within some predefined Delphi scenarios and computing the indices. Scripts for computing KL_t, KF_t, KC_t, CA_t, ICA_t, ICC_t and W_t were obtained from an official package on the R website (Gamer et al., 2010). All scripts were thoroughly tested by applying them on datasets obtained from existing publications and comparing the outcomes in R with the reported values in the corresponding publications. The complete script in R can be obtained from the corresponding author on request.

Simulation of three Delphi scenarios

Three Delphi scenarios were simulated in which the number of objects and the number of experts were set at common values of N = 60 and M = 20. Within all three Delphi scenarios an ordinal rating scale consisting of 5 categories was used, and the number of rounds was set equal at 3.

For each of the three Delphi scenarios, the distribution of object ratings was either a skewed binomial (π = .2), a symmetrical binomial (π = .5), or a uniform distribution. The uniform distribution was used to simulate Delphi scenarios in which the initial disagreement among experts is maximized. Within a single Delphi scenario, all objects followed the same distribution.

For the first round, no systematic agreement among experts was assumed. Consequently, the data for the first round were created by drawing random numbers from one of the three distributions of object ratings, each random draw of a number representing a rating of an object by an expert on the ordinal rating scale (1 to 5). For the second and third rounds, it was assumed that the experts had received feedback about the mean rating of each object in the previous round. Based on this feedback, experts might adjust their initial rating for

each object, conforming to a more or lesser degree to the mean rating of that object. As such, the data for the second and third round were created, based on the algorithm

$$x_{irt}=\text{round}\left(\bar{x}_{i,t-1}\cdot\beta_t+x_{ir,t-1}\cdot\left[1-\beta_t\right]\right),\ t>1.$$

Thus, the rating of object i by expert r in rounds two and three depends on the mean rating of object i over all experts in the previous round (denoted $\bar{x}_{i,t-1}$) and the rating of object i by expert r in the previous round. The extent to which expert r conforms to the mean rating depends on the magnitude of the so-called conformity index β_t. This index determines the level of conformity of all experts on all objects within a Delphi round. Within the three Delphi scenarios the conformity indices β_2 and β_3 were set to 0.5, indicating a moderate conformity level. Because an ordinal rating scale was used, the ratings were rounded to the nearest integer.

Simulation of variations

To determine if results based on the three Delphi scenarios were sensitive to changes in parameter values, variations on the three Delphi scenarios were simulated. Each variation differs from the original Delphi scenario in that only one parameter was decreased or increased to another value. As such, the effect of each parameter on the behaviour of the different indices within and across Delphi rounds could be identified.

First, for each of the three Delphi scenarios, two variations were simulated in which the conformity indices were decreased to $\beta_2 = \beta_3 = 0.2$, indicating a low conformity level, or increased to $\beta_2 = \beta_3 = 0.8$, indicating a high conformity level. Second, for each of the three Delphi scenarios, two variations were simulated in which the number of experts was decreased or increased to M = 10 or M = 50. Finally, for each of the three Delphi scenarios, two variations were simulated in which the number of objects was decreased or increased to N = 12 or N = 120.

Table 2.2 shows the set values per parameter for each of the three Delphi scenarios. Alternative values that were set to simulate variations on these three Delphi scenarios are shown in between parenthesis.

Procedure

All Delphi scenarios and the corresponding variations were replicated 1000 times. For each replication, the level of agreement within all three Delphi rounds was calculated using the included indices. Then, for each index a mean value and corresponding empirical 95%-range

was calculated. The empirical 95%-range was obtained by ordering the values of each index across the 1000 replications from lowest to highest and calculating the difference between the 26^{th} and 975^{th} value.

Table 2.2
Set values per parameter for each of the three Delphi scenarios (variations).

Delphi scenario	Number of objects N (variations)	Number of experts M (variations)	Number of rating categories	Number of rounds	Distribution of object ratings	Conformity index β_t (variations)
1	60 (12 or 120)	20 (10 or 50)	5	3	Skewed	0.5 (0.2 or 0.8)
2	60 (12 or 120)	20 (10 or 50)	5	3	Symmetrical	0.5 (0.2 or 0.8)
3	60 (12 or 120)	20 (10 or 50)	5	3	Uniform	0.5 (0.2 or 0.8)

2.4 Results

Three Delphi scenarios

Table 2.3 shows the mean and 95%-range of each index within each round of the three original Delphi scenarios that were simulated. All three generalizations of Cohen's kappa (KL_t, KF_t and KC_t) and both types of the intra-class correlation coefficient (ICA_t and ICC_t) gave almost identical results. These indices will, therefore, be discussed as one kappa (K_t) and one intra-class correlation coefficient (IC_t) in the remainder of this study.

Regardless of Delphi scenario, all indices except SA_1 gave a mean close to 0 in round 1. Of all indices, the association index CA_1 had the greatest 95%-range in round 1. In rounds 2 and 3, the association indices IC_t and W_t had a greater 95%-range than the agreement indices.

Within Delphi scenario 1, round 2, different indices gave different means. While CA_2 gave the greatest mean, DM_2 still gave a mean very close to 0. Apart from DM_2, K_2 gave the smallest mean in round 2. The agreement index CSA_2 and the association indices IC_2 and W_2 gave almost identical means. All three indices gave greater means than K_2, but smaller means than SA_2. In round 3, DM_3 and K_3 gave the same means as in round 2. The means of all other indices did increase, although not more than 0.04.

Delphi scenario 2 differed slightly from Delphi scenario 1. From round 2 to round 3, the means of all indices, except the consensus index, increased by at least 0.03. The association indices IC_t and W_t suggested the greatest increase of the mean from round 2 to round 3.

Table 2.3
Means (95%-range) of each index per Delphi scenario (based on 1000 replications, N = 60 objects, M = 20 experts, and conformity indices of $\beta_{t=2} = \beta_{t=3} = 0.5$).

Index	Delphi scenario 1 (binomial 0.2)			Delphi scenario 2 (binomial 0.5)			Delphi scenario 3 (uniform)		
	Round 1	Round 2	Round 3	Round 1	Round 2	Round 3	Round 1	Round 2	Round 3
Consensus									
DM_t	0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.07)	(0.07)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Agreement									
SA_t	0.36	0.51	0.53	0.27	0.46	0.51	0.22	0.41	0.49
	(0.03)	(0.06)	(0.05)	(0.03)	(0.04)	(0.03)	(0.02)	(0.03)	(0.02)
CSA_t	0.00	0.15	0.17	0.00	0.19	0.24	0.00	0.19	0.27
	(0.03)	(0.06)	(0.05)	(0.03)	(0.04)	(0.03)	(0.02)	(0.03)	(0.02)
KL_t	0.00	0.07	0.07	0.00	0.08	0.11	0.00	0.09	0.15
	(0.03)	(0.09)	(0.09)	(0.02)	(0.04)	(0.03)	(0.02)	(0.03)	(0.04)
KF_t	0.00	0.07	0.07	0.00	0.08	0.11	0.00	0.09	0.15
	(0.03)	(0.09)	(0.09)	(0.02)	(0.04)	(0.03)	(0.02)	(0.03)	(0.04)
KC_t	0.00	0.07	0.07	0.00	0.08	0.11	0.00	0.09	0.15
	(0.03)	(0.09)	(0.09)	(0.02)	(0.04)	(0.03)	(0.02)	(0.04)	(0.04)
Association									
CA_t	–0.03	0.79	0.82	–0.04	0.85	0.91	–0.03	0.83	0.94
	(0.76)	(0.23)	(0.20)	(0.83)	(0.07)	(0.03)	(0.79)	(0.07)	(0.02)
ICA_t	0.00	0.16	0.20	0.00	0.23	0.35	0.00	0.20	0.42
	(0.04)	(0.16)	(0.18)	(0.04)	(0.09)	(0.08)	(0.04)	(0.08)	(0.07)
ICC_t	0.00	0.16	0.20	0.00	0.23	0.35	0.00	0.20	0.42
	(0.04)	(0.16)	(0.18)	(0.04)	(0.09)	(0.08)	(0.04)	(0.08)	(0.07)
W_t	0.05	0.19	0.21	0.05	0.26	0.38	0.05	0.24	0.45
	(0.03)	(0.15)	(0.16)	(0.04)	(0.09)	(0.08)	(0.04)	(0.08)	(0.08)

Delphi scenario 3 differed slightly from Delphi scenario 2. From round 2 to round 3, the means of all indices, except the consensus index, increased by at least 0.06. The association indices IC_t and W_t again suggested the greatest increase of the mean from round 2 to round 3. Both indices suggested a greater increase of the mean from round 2 to round 3 than from round 1 to round 2.

The different distributions of object ratings influenced the outcomes of the indices and can, at least partly, be explained by differences in the level of chance agreement. This is best illustrated by SA_1 in round 1 (see Table 2.3), with the smallest mean in Delphi scenario 3 (uniform distribution of object rating) and the largest mean in Delphi scenario 1 (skewed binomial distribution of object ratings). Although the different distributions of object ratings did have an impact on the consensus index DM_t, the influence on the agreement and association indices was more obvious. Whereas in Delphi scenario 1 from round 2 to round 3 the means of the agreement indices remained rather stable, they increased in the other two scenarios. The means of the association indices increased from round 2 to round 3

regardless of Delphi scenario. The increase in the mean of the association indices was smallest in Delphi scenario 1.

Variations on the three Delphi scenarios

level of conformity

When the conformity level was changed from $\beta_t = 0.5$ to $\beta_t = 0.2$, keeping other parameters unchanged (data not shown), the means of the indices did not change in rounds 2 and 3. Regardless of Delphi scenario, all indices remained stable over rounds.

Table 2.4 shows the mean and 95%-range of each index within each round of the three Delphi scenarios when the conformity level was changed from $\beta_t = 0.5$ to $\beta_t = 0.8$, keeping other parameters unchanged. Because the conformity level only affected the ratings in rounds 2 and 3, round 1 of each Delphi scenario in table 2.4 shows the same results as in table 2.3.

With a conformity level of $\beta_t = 0.8$, DM_t no longer gave means close to zero in rounds 2 and 3. Regardless of Delphi scenario, this consensus index gave greater means than the agreement indices CSA_3 and K_3 as well as the association indices IC_3 and W_3 in round 3. In all three Delphi scenarios, the 95%-range of DM_t, K_t, IC_t, and W_t increased substantially from round 1 to round 2. From round 2 to round 3, the 95%-range of K_t, IC_t, and W_t increased even further.

Within Delphi scenario 1, round 2, experts on average reached consensus on 79% of all objects. The agreement indices CSA_2 and K_2 as well as the association indices IC_2 and W_2 gave smaller means than DM_2 (between 0.47 and 0.55). From round 2 to round 3 the means of all indices increased by not more than 0.10. In round 3 experts on average reached consensus on 86% of all objects. The indices CSA_3, K_3, IC_3, and W_3 gave smaller means than DM_3 (between 0.57 and 0.61).

Delphi scenario 2 differed slightly from Delphi scenario 1. In round 2, experts on average reached consensus on 48% of all objects. Still, K_2, IC_2, and W_2 gave smaller means than DM_2. In round 3 experts on average reached consensus on 92% of all objects and DM_3 again gave a greater mean than CSA_3, K_3, IC_3, and W_3.

Delphi scenario 3 differed slightly from Delphi scenario 2. In round 2, experts on average reached consensus on 36% of all objects and only K_2 gave a smaller mean than DM_2. However, the mean of DM_3 more than doubled compared with round 2. As such, DM_3 again gave a greater mean than CSA_3, K_3, IC_3, and W_3.

Table 2.4
Means (95%-range) of each index per Delphi scenario (based on 1000 replications, N = 60 objects, M = 20 experts, and conformity indices of $\beta_{t=2} = \beta_{t=3} = 0.8$).

Index	Delphi scenario 1 (binomial 0.2)			Delphi scenario 2 (binomial 0.5)			Delphi scenario 3 (uniform)		
	Round 1	Round 2	Round 3	Round 1	Round 2	Round 3	Round 1	Round 2	Round 3
Consensus									
DM_t	0.00	0.79	0.86	0.00	0.48	0.92	0.00	0.36	0.86
	(0.00)	(0.20)	(0.17)	(0.00)	(0.27)	(0.13)	(0.00)	(0.25)	(0.18)
Agreement									
SA_t	0.36	0.91	0.93	0.27	0.87	0.96	0.22	0.79	0.93
	(0.03)	(0.09)	(0.09)	(0.03)	(0.08)	(0.07)	(0.02)	(0.09)	(0.09)
CSA_t	0.00	0.55	0.57	0.00	0.60	0.68	0.00	0.57	0.71
	(0.03)	(0.09)	(0.09)	(0.03)	(0.08)	(0.07)	(0.02)	(0.09)	(0.09)
KL_t	0.00	0.47	0.57	0.00	0.24	0.49	0.00	0.29	0.61
	(0.03)	(0.19)	(0.35)	(0.02)	(0.27)	(0.41)	(0.02)	(0.20)	(0.38)
KF_t	0.00	0.49	0.59	0.00	0.24	0.52	0.00	0.29	0.63
	(0.03)	(0.18)	(0.33)	(0.02)	(0.27)	(0.35)	(0.02)	(0.20)	(0.35)
KC_t	0.00	0.49	0.59	0.00	0.24	0.52	0.00	0.29	0.63
	(0.03)	(0.18)	(0.33)	(0.02)	(0.27)	(0.35)	(0.02)	(0.20)	(0.35)
Association									
CA_t	−0.04	0.95	0.97	−0.04	0.89	0.96	−0.03	0.92	0.97
	(0.81)	(0.03)	(0.04)	(0.80)	(0.15)	(0.05)	(0.79)	(0.06)	(0.04)
ICA_t	0.00	0.50	0.59	0.00	0.30	0.54	0.00	0.39	0.66
	(0.04)	(0.17)	(0.33)	(0.04)	(0.27)	(0.35)	(0.04)	(0.20)	(0.34)
ICC_t	0.00	0.50	0.59	0.00	0.30	0.54	0.00	0.39	0.66
	(0.04)	(0.17)	(0.33)	(0.04)	(0.28)	(0.35)	(0.04)	(0.20)	(0.34)
W_t	0.05	0.53	0.61	0.05	0.33	0.56	0.05	0.42	0.67
	(0.03)	(0.16)	(0.31)	(0.04)	(0.27)	(0.33)	(0.03)	(0.19)	(0.32)

The different distributions of object ratings influenced the outcomes of the indices. The mean of the consensus index DM_2 in round 2 (see table 2.4) was greater in Delphi scenario 1 (skewed distribution of object ratings) compared with Delphi scenario 2 (symmetrical distribution of object ratings) and Delphi scenario 3 (uniform distribution of object ratings). The different distributions of object ratings had a similar, although less obvious, influence on the agreement indices, except for CSA_t, and the association indices. For all indices, the increase in the mean from round 2 to round 3 was smallest in Delphi scenario 1.

Delphi scenarios in which $\beta_2 < \beta_3$ were also simulated. These scenarios did not provide any new insights and are not reported in this paper.

Number of experts

When the number of experts was changed from M = 20 to M = 10, keeping other parameters unchanged (data not shown), the means of all indices, except CA_t, increased in rounds 2 and

3 of Delphi scenario 1. Changing the number of experts to M = 10 also caused an overall increase of the 95%-range of all indices in virtually all cases. When M = 10, W_t gave a mean of 0.10 in round 1 of all three Delphi scenarios, compared to 0.05 when M = 20.

When the number of experts was changed from M = 20 to M = 50, keeping other parameters unchanged (data not shown), the means of all indices decreased in rounds 2 and 3 of Delphi scenario 1. Changing the number of experts to M = 50 also caused an overall decrease of the 95%-range of all indices in virtually all cases. When M = 50, W_t gave a mean of 0.02 in round 1 of all three Delphi scenarios, compared to 0.05 when M = 20.

Number of objects

When the number of objects was changed from N = 60 to N = 12, keeping other parameters unchanged (data not shown), CA_t gave means close to −0.20 in the first round of all three Delphi scenarios (compared to −0.03 when N = 60). In addition, CA_t had an exceptionally large 95%-range, taking on values between −1.90 and 0.51. Changing the number of objects to N = 12 did not affect the means of the other indices. It did cause an overall increase of the 95%-range of all indices in virtually all cases.

When the number of objects was changed from N = 60 to N = 120, keeping other parameters unchanged (data not shown), the means of the indices were not affected. It did cause an overall decrease of the 95%-range of all indices in virtually all cases.

2.5 Discussion and conclusion

In this study a theoretical distinction was made between consensus, agreement and association indices. The simulation study showed that the indices also differ from each other empirically. Different indices suggested different levels of agreement in the same data.

The indices were studied within the context of a Delphi study, which is unique in the sense that a certain level of conformity might cause the level of agreement to increase across rounds. The consensus, agreement, and association indices differed from each other in how they coped with this level of conformity. Overall, with a moderate conformity level ($\beta_t = 0.5$), the association indices suggested a greater increase in the estimated level of agreement across rounds than the agreement indices, while the consensus index suggested almost no increase at all across rounds. With a high conformity level ($\beta_t = 0.8$), the consensus index, except for CA_t, suggested the greatest increase in the estimated level of agreement across rounds. Consequently, the consensus index, aside from SA_t and CA_t, suggested a greater level of agreement than the agreement indices and the association indices in the third round.

The number of experts affected the level of agreement suggested by the indices. Overall, the larger the number of experts, the smaller the level of agreement as suggested by the indices, with the exception of CA_t. The number of objects only affected the level of agreement as suggested by CA_t. With a small number of objects (N = 12), CA_t suggested a large range of different, and sometimes uninterpretable, agreement levels in the first round. All indices proved to be more or less sensitive to the underlying distribution of object ratings. The sensitivity of K_t to the marginal distribution of the ratings, causing under- or overestimation of agreement, was also established in other studies (Banerjee, 1999; Byrt et al., 1993; Dijkstra & van Eijnatten, 2009; Feinstein & Cicchetti, 1990).

It is impossible to judge which index is most suitable for measuring the level of agreement in Delphi studies. The consensus index is strict in the sense that all experts need to attribute the same rating to an object. This can become problematic when a Delphi study includes a large number of experts. The index can, however, be easily interpreted as a percentage. The agreement indices SA_t and CSA_t are less strict than the consensus index and are also easily interpretable. A drawback of SA_t is that it does not correct for chance agreement. The CSA_t does correct for chance agreement, however, this chance agreement cannot be computed in an actual Delphi study. The agreement index K_t incorporates the correction for chance agreement in its calculation. This also makes the interpretation of K_t more problematic. The use of CA_t as an index for measuring the internal consistency of a group of items has received considerable criticism (Cortina, 1993; Schmitt, 1996; Streiner, 2003). This study showed that CA_t is also not suitable for measuring the level of agreement in Delphi studies. The index suggested such a large increase in the mean across rounds that a high level of agreement was easily achieved after three rounds. Furthermore, CA_t suggested a large range of sometimes uninterpretable values in the first round. As association indices IC_t and W_t can be used to measure the level of agreement in Delphi studies, although both indices are difficult to interpret. In addition, W_t becomes unsuitable in Delphi studies that include a small number of experts as it does not sufficiently correct for chance agreement.

Not all known consensus, agreement and association indices were used in the present study. Only those indices that are suitable for the Delphi study as designed were included. Some indices are irrelevant for Delphi studies as they rely on two experts only, for example Cohen's kappa (Cohen, 1960) or Scott's pi (Scott, 1955). Other indices, like the generalization of Cohen's kappa by Fleiss and Cuzick (Fleiss & Cuzick, 1979), did not suit the chosen ordinal rating scale. Other extensions of Cohen's kappa that allow for objects to be rated by different groups of experts of unequal size (Kraemer, 1980; Landis & Koch, 1977; Uebersax, 1982) did not fit the presented scenarios.

Computer simulations were used to study the behaviour of several indices within different Delphi scenarios. The different numbers of objects and experts were chosen to reflect a realistic range, but of course had to be limited. According to Keeney et al. (2006) there exists

no guidance on the minimum or maximum number of experts to be included in a Delphi study. As such, the number of experts in a Delphi study can range from as few as 6 experts (Eberman & Cleary, 2011) to as many as 3000 experts (Jung-Erceg et al., 2007). However, Delphi studies that include such many experts are the exceptions.

In all simulated Delphi scenarios a scale was used with an ordinal measurement level. However, the Delphi method also allows for the use of scales with a nominal, interval or even ratio measurement level (e.g. Di Zio and Pacinelli, 2011). More research is needed to investigate the performance of consensus, agreement and association indices in Delphi studies that use these other measurement levels.

For each Delphi scenario, the distribution of object ratings was either skewed, symmetrical, or uniform. The convergence of expert opinions across rounds was approached by an algorithm based on the conformity index. For this index, three different values were employed. To what extent the distributions of the object ratings and the chosen values of the conformity index reflect actual Delphi studies is hard to establish, as most studies give limited information or no information at all on what happens in and between Delphi rounds. Recently, some research has focused on which factors influence opinion change within Delphi studies (Bolger et al., 2011; Bolger & Wright, 2011; Ecken et al., 2011; Hussler et al., 2011). More research in this area is needed. Modelling the level of agreement (Banerjee, 1999) might shed more light on how experts reach agreement across the rounds of a Delphi study.

This study has some important implications for future Delphi studies. Delphi studies are often conducted with the aim of achieving consensus among experts. However, many Delphi studies fail to offer an interpretation of the meaning of consensus (Hasson & Keeney, 2011; Powell, 2003). Because different kinds of indices might take on different values, researchers are advised to be clear about what they want to measure: consensus, agreement, or association. Additionally, researchers need to report which index (or indices) they use within their Delphi study and clearly explain their choice. Because it might be tempting to simply choose the index that reports the highest value (like SA_t or CA_t), researchers are advised to make these decisions a priori. Furthermore, researchers need to report the value of the chosen index within every Delphi round in order to provide insight into how the suggested level of agreement has developed across Delphi rounds.

Chapter 3

Exploring research priorities in landscape architecture: An international Delphi study

This chapter is published as:
Meijering, J. V., Tobi, H., van den Brink, A., Morris, F., & Bruns, D. (2015). Exploring research priorities in landscape architecture: An international Delphi study. *Landscape and Urban Planning*, 137, 85-94

Abstract

Many of the world's major challenges require responses that are embedded in landscape planning, design, and management. To date, however, it is unclear which research domains should form the core of a future landscape architecture research agenda. This study explored which domains landscape architecture experts prioritise as most important for landscape architecture as a research discipline and which domains they prioritise as most useful for landscape architecture practice. A Delphi study was conducted with an international sample of landscape architecture experts from academia and professional practice. Results suggest that research into 'human dimensions of planning and design' and 'built environments and infrastructure' is desirable from an academic and practice-oriented view. Additionally, the domains 'global landscape issues' and 'green urban development' seem to be important for landscape architecture as a research discipline. These four domains could thus form the core of a future research agenda. Some differences appeared to exist between academic and professional experts as well as between experts from different continents. This suggests that a future research agenda should allow for refinements according to specific regional needs. For the first time, landscape architecture is now in possession of a foundation upon which a fascinating research agenda may be built. Additionally, pertinent discussions are expected to contribute to the continuing maturation of landscape architecture as a discipline that does not only rely on other established research disciplines, but also builds its own body of knowledge.

Keywords: landscape architecture, Delphi method, research priorities, research agenda, research domains, knowledge areas

3.1 Introduction

Many of the world's major challenges, such as demographic and lifestyle changes, urban and rural transformation, climate change and energy needs, to name only a few, require responses that are embedded in landscape. Landscape architects are experts in providing such responses and, in doing so, have been successful in practically applying methods of landscape planning, design and management. But, one may ask, how and to what extent were such methods also applied to the building of a body of knowledge that is becoming fundamental to informing practice? Compared with other scholarly disciplines, landscape architecture is not always recognised as a research discipline (Deming & Swaffield, 2011; Gobster et al., 2010; LaGro, 1999; Milburn et al., 2001; Tai, 2003). During the last couple of decades, landscape architecture has developed an increasingly stronger research focus but, compared with its long and rich history of professional practice, its own research culture is still underdeveloped (van den Brink & Bruns, 2014). There are ongoing debates about the specificity of landscape architecture theory and methodology (Ward-Thompson, 2010), as well as what constitutes research in relation to design (e.g. Lenzholzer et al., 2013; Milburn & Brown, 2003). Additionally, not all landscape architects who conduct research feel the need to publish in (international) peer-reviewed journals (Gobster et al., 2010).

If landscape architecture is to reach greater academic distinction, it is important that more research is conducted to create a sound evidence base that helps to justify landscape planning, design, and management decisions (e.g. Deming & Swaffield, 2011). As Brown and Corry (2011, p. 328) put it, it is time for "the deliberate and explicit use of scholarly evidence in making decisions about the use and shaping of land". Obviously, developing a sound evidence base will be more effective when there is a clear focus regarding the research domains which should be considered core to landscape architecture. Research domains, also referred to as domains of inquiry or knowledge areas, are overarching themes in which research into specific and related topics occurs. As shown by Deming and Swaffield (2011, p. 25), many different domains exist in landscape architecture, such as 'human and environment relationships', 'built environments', and 'values and ethics' to name only a few. This wide range of domains indicates that landscape architecture research is fragmented. As such, the core of landscape architecture research is still not clearly defined (van den Brink & Bruns, 2014). This is a shortcoming that should be of great concern to landscape architects, academics and professionals alike, because it may restrict potential future contributions to solving pressing landscape challenges. The question, then, is which domains should form the core of a future landscape architecture research agenda? No inquiry into this question seems to exist, apart from Chen (2013) who showed that North American practitioners consider additional research into 'construction techniques', 'water resource management', and 'sustainable design' to be most helpful.

In the current study the Delphi method was used to systematically and interactively explore research priorities in landscape architecture by consulting landscape architecture experts from academia as well as from professional practice. Both groups are vital for a discipline that is highly practice-oriented and in search of an enhanced knowledge base. Within landscape architecture, however, there is a noticeable divide between academia and professional practice (Gobster et al., 2010). Consequently, research domains that are important from an academic perspective may not necessarily be considered useful in professional practice. To bridge the gap between academia and professional practice, this study addressed the following two research questions:

1. Which research domains do landscape architecture experts prioritise as *most important* for landscape architecture as an academic discipline?

2. Which research domains do landscape architecture experts prioritise as *most useful* for landscape architecture practice?

By answering these questions a future research agenda may be developed, one which lays the foundations for evidence based landscape architecture.

3.2 Methods

The Delphi method

The Delphi method was developed in the 1950s by Dalkey and Helmer (1963) and is considered particularly suitable for allowing experts to achieve agreement on certain topics. A Delphi study consists of at least two rounds of inquiry. In round 1, experts give their opinion on the topic of interest using a standardised questionnaire. To prevent group pressure, experts remain anonymous and communication among them is avoided. Instead, the researcher provides controlled opinion feedback in the form of a summary of findings from the previous round. In round 2 experts fill in a more or less adapted version of the first questionnaire. Based on the feedback provided, experts may alter their opinion. This procedure continues until a certain level of agreement among the experts has been achieved or until a pre-specified number of rounds (usually not more than four) has been completed (Hung et al., 2008; Keeney et al., 2006; Landeta, 2006; Linstone & Turoff, 1975; Powell, 2003). In this study the Delphi method was used to allow a sample of landscape architecture experts to achieve agreement on the most important research domains for landscape architecture as a research discipline and the most useful domains for landscape architecture practice.

Expert sample

To ensure that a wide range of views is included in a Delphi study, the recommendation is to assemble a heterogeneous sample of experts (Keeney et al., 2001; Powell, 2003). For the current study it was decided to sample landscape architecture experts from academia and professional practice to acquire both academic- and practice-oriented views. Additionally, experts from different parts of the world may have different views. It was therefore decided to sample landscape architecture experts from six continents: Africa, Asia, Australia, Europe, North America, and South America.

A convenience sample consisting of landscape architecture academics and professionals from different parts of the world was assembled. To prevent selection bias, criteria for including experts need to be established before a Delphi study begins (Keeney et al., 2006). For this study it was decided that experts representing academia should hold a position at an academic institution and, to ensure the inclusion of academics who were actively engaged in research, should have published at least one paper on the subject of landscape architecture in an international peer-reviewed journal between the years 2008 and 2013. It was further decided that experts from professional practice should hold a position at a professional organisation (i.e. private companies and public institutions involved in the practice of landscape architecture) and, to ensure the inclusion of high quality professional experts, should have been jurors or winners of competitions that were administered or promoted by the globally active International Federation of Landscape Architects (IFLA), a sub-group of IFLA (e.g. IFLA Asia-Pacific Region), or a national professional organisation that is affiliated with IFLA (e.g. the Colombian Society of Landscape Architects). To be able to access an adequate number of competitions and corresponding jurors and winners, competitions between the years 2003 and 2013 were included.

Search strategies were developed and applied to find potentially suitable experts from all six continents. Initially, names of potentially suitable academic experts were obtained from the 2012 conference proceedings of the European Council of Landscape Architecture Schools (ECLAS) and the Council of Educators in Landscape Architecture (CELA). However, because these two sources mainly yielded experts from Europe and America, the 2011 conference proceedings of ECLAS and CELA as well as a list of members from the LE:NOTRE network were searched to obtain additional non-European and non-American academic experts. Names of potentially suitable professional experts were found by searching IFLA websites, including websites of various IFLA sub-groups and IFLA affiliated national professional organisations, for information on past competitions and corresponding lists of jurors and winners. Because this information was sometimes unavailable, an additional search strategy was used which involved searching the IFLA websites for members lists. All names of academic and professional experts, acquired through the applied search strategies, were researched online to verify whether they met the selection criteria as discussed above.

Experts were assigned to a continent by identifying the country in which they held a position. For some continents it appeared difficult to obtain suitable experts due to language barriers, the small number of experts detected through the searches, and the unavailability of contact information. Table 3.1 gives an overview of the number of academics and professionals within each continent that was included in the expert sample.

Table 3.1
Number of experts (academics + professionals) within each continent that was included in the expert sample and responded in each round of the Delphi study.

Continent	Number of experts			
	Sample	Round 1	Round 2	Round 3
Africa	15 (2 + 13)	2 (0 + 2)	2 (0 + 2)	1 (0 + 1)
Asia	45 (34 + 11)	10 (10 + 0)	6 (6 + 0)	4 (4 + 0)
Australia	24 (10 + 14)	8 (6 + 2)	4 (3 + 1)	3 (3 + 0)
Europe	90 (55 + 35)	37 (28 + 9)	23 (21 + 2)	23 (21 + 2)
North America	88 (59 + 29)	22 (19 + 3)	15 (14 + 1)	11 (11 + 0)
South America	17 (2 + 15)	7 (2 + 5)	5 (0 + 5)	4 (0 + 4)
Total	279 (162 + 117)	86 (65 + 21)	55 (44 + 11)	46 (39 + 7)

Questionnaire development

For the purpose of devising a Delphi questionnaire, a list of research domains within the field of landscape architecture was drawn up. Existing lists as stated in Deming and Swaffield (2011, p. 25) served as a first inspiration. Four different classes of domains were considered in this study: academic perspectives, subjects of study, competencies of the landscape architect, and other areas of knowledge and expertise. For each of these four classes, three domains and corresponding example research topics were formulated. Appendix 3.1 gives an overview of the 12 domains.

Three Delphi rounds were conducted. For round 1, a questionnaire was developed which was pre-tested using the cognitive interview approach (Willis, 2005) with three landscape architects working within the department of Landscape Architecture of Wageningen University. Based on the pre-tests, the questionnaire was developed further. The final version of the questionnaire consisted of two parts. In the first part experts were presented with the question: “How important or unimportant are the following research domains for Landscape Architecture as a scientific research discipline within the next 5 years?” For each of the twelve domains experts were invited to state their opinion using a 5-point rating scale ranging from ‘very unimportant’ to ‘very important’. Experts were also invited to explain why they considered certain domains as important (for a maximum of three domains) and to suggest additional domains. In the second part of the questionnaire experts were presented

with the question: “How useful or useless are the following research domains for Landscape Architecture practice?” For each of the twelve domains experts were invited to state their opinion using a 5-point rating scale ranging from ‘very useless’ to ‘very useful’. Experts were also invited to explain how they thought domains are useful to practice (for a maximum of three domains) and to suggest additional domains.

For round 2 the questionnaire was similar to the one used during round 1. In the first part, experts were asked to rate the importance of the twelve original research domains and three new domains which were formed based on suggestions provided by the experts during round 1. In round 2 each of the twelve original domains was accompanied by a summary of findings from round 1. As an example, figure 3.1 shows the summary which accompanied the research domain ‘built environments and infrastructure’. This summary consists of a table showing the domain’s median rating and its percentage of ‘very important’ ratings, compared to the most and least important domain, as well as a short account of why experts rated the domain as important. After rating the importance of all domains, experts were invited, once again, to explain why they rated certain domains as important (for a maximum of three domains). For the second part of the questionnaire, where experts were invited to rate the usefulness of the domains for landscape architecture practice, the same format and procedure was followed.

Summary of opinions provided by experts on the research domain **‘built environments and infrastructure’:**

	% 'very important' evaluations	Median evaluation
Most important research domain	77%	5 (very important)
Built environments and infrastructure	62%	5 (very important)
Least important research domain	40%	4 (somewhat important)

Research into the domain 'built environments and infrastructure' is needed because the world is urbanizing. Cities need to become greener by examining opportunities for urban agriculture. Built environments and infrastructure are also important because they affect human health, wellbeing, and food security. Furthermore, built environments and infrastructure have a significant role in the adaptation to and mitigation of climate change.

According to you, how important or unimportant is the research domain 'built environments and infrastructure' for Landscape Architecture *as a scientific research discipline* within the next 5 years?
Please give your opinion by using the scale ranging from ‘very unimportant’ to ‘very important’.

	very unimportant	somewhat unimportant	neither important, nor unimportant	somewhat important	very important	don't know
Built environments and infrastructure Example research topics: urban agriculture, public squares, motor- and expressways	○	○	○	○	○	○

Figure 3.1: screenshot of the questionnaire in the second round of the Delphi study.

In round 3 a different set-up was used. In the first part of the questionnaire, experts received a question about which section(s) of the summary of findings from round 2 they wished to read. The summary consisted of two sections: (1) a ranking of the domains, based on each domain's median and its percentage of 'very important' ratings, and (2) short accounts stating why experts considered each domain as important. By answering the question experts indicated if they wanted to read both, one, or none of the sections. Experts were then asked to select the three most important domains and to explain why they considered these as most important. For the second part of the questionnaire, regarding the usefulness of the domains for landscape architecture practice, the same format and procedure was followed. Figure 3.2 summarises the process of the Delphi study and the content of the three questionnaires.

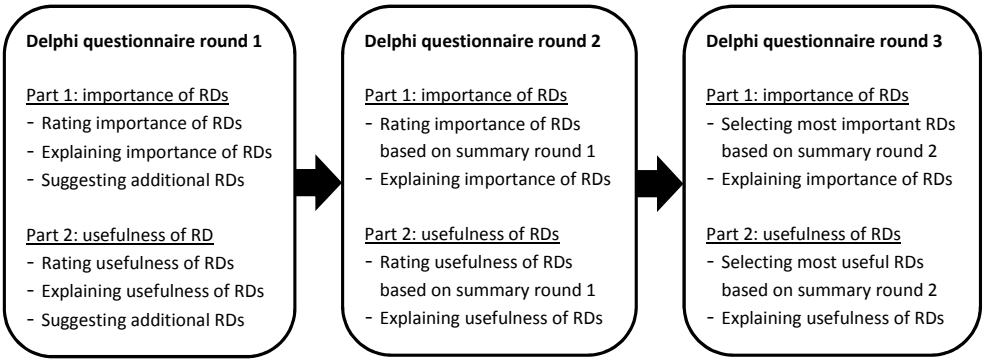

Figure 3.2: Delphi process and contents of the questionnaires on the importance and usefulness of research domains (RD's).

Because experts from different continents were included in the Delphi study, all three questionnaires were written in English and programmed as web-surveys using Qualtrics (Qualtrics, 2015). In April 2013, before the start of round 1, an invitation e-mail was sent to all experts. Shortly thereafter, experts received an e-mail with a link to the first web-survey. In each round experts were given two weeks to respond. Experts who did not respond after one week received a reminder. In every e-mail an attempt was made to motivate experts to participate, such as offering experts who participate in all rounds a forthcoming book and an acknowledgement. Time between rounds was limited to three weeks maximum.

Data analysis

Data collected in each round were analysed to develop a summary of the ratings and explanations provided for each research domain. Per domain, the median importance and usefulness were calculated, as well as the percentage of 'very important' and 'very useful' ratings. Explanations provided by experts on the importance and usefulness of each domain were content analysed by categorising comparable explanations, labelling each category

with a single sentence that captured the underlying explanations, and combining the labels into one short account stating why the domain is important or how it is useful. For example, in round 1 various experts explained that the domain 'built environments and infrastructure' is important because of increased urbanisation, expanding urban areas, the great percentage of people moving to cities, and so on. All these explanations were put in one category which was then labelled with the sentence *"The world is urbanizing"*. Other explanations for the importance of the domain were categorised in the same way. Figure 3.1 illustrates how the labels of all these categories were combined into one short account which was fed back to the experts in round 2.

Suggestions provided by experts for additional research domains that are important for landscape architecture research were content analysed by clustering comparable suggestions into categories. For each category that contained more than four suggestions a new domain was formed that captured the underlying suggestions as fully as possible. Suggestions provided by experts for additional domains that are useful for landscape architecture practice were analysed in the same way. Appendix 3.1 shows the 12 original domains formed by the authors, the three additional domains which were formed as important for landscape architecture research, and the two additional domains which were formed as useful for landscape architecture practice.

To estimate the extent of agreement among experts on the importance and usefulness of the research domains the strict agreement index was used (denoted SA_t, with t denoting the round number) (Meijering et al., 2013). The SA_t is based on the assumption that agreement occurs when, within a group of experts, two experts attribute the same rating to an object. The SA_t expresses the number of agreeing expert pairs as a proportion of the total number of possible expert pairs. The index can take on values between 0 (complete disagreement) and 1 (complete agreement). When a domain has an SA_t above 0.5 it means that more than half of all possible pairs of experts agrees on its importance or usefulness. In this Delphi study, research domains which had a percentage of 'very important' or 'very useful' ratings above 50% and an SA_t above 0.5 were considered most important or most useful.

Data were analysed for the total expert sample and for different expert groups (i.e. academics, professionals, and experts from different continents). Expert groups with less than 15 experts were not separately analysed. This meant that professionals were only treated as a separate group in round 1. In round 1 and 2 groups of Europeans, North Americans, and experts from elsewhere (experts from Africa, Asia, Australia, and South America combined) could be analysed. In round 3 too few North Americans participated and two groups were made: one of Europeans and one of non-Europeans (all experts from outside Europe combined).

To find out whether there was a significant difference between groups regarding the ratings of the domains, various tests were performed. To test for differences between academics and professionals, the Mann-Whitney U-test was used (denoted U). Differences between Europeans and non-Europeans were tested using Pearson's chi-square test (denoted X^2, with 1 degree of freedom). To test for differences between Europeans, North Americans, and experts from elsewhere, the Kruskal-Wallis H-test was used (denoted H, with 2 degrees of freedom). Because the Kruskal-Wallis H-test only indicates whether there is a difference between the three groups, post hoc pairwise comparisons were performed to find out which groups differed. All tests were performed 2-sided. Differences between groups were regarded as significant when the probability of incorrectly identifying a difference (type-I error) was below 5% ($p < 0.05$). When a Kruskal-Wallis H-test was performed, p-values for pairwise comparisons were adjusted so that the overall probability of making a type-I error remained below 5%. For a complete description of the Mann-Whitney U-test, Pearson's chi-square test, and the Kruskal-Wallis H-test, see Field (2013, p. 213-249, 720-746).

3.3 Results: Most important domains

Table 3.1 gives an overview of the number of academics and professionals within each continent that responded in each round of the Delphi study. In the first round 31% of the 279 experts from the initial sample responded. The drop-out rate was 36% in round 2 and 16% in round 3.

Experts' view of most important domains

In round 1 and 2 the median importance of all research domains was either 4 (somewhat important), 4.5, or 5 (very important). Table 3.2 shows a ranking of the research domains based on the percentage of 'very important' ratings in rounds 1 and 2.

In round 1 the percentage of experts that rated a research domain as 'very important' for landscape architecture as a research discipline ranged from 40% (artistic creativity) to 74% (global landscape issues). The domains 'biophysical dimensions of planning and design' and 'global landscape issues' were most often rated as 'very important'.

In round 1, experts suggested 84 additional research domains. These were clustered into 17 categories (including an 'other' category containing unique suggestions). From the three largest categories, containing more than four suggestions, new domains were formed and labelled: 'green urban development', 'measuring landscape performance and impact', and 'landscape architecture education' (see appendix 3.1).

Table 3.2
Ranking of the research domains based on the % of 'very important' ratings in rounds 1 and 2.

Research domains	Total expert sample		Academics		Professionals	
	Round 1 (n = 86) Rank (% 'very important')	Round 2 (n = 55) Rank (% 'very important')	Round 1 (n = 65) Rank (% 'very important')	Round 2 (n = 44) Rank (% 'very important')	Round 1 (n = 21) Rank (% 'very important')	Round 2 (n = 11) Rank (% 'very important')
Human dimensions of planning and design	6 (56%)	**1 (71%)**	4 (58%)	**1 (75%)**	6 (48%)	3 (55%)
Global landscape issues	**1 (75%)**	**2 (69%)**	**1 (74%)**	**2 (73%)**	**2 (76%)**	3 (55%)
Built environments and infrastructure	4 (60%)	**2 (69%)**	4 (58%)	**2 (73%)**	4 (67%)	3 (55%)
Green urban development	-	**2 (69%)**	-	3 (66%)	-	**1 (82%)**
Rural and natural environments	3 (64%)	3 (60%)	3 (62%)	4 (59%)	**3 (71%)**	2 (64%)
Biophysical dimensions of planning and design	**2 (69%)**	4 (58%)	**2 (69%)**	4 (59%)	4 (67%)	3 (55%)
Aquatic environments	5 (57%)	4 (58%)	6 (49%)	5 (57%)	**1 (81%)**	2 (64%)
Landscape architecture education	-	5 (51%)	-	7 (52%)	-	4 (45%)
Values and ethics	8 (48%)	6 (47%)	8 (46%)	8 (48%)	5 (52%)	4 (45%)
Theories	9 (47%)	6 (47%)	5 (51%)	6 (55%)	8 (33%)	6 (18%)
Historic dimensions of planning and design	10 (44%)	7 (44%)	10 (42%)	12 (39%)	5 (52%)	2 (64%)
Artistic creativity	11 (40%)	7 (44%)	11 (31%)	10 (43%)	4 (67%)	4 (45%)
Policy and governance	7 (49%)	7 (44%)	7 (48%)	9 (45%)	5 (52%)	5 (36%)
Tools and technologies	10 (44%)	8 (42%)	9 (45%)	8 (48%)	7 (43%)	6 (18%)
Measuring landscape architecture performance and impact	-	8 (42%)	-	11 (41%)	-	4 (45%)
Range of SA	0.30 - 0.58	0.34 - 0.54	0.27 - 0.57	0.34 - 0.59	0.27 - 0.68	0.20 - 0.67

Figures are based on the question: "How important or unimportant are the following research domains for landscape architecture as a scientific research discipline within the next 5 years?"
Figures with % 'very important' ratings above 50% and an SA above 0.5 are in bold.

In round 2 the percentage of experts that rated the research domains as ‘very important’ for landscape architecture as a research discipline ranged from 42% (measuring landscape architecture performance and impact) to 71% (human dimensions of planning and design). The domains ‘human dimensions of planning and design’, ‘built environments and infrastructure’, ‘global landscape issues’, and ‘green urban development’ were most often rated as ‘very important’.

Table 3.3 shows a ranking of the research domains based on the percentage of experts that selected each domain as 'most important' in round 3.

Table 3.3
Ranking of the research domains based on the % of experts that selected each domain as 'most important' in round 3.

Research domains	Total expert sample Round 3 (n = 46) Rank (% selected)
Human dimensions of planning and design	1 (37%)
Green urban development	2 (33%)
Built environments and infrastructure	3 (30%)
Global landscape issues	3 (30%)
Theories	4 (24%)
Measuring landscape architecture performance and impact	4 (24%)
Rural and natural environments	5 (22%)
Historic dimensions of planning and design	6 (20%)
Biophysical dimensions of planning and design	6 (20%)
Values and ethics	7 (17%)
Artistic creativity	8 (13%)
Policy and governance	9 (9%)
Aquatic environments	10 (7%)
Tools and technologies	10 (7%)
Landscape architecture education	11 (4%)
Development of applied methods and techniques	-
Education of landscape architects	-
Range of SA	0.52 - 0.92

Figures are based on the question: "In your opinion, which of the 15 research domains presented below are the most important for Landscape Architecture as a scientific research discipline within the next 5 years?"

In round 3 the percentage of experts that selected the research domains as ‘most important’ ranged from 4% (landscape architecture education) to 37% (human dimensions of planning and design). Thus, none of the domains was selected by a majority of experts. The domains ‘human dimensions of planning and design’, ‘built environments and infrastructure’, ‘global landscape issues’, and ‘green urban development’, which were rated most often as ‘very important’ in round 2, were also selected most frequently in round 3.

Experts' explanations of most important domains

Several experts stated that the domain 'human dimensions of planning and design' is most important for landscape architecture as a research discipline, because landscape planning and design is done for people. Some experts explained that research into this domain is needed to understand how people perceive and respond to landscapes: *"Planning and design are for people. So, changing attitudes and perceptions of the people towards the landscape in a changing world should be the subject of research."* Others mentioned that knowledge of the domain is required to successfully plan and design landscapes in accordance with people's needs: *"Social studies in landscape architecture are integral to planning and designing most constructed places that succeed in meeting user needs."*

Various experts found the domain 'built environments and infrastructure' most important, because the world is urbanising and landscape architecture needs to influence the future development of built environments and infrastructures: *"The positive impact of landscape architecture in shaping the physical world is essential. As urbanisation continues at pace, it is essential that an integrated design approach drives future development."* Additionally, some experts pointed out that research into this domain is needed to improve the sustainability of cities: *"Cities (…) need to be much more people-friendly, liveable, and offering quality of life. New forms of built environment and new approaches to infrastructure need to be devised."*

Experts who selected the domain 'global landscape issues' as most important stressed that global landscape issues are pressing and landscape architecture has a role in dealing with them: *"With climate change a number of Earth's species could be headed for extinction. It's very important to find an idea how to inhibit or even to stop this rapid process."* Others also explained that research into this domain is important for expanding the landscape architecture body of knowledge and remaining relevant as a discipline.

Some experts explained that research into 'green urban development' is most important because of urbanisation and thus the need to improve the urban environment: *"More and more people will be living in cities in the future. As it is, cities are mostly impervious and have climates of their own. Modifying urban environments to integrate/maintain clean air, water, soil, and food will be necessary."* Others stated that research into the domain is needed to increase the knowledge of urban ecological processes and to understand the social aspects of green urban development. It was also pointed out that research into the domain provides the landscape architecture discipline with possibilities to show its potential and expertise.

Differences between academics and professionals

In round 1 differences appeared between academics and professionals regarding the importance of the research domains 'artistic creativity' ($U = 946$, $p = .005$) and 'aquatic environments' ($U = 973$, $p = .012$). Professionals attached more importance to these two domains than academics. In round 2 and 3 differences between academics and professionals regarding the importance of these domains were no longer statistically significant.

Differences between continents

In round 1, a difference was found between Europeans, North Americans, and experts from elsewhere regarding the importance of the domain 'historic dimensions of planning and design' ($H(2) = 7.83$, $p = .02$). Pairwise comparisons showed a difference between Europeans and North Americans ($p = .016$), with Europeans attaching more importance to the domain (57% 'very important') than North Americans (32% 'very important'). A difference was also found regarding the importance of the domain 'biophysical dimensions of planning and design' ($H(2) = 9.61$, $p = .008$). Pairwise comparisons showed a difference between Europeans and experts from elsewhere ($p = .015$), with Europeans attaching less importance to the domain (54% 'very important') than experts from elsewhere (85% 'very important).

In round 2 a difference was found between Europeans, North Americans, and experts from elsewhere regarding the importance of the domain 'measuring landscape performance and impact' ($H(2) = 6.14$, $p = .047$). Pairwise comparisons showed a difference between Europeans and experts from elsewhere ($p = .047$), with Europeans attaching less importance to the domain (26% 'very important') than experts from elsewhere (59% 'very important'). A difference was also found regarding the importance of the domain 'aquatic environments' ($H(2) = 19.61$, $p < .001$). Pairwise comparisons showed a difference between Europeans and North Americans ($p = .001$) and between Europeans and experts from elsewhere ($p = .001$). Europeans attached less importance to this domain (22% 'very important') than North Americans (87% 'very important') and experts from elsewhere (82% 'very important').

In round 3 differences were found between European and non-European experts regarding the importance of the domains 'historic dimensions of planning and design' ($X^2(1) = 6.77$, $p = .009$), 'theories' ($X^2(1) = 5.86$, $p = .016$), and 'measuring landscape architecture performance and impact' ($X^2(1) = 5.86$, $p = .016$). Europeans selected the domains 'historic dimensions of planning and design' and 'theories' more often (35% and 39% respectively) than non-Europeans (4% and 9% respectively). Non-Europeans selected the domain 'measuring landscape architecture performance and impact' more often (39%) than Europeans (9%).

Agreement on domain importance

In round 1 only the research domains 'biophysical dimensions of planning and design' and 'global landscape issues' had an SA_1 above 0.5. In round 2 the domains 'human dimensions of planning and design', 'built environments and infrastructure', 'global landscape issues', and 'green urban development' had an SA_2 above 0.5.

In round 3 experts had to select the three most important domains instead of rating all domains on a 5-point scale. Because none of the domains was selected by a majority of experts, an SA_3 above 0.5 indicated that more than half of all possible pairs of experts agreed that the corresponding domain was *not* the most important. The SA_3 of the domains 'landscape architecture education', 'tools and technologies', 'aquatic environments', and 'policy and governance' had a value above 0.8, suggesting that experts generally agreed that these domains were the least important for landscape architecture as a research discipline.

3.4 Results: Most useful domains

Experts' view of most useful domains

In round 1 and 2 the median usefulness of all research domains was either 4 (somewhat useful), 4.5, or 5 (very useful). Table 3.4 shows a ranking of the research domains based on the percentage of 'very useful' ratings in rounds 1 and 2.

In round 1 the percentage of experts that rated a research domain as 'very useful' for landscape architecture practice ranged from 30% (theories) to 73% (built environments and infrastructure). The domains 'human dimensions of planning and design', 'biophysical dimensions of planning and design', and 'built environments and infrastructure' were rated most often as 'very useful'.

In round 1, experts suggested 31 additional research domains. These were clustered into seven categories (including an 'other' category containing unique suggestions). From the two largest categories, containing more than three suggestions, new domains were formed and labelled: 'development of applied methods and techniques' and 'education of landscape architects' (see appendix 3.1).

In round 2 the percentage of experts that rated the research domains as 'very useful' ranged from 36% (theories) to 75% (human dimensions of planning and design). The domains 'human dimensions of planning and design' and 'built environments and infrastructure' were rated most often as 'very useful'.

Table 3.4

Ranking of the research domains based on the % of 'very useful' ratings in rounds 1 and 2.

Research domains	Total expert sample		Academics		Professionals	
	Round 1 (n = 86) Rank (% 'very useful')	Round 2 (n = 55) Rank (% 'very useful')	Round 1 (n = 65) Rank (% 'very useful')	Round 2 (n = 44) Rank (% 'very useful')	Round 1 (n = 21) Rank (% 'very useful')	Round 2 (n = 11) Rank (% 'very useful')
Human dimensions of planning and design	**2 (71%)**	**1 (75%)**	**2 (69%)**	**1 (73%)**	**1 (76%)**	**1 (82%)**
Built environments and infrastructure	**1 (73%)**	**2 (71%)**	**1 (72%)**	**1 (73%)**	**1 (76%)**	3 (64%)
Tools and technologies	5 (62%)	3 (60%)	4 (66%)	**2 (70%)**	6 (48%)	7 (18%)
Education of landscape architects	-	3 (60%)	-	3 (61%)	-	4 (55%)
Biophysical dimensions of planning and design	**3 (67%)**	4 (58%)	3 (68%)	4 (59%)	**3 (67%)**	4 (55%)
Aquatic environments	6 (59%)	5 (56%)	7 (55%)	5 (52%)	**2 (71%)**	**2 (73%)**
Rural and natural environments	6 (59%)	6 (55%)	6 (57%)	6 (50%)	3 (67%)	**2 (73%)**
Development of applied methods and techniques	-	6 (55%)	-	3 (61%)	-	6 (27%)
Global landscape issues	4 (64%)	7 (53%)	5 (63%)	6 (50%)	**3 (67%)**	3 (64%)
Historic dimensions of planning and design	7 (53%)	8 (49%)	8 (49%)	9 (41%)	3 (67%)	**1 (82%)**
Artistic creativity	10 (47%)	9 (47%)	10 (37%)	8 (43%)	**1 (76%)**	3 (64%)
Policy and governance	8 (52%)	10 (45%)	8 (49%)	7 (48%)	4 (62%)	5 (36%)
Values and ethics	9 (50%)	11 (44%)	9 (48%)	9 (41%)	5 (57%)	4 (55%)
Theories	11 (30%)	12 (36%)	11 (28%)	10 (36%)	7 (38%)	5 (36%)
Range of SA	0.33 - 0.59	0.36 - 0.59	0.28 - 0.57	0.34 - 0.58	0.31 - 0.62	0.27 - 0.67

Figures are based on the question: "How useful or useless are the following research domains for landscape architecture practice?"

Figures with % 'very useful' ratings above 50% and an SA above 0.5 are in bold.

Table 3.5 shows a ranking of the research domains based on the percentage of experts that selected each domain as 'most useful' in round 3.

Table 3.5
Ranking of the research domains based on the percentage of experts that selected each domain as 'most useful' in round 3.

Research domains	Total expert sample Round 3 (n = 46) Rank (% selected)
Human dimensions of planning and design	1 (48%)
Built environments and infrastructure	2 (41%)
Tools and technologies	3 (37%)
Biophysical dimensions of planning and design	4 (26%)
Education of landscape architects	4 (26%)
Development of applied methods and techniques	5 (24%)
Aquatic environments	6 (15%)
Policy and governance	6 (15%)
Artistic creativity	6 (15%)
Global landscape issues	7 (13%)
Rural and natural environments	8 (11%)
Historic dimensions of planning and design	9 (9%)
Theories	9 (9%)
Values and ethics	10 (0%)
Green urban development	-
Measuring landscape architecture performance and impact	-
Landscape architecture education	-
Range of SA	0.49 - 1.00

Figures are based on the question: "In your opinion, which of the 14 research domains presented below are the most useful for Landscape Architecture practice?"

In round 3 the percentage of experts that selected a research domain as 'most useful' for landscape architecture practice ranged from 0% (values and ethics) to 48% (human dimensions of planning and design). Thus, none of the domains was selected by a majority of experts. The domains 'human dimensions of planning and design' and 'built environments and infrastructure', which were rated most often as 'very useful' in round 2, were also selected most frequently in round 3.

Experts' explanations of most useful domains

Several experts stated that the domain 'human dimensions of planning and design' is most useful for landscape architecture practice, because landscape design is mostly done for humans. Some experts explained that research into this domain provides landscape architects with knowledge about how people perceive, value, use, and are affected by

landscapes: *"It gives insight in the perception and possible use and value of individuals and groups concerning their environment."* Others mentioned that research into the domain helps landscape architects to design landscapes that are meaningful to people and respond to their needs: *"Research in this field may inform about best ways to create practical solutions that are fully in line with human needs (e.g. by providing knowledge on how to integrate people into planning or on socially stratified perceptions, values and needs)."* It was also suggested that knowledge of the domain is useful for obtaining people's acceptance of design proposals.

Various experts found the domain 'built environments and infrastructure' most useful because it is a large field in which many landscape architects are active and can make an impact. Some experts explained that research into the domain will provide practice with new knowledge, ideas, and solutions necessary to meet future challenges: *"New challenges are very important (e.g. demographical change) and for this research provides new solutions and ideas."* It was also suggested that research is necessary to find out how to integrate the rural landscape into the urban landscape*: "The expansiveness of Latin American cities (…) makes it necessary to analyse and study how urban agriculture might be included into the built environment and infrastructure corridors".*

Differences between academics and professionals

In round 1 a difference was found between academics and professionals regarding the usefulness of the research domain 'artistic creativity' ($U = 942$, $p = .003$). Professionals found the domain more useful than academics. In round 2 and 3 differences between academics and professionals regarding the usefulness of the domains were no longer statistically significant.

Differences between continents

In round 1 a difference was found between Europeans, North Americans, and experts from elsewhere regarding the usefulness of the domain 'historic dimensions of planning and design' ($H(2) = 6.41$, $p = .04$). Pairwise comparisons showed a difference between Europeans and North Americans ($p = .041$). Europeans found the domain 'historic dimensions of planning and design' more useful (59% 'very useful') than North Americans (36% 'very useful').

In round 2 a difference was found between Europeans, North Americans, and experts from elsewhere regarding the usefulness of the domain 'aquatic environments' ($H(2) = 14.99$, $p = .001$). Pairwise comparisons showed a difference between Europeans and North Americans

(p = .001) and between Europeans and experts from elsewhere (p = .01). Europeans found the domain less useful (26% 'very useful') than North Americans (87% 'very useful') and experts from elsewhere (71% 'very useful').

In round 3 a difference was found between European and non-European experts regarding the usefulness of the domain 'human dimensions of planning and design' ($X^2(1) = 8.71$, p = .003). The domain was selected by 70% of Europeans, while it was selected by 26% of non-Europeans.

Agreement on domain usefulness

In round 1 the research domains 'human dimensions of planning and design', 'biophysical dimensions of planning and design', and 'built environments and infrastructure' had an SA_1 above 0.5. In round 2 only the domains 'human dimensions of planning and design' and 'built environments and infrastructure' had an SA_2 above 0.5. In round 3 the SA_3 of the domains 'values and ethics', 'theories', and 'historic dimensions of planning and design' had a value above 0.8, suggesting that experts generally agreed that these domains were the least useful for landscape architecture practice.

3.5 Discussion and conclusion

Study outcomes

This study seems to be the first to provide insights, on a global scale, as to which research domains should form the core of a future landscape architecture research agenda according to landscape architecture experts. A complex picture of a maturing field emerges. Results suggest that the emphasis for landscape architecture as a field of academic research should be on 'human dimensions of planning and design', 'built environments and infrastructure', 'global landscape issues', and 'green urban development'. The domains 'human dimensions of planning and design' and 'built environments and infrastructure' also seem to be the most useful for landscape architecture practice. The need for research into 'human dimensions of planning and design' coincides with previous studies in which similar domains were identified as being majorly significant within the academic journals 'Landscape and Urban Planning' (Gobster, 2014) and 'Landscape Journal' (Powers & Walker, 2009), both of which are considered to be top journals among North-American landscape architecture academics (Gobster et al., 2010).

Results from the current study suggest that there are some differences between expert groups. Apart from differences between academics and professionals, the continent to

which an expert belongs had an impact on which research domains were considered important or useful. For example, European experts selected the domain 'historic dimensions of planning and design' as most important for landscape architecture research more often than North American experts. This may be explained by particular regional characteristics such as geography, climate, history and socio-economic context, as well as specific professional cultural idiosyncrasies.

Landscape architecture is a field that requires a broad knowledge base (Bruns et al., 2010). This is reflected in the current study as not one research domain was selected by a majority of experts as most important or useful. This limited agreement may only in part be explained by differences between academics and professionals and by differences between experts from different continents. Another likely explanation may be the interdisciplinary nature of landscape architecture, where a number of sub-specialisations are developing along the lines of landscape design, planning, and management. One must also consider that some experts were educated as engineers, while others were educated as designers or artists.

Study limitations

The results of this study are based on a convenience sample. Strict search strategies and selection criteria were used to avoid selection bias and to make the sample selection as transparent as possible. The resulting expert sample contained more academics than professionals, and more experts from Europe and North America than from Africa, Asia, Australia, and South America. Although it would be interesting to know how closely this sample represents the world population of landscape architects, this cannot be ascertained. It may be expected, however, that there are some deviances. Nonetheless, landscape architecture experts from academia and professional practice as well as from all six continents were represented in all rounds of the Delphi study.

There was non-response in the first round of the Delphi study and some drop-out occurred in subsequent rounds. It is difficult to determine whether the non-response and drop-out rates in this study are favourable or unfavourable in comparison to other Delphi studies, because many researchers failed to report them (Boulkedid et al., 2011). Low response and high drop-out rates are well-known limitations of the Delphi method (Hung et al., 2008; Keeney et al., 2006). A meta-analysis of 39 ordinary web-surveys showed an average response rate of 34% (Shih & Fan, 2008). Because experts in this Delphi study received an invitation in which they were asked to participate in a study consisting of not one, but three web-surveys, the response rate of 31% in the first round can be regarded as satisfactory.

Experts were invited to respond to and comment on a list of research domains. The drawing up of this list for the first round of the Delphi study was a critical task. While this list was

inspired by existing ones (see: Deming and Swaffield, 2011, p. 25), it may not fully reflect the range of domains considered prevalent in landscape architecture. Therefore experts were invited to make suggestions for additional domains. These were collected and consequently used to expand the list for the second and third round questionnaire. For each listed domain (except for the domain 'artistic creativity') three specific example research topics were formulated to support the interpretation of the domain. Nevertheless, considering the backgrounds of the experts and the broad formulation of the domains, experts may still have interpreted domains differently.

Implications and recommendations

This study offers insights into a future landscape architecture research agenda. Research into 'human dimensions of planning and design' and 'built environments and infrastructure' seems to be desirable from an academic and practice-oriented view. Both domains could thus form the core of a future research agenda, possibly complemented with the domains 'global landscape issues', and 'green urban development'. It is important to realise, however, that results were not unequivocal. Differences emerged that relate to academic and professional backgrounds as well as regional contexts. This suggests that a future research agenda should allow for refinements according to specific regional needs. As regional needs were only briefly explored in this study, and because no pertinent results appear to exist from other studies, additional research is necessary to identify and explore regional research specifics in more detail. The Delphi method may be applied in regional contexts to find out which domains could possibly be added to the most important and useful domains as identified in this study, and to determine how the domains could be further specified. Umbrella organisations, such as ECLAS in Europe, CELA in North America, and the world-region subgroups of IFLA, may wish to play a leading role in determining such regional research specifics and developing contextualised research agendas accordingly.

Differences found between the various expert groups and the limited agreement among experts seems to indicate that landscape architecture as an academic discipline is becoming more diverse, and that it is developing sub-specialisations. Such processes are typical signs of maturing disciplines. Brown and Corry (2011), for instance, compared landscape architecture with medical science as a mature evidence based discipline having many sub-specialisations. What will be important in the process of maturation of landscape architecture as an academic discipline is the celebration of diversity within the field. It can be seen from this study that there is no single globally most important or useful research domain. Still, the traditional comprehensiveness of landscape architecture may be challenged, giving a new impetus to continue developing demand-driven and, at the same time, contextualised research agendas. Critical discussions may also stir a broadening of scope and cooperation between landscape architecture and disciplines that already have a

longstanding research tradition. For landscape architecture to be recognised as a partner that contributes to interdisciplinary research it remains important to be specific about what landscape architecture actually is and what forms its intellectual core.

Appendix 3.1

Overview of research domains (including example research topics) presented to the experts.

Research domains formed by authors

Academic perspectives
- Historic dimensions of planning and design
  (Landscape Architecture history, cultural heritage preservation, genius loci)
- Human dimensions of planning and design
  (place attachment, landscape perception, human-environment interaction)
- Biophysical dimensions of planning and design
  (agroecology, green infrastructure, soil erosion)

Subjects of study
- Rural and natural environments
  (agriculture, nature preservation, multifunctional landscapes)
- Built environments and infrastructure
  (urban agriculture, public squares, motor- and expressways)
- Aquatic environments
  (lake restoration, coastal management, urban water fronts)

Competencies of the landscape architect
- Theories
  (design theories, planning theories, milieu theories)
- Tools and technologies
  (Geo Information Systems, social media, visioning/modelling tools)
- Artistic creativity

Other areas of knowledge and expertise
- Values and ethics
  (landscape values, multiculturalism, land ethic)
- Global landscape issues
  (climate change, energy needs, urbanization)
- Policy and governance
  (landscape governance, policy affecting landscapes, commissioner and landscape architect relationship)

Research domains formed based on experts' suggestions (as important for landscape architecture research)

- Green urban development
  (urban ecology and biodiversity, effects of street tree planting, urban waste treatment)
- Measuring landscape architecture performance and impact
  (indicator construction, visual impact assessment, quantification of costs and benefits)
- Landscape architecture education
  (teaching methods, curriculum development, education of the population)

Research domains formed based on experts' suggestions (as useful for landscape architecture practice)

- Development of applied methods and techniques
  (methods for research within practice, decision making support, creativity as a research tool)
- Education of landscape architects
  (curriculum development, improving research skills, raising awareness of careful landscape use)

Chapter 4

The effect of controlled opinion feedback on Delphi features: Mixed messages from a real-world Delphi experiment

This chapter is published as:
Meijering, J. V., Tobi, H. (2016). The effect of controlled opinion feedback on Delphi features: Mixed messages from a real-world Delphi experiment. *Technological Forecasting & Social Change*, 103, 166-173

Abstract

A real-world Delphi experiment was conducted to investigate the effect of two controlled opinion feedback conditions on the drop-out rate, experts' degree of opinion change, and the increase in the level of agreement among experts. Additionally, experts' perceived usefulness of feedback was explored. In the first and second Delphi round experts received a questionnaire which consisted of two sections. Within each section experts were asked to rate several items. In round 2, experts in one condition received feedback consisting of summary statistics and rationales (S&R condition), whereas experts in the other condition received rationales only (R condition). Results showed that drop-out of experts was greater in the S&R condition than in the R condition. No difference between conditions was found concerning experts' degree of opinion change. The increase in the level of agreement across the items in the second section of the questionnaire differed significantly between conditions. This difference was mainly due to a decrease in agreement in the R condition, suggesting that feedback of rationales may increase disagreement among experts. In round 3 experts preferred to receive both summary statistics and rationales, although they tended to perceive rationales as somewhat more useful than summary statistics.

Keywords: Delphi experiment, controlled opinion feedback, drop-out, opinion change, agreement, feedback perception

4.1 Introduction

The Delphi method, originally developed by Dalkey and Helmer (1963), is a structured data-collection process that is often used to allow experts to achieve a certain level of agreement on a particular topic (Keeney et al., 2006). However, the method has various other uses such as maximizing the range of expert opinions (Banwell et al., 2005; Landeta & Barrutia, 2011; Pätäri, 2010; Steinert, 2009; van de Linde & van der Duin, 2011). Any Delphi study consists of at least two rounds. In each round experts are independently questioned about their opinion on the topic of interest by means of a standardized questionnaire. To avoid undue influence of dominant experts and group pressure, experts are anonymous and are not allowed to communicate with each other. Instead, the researcher provides controlled opinion feedback to the experts in the form of a summary of the results from the previous round. Based on this feedback, experts may choose to revise their opinion in the next round. A Delphi study usually ends when a desired level of agreement has been achieved or when a certain stability in experts' responses has been reached (Dalkey et al., 1969; Keeney et al., 2006; Landeta, 2006; Linstone & Turoff, 1975; Rowe & Wright, 1999).

Controlled opinion feedback is an essential part of the Delphi method. Still, no evidence based guidelines exist on how to provide feedback. As a result, Delphi studies differ in the kind of feedback provided. Typically, a distinction is made between summary statistics, which show the majority opinion, and rationales, which show why experts hold certain opinions (Rowe et al., 2005). A systematic review on the use of the Delphi method for selecting healthcare quality indicators, showed that most Delphi studies provided feedback consisting of summary statistics only (Boulkedid et al., 2011). Various researchers criticised this kind of feedback as being insufficiently informative and they proposed to feed back rationales as well (Bolger & Wright, 2011; Murphy et al., 1998; Rowe et al., 1991). Lately, some even suggested that feedback should solely consist of rationales to prevent experts from simply changing their opinion in the direction of the majority (Bolger et al., 2011).

Although the provision of feedback in the form of both summary statistics and rationales is advocated by some, there seems to be little empirical evidence in support of this claim. In response to the debate about controlled opinion feedback, the current Delphi experiment aimed to investigate the effect of feeding back rationales with and without summary statistics on various Delphi features. These features include the drop-out rate, experts' degree of opinion change, and the increase in the level of agreement among experts. Additionally, experts' perception of the usefulness of feedback was explored.

4.2 Theory and hypotheses

Research into the effect of controlled opinion feedback

Few experiments investigated the effect of different kinds of controlled opinion feedback. Generally, these experiments aimed to measure the effect of feeding back either summary statistics or rationales, sometimes in addition to summary statistics, on experts' degree of opinion change and their forecast accuracy (i.e. the correspondence between experts' judgments and a verifiable true value (Woudenberg, 1991)).

Concerning the effect of feedback on experts' degree of opinion change, experiments produced rather similar results: no significant difference in the degree of opinion change was found between study participants who received summary statistics and those who received rationales (Rowe & Wright, 1996; Rowe et al., 2005) or rationales in addition to summary statistics (Bolger et al., 2011). Concerning the effect of feedback on experts' forecast accuracy, experiments produced mixed results. Best (1974) showed that study participants who received rationales in addition to summary statistics were significantly more accurate on one out of two questions than those who received summary statistics only. Rowe and Wright (1996) found that the improvement in accuracy across rounds did not differ significantly between study participants who received rationales and those who received summary statistics. However, in a replication of the experiment Rowe et al. (2005) discovered that only those study participants who received summary statistics showed a significant improvement in accuracy across rounds. Finally, Bolger et al. (2011) found no significant difference in accuracy improvement between study participants who received summary statistics and those who received rationales in addition to summary statistics.

From the experiments mentioned above it may be concluded that the advantages of feeding back rationales have not been convincingly demonstrated in terms of increased opinion change and forecast accuracy. This may be due to the design characteristics of the experiments: study participants consisted of university students and staff who had to answer rather trivial questions. As such, high quality rationales were perhaps not elicited. Several researchers criticized the oversimplification of these so called laboratory Delphi experiments, dismissing them as largely inappropriate (Rowe & Wright, 1999; Rowe et al., 1991).

Based on a review of numerous studies Woudenberg (1991) concluded that feeding back summary statistics induces conformity to the majority opinion. This is particularly troublesome as the original idea of providing controlled opinion feedback is that it counteracts group pressure. Furthermore, pressure to conformity may impede experts to achieve a genuine agreement (Hung et al., 2008; Woudenberg, 1991). Bolger et al. (2011) concluded that study participants tended to ignore feedback of rationales and merely used

summary statistics to change their opinion. Therefore, they suggested eliminating any information concerning the majority opinion from feedback and solely present rationales. Empirical confirmation for this suggestion is lacking.

The current study took the reviewed literature into account by conducting an experiment within a real-world Delphi study in which actual experts were asked their opinion on issues that were relevant to them. Experts were presented with either the recommended feedback consisting of summary statistics and rationales (S&R condition) or the feedback as suggested by Bolger et al. (2011) consisting of rationales only (R condition).

Hypotheses on the effect of controlled opinion feedback on Delphi features

The feedback given in the S&R and R condition may influence Delphi features in several different ways. First, the two feedback conditions may have a different effect on the drop-out rate. Drop-out of experts is recognized as a serious methodological issue in Delphi studies (Hung et al., 2008; Keeney et al., 2006; Landeta, 2006; Powell, 2003). Nevertheless, no study could be found that examined the effect of different feedback conditions on the drop-out rate. Therefore, the following hypothesis was tested against the null-hypothesis (no difference):

H_1: There is a difference between the S&R and R condition concerning the drop-out rate.

Second, the two feedback conditions may have a different effect on experts' degree of opinion change. The Delphi experiments mentioned earlier found no significant difference in the degree of opinion change between study participants who received summary statistics and those who received rationales, whether or not in addition to summary statistics (Bolger et al., 2011; Rowe & Wright, 1996; Rowe et al., 2005). The current study differs from these experiments in two important ways. First, feedback was manipulated by providing rationales either with or without summary statistics. Second, the experiment was conducted in a real-world setting. Because the effect of feedback on experts' degree of opinion change has not been investigated in such a real-world experiment, the following hypothesis was tested against the null-hypothesis (no difference):

H_2: There is a difference between the S&R and R condition concerning experts' degree of opinion change.

Third, the two feedback conditions may have a different effect on the level of agreement among experts. This is of particular importance, because the current Delphi study involved not a forecasting task, but a policy formation task (Rowe & Wright, 1996, p. 75): a task "(…) where subjective opinions and views are sought because objective optimal solutions are

difficult to specify". In a policy formation task experts try to find common grounds concerning for example the indicators needed to measure a particular concept, the guidelines to be incorporated in a new protocol, or the content of a future research agenda. In such tasks true values do not exist. Consequently, determining the accuracy of experts' opinions is impossible. Alternatively, it is essential to determine the level of agreement among experts.

Although many Delphi studies involve a policy formation task, remarkably little is known about the effect of different feedback conditions on the level of agreement among experts. Gowan and McNichols (1993) showed that experts who received feedback in the form of computer-generated if-then rules achieved a greater level of agreement than experts who received either of two kinds of summary statistics. While it is generally assumed that in a Delphi study the level of agreement among experts increases across rounds, no evidence could be found indicating that the increase in the level of agreement would be greater in one of the two conditions under study. Therefore, the following hypothesis was tested against the null-hypothesis (no difference):

H_3: There is a difference between the S&R and R condition concerning the increase in the level of agreement among experts.

Finally, experts may perceive feedback as more or less useful. Although some research examined experts' satisfaction with the Delphi method as a whole (see for example Boje and Murnighan, 1982) no study could be found that specifically investigated experts' perception of the usefulness of feedback. Finding out how experts perceived feedback may help to explain other Delphi features. For example, a limited increase in the level of agreement across rounds may be expected when experts perceived the feedback as rather useless. Therefore, experts' perception of the usefulness of the feedback as provided in the S&R and R condition, as well as experts' perception of the usefulness of summary statistics and rationales as separate feedback components, was further explored.

4.3 Methods

Context of experiment

The Delphi study in which this experiment was introduced, was conducted within the field of landscape architecture (see Meijering et al., 2015 or chapter 3 of this thesis for a complete description of this study). The objective of the Delphi study was twofold: to explore which research domains landscape architecture experts prioritize as most important for landscape architecture as a scientific research discipline and which research domains they prioritize as most useful for landscape architecture practice.

Expert sample

An international sample of landscape architecture experts active in academia and professional practice was assembled. These experts were considered important, because landscape architecture is an academic discipline that is highly practice-oriented. As such, academics and professionals have much in common: they have a similar educational background and frequently meet at international design competitions. Additionally, many academics are also active in practice, while professionals often teach design courses at universities.

Search strategies were used to find potentially suitable experts. Names of academics were obtained from the LE:NOTRE network as well as conference proceedings of the European Council of Landscape Architecture Schools (ECLAS) and the Council of Educators in Landscape Architecture (CELA). Names of professionals were obtained through the website of the International Federation of Landscape Architects (IFLA) as well as the websites of various IFLA subgroups and professional landscape architecture organizations affiliated to IFLA. Selection criteria were used to reduce a selection bias and to assure the inclusion of actual experts. Academics were only included in the sample if they held a position at an academic institution and published at least one paper on the subject of landscape architecture in an international peer-reviewed journal. Professionals were only included if they held a position at a professional organisation (i.e. private companies and public institutions involved in the practice of landscape architecture) and had been jurors or winners of competitions that were administered or promoted by IFLA (including the various subgroups of IFLA and the professional landscape architecture organisations affiliated to IFLA). The final sample included 279 landscape architecture experts of whom 162 academics and 117 professionals (see Meijering et al., 2015 or chapter 3 of this thesis for a complete description of the expert sample).

Study design

The Delphi study was conducted online and consisted of three rounds. In round 1, experts received a questionnaire that consisted of two sections. In the first section, experts were asked to rate the importance of twelve research domains for landscape architecture as a scientific research discipline within the next five years. Experts gave their opinion using a 5-point scale ranging from ‘very unimportant’ to ‘very important’. In the second section, experts were asked to rate the usefulness of the same twelve research domains for landscape architecture practice. Experts gave their opinion using a 5-point scale ranging from ‘very useless’ to ‘very useful’. After each section, experts were offered the option to explain for up to three domains why they rated them as important or how they considered them to be useful.

Summary of opinions provided by experts on the research domain **'human dimensions of planning and design':**

	% 'very important' evaluations	Median evaluation
Most important research domain	77%	5 (very important)
Human dimensions of planning and design	57%	5 (very important)
Least important research domain	40%	4 (somewhat important)

The research domain 'human dimensions of planning and design' has been studied intensively and represents a large part of the Landscape Architecture body of knowledge. According to some experts, Landscape Architecture revolves around landscapes made for people (e.g. cities) and human dimensions need to be taken into account for the successful planning, design, and management of landscapes. It is also important to understand how landscapes shape human life (e.g. with regard to place attachment, quality of life, social interaction, values) and vice versa. A stronger emphasis on the human dimensions of planning and design is needed to solve the challenges of our time, especially urbanization.

According to you, how important or unimportant is the research domain 'human dimensions of planning and design' for Landscape Architecture as a scientific research discipline within the next 5 years?
Please give your opinion by using the scale ranging from 'very unimportant' to 'very important'.

	very unimportant	somewhat unimportant	neither important, nor unimportant	somewhat important	very important	don't know
Human dimensions of planning and design Example research topics: place attachment, landscape perception, human-environment interaction	○	○	○	○	○	○

Figure 4.1: screenshot of feedback presented in the S&R condition.

Experts who completed round 1 were randomly assigned to one of two controlled opinion feedback conditions. In round 2 experts received a questionnaire that was similar to the one used in the previous round. This time, however, each research domain was presented with controlled opinion feedback. In the S&R condition feedback consisted of summary statistics and rationales. The summary statistics showed a domain's median rating and its percentage 'very important' or 'very useful' ratings, compared to the most and least important or useful domain. The rationales summarized in several sentences why experts considered a domain as important or how they considered it to be useful (see figure 4.1). In the R condition the summary statistics were removed from the feedback (see figure 4.2).

In round 2 experts in both feedback conditions were asked once again to rate the research domains on their importance and usefulness (in the first and second section of the questionnaire respectively). As in the previous round, they were also offered the option to explain for up to three domains why they rated them as important or how they considered them to be useful. Finally, experts were asked how useful or useless the provided feedback was for their evaluation of the importance and usefulness of the domains.

Summary of opinions provided by experts on the research domain **'human dimensions of planning and design':**

The research domain 'human dimensions of planning and design' has been studied intensively and represents a large part of the Landscape Architecture body of knowledge. According to some experts, Landscape Architecture revolves around landscapes made for people (e.g. cities) and human dimensions need to be taken into account for the successful planning, design, and management of landscapes. It is also important to understand how landscapes shape human life (e.g. with regard to place attachment, quality of life, social interaction, values) and vice versa. A stronger emphasis on the human dimensions of planning and design is needed to solve the challenges of our time, especially urbanization.

According to you, how important or unimportant is the research domain 'human dimensions of planning and design' for Landscape Architecture as a scientific research discipline within the next 5 years?
Please give your opinion by using the scale ranging from 'very unimportant' to 'very important'.

	very unimportant	somewhat unimportant	neither important, nor unimportant	somewhat important	very important	don't know
Human dimensions of planning and design Example research topics: place attachment, landscape perception, human-environment interaction	○	○	○	○	○	○

Figure 4.2: screenshot of feedback presented in the R condition.

In the third and final round experts were asked to select the three most important and three most useful research domains (in the first and second section of the questionnaire respectively). Before making each selection, however, experts could choose to receive summary statistics, rationales, both summary statistics and rationales, or neither. In this way, additional insight was gained into experts' perception of the usefulness of feedback. Experts were also asked to rate the usefulness of the chosen feedback components for selecting the most important and most useful domains on a 5-point scale ranging from 'very useless' to 'very useful'.

Analysis

Drop-out rates with 95% confidence intervals were calculated in round 2 and 3 for both the S&R and R condition. The 95% confidence intervals were calculated using the Wilson method (Wilson, 1927). This method is recommended when sample sizes are relatively small, e.g. 40 or less (Brown et al., 1998). Assuming the S&R condition as the default, a difference in drop-out rate was considered significant when the 95% confidence interval of the S&R drop-out rate did not include the point estimate of the R drop-out rate.

The degree of opinion change from round 1 to round 2 was calculated for each expert on each research domain within both sections of the questionnaire by taking the absolute difference between the rating in round 1 and the rating in round 2. Next, each expert's

median degree of opinion change across all domains within a section was calculated. Depending on the initial ratings, an expert could change his or her ratings from round 1 to round 2 up to four points (e.g. from 1 'very unimportant' to 5 'very important'). Across all research domains within a section, an expert could have a median degree of opinion change of up to 4. Because of the limited range of values obtained, significant differences between the S&R and R condition were tested using Pearson's chi-square test (denoted X^2, with corresponding degrees of freedom).

The level of agreement among experts was calculated using the Strict Agreement Index (denoted SA_t, with round t), because this index is easy to interpret. The index is based on agreeing expert pairs, where two experts only agree if they attribute the same rating to an object (i.e. research domain). The SA_t expresses the level of agreement among experts as the proportion of agreeing expert pairs within the total number of possible expert pairs (see Meijering et al., 2013 or chapter 2 of this thesis). Theoretically, the index can take on any value between 0 and 1. The SA_t was calculated in round 1 and 2 for the S&R and R condition across all twelve domains within both sections of the questionnaire. The change in the level of agreement from round 1 to round 2 (denoted SA_{2-1}) was calculated by $SA_2 - SA_1$. To investigate whether the SA_{2-1} differed between the two conditions, the SA_{2-1} in the R condition (denoted $SA_{2-1,R}$) was subtracted from the SA_{2-1} in the S&R condition (denoted $SA_{2-1,S\&R}$). Corresponding 95% bootstrap confidence intervals were calculated using the bias-corrected and accelerated method (Efron & Tibshirani, 1993) based on 2000 bootstrap samples. To verify whether significant differences were not due to the particular agreement index chosen, the same calculations were performed using Light's kappa (denoted LK_t) (Light, 1971). This index is also based on agreeing expert pairs, but takes into account that some agreement occurs by chance. The LK_t is highly sensitive to the marginal distribution of the ratings and may overestimate the chance agreement (Feinstein & Cicchetti, 1990). Theoretically, LK_t can take on any value between -1 and 1, with 0 indicating that there is no agreement beyond chance agreement.

Finally, to test whether there was a significant difference between the S&R and R condition concerning the perceived usefulness of feedback in round 2, the Mann-Whitney U test (denoted U) was used. In round 3 expert's perceived usefulness of feedback components was explored using descriptive statistics.

4.4 Results

Drop-out rate

In the first round of this Delphi study 31% of the 279 experts from the initial sample responded (n = 86). The drop-out rate was 36% in round 2 and 16% in round 3. Table 4.1

shows the drop-out rates in the S&R and R condition in round 2 and 3 with corresponding 95% confidence intervals. In round 2 drop-out rates in the S&R and R condition were almost similar whereas in round 3 there was a significant difference in drop-out rate between the two conditions. Mixed results were thus found. The null-hypothesis (no difference) cannot be rejected in favour of H_1 based on the drop-out rates in round 2, but should be rejected in favour of H_1 based on the drop-out rates in round 3.

Table 4.1
Drop-out rates in round 2 and 3 in the S&R and R condition (95% confidence interval).

Round number	S&R condition	R condition
Round 2	35% (22% - 50%)	37% (24% - 52%)
Round 3	25% (13% - 43%)	7% (2% - 23%)

After round 1 the S&R and R condition each contained 43 experts.

Opinion change

Generally, the degree of opinion change from round 1 to round 2 was rather limited (see Table 4.2). In both conditions and in both questionnaire sections the majority of experts had a median degree of opinion change of either 0 or 0.5. This means that the majority of experts did not change their opinion on at least half of the research domains. Pearson's chi-square, with the column categories 1 through 4 collapsed to 'equal or larger than 1', showed that there was no significant difference between the S&R and R condition concerning experts' median degree of opinion change in section 1 ($X^2(2) = 0.073$, $p = .964$) and section 2 ($X^2(2) = 1.842$, $p = .398$). Based on these results, the null hypothesis (no difference) cannot be rejected in favour of H_2.

Table 4.2
Frequency distribution of experts' median degree of opinion change.

Section	Condition	n with median degree of opinion change								
		0	0.5	1	1.5	2	2.5	3	3.5	4
1	S&R (n=28)	12	6	9	0	1	0	0	0	0
	R (n=27)	12	5	9	1	0	0	0	0	0
2	S&R (n=28)	13	7	7	0	0	0	1	0	0
	R (n=27)	14	3	10	0	0	0	0	0	0

Agreement

Table 4.3 and 4.4 show the level of agreement among experts in the S&R and R condition within both sections of the questionnaire. Based on the SA_t and LK_t similar results were obtained. Remarkably, the level of agreement among experts did not increase in most

instances. Within section 1, a slight decrease in the level of agreement was found among experts within both the S&R and R condition. The 95% confidence intervals of the $SA_{2-1,S\&R}$ and $SA_{2-1,R}$ show that this change within both conditions was not significant. Nor was there a significant difference between the conditions concerning the change in the level of agreement, as can be seen from the 95% confidence interval of $SA_{2-1,S\&R} - SA_{2-1,R}$.

Within section 2, an insignificant increase in the level of agreement was found among experts in the S&R condition. In contrast, the level of agreement among experts in the R condition decreased significantly. Consequently, there was a significant difference between the S&R and R condition concerning the change in the level of agreement across rounds. Again, mixed results were found. The null hypothesis (no difference) cannot be rejected in favour of H_3 based on the results of section 1, but should be rejected in favour of H_3 based on the results of section 2.

Table 4.3

SA_t (95% confidence interval) across experts in the S&R and R condition within the first and second questionnaire section.

Section	S&R condition (n = 28)			R condition (n = 27)			$SA_{2-1,S\&R} - SA_{2-1,R}$
	SA_1	SA_2	SA_{2-1}	SA_1	SA_2	SA_{2-1}	
1	0.4140	0.4083	–0.0057 (–0.0773, 0.0725)[a]	0.4349	0.4309	–0.0040 (–0.0845, 0.0827)[a]	–0.0017 (–0.1159, 0.1108)[a]
2	0.3818	0.4519	0.0701 (–0.0022, 0.1726)[a]	0.5207	0.4352	–0.0856 (–0.1693, –0.0002)[a]	0.1556 (0.0490, 0.2823)[a]

[a] 95% confidence interval based on 2000 bootstrap samples.

Table 4.4

LK_t (95% confidence interval) across experts in the S&R and R condition within the first and second questionnaire section.

Section	S&R condition (n = 28)			R condition (n = 27)			$LK_{2-1,S\&R} - LK_{2-1,R}$
	LK_1	LK_2	LK_{2-1}	LK_1	LK_2	LK_{2-1}	
1	0.0506	0.0327	–0.0179 (–0.0755, 0.0392)[a]	0.0141	0.0007	–0.0134 (–0.0535, 0.0304)[a]	–0.0045 (–0.0842, 0.0618)[a]
2	0.0326	0.0571	0.0245 (–0.0367, 0.0938)[a]	0.0753	–0.0101	–0.0854 (–0.1695, –0.0280)[a]	0.1099 (0.0156, 0.2139)[a]

[a] 95% confidence interval based on 2000 bootstrap samples.

Experts' perception of feedback usefulness

In the R condition 44% of the experts rated the feedback in round 2 as somewhat useful and 41% rated it as very useful, while 69% of the experts the S&R condition rated the feedback in round 2 as somewhat useful and 19% rated it as very useful. The difference between the conditions concerning the perceived usefulness of feedback was not significant (U = 294.5, p = .258).

In round 3 experts' perception of the usefulness of feedback components was further explored. Table 4.5 shows the percentage of experts within the S&R and R condition that chose both, one, or none of the feedback components before starting with the first and second section of the questionnaire.

Table 4.5
Percentage of experts within the S&R and R condition that choose both, one, or none of the feedback components before starting with the first and second section of the questionnaire.

Feedback components chosen	Section 1		Section 2	
	S&R (n = 21)	R (n = 25)	S&R (n = 21)	R (n = 25)
Both components	81%	56%	71%	60%
Rationales only	10%	8%	10%	4%
Summary statistics only	10%	20%	14%	16%
None of the components	0%	16%	5%	20%

Before starting with the first section of the questionnaire, the vast majority of experts in the S&R condition chose to continue to receive both feedback components. In the R condition only a few experts chose rationales only, while the majority chose to receive summary statistics in addition to rationales and 36% chose not to receive rationales or any feedback at all. After completing section 1, 55% of those experts who chose both feedback components rated the summary statistics as somewhat or very useful and 61% rated the rationales as somewhat or very useful (not in table).

Before starting with the second section of the questionnaire, 89% of all experts chose to receive the same feedback as in the previous section (not in table). Similar results were found concerning the perceived usefulness of the feedback components: 53% of those experts who chose both feedback components rated the summary statistics as somewhat or very useful and 60% rated the rationales as somewhat or very useful (not in table).

4.5 Discussion and conclusion

This real-world Delphi experiment showed that two controlled opinion feedback conditions, one consisting of solely rationales and one consisting of both summary statistics and

rationales, influenced some Delphi features differently. The two conditions had a different effect on the drop-out rate in the final round and, in one section of the questionnaire, on the change in the level of agreement among experts. In this section of the questionnaire the level of agreement decreased significantly when solely rationales were fed back. The two conditions did not have a different effect on experts' degree of opinion change nor on their perception of the usefulness of the feedback. Some of the results contained mixed messages which require further elaboration and research.

The S&R condition had a higher drop-out rate than the R condition in the final round of this Delphi study, suggesting that feedback of summary statistics in addition to rationales may eventually increase drop-out of experts. A possible explanation for this result may be that the feedback in the S&R condition was lengthier than in the R condition. Another possible explanation is given by Bardecki (1984) who showed that experts with a minority opinion drop-out more often than experts with a majority opinion. In the current study only experts in the S&R condition were able to determine to what extent their opinion differed from the majority opinion, which may have resulted in additional drop-out of experts with a minority opinion. It is remarkable, however, that a difference in drop-out rate occurred in round 3, whereas the two feedback conditions only differed from each other in round 2. Possibly, experts in the S&R condition who found the feedback too lengthy or realized that they had a minority opinion tended to complete round 2, but did not participate anymore in round 3. Alas, this possibility could not be substantiated by the data in a post-hoc analysis.

Concerning experts' degree of opinion change, no difference was found between experts who received rationales and those who received summary statistics in addition to rationales. This result is in line with earlier Delphi experiments (Bolger et al., 2011; Rowe & Wright, 1996; Rowe et al., 2005). In accordance with the experiment of Bolger et al. (2011), experts' degree of opinion change in this study was generally limited. A possible explanation for the limited opinion change is provided by Yaniv (2004) who showed that when study participants make a final judgment they tend to discount advice from others in favour of their own initial opinion. Bolger and Wright (2011) suggested to reduce this so called egocentric discounting in Delphi studies by feeding back high quality rationales. To acquire high quality rationales they recommended recruiting actual experts on the topic of interest, increasing their involvement and motivation, and facilitating them to provide causal explanations for their opinions. Although the current study followed these recommendations, high quality rationales were by no means always acquired. Also, the rationales were fed back as rather short summaries which may not have sufficiently reduced egocentric discounting. Perhaps that more extensive feedback of rationales is more effective, however, caution is advised when feeding back large amounts of information as lengthier feedback may increase drop-out rates.

This study showed that agreement among experts in Delphi studies does not necessarily increase across rounds. In most cases there was no significant change in level of agreement among experts. In the R condition, however, the level of agreement among experts decreased significantly in the second section of the questionnaire. This suggests that, while feedback of summary statistics may induce conformity (Woudenberg, 1991), feedback of rationales may cause experts to disagree rather than agree.

In the second section of the questionnaire there was also a significant difference between the S&R and R condition concerning the change in the level of agreement. This difference may be explained by the opposite effect that each of the two feedback components supposedly had. In the S&R condition this may have resulted in an insignificant increase in the level of agreement. In the R condition the influence of the rationales, in the absence of any summary statistics, may have resulted in a significant decrease in the level of agreement.

The two feedback conditions thus had a different effect on the change in agreement, but only in the second section of the questionnaire. This may be partly explained by the different questions that were asked, given the composition of the expert sample. In section 1 experts were asked to rate twelve research domains on their importance for landscape architecture research, while in section 2 they were asked to rate the same twelve domains on their usefulness for landscape architecture practice. By far most experts in the expert sample were academics who may have had a more pronounced opinion about the importance of the domains for landscape architecture research than their usefulness for practice. Consequently, in section 1 there may have been less room for an effect of the feedback conditions on the change in agreement compared to section 2.

In general, experts in both the S&R and R condition perceived the provided feedback as useful for rating the importance and usefulness of the research domains. This result does not seem to correspond with the limited degree of opinion change nor the limited change in the level of agreement across rounds. Apparently, although experts found the feedback useful, they were not so much influenced by it. Results of round 3 showed that when given a choice experts preferred to receive both summary statistics and rationales. Experts who chose to receive both feedback components seemed to perceive the rationales as somewhat more useful than the summary statistics.

This study examined several important effects of controlled opinion feedback which until now appear to have been largely overlooked in Delphi literature. It is remarkable that so few studies were found that investigated the effect of different feedback conditions on drop-out rates, the level of agreement among experts, and experts' perception of the usefulness of feedback. Although the current study is a controlled experiment, there is no absolute certainty that the observed effects were actually caused by the different feedback

conditions. It is possible that the results found are due to chance. Because the experiment was conducted within a real-world Delphi study, with actual experts who gave their opinion on relevant questions, the results may be better generalizable to other real-world Delphi studies than results obtained from laboratory Delphi experiments. Still, the current experiment was conducted within a specific type of Delphi study that involved a policy-formation task and aimed at achieving agreement among experts. The results of the experiment may thus have limited value for other types of Delphi studies, such as those that aim for dissensus.

Future research into the effect of feedback in Delphi studies ought to widen its scope and include more than just experts' degree of opinion change and forecast accuracy. Preferably, future Delphi experiments will also examine drop-out rates, the level of agreement among experts, and experts' perception of the usefulness of feedback. Drop-out threatens any Delphi study (Hung et al., 2008; Keeney et al., 2006; Landeta, 2006; Powell, 2003) and as such, it deserves more attention. For Delphi studies that involve a policy-formation task and aim to achieve agreement among experts the study of agreement is pivotal. In this regard it is important to note that there are different measures of consensus, agreement and association (see Meijering et al., 2013 or chapter 2 of this thesis). Many Delphi studies aim for consensus, but actually measure agreement or association. Finally, considering that controlled opinion feedback is a key characteristic of the Delphi method, it seems important to better understand how experts perceive and use it when reconsidering their initial evaluations.

The results of this study suggest several specific directions for future research. First of all, research is needed to clarify the mixed results which were found in the current study. For the development of evidence based guidelines on the provision of controlled opinion feedback it is important to know whether feedback of summary statistics induces drop-out of experts and whether feedback of rationales enhances disagreement among experts. Additionally, research is needed to learn how feedback of rationales influences the level of agreement via opinion change. In this regard knowledge is needed on how to elicit and feed back rationales in a way that sufficiently reduces egocentric discounting, without overburdening experts with information. Finally, future research may reveal how experts in Delphi studies perceive different feedback components and how these perceptions may influence other Delphi features.

Researchers who want to design a Delphi study with a policy-formation task need to make some important considerations regarding the provision of feedback. Although experts seem to prefer to receive summary statistics in addition to rationales, feeding back both components may increase drop-out rates. This is clearly undesirable, especially when drop-out occurs among experts with a minority opinion. On the other hand, feeding back solely rationales may result in more disagreement rather than more agreement among experts.

This is undesirable when the Delphi study aims to achieve agreement among experts. Then, feedback of summary statistics may be necessary to induce a certain conformity among experts. Of course, the Delphi method should not be used to force experts to reach an agreement they do not truly support. Moreover, as noted in the introduction of this paper, not all Delphi studies focus on agreement. Instead, some Delphi researchers employed a so called policy or dissensus Delphi to maximize the range of expert opinions (Banwell et al., 2005; Landeta & Barrutia, 2011; Pätäri, 2010; Steinert, 2009; van de Linde & van der Duin, 2011). For these dissensus Delphis feeding back solely rationales may be considered.

Bolger et al. (2011) suggested to eliminate summary statistics from feedback. Based on the current study a profound conclusion on the exclusion of summary statistics from feedback cannot be made. For now, it is of particular importance that more Delphi experiments are conducted, preferably in a real-world context, that shed more light on the effect of controlled opinion feedback on various Delphi features.

Chapter 5

Identifying the methodological characteristics of European green city rankings

This chapter is published as:
Meijering, J. V., Kern, K., & Tobi, H. (2014). Identifying the methodological characteristics of European green city rankings. *Ecological Indicators*, 43, 132-142

Abstract

City rankings that aim to measure the environmental sustainability of European cities may contribute to the evaluation and development of environmental policy of European cities. The objective of this study is to identify and evaluate the methodological characteristics of these city rankings. First, a methodology was developed to systematically identify methodological characteristics of city rankings within different steps of the ranking development process. Second, six city rankings (European Energy Award, European Green Capital Award, European Green City Index, European Soot-free City Ranking, RES Champions League, Urban Ecosystem Europe) were examined. Official websites and any methodological documents found on those websites were content analysed using the developed methodology. Interviews with representatives of the city rankings were conducted to acquire any additional information. Results showed that the city rankings varied greatly with respect to their methodological characteristics and that all city rankings had methodological weaknesses. Developers of city rankings are advised to use the methodology developed in this study to find methodological weaknesses and improve their ranking. In addition, developers ought to be more transparent about the methodological characteristics of their city rankings. End-users of city rankings are advised to use the developed methodology to identify and evaluate the methodological characteristics of city rankings before deciding to act on ranking results.

Keywords: methodology, ranking, index, indicator, city, environmental sustainability

5.1 Introduction

Today many different rankings exist which are often used as a tool for influencing national and international policy debates (Kern, 2008). All these rankings consist of two or more objects that have been ordered based on their performance on certain attributes. Rankings are transitive in the sense that if object A is ranked higher than object B, and object B is ranked higher than object C, then object A is ranked higher than object C (Jones, 1971). The ordered objects are usually given ascending rank numbers (starting with 1 for the highest ranked object). These rank numbers indicate if an object performs better or worse than another object, but they do not provide any information concerning the extent to which an object performs better or worse than another object. This means that even when the ranks of two objects are wide apart, the difference between their actual attribute values may still be very small (Jones, 1971).

Ratings differ from rankings in that each object is assigned an actual attribute value on some predefined scale (Lange, 2010). As such, ratings do provide information concerning the extent to which an object performs better or worse than another object. Ratings can easily be converted to rankings by ordering the objects based on their attribute values and replacing the attribute values with rank numbers. Both rankings and ratings operationalize the performance of objects on an attribute using an indicator system (Lange, 2010). Such a system consists of several indicators each measuring the performance of the objects on one aspect of the attribute. For each object, the measured indicator values are aggregated to calculate one composite index value that reflects the attribute value. By summarizing the performance of an object into a composite index value and corresponding rank number, rankings make it easy to discern how well an object performs in comparison to the other objects included in the ranking. However, multiple studies showed that composite indices, and thus rankings, have some methodological issues concerning among others: the definition of the ranking attribute (Wilson et al., 2007), the selection of indicators (Maretzke, 2006), the aggregation of indicators into a composite index (Jacobs et al., 2005; Schwengler & Binder, 2006), data availability (Almeida et al., 2001), and data quality (Ochel & Röhn, 2008). Various methodological choices within the ranking development process can severely influence the final ranking outcome.

Different kinds of objects are ranked based on different kinds of attributes. The World University Rankings (Quacquarelli Symonds, 2013) ranks universities based on their performance on 30 subject areas, while the Environmental Performance Index (Nesshöver et al., 2007) ranks countries based on the performance of their environmental policy. Since the late 1980s different rankings emerged that rank cities based on their quality of life, business climate, or market potential (Chapman & Pike, 1992).

Rankings are subject of scientific and political debate (Buela-Casal et al., 2007). Because of methodological issues, controversies exist about the extent to which rankings reflect the actual performance of objects on the ranking attribute (Ham et al., 2004). Rankings oversimplify the performance of objects, causing misinterpretation and misuse by unwary end-users (Espeland & Sauder, 2007; Taylor, 2011). Furthermore, they may incite objects to manipulate data (Espeland & Sauder, 2007; Rauhvargers, 2011). Nonetheless, city rankings may contribute to the evaluation and development of urban policy. According to Schönert (2003) city rankings help trigger a discussion process about regional development strategies and stimulate cities to learn from each other. Grabow (2006) stated that city rankings may aid cities in making strategic decisions, while Besecke and Herkommer (2007) argued that city rankings give insight into the strengths and weaknesses of cities and may therefore be used for city planning and development.

In the European Union almost 75% of the population lives in cities. Therefore, the European Union is committed to make its cities more sustainable (European Commission, 2010a). Multiple city rankings have been developed that specifically focus on measuring the environmental sustainability of European cities. Take for example the European Green Capital Award, the European Green City Index and Urban Ecosystem Europe. These European city rankings may contribute to the evaluation and development of environmental policy of European cities. Like rankings in general, city rankings have some methodological issues (Besecke & Herkommer, 2007; Grabow, 2006; McManus, 2012; Schönert, 2003). A serious problem of city rankings is that their methodological characteristics are rarely considered (Giffinger et al., 2007a). This includes city rankings that aim to measure the environmental sustainability of European cities. Because of methodological differences, a city may have a high position in one ranking and simultaneously a low position in another ranking. For example, Vienna ranked fourth place (out of 30 cities) in the European Green City Index 2009 and thirteenth place (out of 35 cities) in the European Green Capital Award 2010 (which was also published in 2009). The sensitivity of city ranking outcomes to methodological choices obviously poses problems for urban environmental policy makers. The objective of the current study is therefore to identify and evaluate the methodological characteristics of existing city rankings that aim to measure the environmental sustainability of European cities. The knowledge from this research may be used to help urban policy makers deal with European city rankings and to improve the methodological quality of (future) city rankings.

5.2 Literature review on city ranking methodology

The development of a city ranking consists of several phases: the decomposition of the ranking attribute into indicators, the aggregation of indicators into a composite index, the selection of cities, the data collection, and the reporting. Literature was reviewed to identify

methodological issues within each phase and to develop a methodology for systematically identifying the methodological characteristics of city rankings. The literature that was reviewed is discussed below.

Decomposition of the ranking attribute

To measure a city's performance on a complex ranking attribute, the ranking attribute needs to be decomposed into indicators. For example, to measure a city's environmental sustainability, indicators concerning air quality (e.g. annual daily mean of PM_{10} emissions), energy consumption (e.g. annual energy consumption in gigajoules per resident), and waste production (e.g. annual waste collected in kilograms per resident) could be used. Ideally, the decomposition of a ranking attribute into indicators is based on a theoretical framework (Giovannini et al., 2008; Ham et al., 2004). Such a framework should provide a clear definition of the ranking attribute, including its underlying categories and criteria for selecting indicators (Giovannini et al., 2008). Remarkably, rankings sometimes do not provide a (clear) definition of their ranking attribute (see Nesshöver et al. (2007) for an example). This obviously complicates the justification of the selected indicators.

The selection of indicators may also be justified by the use of stakeholders or experts (Morse & Fraser, 2005; Singh et al., 2009). Experts can be acquired from within the organization that initiated the ranking (and its project partners) or they can be acquired from outside. It is important to make this distinction, because the use of internal experts may more easily bias the indicator selection than the use of external and perhaps more independent experts. The selection of indicators may also depend on political and practical considerations. Developers of rankings may choose to align their selection of indictors with certain policy frameworks or discard the use of indicators for which data are not readily available. Especially in the context of international rankings, data availability is a severe selection criterion due to the scarcity of internationally comparable data (Giovannini et al., 2008). When comparing and ranking cities across countries, data availability also poses a problem (Kahn, 2006; Türksever & Atalik, 2001).

Usually, developers of rankings can choose from a wide range of indicators. For example, when measuring CO_2 emissions many different indicators could be selected (e.g. total CO_2 emissions in tonnes per resident, total CO_2 emissions in grams per unit of cities' gross domestic product). As such, developers of rankings need to select a limited number of indicators that still captures the meaning of the ranking attribute as a whole (Grabow, 2006). It is important that developers justify their specific selection of indicators, because it can severely influence the final ranking outcome (Lun et al., 2006; Maretzke, 2006).

With regard to the methodological characteristics of city rankings, a clear definition of the ranking attribute is pivotal. Without such a definition it is impossible to determine if appropriate indicators were selected. Some definitions of urban environmental sustainability and closely related concepts are provided in the literature. According to Kahn (2006, p. 4) green cities have clean air and water, are resilient in the face of natural disasters, run a low risk of major infectious disease outbreaks, encourage green behaviour, and have a relatively small ecological impact. Goodland (1999, p. 715) defined environmental sustainability as the "maintenance of natural capital", with natural capital as a provider of inputs (e.g. air, water, energy) and as a sink of waste emissions (e.g. greenhouse gases). Others defined urban environmental sustainability by decomposing it into categories such as air, water, energy, and solid waste (Shane & Graedel, 2000; Yu & Wen, 2010). Although there are similarities between definitions, one widely accepted definition of urban environmental sustainability does not yet exist. Therefore, European green city rankings should clearly state and define their ranking attribute. Additionally, they need a theoretical framework or emerging theory (acquired from expert opinion) to justify their specific selection of indicators.

Aggregation of indicators

When aggregating indicators into a composite index, it is important to consider their measurement level (nominal, ordinal, interval or ratio) as this has implications for the legitimacy of any subsequent mathematical operation (Coste et al., 1995). In this regard it is especially important to consider the applied normalization technique. There are numerous ways to normalize indicators (Giovannini et al., 2008; Singh et al., 2009). Commonly used techniques are ranking, ordinal categorization and re-scaling. Ranking and ordinal categorization are usually applied when qualitative data were collected. For each indicator, the collected data are evaluated (e.g. by experts) on the basis of which objects are ranked or assigned to ordinal categories. Ranking and ordinal categorization can also be applied on indicators with an interval or ratio measurement level, however, information regarding relative distances between objects is lost. Re-scaling is usually applied when indicators have an interval or ratio measurement level. When using this technique, the scales of indicators are mathematically transformed to have an identical range (e.g. 0-100) and information regarding relative distances between objects is retained.

When aggregating (normalized) indicators into a composite index, weights may serve to reflect the importance of each indicator in measuring the ranking attribute. There are numerous ways to determine the weights of indicators. Although the easiest way is to use equal weighting, the assumption that all indicators are equally important is unappealing (Kahn, 2006). Ideally, indicator weights are based on a theoretical framework (Ham et al., 2004; Ochel & Röhn, 2008), but it is also possible to determine weights based on expert and

stakeholder opinion (Giovannini et al., 2008; Grabow, 2006; Kahn, 2006; Mayer, 2008; Morse & Fraser, 2005).

Most commonly, indicators are aggregated into a composite index using either the additive or multiplicative method (Besecke & Herkommer, 2007; Grabow, 2006; Maretzke, 2006; Schwengler & Binder, 2006). Whereas the additive method allows an object to compensate a low value on one indicator with a high value on another indicator, the multiplicative method partly prevents such a compensation.

With regard to the methodological characteristics of city rankings, it is important to consider the measurement level of indicators as it determines which subsequent normalization techniques are appropriate. In addition, it is important to determine which normalization, weighting and aggregation techniques were used, because multiple studies showed that these techniques can severely influence the final ranking outcome (Floridi et al., 2011; Jacobs et al., 2005; Lun et al., 2006; Maretzke, 2006; Schwengler & Binder, 2006).

Selection of cities

To prevent rankings from comparing apples with oranges, it is recommended to select a relatively homogeneous group of objects (Ochel & Röhn, 2008). For this, selection criteria are needed. When city rankings apply selection criteria, they are actually defining their target population. Most city rankings start from a certain geographic scope (e.g. only European cities). Population size is often used as an additional selection criterion (Giffinger et al., 2007a). Hendriksen et al., (2010) proposed to build a city typology based on cities' population density, economic character, wealth, climate, and history. Such a typology may be useful for establishing a homogeneous group of cities.

If the defined target population contains many cities, a final selection of cities needs to be made. In this case it is possible to draw a random sample of cities from the target population. However, cities within the target population are often purposively sampled based on practical considerations like data availability and the perceived importance of the cities for the target audience of the ranking (Giffinger et al., 2007a; Taylor, 2011). Some rankings use convenience sampling by allowing cities within the target population to apply for participation.

For city rankings it is important to be clear on the applied selection criteria. These criteria determine to what extent the target population of cities can be regarded as a homogeneous group of objects. In addition, it is important to describe the subsequent sampling strategy as this gives end-users of city rankings insight into the extent to which the sample of cities reflects the target population.

Data-collection

Data can be collected from different sources. Obvious sources for data on cities are local city authorities. Official institutions like national statistical offices and international organizations (e.g. Eurostat, European Environment Agency) also provide data on cities in the form of (publicly available) databases and reports. Various expert and stakeholder groups are another data source.

According to Ochel and Röhn (2008), the quality of data collected from official institutions and international organizations is largely unproblematic, although there are limitations concerning, among others, the actuality and accuracy of the data. They argued that data collected from experts are usually more up to date and may represent reality better. On the other hand, they stated that the quality of expert data may be more problematic as experts may be biased by sentiments. Nevertheless, because data from any source may be biased, it is deemed important that city rankings have a procedure for checking the quality of the collected data. A possible procedure could be the use of triangulation which involves comparing data collected from one source with data collected from another source to see whether they corroborate (Silverman, 2011).

Giffinger et al., (2007a) studied several city rankings and concluded that most of them collected data through desk research. Other methods mentioned with regard to rankings, are: interviews (Grabow, 2006; Schönert, 2003), surveys (Besecke & Herkommer, 2007; Hendriksen et al., 2010), and participant observations (Besecke & Herkommer, 2007).

When data have been collected, two issues possibly need to be dealt with. First, data collected on an indicator may be qualitative in nature. With regard to rankings, qualitative data always have to be quantified. This may be done by the initiator of the ranking (or its project partners) or some external party (e.g. expert group). In any case, a list of criteria should be used to make sure all objects are evaluated in a similar way. Second, there may be missing data. Missing data can be handled in different ways: it can be supplemented based on data collected from additional sources, it can be estimated using imputation techniques (Giovannini et al., 2008), or it can become part of the evaluation process.

With regard to the methodological characteristics of city rankings, it is important to determine how and from what source(s) data were collected. Additionally, it is important to know how any qualitative data were quantified, how missing data were handled, and how the quality of data was checked. All these characteristics provide insight into the quality of the collected data on which a city ranking is based.

Reporting

Taylor (2011) showed that many economic city rankings are not transparent about their methodology. It is important, however, that end-users of rankings understand how rankings were developed and how this may have influenced the final result (Mayer, 2008). This implies that a ranking should report as much information as possible about its methodological characteristics. Lun et al. (2006) argued that every ranking should report a sensitivity analysis which shows how robust the ranking is to changes made to certain methodological characteristics like the number of indicators and the applied normalization, weighting, and aggregation techniques.

A ranking of objects on an overall ranking attribute does not give detailed insight into the strengths and weaknesses of objects. Hence, reporting an overall ranking hardly provides meaningful information useful for practice (Nesshöver et al., 2007). A report in which objects are ranked on separate dimensions and indicators may be more useful (e.g. for urban policy makers). When many different cities are included in a ranking, it would also make sense to report (decomposed) results per city type to provide a meaningful comparison between cities. Again, a city typology might prove useful in this regard.

A disadvantage of rankings is that they suggest differences between objects that may not exist. Objects in the middle of a ranking usually differ little from each other when it comes to their composite index values (Grabow, 2006). Therefore, reporting the composite index value (i.e. the rating) next to the rank of each object enhances insight into the differences between objects. Another option is to classify objects in clearly distinguishable groups such as forerunners, mediocrities, and laggards (Grabow, 2006; Lun et al., 2006; Maretzke, 2006).

With regard to the methodological characteristics of city rankings, it is important to determine the transparency of the procedure followed and whether a sensitivity analysis was performed. In addition, it is important to determine to what extent the overall ranking result is decomposed (to dimension/indicator level and into different city types) and which alternatives to the ranking of objects are reported (e.g. a rating or classification of objects). The report determines the methodological transparency of a city ranking and thus its usefulness for end-users.

5.3 Methods

The developed methodology, for systematically identifying the methodological characteristics of city rankings, was applied on several city rankings. City rankings were included in this study if they satisfied all of the following criteria: having a European scope (ranking cities from at least two different European countries), being based on an indicator

system (consisting of at least two indicators), having a primary focus on measuring the environmental sustainability of cities or a more specific aspect of the environmental sustainability of cities (e.g. air quality), publishing a publicly available ranking of cities (either on their website or in a separate report), publishing a ranking within the years 2007 through 2012.

City rankings were found by searching on the internet, reading literature and talking to experts on urban sustainability and urban policy makers. Based on these sources, a list of potentially suitable city rankings was established. Following further investigation, several city rankings were excluded from the list as they were not primarily focussed on measuring environmental sustainability (Cushman & Wakefield, 2010; Giffinger et al., 2007b; Mercer, 2012) or did not publish a publicly available ranking of cities (Climate Alliance, 2012; Foundation of Environmental Education, 2013).

Table 5.1 gives an overview of the six city rankings that were included in this study. Different kinds of institutions are involved in the development of these city rankings: energy companies, government agencies, city network organizations, commercial businesses, research institutes, and NGOs. Some of the city rankings in Table 5.1 were conducted multiple times, while other city rankings were conducted only once. When a city ranking was conducted more than once within the years 2007 through 2012, data were collected on the most recent version within that period.

For each included city ranking the official website, the ranking report, as well as any additional methodological background documents that could be found on the website were studied. In addition, semi-structured interviews were conducted with representatives of the selected city rankings to get insight in any methodological information that could not be obtained otherwise. A summary of each interview was sent to the corresponding representative(s) with the question to check it on errors and completeness. Representatives of four of the six city rankings replied (representatives of the European Green City Index and the European Soot-free City Ranking did not reply).

An overview of the interviewees and the websites, reports, and methodological background documents that were included in the analysis is found in appendix 5.1. Data were analysed by means of content analysis. The developed methodology provided codes which were used to identify the methodological characteristics of the city rankings in the data.

Table 5.1
Overview of six city rankings included in the current study.

Ranking	Main initiator(s)	Main project partner(s)	Times published[a]	Version of ranking studied
1. European Energy Award (EEA)	Partners from Switzerland (Brandes Energie), Austria (Energieinstitut Vorarlberg), Germany (Beratungs- und Service-gesellschaft Umwelt), and Poland (KESCO Energy).	The international office EEA, Forum EEA, various national and regional EEA offices responsible for organizing the EEA in their nation/region.	11	2012
2. European Green Capital Award	European Commission (DG Environment) based on the idea of Mr. Jüri Ratas, former mayor of Tallinn and the Tallinn memorandum.	European Parliament, European Environment Agency, European Environmental Bureau, Committee of the Regions, Covenant of Mayors, ICLEI[b], RPS[c] Group (manages the Secretariat on behalf of DG Environment).	3	2012
3. European Green City Index	Corporate communications department of Siemens Headquarters.	Economist Intelligence Unit.	1	2009
4. European Soot-free City Ranking	Bund für Umwelt- und Naturschutz Deutschland.	European Environmental Bureau, Deutsche Umwelthilfe, Naturschutzbund Deutschland, Verkehrsclub Deutschland, ClimateWorks.	1	2011
5. Renewable Energy Systems (RES) Champions League	Comité de Liaison Energies Renouvelables.	Solarthemen, Association of Bulgarian Energy Agencies, League of Ecological Alternatives, Energy Club Environmental Association, Polish Network "Energie-Cites", Legambiente, Deutsche Umwelthilfe, Climate Alliance, European Commission.	Constantly updated on website	2012
6. Urban Ecosystem Europe	Ambiente Italia.	Dexia, Legambiente, ICLEI[b], Comité '21, Climate Alliance, Union of the Baltic Cities, MedCities, National Association of Italian Cities, Agenda 21 Locali Italiane.	2	2007

[a] Until the end of the year 2012.
[b] International Council for Local Environmental Initiatives.
[c] Rural Planning Services.

5.4 Results

Decomposition of the overall ranking attribute

Five city rankings reported a ranking on an overall ranking attribute (see Table 5.2). Most of these city rankings (numbered in all tables: 1, 2, 4, 5) did not provide a clear definition of the overall ranking attribute on the official website, in the ranking report, nor in methodological background documents. Only the European Green City Index stated in the methodology section of the report what the ranking as a whole measured. During the interviews, most representatives of the city rankings were able to provide a description of the overall ranking attribute. The representative of the European Green Capital Award was unable to provide a description of the overall ranking attribute and stated that the overall ranking is a technical combination of the twelve separate indicators. During the interview, it was also explained that a jury assessed the three highest ranked cities "on their ability to be green role models for other cities" and then selected in July 2012 the winner of the European Green Capital Award 2014.

Urban Ecosystem Europe was the only city ranking that did not report a ranking on an overall ranking attribute. During the interview the representative of Urban Ecosystem Europe stated that Ambiente Italia decided not to rank cities on an overall ranking attribute, because the communication of a single index could be misleading from a technical point of view.

All city rankings that reported a ranking on an overall ranking attribute, decomposed this attribute into a number of indicators. The number of indicators used in the different city rankings ranged from four to 79. Four city rankings grouped the indicators into a smaller number of categories. None of the city rankings provided a clear explanation on the official website, in the ranking report, nor in methodological background documents about how the indicators were selected.

Additional information about the selection of indicators was asked during the interviews. It turned out that some of the city rankings (1, 2, 5, 6) based the selection of indicators at least partly on a former national city ranking or previously developed indicator framework. Experts were also used for the selection of indicators. Aside from the use of internal experts, acquired from within the initiating institution of a city ranking or its project partners, external experts were (also) used by some city rankings (1, 3, 6). None of the city rankings based the selection of indicators on a theoretical framework.

Table 5.2
Methodological characteristics of the six city rankings regarding the decomposition of the overall ranking attribute.

Ranking	Overall ranking attribute	Number of categories	Number of indicators	Basis of indicator selection
1. European Energy Award (EEA)	Cities were ranked based on their energy policy performance.[a]	6[a]	79	- Expert panel (including experts from municipalities, energy agencies and other energy experts) discussed former assessment tool developed for Switzerland and established initial indicator list for EEA.[a] - Initial indicator list was updated (period 2009-2012) during meetings in which people from national and regional EEA offices discussed feedback acquired from EEA consultants, EEA auditors, and municipalities.[a]
2. European Green Capital Award	The overall ranking was a technical combination of 12 separate indicators.[a]	Not available.	12	- Indicators were inspired by the 10 European Common Indicators and indicators part of the Aalborg Commitment. - Indicators were based on the 6th Environmental Action Program and the Thematic Strategy on the Urban Environment.[a] - Indicators were updated every year based on feedback from expert panel and participating cities, and publications from the European Commission.[a]
3. European Green City Index	The European Green City Index measured the current environmental performance of major European cities, as well as their commitment to reducing their future environmental impact by way of ongoing initiatives and objectives.	8	30	- Economist Intelligence Unit made initial selection of indicators which was commented on by Siemens (by discussing the indicators internally with their experts).[a] - Based on a meeting with an independent expert panel (consisting of urban sustainability experts) the initial selection of indicators was discussed and a final selection was made.[a]

[a] Information based on interviews.

Table 5.2 (continued)
Methodological characteristics of the six city rankings regarding the decomposition of the overall ranking attribute.

Ranking	Overall ranking attribute	Number of categories	Number of indicators	Basis of indicator selection
4. European Soot-free City Ranking	Cities were ranked based on the efforts made to improve the air quality (with a specific focus on black carbon and fine particles).[a]	Not available.	9	- Experts from within the Bund für Umwelt und Naturschutz Deutschland (BUND) and other partner NGOs selected indicators.[a] - Former head of the department for transport of the German Environmental Agency consulted BUND on the indicators.[a]
5. Renewable Energy Systems (RES) Champions League	On the combined solar ranking, cities were ranked based on their solar energy usage (use of solar photovoltaic & solar thermal technologies).[a]	3	4	- CLER[b], Solarthemen, League of Ecological Altrnatives, and Climate Alliance decided which renewable technologies to include in the ranking.[a] - Representatives and experts from all project partners decided on ranking rules.[a] - Rules used in the German Solarbundesliga were largely adopted.[a]
6. Urban Ecosystem Europe	There was no overall ranking attribute. Individual indicators were not aggregated into an overall ranking attribute.	6	25	- Expert panel (including politicians and experts from ten Italian local city authorities) agreed on indicators of the Italian edition of the Urban Ecosystem.[a] - Indicators from the Italian edition were selected for the European edition based on data availability.[a] - Indicators were also selected based on various pre-existing indicator frameworks (e.g. Urban Audit, 10 European Common Indicators) and policy frameworks (Aalborg Commitments, Thematic Strategy on Urban Environment, the Leipzig Charter).

[a] Information based on interviews.
[b] Comité de Liason Energies Renouvelables.

Aggregation of indicators

Most of the city rankings (1, 2, 3, 4, 6) contained indicators that required the collection of qualitative data, whether or not supplemented by quantitative data (see Table 5.3). These indicators had a nominal measurement level. Some city rankings (3, 5, 6) contained indicators that required the collection of quantitative data expressed as a single number (having a natural zero value). These indicators had a ratio measurement level. Some indicators of the Urban Ecosystem Europe consisted of multiple items which required the collection of a combination of nominal, ordinal, interval, and ratio data. For the purpose of this study these indicators were classified as having a nominal (multi-item) measurement level.

Different normalization techniques were applied: ranking (2), ordinal categorization (3, 4), and re-scaling (3, 6). The developers of the European Energy Award and the RES Champions League normalized their indicators by attributing points to indicators. Both city rankings incorporated a weighting of indicators into the normalization technique by varying the maximum amount of points each indicator could yield.

Three city rankings (1, 3, 5) weighted their indicators. Of these three, the European Energy Award and the European Green City Index based the weighting of indicators on feedback from the same experts who contributed to the selection of indicators. The RES Champions League based the weighting of indicators on the experience of their partner organization Solarthemen.

All city rankings, except for the Urban Ecosystem Europe, added the normalized (and weighted) indicator values to obtain the composite index value. To take into account different city conditions, the European Energy Award divided the composite index value of each city by the maximum amount of points each city theoretically could obtain. The RES Champions League calculated bonus points for each city, based on the specific combination of points assigned to the solar photovoltaic and solar thermal category, which were added to the composite index value.

Selection of cities

The number of cities included in the different city rankings ranged from 17 to more than 3675 (see Table 5.4). City selection criteria varied across city rankings. Aside from a geographic scope, the European Energy Award and RES Champions League did not apply any selection criteria. All other city rankings used at least the number of inhabitants as an additional selection criterion. None of the city rankings used a city typology.

Table 5.3
Methodological characteristics of the six city rankings regarding the aggregation of indicators.

Ranking	Measurement level indicators	Normalization technique	Weighting technique	Aggregation technique
1. European Energy Award (EEA)	Nominal[a]	For each indicator an EEA consultant determined the maximum number of points (between 2 and 10) a city theoretically could obtain (this could differ per city).	Because of the normalization method, indicators had different weights.[a]	- Normalized indicator values were added to obtain a composite index value. - Percentage score was calculated by dividing the composite index value by the maximum number of points a city theoretically could obtain.[a]
2. European Green Capital Award	Nominal	Ranking (on a scale of 1 to 18 by an expert panel).	Indicators had equal weights	Normalized indicator values were added to obtain a composite index value.
3. European Green City Index	Nominal and Ratio	- For nominal indicators: ordinal categorization (on a scale of 0-10 by analysts of the Economic Intelligence Unit). - For ratio indicators: re-scaling (on a scale of 0-10).	- Across categories, indicators had different weights. - Within categories indicators had equal weights.[b] - All categories had equal weights.	- Normalized indicator values were added to obtain category values (which were then rebased onto a scale of 0-10). - Rebased category values were added to obtain a composite index value (which was then rebased onto a scale of 0-100).
4. European Soot-free City Ranking	Nominal	Ordinal categorization (on a scale of 1 to 5 by staff of the Bund für Umwelt und Naturschutz Deutschland).	Indicators had equal weights.	- Normalized indicator values were added to obtain a composite index value. - Percentage score was calculated by dividing the composite index value by 45.
5. Renewable Energy Systems (RES) Champions League	Ratio	- For each indicator a number of points was obtained (based on a per capita ratio). - There was no maximum on the number of points that could be obtained on an indicator.	Because of the normalization method, indicators had different weights.	- Normalized indicator values were added to obtain a composite index value. - Bonus points were added when a balanced mix of solar photovoltaic and thermal installations was present.
6. Urban Ecosystem Europe	Ratio and nominal (multi-item)	For 9 ratio indicators: re-scaling (to a scale of 0-10)	Not available.	Not available.

[a] Information based on interviews.
[b] Except for the indicator 'Size of non-car transport network' which received a lower weight than the other indicators within the category 'Transport'.

Table 5.4
Methodological characteristics of the six city rankings regarding the selection of cities.

Ranking	Number of cities	City selection criteria	Sampling strategy
1. European Energy Award (EEA)	500+	All cities, within a state or region which had acquired a licence in the EEA programme, could participate.	Convenience sampling.
2. European Green Capital Award	18	- Cities had to be located in European Union member states, candidate countries, or European Economic Area countries. - Cities had to have more than 200,000 inhabitants (in countries without such cities, the largest city could participate). - Past winners could not participate for a period of ten years after they won the award.	Convenience sampling.
3. European Green City Index	30	Within each of Europe's 30 largest countries (excluding Russia) the largest city was selected (based on population size).[a]	Not available.
4. European Soot-free City Ranking	17	- Cities had to be capital or major cities in Western Europe with more than 300,000 inhabitants. - Cities had to have high exceedances of PM_{10} emission levels. - Cities had to be candidates for "best practice" of soot-reduction measures. - Cities had to house local NGO partners from the European Environmental Bureau network.	Purposive sampling of cities that were good or bad examples of how to deal with air pollution, and were expected to generate enough media attention.[a]
5. Renewable Energy Systems (RES) Champions League	3675+	- Cities had to be located in a country that had a national League (Bulgaria, Czech Republic, Italy, France, Germany, Hungary or Poland). - National Leagues could have additional selection criteria.[a]	Convenience sampling.
6. Urban Ecosystem Europe	32	- Cities had to be located in EU member states, candidate countries, or neighbouring countries. - Cities had to be capital or medium-big cities (more than 150,000 inhabitants).	Purposive sampling of cities that were part of a European city network or in which Ambiente Italia had personal contacts (taking into account a sufficient spread in geographical area and population size).[a]

[a] Information based on interviews.

None of the city rankings randomly selected cities from their target population. Three city rankings (1, 2, 5) applied convenience sampling by allowing any city to participate which met the selection criteria. The European Soot-free City Ranking and the Urban Ecosystem Europe on the other hand purposefully selected cities from all the cities which met the selection criteria. The European Green City Index applied strict selection criteria which made subsequent sampling of cities unnecessary.

Data-collection

All city rankings used local city authorities as a data source (see Table 5.5). In addition, four city rankings (1, 2, 3, 4) employed certain stakeholder or expert groups to evaluate and quantify the qualitative data. These expert or stakeholder groups provided quantitative data used to construct the city ranking and can thus be considered as data sources as well. All city rankings, except for the European Green City Index, used some kind of standardized questionnaire to collect data from their data source(s).

With three city rankings (1, 3, 4), the evaluation and quantification of qualitative data were done using a guide or list of criteria. The experts who evaluated the data collected for the European Green Capital Award, however, did so according to their own personal discretion.

With some city rankings (3, 4, 6) the initiator of the ranking (or the project partner responsible for the data-collection) actively supplemented missing data by collecting data from other sources. Project partners of the RES Champions League did not actively supplement missing data, however, (other) sources were allowed to provide additional data at any time. With the European Energy Award and the European Green Capital Award missing data were not supplemented or estimated in any way, but became part of the evaluation process. All city rankings, except for the European Green City Index, performed some activities to check the quality of the collected data.

Reporting

Table 5.2 to 5.5 show that none of the city rankings provided complete information about their methodological characteristics. Additional information had to be obtained from interviews. In particular, information about the overall ranking attribute and the selection of indicators was missing. In addition, none of the rankings, that reported a ranking on an overall ranking attribute, mentioned something about the robustness of the ranking and its sensitivity to the applied normalization, weighting, and aggregation techniques.

Table 5.5
Methodological characteristics of the six city rankings regarding the data-collection.

Ranking	Data source(s) used	Method(s) used	Evaluation of qualitative data	Handling of missing data	Data quality check
1. European Energy Award (EEA)	- Energy teams (including representatives of local city authorities, political actors, external energy experts, committed citizens). - EEA consultants.	Standardized questionnaire (filled in by EEA consultants).	EEA consultants evaluated qualitative data using an assessment guidance (which could differ from country to country).[a]	- Missing data were not supplemented in any way.[a] - When data on an indicator were missing, a city received zero points on that indicator.[a]	- Only accredited EEA consultants were allowed to fill in the standardized questionnaire. - When cities achieved a score of 50%, external EEA auditors verified the collected data.
2. European Green Capital Award	- Local city authorities. - Expert panel.	Standardized questionnaire (filled in by local city authorities).	- Experts evaluated qualitative data according to their own personal discretion. - Experts did not use a list of criteria.[a]	- Missing data were not supplemented in any way. - When data on an indicator were missing, it had a negative effect on the ranking of a city on that indicator.[a]	- Experts could ask cities for clarifications. - The evaluation and ranking of cities by each expert was peer reviewed by another expert.
3. European Green City Index	- (Publicly available) data provided by local city authorities. - National sources (e.g. national statistical offices and environmental bureaux). - Analysts from the Economic Intelligence Unit (EIU).	Desk research (performed by in-house and external contributors from EIU).	EIU analysts evaluated qualitative data using a list of criteria.	Missing data were supplemented by producing estimates from national sources (e.g. national statistical offices and environmental bureaux).	None.

[a] Information based on interviews.

Table 5.5 (continued)
Methodological characteristics of the six city rankings regarding the data-collection.

Ranking	Data source(s) used	Method(s) used	Evaluation of qualitative data	Handling of missing data	Data quality check
4. European Soot-free City Ranking	- Local city authorities. - Cities' air quality and transport action plans. - Staff of the Bund für Umwelt und Naturschutz Deutschland (BUND). - Local NGO partners.	Standardized questionnaire (filled in by local city authorities, local NGO partners or BUND staff).	BUND staff evaluated qualitative data using a list of criteria.	When local city authorities did not respond, cities were evaluated by BUND staff (through desk research) and local NGO partners.	Local NGO partners commented on the questionnaire filled in by local city authorities and gave feedback on individual city gradings and a draft ranking.
5. Renewable Energy Systems (RES) Champions League	Anyone who could make the requisite figures seem plausible.	Standardized questionnaire.	Not available.	- When data on an indicator were missing, a city received zero points on that indicator. - Different data sources were allowed to provide additional data at any time.[a]	The organizations responsible for the national Leagues needed to monitor, correct and approve the data for release.
6. Urban Ecosystem Europe	Local city authorities.	Standardized questionnaire (filled in by local city authorities)	Qualitative data were not evaluated.	- If possible, missing data were supplemented by Ambiente Italia using publicly available data. - Local city authorities needed to confirm this data.	- Ambiente Italia checked data sent by cities with publicly available data (e.g. Urban Audit database). - Ambiente Italia also checked data based on their own experience.[a]

[a] Information based on interviews.

All city rankings, except for the Urban Ecosystem Europe, reported composite index values (i.e. ratings) next to the overall ranks of cities to give insight into the size of the differences between cities. None of the city rankings classified cities into groups such as forerunners, mediocrities, and laggards.

Most city rankings (2, 4, 5, 6) reported city rankings and ratings on individual indicators. The European Green City Index reported city rankings and ratings on each of its eight categories. The RES Champions League is the only city ranking that reported separate rankings for different types for cities (based on the number of inhabitants). The Urban Ecosystem Europe is the only city ranking that reported city profiles (showing the rating of each city on nine indicators in comparison to the median rating).

5.5 Discussion

Based on literature a methodology for identifying the methodological characteristics of city rankings was developed and applied on six city rankings that aimed to measure the environmental sustainability of European cities. Results showed that the city rankings varied greatly with respect to their methodological characteristics and that all city rankings had important methodological weaknesses.

Most remarkably, none of the city rankings that ranked cities on an overall ranking attribute provided a clear definition of the ranking attribute, which makes it hard to establish on what cities were actually ranked. This may in part be due to the difficulty of defining and operationalizing intangible concepts such as 'urban environmental sustainability'. Kuik and Gilbert (1999) stated that the operationalization of the related concept 'sustainable development' needs an interdisciplinary approach. They also pointed out that this concept is highly political which makes it unlikely that a common operationalization will be developed in the near future. The same may be said of the concept 'urban environmental sustainability'. Tobi (2014) argued that the operationalization of interdisciplinary concepts may be difficult but not impossible if all involved disciplines are willing to collaborate. Indeed, according to Türksever and Atalik (2001) basic consensus about the meaning of the related concept 'urban quality of life' was also achieved. In much the same way it is possible that future research will result in a definition and operationalization of 'urban environmental sustainability' that is supported by a wide range of scientific disciplines and political views.

Theoretical frameworks with corresponding indicators could be useful to developers of European green city rankings (Olewiler, 2006; Tanguay et al., 2010), but were not used by any of the examined city rankings to justify the selection of indicators. Most of the city rankings justified the selection of indicators by using experts. Although the use of experts may have contributed to the credibility of these city rankings, questions need to be asked

about how experts were defined and selected, to what extent experts were unbiased (especially when internal experts were used), and to what extent the selection of indicators by experts followed a systematic approach.

The Delphi method may provide an opportunity for developers of European green city rankings to develop their own emerging theory and to justify their specific selection of indicators through expert opinion. The aim of the Delphi method is to allow experts to achieve agreement on a certain topic (Linstone & Turoff, 1975). An important characteristic of the Delphi method is that expert participation is anonymous. This prevents bias as a result of group pressure and existing power relations. Using the Delphi method it may be possible for experts to achieve agreement on the definition and operationalization of 'urban environmental sustainability', including a selection of the most important indicators needed to measure the concept. Research is needed to find out whether the Delphi method is indeed suitable for this purpose.

Although indicators should ideally be selected based on their suitability to measure the overall ranking attribute, the lack of internationally comparable quantitative data limits the choice of indicators (Giovannini et al., 2008). Specifically with regard to European city rankings it may be difficult to find indicators for which comparable quantitative data are available. In response to this problem developers often select indicators which require the collection of qualitative data (Giovannini et al., 2008). This might explain why most of the examined city rankings also included this kind of indicators. Still, qualitative data need to be evaluated and quantified (e.g. by experts) for ranking purposes, which in turn may introduce bias.

A methodological weakness of all composite indices and rankings is the normalization, weighting, and aggregation of their underlying indicators. Although several studies showed that the applied techniques may severely distort the final results (Floridi et al., 2011; Jacobs et al., 2005; Lun et al., 2006; Maretzke, 2006; Schwengler & Binder, 2006), the examined city rankings hardly substantiated their choice for certain techniques. To obtain composite index values, the city rankings simply added the normalized and weighted indicator values, allowing cities to compensate low values on some indicators with high values on other indicators. A proper substantiation of applied techniques is also necessary to reduce the potential risks of lobbyism and developers manipulating ranking results. Again, the Delphi method may provide an opportunity, for example by using experts to establish indicator weights.

The examined city rankings used insufficient criteria to obtain a homogeneous target population of cities. Consequently, city rankings included many different types of cities (e.g. coastal and inland cities) that perhaps ought not be compared with each other in terms of environmental sustainability. Türksever and Atalik (2001) stated that it is important to take

cities' climatic, topographic and physical welfare features into account when measuring their quality of life. The same may be said when measuring cities' urban environmental sustainability. To prevent European green city rankings from comparing apples with oranges, a European city typology may be useful. Although such a typology exists (Seidel-Schulze et al., 2009), more research is needed to develop a European city typology that distinguishes different kinds of cities based on features that are important when measuring urban environmental sustainability.

Local city authorities appeared the most important source for data on urban environmental sustainability. All examined city rankings, except Urban Ecosystem Europe, also collected data from other sources. From all sources that have an interest in city rankings, biased data may be expected. Whenever qualitative data were evaluated by experts or other stakeholders, the possibility of bias was also introduced. Most city rankings checked the quality of the collected data using triangulation. Nevertheless, it remained unclear to what extent these checks really contributed to better data quality or introduced other bias (e.g. when local NGO partners were allowed to check data provided by local city authorities).

The reporting of the examined green city rankings left much to be desired for (Taylor, 2011). None of the green city rankings reported complete information on their methodological characteristics. Representatives of the city rankings were willing to be interviewed and were able to provide additional information on most methodological characteristics. A general shortcoming of rankings, relative to composite indicator systems, is that the composite index values of all objects are transformed into ordinal (rank) numbers and information about relative differences between objects is lost. Therefore, it is good to note that all the city rankings reported composite index values next to the ranks. Nevertheless, none of the city rankings reported a sensitivity analysis as has been suggested in the literature (Floridi et al., 2011; Lun et al., 2006; Singh et al., 2009). Without a sensitivity analysis end-users of rankings are unable to judge the robustness of the results (Lun et al., 2006).

5.6 Conclusion

The results of the current study are based on six city rankings that aim to measure the environmental sustainability of European cities and cannot be generalized to other types of city rankings or rankings of other objects. However, all rankings which are based on an indicator system do have similar methodological issues. The methodology as developed in this study can therefore be used in future research aimed at identifying the methodological characteristics of other rankings as well. As such, the methodology may enhance insight into the methodological quality of rankings in general and monitor their future methodological development.

Developers of city rankings are advised to use the methodology developed in the current study to identify methodological weaknesses in their ranking and make improvements where possible. For developers of European green city rankings several specific recommendations can be made. Most importantly, developers are advised to make use of existing research into the definition, operationalization, and measurement of 'urban environmental sustainability' and closely related concepts (such as presented here). Theoretical frameworks including indicators for measuring the concept, exist which could be useful. Alternatively, developers may use external urban sustainability experts as well as end-users to come up with their own tailored definition and operationalization of the concept with corresponding indicators. When selecting cities, developers ought to use selection criteria beyond geographical scope and population size. If possible, they ought to use a city typology to make sure similar types of European cities are compared on their environmental sustainability in a meaningful way. Although local city authorities are an important source for data on the environmental sustainability of cities, developers ought to verify data collected from this source. Preferably by using a source that has no interest in the ranking (e.g. national statistical offices, environmental bureaux, the urban audit database). If indicators are used that require the collection of qualitative data, developers are advised to use external experts, in the field of urban environmental sustainability, to evaluate and quantify these data according to pre-defined criteria. As for any ranking, developers of European green city rankings ought to provide full information on all methodological characteristics, preferably in a separate methodological background document that is freely accessible. In addition, developers ought to conduct and report a sensitivity analysis.

Finally, end-users of city rankings (e.g. urban policy makers and the media) are advised to use the methodology developed in the current study to identify and evaluate the methodological characteristics of city rankings. In particular urban environmental policy makers ought to find out how European green city rankings defined and operationalized the overall ranking attribute, whether the specific sample of cities allows meaningful comparisons to be made, and to what extent the results may be distorted and biased. In doing so, they may be able to identify the ranking which is most meaningful for evaluating and monitoring the environmental sustainability of their city. This implies that the usefulness of any ranking largely depends on how it reports its methodology. End-users ought to be careful using results of city rankings that provide insufficient information about methodological characteristics and do not provide any information about the robustness of the results.

Appendix 5.1

Ranking	Representative(s) interviewed (institution)	Website[a]	Documents[b]
1. European Energy Award (EEA)	Mrs. Kornmann (International office European Energy Award).	www.european-energy-award.org	- Optimising and successfully implementing Municipal Energy and Climate Protection Activities
2. European Green Capital Award	Secretariat European Green Capital Award (Rural Planning Services (RPS) group).	http://ec.europa.eu/environment/europeangreencapital/index_en.htm	- Expert Evaluation Panel – Synopsis Technical Assessment Report European Green Capital Award 2014. - Jury Report for the European Green Capital Award 2014. - The Expert Panel's Evaluation Work & Final Recommendations for the European Green Capital Award of 2010 and 2011.
3. European Green City Index	Mrs. Stelzner (Corporate communications department of Siemens Headquarters).	www.siemens.com/entry/cc/en/greencityindex.htm	- European Green City Index. Assessing the environmental impact of Europe's major cities.
4. European Soot-free City Ranking	Mr. Fellerman (Bund für Umwelt und Naturschutz Deutschland) & Mrs. Saar (Deutsche Umwelthilfe).	http://sootfreecities.eu	- European City Ranking. Best practices for clean air. Background Information.
5. Renewable Energy Systems (RES) Champions League	Mr. Witt (Solarthemen).	www.res-league.eu	- Starting a National League for Renewable Energy. Joining the European RES Champions League as a Partner. A Methodological Guide. - Res Champions League Rules
6. Urban Ecosystem Europe	Mr. Bono (Ambiente Italia).	http://informed-cities.iclei-europe.org/index.php?id=7597	- Report 2007 Urban Ecosystem Europe. An integrated assessment on the sustainability of 32 European cities.

[a] All websites were accessed before December 31, 2012.

[b] All documents were downloaded from the website before December 31, 2012.

Chapter 6

Defining and measuring urban sustainability in Europe: A Delphi study on identifying its most relevant components

This chapter is submitted as:
Meijering, J. V., Tobi, H. & Kern, K. Defining and measuring urban sustainability in Europe: A Delphi study on identifying its most relevant components.

Abstract

Context:
How urban sustainability is defined influences the results of urban sustainability rankings. Various efforts have been made to define the concept and to operationalize it into specific components (e.g. air quality, inequality, employment). Consequently, numerous different components are currently being used without agreement on which components are most relevant for defining and measuring urban sustainability.

Objectives:
This study identified which components experts find most relevant for defining and measuring urban sustainability in a European context. The study thereby provides insight into what the concept actually entails. This may facilitate the development of future urban sustainability rankings.

Methods:
A European sample of 419 urban sustainability experts was invited to participate in a three-round Delphi study. In each round experts were asked to evaluate and comment on the relevance of various components of urban sustainability.

Results:
The following seven components were identified as most relevant: air quality, governance, energy consumption, non-car transportation infrastructure, green spaces, inequality, and CO_2 emissions. Five of these components are part of the environmental dimension of urban sustainability, which suggests that urban sustainability is still perceived as mainly an environmental concept. Based on experts' evaluations of the components weights could be established that reflect the relative relevance of each component for measuring urban sustainability.

Conclusions:
This study provides an expert-based framework in which urban sustainability is operationalized into several weighted components. This framework may be used by future developers of urban sustainability rankings to properly define the concept and to select appropriate indicators.

Keywords: Delphi method, urban sustainability, definition, measurement, operationalization, components, dimensions, categories, themes, indicators

6.1 Introduction

With more than two-thirds of Europeans living in urban areas, Europe is one of the most urbanized continents in the world (European Commission, 2011). Although cities are considered to be the engines of the European economy, they are also home to many problems such as unemployment, poverty, and environmental pollution to name only a few (European Commission, 2011). The European Union is therefore committed to making its cities more sustainable (European Commission, 2010a). In 2008 the European Commission initiated the European Green Capital Award: a competition in which European cities are evaluated and ranked according to their environmental standards and commitment to future environmental improvement and sustainable development (European Commission, 2015). Since then, other institutions have also developed and published some form of urban sustainability ranking. Examples are the European Green City Index (Siemens, 2009) and the Sustainable Cities Index (Arcadis, 2015).

City rankings may be useful for urban governance, in particular urban planning and development (Besecke & Herkommer, 2007), but should also be used with caution because of methodological issues (Meijering et al., 2014). City rankings are often based on an indicator system. This means that the ranking attribute on which cities were finally ranked (e.g. urban sustainability) was operationalized into various indicators that each measure a specific aspect of the ranking attribute. For each city, data were collected on the indicators and then aggregated into a composite index value and corresponding rank number. Although rankings thus developed ought to reflect the performance of cities on the ranking attribute, they may be very sensitive to various methodological choices made, such as the techniques used to normalize and weigh indicators (Floridi et al., 2011; Jacobs et al., 2005; Lun et al., 2006).

A fundamental choice concerns the definition of the ranking attribute. How urban sustainability is defined influences the selection of indicators and thereby ranking results (McManus, 2012). In this regard it is problematic that many different definitions of urban sustainability exist (for an overview see: Huang et al., 2015). Since the report of the Brundtland Commission (World Commission on Environment and Development, 1987), it is widely accepted that sustainability consists of three pillars or dimensions: environmental, economic, and social sustainability (Hassan & Lee, 2015; Huang et al., 2015; Tanguay et al., 2010). Still, these three dimensions are very abstract and open to a wide range of interpretations. To help define and measure urban sustainability, various efforts have been made to divide the three dimensions into more specific components, also referred to as themes or categories (see for example: Huang et al., 1998; Michael et al., 2014; Tanguay et al., 2010). So far, these efforts all ended up with a different mix of components. As a result, many different components of urban sustainability are currently being used without agreement on which components are most relevant for defining and measuring the concept.

Meijering et al. (2014) suggested that agreement on the definition and operationalization of urban environmental sustainability may be achieved by using the Delphi method: a structured data-collection method that aims to facilitate a group of experts in achieving agreement on a topic (see also chapter 5 of this thesis). The method has indeed been frequently used to develop definitions and operationalizations of various concepts such as 'team effectiveness' (Lohuis et al., 2013) and 'acute respiratory distress syndrome' (Ferguson et al., 2005). In the current study the Delphi method was used to identify which components experts find most relevant for defining and measuring urban sustainability in a European context, thereby providing insight into what the concept actually entails and facilitating the development of future urban sustainability rankings. The study was restricted to the European context as urban sustainability is a place-dependent concept (Hassan & Lee, 2015) and may thus be defined and measured differently in different parts of the world.

6.2 Methods

The Delphi method

The Delphi method, developed in the 1950s by Dalkey and Helmer (1963), consists of at least two rounds of data collection. In the first round experts are independently questioned about their opinion on the topic of interest, usually by means of a standardized questionnaire. To prevent group pressure and inadvertent influence of dominant individuals, experts participate anonymously and do not directly communicate with each other. Instead, the study moderator provides experts with so called controlled opinion feedback: a summary of the findings from the previous round. Based on this feedback experts are allowed to reconsider and change their opinion in the second round. This process continues until a pre-specified number of rounds has been completed, a certain level of agreement has been achieved, or experts' opinions have stabilized (Diamond et al., 2014). In the current Delphi study a sample of urban sustainability experts was questioned in three subsequent rounds about which components are most relevant for defining and measuring urban sustainability in a European context.

Expert sample

Urban sustainability experts were considered to be people whose work is related to urban sustainability as inferred from the institution they work for, their position within that institution, their job description, or work related activities (i.e. participating in urban sustainability conferences or projects). With regard to Delphi studies it is recommended to compile a heterogeneous panel of experts to assure the inclusion of a diverse range of views

(Hussler et al., 2011; Powell, 2003). For the current study it was therefore decided to search for urban sustainability experts from four different types of institutions: academia, business, civil society (i.e. NGOs, non-profit, and community-based organisations that pursue charitable or member-oriented goals), and government. Furthermore, it was decided to search for experts from various European countries.

A convenience sample was assembled from various sources. Several conferences on urban sustainability that took place in Europe in 2013 or 2014 formed a first major source of experts. Initially, contributors (i.e. presenters and authors of accepted abstracts as mentioned in the conference program or proceedings) to the following three conferences were regarded as potentially suitable experts: The Sustainable City Conference 2014, The Urban Sustainability and Resilience Conference 2014, and The PLEA Conference 2013 (on sustainable architecture and urban design). Because the three conferences mainly yielded experts from academia, additional experts were acquired from two conferences targeted at a more diverse audience: The Future Cities Forum 2014 and The Reference Framework for European Sustainable Cities Conference 2013. Names of potentially suitable experts were researched online to acquire additional background information (i.e. the institution they work for, their position within that institution, their e-mail address) and to verify whether they held a position in an institution located in a European country.

The Joint Programming Initiative Urban Europe formed another major source of experts. This program was established in 2010 by the European Commission and aims to "Enhance the capacities and knowledge on transition towards more sustainable, resilient and liveable urban developments" (Robinson et al., 2015, p. 5). By means of two calls for proposals the program selected and funded 20 projects. The coordinators of these projects and their project partners as listed on the website were regarded as potentially suitable experts. Their names were researched online to acquire additional background information and to verify whether they held a position in an institution located in a European country.

Finally, by searching on the internet and talking to experts, many institutions were found that are active in the field of urban sustainability. For example, developers of urban sustainability rankings (e.g. Arcadis), partners of sustainable and smart city conferences (e.g. Accenture), EU funded research projects (e.g. TRANSFORM), partnerships (e.g. Climate-KIC), and consortia (e.g. Amsterdam Institute for Advanced Metropolitan Solutions). These institutes were contacted by telephone to find out whether they had urban sustainability experts who are willing to participate in the Delphi study. In some cases names and contact details of one or more experts were directly acquired. In other cases a contact person agreed to forward a ready-made invitation to potentially suitable experts or to put the invitation in a newsletter or on a website. In this invitation the objective and procedure of the Delphi study was shortly explained. Furthermore, experts were asked to register their participation by sending an e-mail to the research team with the following information: their

name, the name of the European country where they work, the name of the institution they work for, their job title, and how their work relates to urban sustainability. Experts who responded to the invitation were included in the sample. Based on all sources a sample of 419 experts from all over Europe was assembled.

Questionnaire development

The questionnaire for the first round of the Delphi study was developed based on several existing urban sustainability rankings. By entering a search query in three different search engines on the internet (Google, Bing, Yahoo) on two different computers, eight rankings were found that rank European cities on their sustainability or another closely related ranking attribute (see table 6.1). The components and the corresponding indicators of these eight rankings were identified and put together in one list. This list was content analysed by categorizing components of different rankings (e.g. air pollution, ambient air quality) and labelling each category with a single component name (e.g. air quality). As such, a total of 38 components was identified of which 22 components recurred in at least three of the eight rankings. These 22 components were given a single-sentence explanation and presented in a concept version of the questionnaire. This concept was pre-tested using the cognitive interview approach (Willis, 2005) with three urban sustainability experts and one research methodologist from Wageningen University. Based on the pre-tests, the questionnaire and the list of components were reviewed and refined (one component was removed from the list). The 21 components included in the final version of the questionnaire are listed in appendix 6.1.

Table 6.1
Overview of urban sustainability rankings used to construct the initial list of components.

Ranking	Initiator/developer	Edition
Cities of Opportunity	PricewaterhouseCoopers	2014
European Green Capital Award	European Commission	2016
European Green City index	Siemens/Economist Intelligence Unit	2009
European Smart Cities	Vienna University of Technology (Department of Spatial Planning)	2014
Networked Society City Index	Ericsson	2014
Sustainable Cities Index	Arcadis/Centre for Economics and Business Research	2015
The Smartest Cities in the World	Boyd Cohen	2014
The Sustainable Cities Index	Forum for the future	2010

In the questionnaire experts were asked to express their agreement or disagreement with two statements about their work in relation to urban sustainability, using a 7-point scale ranging from ‘strongly disagree’ to ‘strongly agree’. Next, experts were presented with the 21 components in a random order and were asked the following question: “Based on your expertise, how relevant are the following components for defining and measuring ‘Urban Sustainability’ in a European context?” For each component experts could give their answer

on a 10-point scale, ranging from 1 'not relevant at all' to 10 'entirely relevant'. Subsequently, experts were asked to explain for up to three components (they had rated with at least an 8) why they evaluated them as relevant. Additionally, experts were invited to suggest up to three relevant components that they missed on the list of 21 components. The questionnaire was concluded with several background questions to verify whether experts were assigned to the correct institution type and European country.

For the second Delphi round a questionnaire was developed which was similar to the previous one. Again, experts were asked to evaluate the relevance of the 21 components for defining and measuring urban sustainability. This time, each component was accompanied by a summary of the findings from the first round. This summary consisted of a table with summary statistics showing the component's median evaluation, interquartile range, and the percentage of evaluations equal to or greater than 8, compared to the most and least relevant component. The summary also provided a short explanation of why experts evaluated the component as relevant. As an example, figure 6.1 shows the summary that accompanied the component 'green spaces'. Next, experts were asked to evaluate the relevance of six additional components which were added based on experts' suggestions given in the first round (see appendix 6.1). Finally, experts were once again invited to explain for up to four components (rated with at least an 8) why they evaluated them as relevant.

Summary statistics of the component '**green spaces**' in comparison to the most and least relevant components:

	Median evaluation [1]	Interquartile range [2]	% of evaluations ≥ 8
Most relevant component	9	8 - 10	88%
Green spaces	9	8 - 10	81%
Least relevant component	6	5 - 8	28%

Note: experts evaluated components on a scale which ranged from 1 'not relevant at all' to 10 'entirely relevant'.
[1] When all evaluations of a component are ordered from lowest to highest, the median is the middle evaluation.
[2] The interquartile range contains the middle 50% of evaluations.

Summary of experts' explanations regarding why the component is relevant:
Green spaces provide many benefits for inhabitants of cities, such as an improved well-being and opportunities for recreation and community building. Furthermore, green spaces offer potential solutions for various problems related to among others climate change, water runoff, and the urban heat island effect. Green spaces are also relevant for nature and biodiversity within a city.

Based on your expertise, how relevant is the component 'green spaces' for defining and measuring 'Urban Sustainability' in a European context?
Please give your answer by using the scale which ranges from 1 'not relevant at all' to 10 'entirely relevant'.

	not relevant at all									entirely relevant	
	1	2	3	4	5	6	7	8	9	10	don't know
Green spaces: the amount of nature and parks within a city	○	○	○	○	○	○	○	○	○	○	○

Figure 6.1: screenshot of the questionnaire administered in the second Delphi round.

The questionnaire for the third and final Delphi round had a different set-up. First, an overall summary of the findings from the second round was presented. This summary consisted of two parts: (1) a table with summary statistics showing each component's median evaluation, interquartile range, and percentage of evaluations equal to or greater than 8, and (2) short explanations of why experts evaluated each component as relevant. Experts were asked to read the summary carefully and to select, from the list of 27 components, five components which are most relevant for defining and measuring urban sustainability. To acquire insight into the relative relevance of the five selected components, experts were also asked to distribute 100 points across these components. Finally, experts were invited to explain why they selected the five components as most relevant.

All three questionnaires were written in English, checked by a native English speaker, and programmed as web-surveys using the online survey builder Qualtrics (Qualtrics, 2015). In May 2015 an invitation e-mail was sent to all 419 experts in the sample. One week later experts received an e-mail with a link to the first online questionnaire. The e-mails with a link to the second and third questionnaire were sent in June and September 2015 respectively. In each round of the Delphi study experts were given up to three weeks to complete the questionnaire. Experts who did not respond received a maximum of two reminders.

Analysis

After the first and second round of the Delphi study the following statistics were calculated for each component: the median evaluation, the interquartile range, and the percentage of evaluations equal to or greater than 8. After the third round the percentage of experts that selected each component as most relevant was calculated as well as the total number of points attributed to each component.

To identify the most relevant components in the first and second Delphi round, a 95% confidence interval was calculated for each component around the percentage of evaluations equal to or greater than 8. A component was regarded as one of the most relevant when the lower limit of that 95% confidence interval was greater than 50%, implying that significantly more than half of all experts evaluated the component with at least an 8. To identify the most relevant components in the third round, a 95% confidence interval was calculated for each component around the percentage of experts that had selected it. A component was regarded as one of the most relevant when the lower limit of the 95% confidence interval was greater than 19%, the percentage of experts that would have selected the component based on chance alone. All confidence intervals were calculated using the Exact method (Newcombe, 1998).

The level of agreement among experts regarding the relevance of each component was estimated using the Strict Agreement index (denoted SA) (Meijering et al., 2013). This index is easy to interpret as it expresses the number of agreeing expert pairs as a proportion of the total number of possible expert pairs, whereby two experts only agree if they evaluated an item (i.e. component) using the same point on the rating scale. Originally, the SA was developed for a 5-point scale whereas in the current Delphi study a 10-point scale was used in the first and second round. Therefore, when calculating the SA in these two rounds the agreement criterion of the index was relaxed in such a way that two expert agree if their evaluations did not differ by more than one point on the rating scale. Theoretically, the SA can take on any value between 0 (none of the experts agreed with each other) and 1 (all experts agreed with each other). Preferably, the SA of the most relevant components as identified in this Delphi study is greater than 0.5 as this implies that more than half of all expert pairs agreed on the evaluation of these components.

The robustness of the results regarding the most relevant components was examined by testing whether there were significant differences between experts from different institutions and experts from different European regions. As most experts in the sample were from academia and the number of experts from business, civil society, and government was limited, a distinction was made between experts from academia and experts from other institutions. Charron et al. (2015) showed that countries in Western Europe (i.e. covering the United Kingdom (UK), Ireland, the three Benelux countries, France, Germany, Austria and Switzerland) and Northern Europe generally score better than countries in Southern Europe and in Central and Eastern Europe (CEE) regarding the quality of governance, a concept that is related to countries' social and economic development as well as their environmental conditions. Moreover, in most city rankings (e.g. European Green City Index, European Green Capital Award) cities in Northern and Western Europe generally score better than cities in Southern Europe and in CEE. Therefore, a distinction was made between experts from Northern and Western Europe on the one hand (in the following called North-West Europe) and Southern Europe and CEE on the other (in the following called South-East Europe). In the first and second Delphi round differences between groups were tested using the Mann-Whitney test (denoted U), whereas in the third round differences were tested using Fisher's exact test (denoted F) (Lindgren, 1976). Because of multiple testing, only differences with a p-value less than 0.01 were considered significant.

Content analysis (Gray, 2004) was performed on answers given on open questions. Suggestions for additional components relevant for defining and measuring urban sustainability, were analysed by categorizing comparable suggestions. For each category that contained more than four suggestions a new component with a corresponding single-sentence explanation was formulated that captured the underlying suggestions as fully as possible. Appendix 6.1 shows the resulting new components included in the second and third round of the Delphi study. Explanations on the relevance of each component were

analysed by categorising comparable explanations, labelling each category with a single sentence that captured the underlying explanations as fully as possible, and combining the labels into one short explanation stating why the component was considered to be relevant. Figure 6.1 shows an explanation for the component 'green spaces'.

6.3 Results

Response and drop-out rates

Table 6.2 shows the number of experts that responded in each round of the Delphi study, specified to institution type and European region. In the first round 33% of the 419 experts in the panel responded. Of these experts 93% agreed with the statement "It is important to me that my work relates to urban sustainability" and 86% agreed with the statement "In my work I spend most of my time on urban sustainability issues", indicating that there was a high level of commitment among the respondents with regard to urban sustainability. In the second and third round 29% and 27% of the experts dropped-out respectively. In all three rounds more experts from academia than experts from other institutions responded. Furthermore, more experts from North-West Europe (including experts from Austria, Belgium, Denmark, Finland, France, Germany, Luxembourg, Netherlands, Norway, Sweden, Switzerland, United Kingdom) than experts from South-East Europe (including experts from Albania, Greece, Italy, Lithuania, Poland, Portugal, Romania, Spain, Turkey) responded.

Table 6.2
Number of respondents in each round of the Delphi study.

Round	Total response	Response per institution type[a]		Response per European region[c, d]	
		Academia	Other[b]	North-West	South-East
1	139	86	46	86	43
2	99	59	35	61	31
3	72	47	22	39	26

[a] Seven expert were not classifiable due to conflicting information.
[b] Expert from business, civil society and government.
[c] Based on (Charron et al., 2015) with Switzerland categorized in North-West and Albania in South-East Europe.
[d] Ten experts were not classifiable as they stated to work on a European level or did not specify the correct country.

The first Delphi round

Table 6.3 shows per component the percentage of evaluations equal to or greater than 8 and the corresponding value of the SA in the first and second Delphi round. In the first round the percentage of evaluations equal to or greater than 8 ranged between 28% (international embeddedness) and 88% (non-car transportation infrastructure). Fourteen components

were evaluated with at least an 8 by significantly more than half of all experts and were thus identified as most relevant for defining and measuring urban sustainability. Of these fourteen components the following ten also had a SA greater than 0.5: non-car transportation infrastructure, air quality, CO_2 emissions, energy consumption, green spaces, health, solid waste, climate resilience, waste water treatment, and water usage.

Table 6.3
Percentage of expert evaluations ≥ 8 (95% Confidence Interval) and SA per component in the first and second Delphi round.

Component	Round 1 (n = 139)		Round 2 (n = 99)	
	% evaluations ≥ 8 (95% CI)	SA	% evaluations ≥ 8 (95% CI)	SA
Air quality	87 (80 - 92)[a]	0.64	89 (81 - 94)[a]	0.66
CO_2 emissions	87 (80 - 92)[a]	0.61	89 (81 - 94)[a]	0.58
Non-car transportation infrastructure	88 (81 - 93)[a]	0.61	89 (81 - 94)[a]	0.66
Energy consumption	83 (75 - 89)[a]	0.59	88 (80 - 94)[a]	0.68
Governance	n.a.	n.a.	88 (80 - 94)[a]	0.65
Green spaces	81 (74 - 87)[a]	0.58	87 (78 - 93)[a]	0.67
Health	82 (75 - 88)[a]	0.59	85 (76 - 91)[a]	0.63
Solid waste	81 (74 - 87)[a]	0.54	80 (70 - 87)[a]	0.57
Climate resilience	80 (72 - 86)[a]	0.55	78 (68 - 85)[a]	0.54
Waste water treatment	82 (74 - 88)[a]	0.55	77 (67 - 85)[a]	0.59
Water usage	82 (74 - 88)[a]	0.53	75 (65 - 83)[a]	0.58
Education	61 (53 - 70)	0.41	71 (61 - 80)[a]	0.51
Civic engagement	71 (62 - 78)	0.49	70 (60 - 79)[a]	0.53
Local resources	n.a.	n.a.	70 (60 - 79)[a]	0.51
Housing	66 (58 - 74)	0.48	68 (58 - 77)[a]	0.61
Inequality	73 (65 - 80)	0.48	68 (58 - 77)[a]	0.51
Employment	59 (50 - 67)	0.44	66 (55 - 75)[a]	0.56
Noise pollution	n.a.	n.a.	65 (54 - 74)	0.47
Safety	67 (59 - 75)	0.44	64 (54 - 73)[a]	0.54
Cultural capacity	n.a.	n.a.	59 (48 - 68)	0.50
Smart infrastructure	61 (52 - 69)	0.41	58 (48 - 68)	0.47
Biodiversity	n.a.	n.a.	56 (46 - 66)	0.46
Economic productivity	48 (40 - 57)	0.45	56 (45 - 66)	0.53
Urban microclimate	n.a.	n.a.	55 (44 - 65)	0.44
Business climate	47 (39 - 56)	0.43	38 (29 - 49)	0.55
International embeddedness	28 (21 - 36)	0.35	37 (27 - 47)	0.40
Entrepreneurship	35 (27 - 43)	0.42	36 (27 - 46)	0.45

[a] Components of which the % evaluations ≥ 8 is significantly greater than 50% and the SA is greater than 0.5.

Significant differences were found between experts from academia and experts from other institutions regarding the relevance of the following seven components: non-car transportation infrastructure ($U = 1395.5$, $p < .01$), air quality ($U = 1495$, $p < .01$), CO_2 emissions ($U = 1429.5$, $p <. 01$), energy consumption ($U = 1055.5$, $p < .01$), solid waste ($U = 1396$, $p < .01$), waste water treatment ($U = 1210.5$, $p < .01$), and water usage ($U = 1103$, $p <$

.01). Experts from academia tended to evaluate these components as more relevant than experts from other institutions. Significant differences were also found between experts from North-West Europe and South-East Europe regarding the relevance of the components 'air quality' ($U = 2341$, $p < .01$) and 'health' ($U = 2407$, $p < .01$), Experts from South-East Europe tended to evaluate these components as more relevant than experts from North-West Europe.

Experts suggested 204 additional components relevant for defining and measuring urban sustainability. The content analysis yielded 43 different categories (including an 'other' category with unique suggestions), of which six contained more than four suggestions. For each of these six categories a component label and explanation was formulated, resulting in the following new components: biodiversity, cultural capacity, governance, local resources, noise pollution, urban microclimate (see appendix 6.1).

The second Delphi round

In the second round the percentage of evaluations equal to or greater than 8 ranged between 36% (entrepreneurship) and 89% (non-car transportation infrastructure, air quality, CO_2 emissions). Nineteen components were identified as most relevant of which 18 also had a SA greater than 0.5. These included, in addition to the ten components that already had a SA greater than 0.5 in the previous round, the following eight components: governance, education, civic engagement, local resources, inequality, housing, employment, and safety.

Differences between experts from academia and experts from other institutions regarding the relevance of the components were no longer statistically significant. Significant differences were found between experts from North-West Europe and South-East Europe regarding the relevance of the components 'green spaces' ($U = 1322.5$, $p < .01$), 'education' ($U = 1249$, $p < .01$), and 'noise pollution' ($U = 1324.5$, $p < .01$). Experts from South-East Europe tended to evaluate these components as more relevant than experts from North-West Europe.

The third Delphi round

Table 6.4 shows the percentage of experts that selected each component as most relevant for defining and measuring urban sustainability, the corresponding value of the SA, and the total number of points that experts attributed to each component in the third round. Remarkably, none of the components was selected by a majority of experts. As a result, the high values of the SA are mainly due to a majority of expert that did *not* consider the components to be most relevant.

Table 6.4

Percentage of experts that selected each component as most relevant (95% Confidence Interval), the SA and total number of points that was attributed to each component in the third Delphi round.

Component	Round 3 (n = 72)		
	% selected (95% CI)	SA	Number of points
Air quality	46 (34 - 58)[a]	0.50	667
Governance	40 (29 - 53)[a]	0.51	676
Energy consumption	40 (29 - 53)[a]	0.51	585
Non-car transportation infrastructure	33 (23 - 45)[a]	0.55	462
CO_2 emissions	32 (21 - 44)[a]	0.56	470
Inequality	32 (21 - 44)[a]	0.56	452
Green spaces	32 (21 - 44)[a]	0.56	446
Health	25 (16 - 37)	0.62	391
Climate resilience	25 (16 - 37)	0.62	372
Solid waste	22 (13 - 34)	0.65	232
Civic engagement	18 (10 - 29)	0.70	247
Local resources	17 (9 - 27)	0.72	265
Biodiversity	17 (9 - 27)	0.72	260
Education	14 (7 - 24)	0.76	185
Employment	13 (6 - 22)	0.78	164
Water usage	13 (6 - 22)	0.78	128
Economic productivity	11 (5 - 21)	0.80	195
Smart infrastructure	11 (5 - 21)	0.80	160
Waste water treatment	11 (5 - 21)	0.80	136
Housing	10 (4 - 19)	0.82	115
Noise pollution	8 (3 - 17)[b]	0.85	119
Safety	7 (2 - 15)[b]	0.87	145
Urban microclimate	7 (2 - 15)[b]	0.87	113
Entrepreneurship	6 (2 - 14)[b]	0.89	85
Cultural capacity	6 (2 - 14)[b]	0.89	60
Business climate	4 (1 - 12)[b]	0.92	50
International embeddedness	1 (0 - 7)[b]	0.97	20

Note: based on chance alone, each component would have been selected by 19% of experts.

[a] Components which were selected significantly more often than expected based on chance alone.

[b] Components which were selected significantly less often than expected based on chance alone.

The following seven components were selected significantly more often as most relevant than what may be expected based on chance alone: air quality, governance, energy consumption, non-car transportation infrastructure, green spaces, inequality, and CO_2 emissions. These components also received the most points. Although the component 'air quality' was selected more often as most relevant than the component 'governance', it received less points. Likewise, the component 'non-car transportation infrastructure' was selected more often than 'green spaces', but also received less points.

Experts provided various explanations of why they selected a component as most relevant. They stressed that air quality is relevant for the health, well-being, and quality of life of inhabitants: "*Air quality is related to health, which is important for people's wellbeing.*" They also pointed out that air quality is related to many other components of urban sustainability, such as biodiversity, green spaces, and CO_2 emissions.

With regard to the relevance of the component 'governance', experts explained that city governments determine the extent to which cities develop in a sustainable way: *"Governance is a key item in developing sustainability. Without political support, plans and projects cannot reach results."* City governments can support a sustainable development among others by adapting legislation, providing resources, involving stakeholders, as well as by planning and managing city development.

Experts selected the component 'energy consumption' as most relevant, because energy is needed to sustain life and because the use of energy impacts the environment. Cities' energy consumption is related to various environmental issues, such as CO_2 emissions, climate change, and the use of natural resources. It also determines cities' self-sufficiency and dependency on non-renewable energy. Experts stressed that renewable energy consumption is crucial for a sustainable urban development: *"The degree of renewability of the overall energy consumption is a crucial element to achieve sustainability."*

Experts explained that a non-car transportation infrastructure is most relevant, because it reduces the use and negative impacts of cars and improves the sustainability of urban mobility: "A *sustainable city should be measured by its mobility infrastructure and in particular to non-car mobility which is the most environment friendly.*" A non-car transportation infrastructure also improves the environment in terms of among others pollution, CO_2 emissions, energy and resource consumption, noise pollution, congestion, quality of life, and space use.

Regarding the relevance of the component 'CO_2 emissions', experts explained that CO_2 emissions impact the environment as a whole and affect climate change: *"The level of industrial activity in urban Europe is very high. These industrial activities generate a lot of CO_2 emissions, which contribute in no small way to climate change and ultimately affect urban sustainability."* As emitters of high levels of CO_2, cities need to mitigate and fight climate change. Experts also explained that CO_2 emissions are related to other urban sustainability issues, such as energy consumption, transportation, as well as inhabitant's health and quality of life.

Experts stated that inequality, as a social aspect of sustainability, is a relevant component that is often neglected: *"Societies that aim to be environmentally sustainable should not forget the social aspects of sustainability, of which reducing inequality is the most*

important." Inequality creates many social problems like social exclusion as well as tensions and conflicts between different groups of people. A sustainable, functioning, and inclusive city needs to provide for all its inhabitants, among others by sharing benefits and offering equal opportunities.

Lastly, experts explained that green spaces are most relevant, because they improve a city's general quality of life as well as the well-being of its inhabitants: "*Green spaces in fact improve the well-being of inhabitants.*" They provide many benefits to a city in terms of among others biodiversity, air quality, health, and urban microclimate. Green spaces are also important for recreation, relaxation, and as meeting places for people. Finally, green spaces support and raise people's awareness regarding environmental topics.

In addition to the most relevant components, table 6.4 also shows the seven components that were selected significantly less often than what may be expected based on chance alone: noise pollution, safety, urban micro climate, entrepreneurship, cultural capacity, business climate, and international embeddedness. The corresponding high values of the SA indicate that experts generally agreed that these components are *not* the most relevant for defining and measuring urban sustainability. The components also received the least points, except for the component 'safety'. This component was selected by few experts who gave it relatively many points.

With regard to the percentage of experts that selected each component as most relevant, no significant differences were found between experts from academia and experts from other institutions. Likewise, no significant differences were found between experts from North-West and South-East Europe.

6.4 Discussion

Study outcomes

In this study the Delphi method was used to identify which components experts find most relevant for defining and measuring the ranking attribute urban sustainability in a European context. The following seven components were identified as most relevant and may therefore be regarded as central to defining and measuring urban sustainability: air quality, governance, energy consumption, non-car transportation infrastructure, green spaces, inequality, and CO_2 emissions. Remarkably, none of these components was selected as most relevant by a majority of experts. This lack of agreement reflects the ambiguity surrounding the definition and measurement of urban sustainability, which has also been discussed in the literature (Ameen et al., 2015; Tanguay et al., 2010).

Five of the seven components that were identified as most relevant are part of the environmental dimension of urban sustainability (air quality, energy consumption, non-car transportation infrastructure, green spaces, and CO_2 emissions). Although sustainability was initially considered to be mainly an environmental concept (Ameen et al., 2015), researchers have stressed the importance of especially the social dimension and to some extent also the economic dimension of urban sustainability (Ameen et al., 2015; Dempsey et al., 2011; Hassan & Lee, 2015; Lorr, 2012; Michael et al., 2014). Nonetheless, in this Delphi study only the social component 'equality' and none of the economic components ended up among the most relevant components. These findings suggest that urban sustainability is still perceived as mainly an environmental concept.

The results of this Delphi study also suggest that most environmental policy sub-fields such as water pollution, waste management, and biodiversity have become somewhat less relevant in favour of other issues such as climate mitigation, renewable energy, transportation and green spaces. The only exception here is air pollution which is a classical field of environmental policy that still seems to be placed high on the agenda. It appears that social issues, in particular those related to health and inequality, have become more relevant, though not yet as relevant as environmental issues. On the other hand, it seems that economic issues are still not considered to be as relevant as social or environmental issues. This suggests that the experts in this Delphi study did not (yet) fully support the EU's green growth agenda which is mainly focused on developing a smart, sustainable, and inclusive economy (European Commission, 2010b). The component 'governance', which was included in the Delphi study based on experts' suggestions, may be difficult to place under one of the three dimensions. Whereas some researchers placed it among the social components (Tanguay et al., 2010), others included it as a separate fourth dimension (Shen et al., 2011) or see the three dimensions as embedded in a framework of governance (Petschow et al., 2005).

Limitations

For the first round of the Delphi study a list of 21 components was constructed. Although this list was based on various existing urban sustainability rankings, it is possible that certain potentially relevant components were not included. Therefore, experts were invited to suggest additional components. Based on these suggestions six components were added to the list. For each component on the list a one-sentence explanation was formulated to facilitate interpretation. The specific way in which these explanations were formulated may have influenced the results.

A convenience sample of European urban sustainability experts was assembled. To avoid a selection bias, selection criteria and different search strategies were applied. Nonetheless,

the sample contained more experts from academia than experts from other institutions as well as more experts from North-West Europe than experts from South-East Europe. Consequently, the majority of the response in each round consisted of experts from academia and experts from North-West Europe. Although no lasting significant differences between expert groups were found, questions may be asked about whether or not similar results would have been obtained with a different composition of experts. As the study was explicitly restricted to a European context, caution is advised when using the results for defining and measuring urban sustainability in other parts of the world (see: Science Communication Unit, 2015).

A low response rate and high drop-out rates are well-known limitations of the Delphi method (Hung et al., 2008; Keeney et al., 2006) that may also be a cause for concern in this study. In the first Delphi round 33% of the invited experts completed the questionnaire, which is comparable to the average response rate of ordinary web-surveys (34%) (Shih & Fan, 2008). Considering that experts were invited to participate in not one, but three questionnaire rounds, the response rate may be regarded as satisfactorily. Whether the drop-out rates in the second (29%) and third (27%) Delphi round are favourable or unfavourable in comparison to other Delphi studies is difficult to determine as many Delphi studies do not report them (Boulkedid et al., 2011).

Implications

The results of this study may serve to develop a conceptual framework in which urban sustainability is conceived as a multi-dimensional concept (i.e. environment, society, economy, and perhaps also governance), with each dimension consisting of several specific components that are measurable by means of indicators. Such a framework, also referred to as a domain-issue-based (Maclaren, 1996) or theme-oriented (Huang et al., 2015) framework, would then contain at least the seven components which were identified as most relevant. As most of these components are part of the environmental dimension of urban sustainability, several social and economic components may need to be added that in the final round of this Delphi study were selected by at least 10% of the experts (e.g. health, education, employment, economic productivity). The total number of points that experts attributed to the components in the final Delphi round may be used as a basis to determine the weights that reflect the relative relevance of each component for measuring urban sustainability.

Selecting indicators for each component remains a daunting task as a large pool of indicators is available (Huang et al., 2015). Systematic procedures and criteria for selecting a parsimonious list of urban sustainability indicators are available (see for example: Maclaren, 1996; Tanguay et al., 2010). Alternatively, indicators may be selected by means of experts

(Giovannini et al., 2008). Perhaps here lies another opportunity for using the Delphi method, for example to identify indicators for the component 'governance'. It would also be interesting to replicate the current study in other parts of the world to find out which relevant components are place-dependent and which components are universally relevant for defining and measuring urban sustainability. For example, it may be assumed that water issues are more relevant in other parts of the world which are confronted with water scarcity on a daily basis.

6.5 Conclusion

Meijering et al. (2014) suggested to use the Delphi method to define and operationalize urban environmental sustainability. This study showed that the Delphi method indeed seems to be useful for this purpose. Although high levels of agreement among experts were not always obtained, the method did prove useful for identifying the most (and least) relevant components for defining and measuring urban sustainability in a European context. This study thereby provides an expert-based framework that may be used by future developers of urban sustainability rankings to properly define the ranking attribute and to select appropriate indicators.

Acknowledgements

We would like to acknowledge the following experts who participated in all three rounds of this Delphi study and gave their valuable opinions: Adriana Galderisi, Agnes Franzen, Ahmed Khoja, Albert Edman, Ana Poças Ribeiro, Angeliki Chatzidimitriou, Anicenta Bubak, Anne-Francoise Marique, Anónio Ramos, Antonio Gagliano, Arie Voorburg, Aurore Cambien, Bastiaan Kees Zoeteman, Blanca Pedrola Vidal, Carlos Borrego, Carmelina Cosmi, Carolina Mateo Cecilia, Conny Weber, Corrado Diamantini, Diogo Alarcão, Dominic Stead, Dominique van Ratingen, Donát Rabb, Emma Terama, Enrico Pisoni, Eva Pangerl, Fabiana Morandi, Francesca Pagliara, Frans van der Reep, Gerald Krebs, Giulia Sonetti, Greg Keeffe, Haldun Süral, Hyunjung Lee, Ian Skinner, Irina Aura Istrate, Jan Dictus, Janine Hogendoorn, Jaques Teller, Jeroen Nagel, Kornilia Maria Kotoula, Kwasi Gyau Baffour Awuah, Lorenzo Chelleri, Luca Bertolini, Luca Coscieme, Lucjan Goczol, Marcin Czyz, Margaretha Breil, Maria Loloni, Mariano Gallo, Marina Rigillo, Marleen Lodder, Martine Laprise, Martyna Surma, Max Grünig, Monica Salvia, Niels van Geenhuizen, Paul Swagemakers, Peter Bröde, Riikka Holopainen, Romano Fistola, Rosa Anna La Rocca, Sophie Jongeneel, Timothy Lee, Tobias Emilsson, Walter Unterrainer, Wim de Haas, as well as five other experts who wished to remain anonymous.

Appendix 6.1

Overview of components evaluated by experts.

Components based on a content analysis of eight existing urban sustainability rankings

- Air quality: the extent to which the air in a city contains pollutants
- Business climate: the extent to which a city is suitable for doing business, for example in terms of taxes, regulations, and corruption
- Civic engagement: the extent to which a city's inhabitants are being engaged in urban policy and politics
- Climate resilience: the extent to which a city is resilient to the potentially harmful effects of climate change
- CO_2 emissions: the amount of a city's carbon dioxide emissions
- Economic productivity: the extent of a city's economic productivity
- Education: the education level of a city's inhabitants
- Employment: the extent to which the inhabitants of a city are economically (in)active
- Energy consumption: the amount of renewable and non-renewable energy that is consumed within a city
- Entrepreneurship: the amount of business start-ups within a city
- Green spaces: the amount of nature and parks within a city
- Health: the health of a city's inhabitants
- Housing: the cost and quality of housing in a city
- Inequality: the extent of differences between groups of inhabitants, for example in terms of income, access to education, and political participation
- International embeddedness: the extent to which a city participates in international networks and hosts international events
- Non-car transportation infrastructure: the size, quality, and use of a city's non-car transportation infrastructure (for trains, the metro, buses, cycling, walking)
- Safety: the amount of violent crimes within a city and the extent to which inhabitants feel safe
- Smart infrastructure: the extent to which technology is used within a city to improve among others government services, commuter traffic, and the energy efficiency of buildings
- Solid waste: the amount of solid waste that is produced and recycled within a city
- Wastewater treatment: the amount of wastewater that is collected and treated within a city
- Water usage: the amount of water that is used within a city as a result of water consumption and leakages in the water distribution system

New components based on a content analysis of experts' suggestions

- Biodiversity: the diversity of plant and animal species in a city
- Cultural capacity: the availability of cultural facilities and activities in a city
- Governance: the extent to which the organization and political composition of the city government supports a sustainable urban development
- Local resources: the extent to which a city produces and uses local resources such as food and energy
- Noise pollution: the extent to which inhabitants of a city are exposed to bothersome noise
- Urban microclimate: the extent to which the climate in a city is comfortable for its inhabitants, for example in terms of temperature and humidity

Chapter 7

Feeding back experts' own initial ratings in Delphi studies: Effects on opinion change and the level of agreement

This chapter is submitted as:
Meijering, J. V., & Tobi, H. Feeding back experts' own initial ratings in Delphi studies: Effects on opinion change and the level of agreement.

Abstract

Feeding back experts' own initial ratings in Delphi studies has been recommended, although empirical evidence is lacking. This study aimed to provide insight into the effects of feeding back experts' own initial ratings on three different outcome measures: (1) the percentage of questionnaire items on which experts changed their opinion, (2) the degree to which experts changed their ratings towards the majority opinion, and (3) the increase in the level of agreement among experts. With regard to the second outcome measure, two conformity indices were developed. Within a real-world Delphi study about the definition and measurement of urban sustainability, experts were randomly assigned to either a condition in which initial ratings were excluded from feedback (EX condition) or a condition in which initial ratings were included in feedback (IN condition). Results showed that experts in the EX condition changed their opinion relatively more often than experts in the IN condition. Results also suggested that experts in the EX condition changed their ratings to a greater degree towards the majority opinion than experts in the IN condition, although differences were not always statistically significant. No difference between conditions was found regarding the increase in the level of agreement among experts. Based on this experiment, feeding back experts' own initial ratings seems to be justified as it may increase the reliability of experts' ratings. The two conformity indices that were developed provided important insights into the degree to which experts changed their ratings towards the majority opinion and may thus be useful for future Delphi studies.

Keywords: Delphi experiment, controlled opinion feedback, initial ratings, opinion change, conformity, agreement

7.1 Introduction

The Delphi method, developed in the 1950s (Dalkey & Helmer, 1963), was generally defined by Linstone and Turoff (1975) as: "a method for structuring a group communication process so that the process is effective in allowing a group of individuals, as a whole, to deal with a complex problem" (p. 3). They further explained that the method is particularly useful when the problem at hand "does not lend itself to precise analytical techniques but can benefit from subjective judgements on a collective basis" (p. 4). Since its public introduction in the 1960s, many different types of Delphi designs have been developed that usually aim to allow a group of experts to achieve agreement on a particular topic (Hasson & Keeney, 2011), such as the dimensions of a concept (Zill et al., 2015) or a list of quality criteria (Verhagen et al., 1998). Regardless of type and aim, the Delphi method has some defining characteristics (Dalkey et al., 1969; Keeney et al., 2006; Rowe & Wright, 1999). A Delphi study consists of at least two rounds of data-collection. In each round experts give their opinion on the topic of interest, commonly by rating a number of items (pre-selected or developed during a first Delphi round) using a standardized questionnaire. Experts do not directly communicate with each other, but instead receive so called controlled opinion feedback: a summary of the findings from the previous round. In light of this feedback experts may reconsider and change their opinion in the next round.

Although controlled opinion feedback is a crucial aspect of any Delphi study, there is debate about what information should be fed back. Often feedback solely consists of summary statistics, showing per questionnaire item a measure of location (e.g. mean, median) and dispersion (e.g. standard deviation, interquartile range). This type of feedback has been criticized because it may induce experts to simply conform their opinion to the majority opinion, creating an artificial agreement (Woudenberg, 1991). Therefore, various researchers recommended to also feed back rationales that give experts insight into why items were rated in a certain way (Bolger et al., 2011; Murphy et al., 1998; Rowe et al., 1991). A related issue of debate concerns the feedback of each expert's own initial ratings from the previous round. Although feeding back this information has been recommended (Keeney et al., 2006; Murphy et al., 1998), a systematic review of Delphi studies (regarding the selection of healthcare quality indicators) by Boulkedid et al., (2011) showed that a minority of studies (39%) reported the feedback of experts' own initial ratings. According to Boulkedid et al. (2011) feedback of initial ratings is necessary, because it informs experts about their position relative to the rest of the group, and thus, assists them in making decisions in future Delphi rounds. However, feeding back experts' own initial ratings may also have disadvantages. Bolger and Wright (2011) stated that opinion change in Delphi studies is less than it could be, because people tend to discount advice from others and favour their own initial opinion. This so called egocentric discounting may be reinforced by feeding back experts' own initial ratings, thereby impeding experts to change their opinion and achieve agreement. Despite the debate about what to feed back, only few experiments

into the effects of different types of controlled opinion feedback have been conducted. Most of these experiments focussed on the effect of feeding back summary statistics, rationales, or some combination of the two, on experts' degree of opinion change, forecast accuracy, or level of agreement (Best, 1974; Bolger et al., 2011; Gowan & McNichols, 1993; Meijering & Tobi, 2016; Rowe & Wright, 1996; Rowe et al., 2005). Experiments into the effect of feeding back experts' own initial ratings could not be found at all.

The current paper seeks to fill the identified knowledge gap by describing an experiment that aimed to provide insight into the effects of feeding back experts' own initial ratings on three different outcome measures: (1) the percentage of questionnaire items on which experts changed their opinion (i.e. by giving different ratings in the second round), (2) the degree to which experts changed their ratings towards the majority opinion, and (3) the increase in the level of agreement among experts. With regard to the second outcome measure, an index was sought that estimates the degree to which experts in Delphi studies changed their ratings towards the majority opinion as presented in the controlled opinion feedback. As no such index could be found, this paper also describes the development and application of two proposed conformity indices.

For each of the three outcome measures a hypothesis was formulated and tested. Considering the possibility that feedback of initial ratings reinforces egocentric discounting (Bolger & Wright, 2011), it may be expected that experts who received their own initial ratings changed their opinion relatively less often and changed their ratings to a lesser degree towards the majority opinion as compared to experts who did not receive their own initial ratings. As no empirical evidence in support of this expectation could be found, the following two hypotheses were tested against the null-hypothesis (no difference):

H_1: there is a difference between experts who received their own initial ratings and those who did not regarding the percentage of questionnaire items on which they changed their opinion.

H_2: there is a difference between experts who received their own initial ratings and those who did not regarding the degree to which they changed their ratings towards the majority opinion.

If feedback of initial ratings impedes experts to change their ratings towards the majority opinion, it may also be expected that it impedes them to achieve agreement. In lack of any empirical evidence, the following hypothesis was tested against the null-hypothesis (no difference):

H_3: the increase in the level of agreement among experts who received their own initial ratings differs from the increase in the level of agreement among experts who did not receive their own initial ratings.

7.2 Materials and methods

Context of experiment

The current experiment was conducted within a real-world Delphi study that aimed to allow a panel of urban sustainability experts to achieve agreement on the components (e.g. air quality, inequality, entrepreneurship) that are most relevant for defining and measuring urban sustainability in a European context (see chapter 6 of this thesis). In total, three rounds of data-collection were conducted of which round one and two are relevant with regard to the experiment.

Expert sample

A convenience sample was compiled that consisted of European urban sustainability experts from four different types of institutions: academia, business, civil society (i.e. NGOs, non-profit, and community-based organisations that pursue charitable or member-oriented goals), and government. Most names of experts were obtained from the programs and proceedings of several urban sustainability conferences that were held in Europe in 2014 and 2013 (e.g. The Sustainable City Conference 2014) and projects funded by the Joint Programming Initiative Urban Europe, a program established by the European Commission that aims to create attractive, sustainable and economically viable urban areas (Robinson et al., 2015). Additionally, by searching on the internet and talking to experts, various institutions and projects active in the field of urban sustainability were found (e.g. Arcadis, Climate-KIC, Amsterdam Institute for Advanced Metropolitan Solutions). These institutes and projects were contacted by telephone and asked whether they had urban sustainability experts who are willing to participate in the Delphi study. All in all, a final sample consisting of 419 experts from 26 European countries was acquired.

Study design

For each round of the Delphi study a standardized questionnaire was developed which was written in English, checked by a native English speaker, and programmed as a web-survey using the online survey builder Qualtrics (Qualtrics, 2015). In May 2015 all 419 experts in the

sample received an e-mail with a link to the first online questionnaire. In this questionnaire experts were asked to rate the relevance of 21 components for defining and measuring urban sustainability in a European context. Experts gave their opinion on a 10-point scale, ranging from 1, labelled 'not relevant at all', to 10, labelled 'entirely relevant'. Thereafter, experts were invited to explain for up to three components (rated with an 8 or higher) why they rated them as relevant.

Experts who completed the first questionnaire were randomly assigned to either the condition in which initial ratings were excluded from the controlled opinion feedback (EX condition) or the condition in which initial ratings were included in the controlled opinion feedback (IN condition). In June 2015 these experts received an e-mail with a link to the second online questionnaire which was very similar to the previous one. Once again experts were asked to rate the relevance of the same 21 components for defining and measuring urban sustainability, on the same 10-point scale. This time, each component was complemented with controlled opinion feedback that consisted of two parts: (1) a table with summary statistics, showing the component's median rating, interquartile range, and percentage of ratings equal or greater than 8, in comparison to the most and least relevant component, and (2) a summary of rationales of experts who rated the component as relevant. In addition, experts in the IN condition were presented with their own initial rating of the component in the previous round (see figure 7.1 and figure 7.2).

Summary statistics of the component '**green spaces**' in comparison to the most and least relevant components:

	Median evaluation [1]	Interquartile range [2]	% of evaluations ≥ 8
Most relevant component	9	8 - 10	88%
Green spaces	9	8 - 10	81%
Least relevant component	6	5 - 8	28%

Note: experts evaluated components on a scale which ranged from 1 'not relevant at all' to 10 'entirely relevant'.
[1] When all evaluations of a component are ordered from lowest to highest, the median is the middle evaluation.
[2] The interquartile range contains the middle 50% of evaluations.

Summary of experts' explanations regarding why the component is relevant:
Green spaces provide many benefits for inhabitants of cities, such as an improved well-being and opportunities for recreation and community building. Furthermore, green spaces offer potential solutions for various problems related to among others climate change, water runoff, and the urban heat island effect. Green spaces are also relevant for nature and biodiversity within a city.

Based on your expertise, how relevant is the component 'green spaces' for defining and measuring 'Urban Sustainability' in a European context?
Please give your answer by using the scale which ranges from 1 'not relevant at all' to 10 'entirely relevant'.

	not relevant at all									entirely relevant	
	1	2	3	4	5	6	7	8	9	10	don't know
Green spaces: the amount of nature and parks within a city	○	○	○	○	○	○	○	○	○	○	○

Figure 7.1: screenshot of feedback presented in the EX condition.

Summary statistics of the component '**green spaces**' in comparison to the most and least relevant components:

	Median evaluation [1]	Interquartile range [2]	% of evaluations ≥ 8
Most relevant component	9	8 - 10	88%
Green spaces	9	8 - 10	81%
Least relevant component	6	5 - 8	28%

Note: experts evaluated components on a scale which ranged from 1 'not relevant at all' to 10 'entirely relevant'.
[1] When all evaluations of a component are ordered from lowest to highest, the median is the middle evaluation.
[2] The interquartile range contains the middle 50% of evaluations.

Summary of experts' explanations regarding why the component is relevant:
Green spaces provide many benefits for inhabitants of cities, such as an improved well-being and opportunities for recreation and community building. Furthermore, green spaces offer potential solutions for various problems related to among others climate change, water runoff, and the urban heat island effect. Green spaces are also relevant for nature and biodiversity within a city.

In the previous round you evaluated the relevance of the component 'green spaces' with a(n): **7**

Based on your expertise, how relevant is the component 'green spaces' for defining and measuring 'Urban Sustainability' in a European context?
Please give your answer by using the scale which ranges from 1 'not relevant at all' to 10 'entirely relevant'.

	not relevant at all									entirely relevant	
	1	2	3	4	5	6	7	8	9	10	don't know
Green spaces: the amount of nature and parks within a city	○	○	○	○	○	○	○	○	○	○	○

Figure 7.2: screenshot of feedback presented in the IN condition.

Analysis

All experts who completed the first and second questionnaire were included in the analysis. The percentage of opinion changes was calculated by expressing the number of components on which an expert changed his or her rating as a percentage of the total number of components that the expert rated in both rounds. Significant differences between the EX and IN condition were tested using the independent samples Mann-Whitney test (denoted U) (Mann & Whitney, 1947). This tests is non-parametric, meaning that it does not require that the percentage of opinion changes within the EX and IN condition is normally distributed nor that there is homogeneity of variances.

Two conformity indices C_1 and C_2 were developed to represent the degree to which experts in Delphi studies changed their ratings towards the majority opinion as presented in the controlled opinion feedback. Both indices are based on the change in the distance between an expert's rating of an item and the median rating of that item in the previous round. Let X_{ijr} denote the rating X of component i by expert j in round r. Furthermore, let M_{i1} denote the median of component i across all experts in round 1, which was fed back in round 2 to inform experts on the majority opinion. Finally, S_{min} and S_{max} denote the minimum and maximum value of the rating scale. In the current study $S_{min} = 1$, $S_{max} = 10$.

The definition of C_1 is as follows:

$$C_1 = \begin{cases} 0, & \text{for } |X_{ij2} - M_{i1}| = |X_{ij1} - M_{i1}| \\ 1 - \dfrac{|X_{ij2} - M_{i1}|}{|X_{ij1} - M_{i1}|}, & \text{for } |X_{ij2} - M_{i1}| < |X_{ij1} - M_{i1}| \\ \dfrac{|X_{ij2} - M_{i1}| - |X_{ij1} - M_{i1}|}{\min\{S_{min} - M_{i1},\, M_{i1} - S_{max}\} + |X_{ij1} - M_{i1}|}, & \text{for } |X_{ij2} - M_{i1}| > |X_{ij1} - M_{i1}| \end{cases}$$

The index C_1 takes on values in between -1 and 1 with 0 reflecting no change from round 1 to round 2 in the distance between an expert's rating and M_{i1}. The index C_1 is greater than 0 if an expert's distance to M_{i1} decreased from round 1 to round 2. The maximum value 1 is reached when an expert changed his or her initial rating to M_{i1}. When an expert's distance to M_{i1} increased from round 1 to round 2, C_1 takes on a negative value. The minimum value of -1 is reached when the distance to M_{i1} becomes as large as possible, reflecting an expert's greatest possible degree of non-conformity to the majority opinion.

Whereas the index C_1 only takes on values in between -1 and 1, the range of values of C_2 is sensitive to the width of the rating scale in such a way that the index takes on a minimum value of $-(S_{max} - S_{min})$ and a maximum value of $(S_{max} - S_{min})$.

The definition of C_2 is as follows:

$$C_2 = \begin{cases} 0, & \text{for } |X_{ij2} - M_{i1}| = |X_{ij1} - M_{i1}| \\ \left(1 - \dfrac{|X_{ij2} - M_{i1}|}{|X_{ij1} - M_{i1}|}\right) \cdot (S_{max} - S_{min}), & \text{for } |X_{ij2} - M_{i1}| < |X_{ij1} - M_{i1}| \\ 1 - \dfrac{|X_{ij2} - M_{i1}|}{|X_{ij1} - M_{i1}|}, & \text{for } |X_{ij2} - M_{i1}| > |X_{ij1} - M_{i1}| \text{ and } X_{ij1} \neq M_{i1} \\ -(S_{max} - S_{min}) \cdot \dfrac{|X_{ij2} - M_{i1}|}{\max\{|M_{i1} - S_{min}|,\, |M_{i1} - S_{max}|\}}, & \text{for } X_{ij1} = M_{i1} \neq X_{ij2} \end{cases}$$

If the distance to M_{i1} increased from round 1 to round 2, a distinction is made between experts whose rating in round 1 was unequal to M_{i1} (see condition 3) and experts whose rating in round 1 was equal to M_{i1} (see condition 4). The index C_2 only takes on its minimum value $-(S_{max} - S_{min})$ if an expert's rating in round 1 equalled M_{i1} and the distance between X_{ij2} and M_{i1} was as large as possible.

The indices C_1 and C_2 were calculated for each expert on each component. Then, *per* expert the following location and dispersion statistics were calculated across components for both

indices: the mean, median, standard deviation, and interquartile range. Cumulative frequency graphs were made to examine the distribution of these statistics within the EX and IN condition. Differences between conditions were tested using the independent samples Mann-Whitney test as well as the bootstrapped independent samples t-test, based on 10.000 bootstrap samples (Efron & Tibshirani, 1993). Like the Mann-Whitney test the bootstrapped independent samples t-test is non-parametric.

The level of agreement among experts in round r was estimated using the Strict Agreement index (SA_r) which was originally developed for an ordinal 5-point Likert scale (Meijering et al., 2013). The index expresses the number of agreeing expert pairs as a proportion of the total number of possible expert pairs, where two experts are said to agree if they attributed the same rating to an item. Because experts in the current Delphi experiment rated items on a 10-point scale, the agreement criterion was relaxed: here two experts were said to agree if their ratings did not differ by more than one point on the rating scale. The SA_r takes on values between 0 (indicating no agreement at all) and 1 (indicating complete agreement or consensus). Differences between the EX and IN condition regarding the increase in the level of agreement from round 1 to round 2 were tested using the following procedure. First, the level of agreement was calculated across all experts and all components in both round 1 and 2 for each condition c, resulting in $SA_{1,c}$ and $SA_{2,c}$. Then, for each condition the increase in the level of agreement from round 1 to 2 was calculated by subtracting $SA_{1,c}$ from $SA_{2,c}$, resulting in $SA_{2-1,c}$. The difference between the two conditions regarding the increase in the level of agreement was calculated as the difference between $SA_{2-1,IN}$ and $SA_{2-1,EX}$. Corresponding 95% bias-corrected and accelerated (BCa) confidence intervals were calculated based on 10.000 bootstrap samples (Efron & Tibshirani, 1993). Because SA_r does not take chance agreement into account, the same procedure was followed using Light's kappa (denoted LK_r) (Light, 1971). This index takes on values between −1 and 1, with 0 indicating that there is only chance agreement. Like other kappa statistics, LK_r overestimates chance agreement when the marginal distribution of the ratings is skewed (Feinstein & Cicchetti, 1990).

7.3 Results

Percentage of opinion changes

A total of n = 99 experts completed the first and second round of the Delphi study, of which n = 49 were randomly assigned to the EX condition and n = 50 to the IN condition. On average, the percentage of items on which experts changed their opinion was 69% in the EX condition and 49% in the IN condition. The Mann-Whitney test showed that this difference was statistically significant ($U = 576.5$, $p < .001$). Based on this result the null-hypothesis (no difference) is rejected in favour of H_1.

Opinion change towards majority opinion

Figure 7.3 depicts the cumulative distributions of the mean, median, standard deviation, and interquartile range of the conformity index C_1 in the EX and IN condition. With regard to the two location statistics, the lines of the conditions clearly diverge in the middle part, indicating that in comparison to the IN condition the EX condition contained relatively more experts with greater values on the mean and median conformity index. With regard to the two dispersion statistics, the lines hardly show any overlap, indicating that the variation of the conformity index across components was greater in the EX condition than in the IN condition.

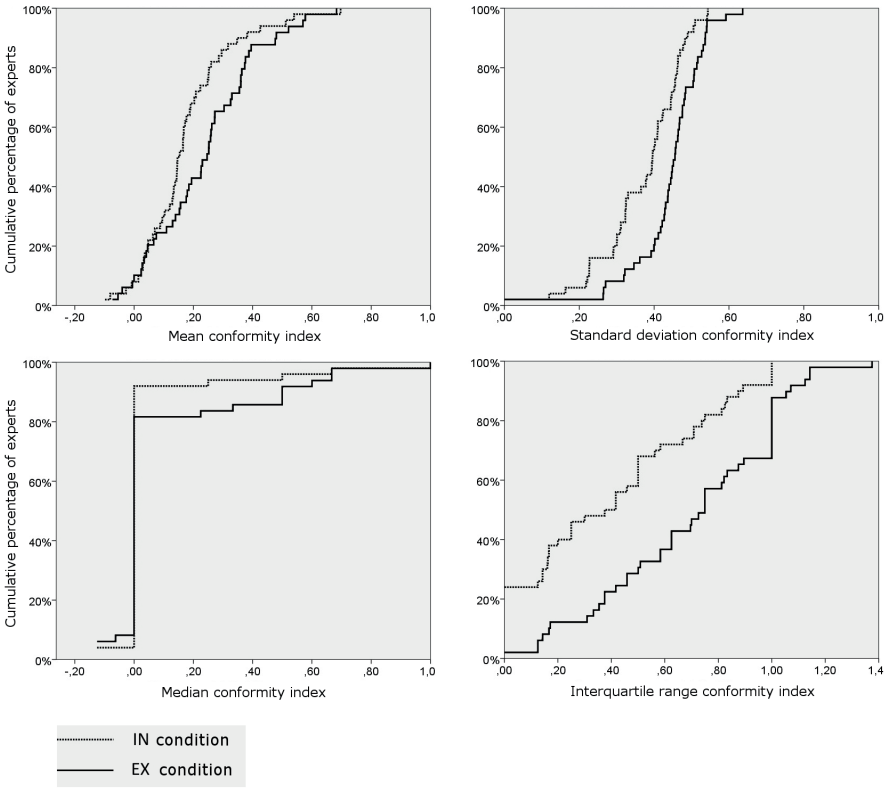

Figure 7.3: cumulative frequency graphs of the mean, median, standard deviation, and interquartile range of the conformity index C_1 in the EX and IN condition.

Table 7.1 shows the results of the bootstrapped independent samples t-test and the independent samples Mann-Whitney test. With regard to the mean C_1, only the Mann-Whitney test showed a significant difference between the EX and IN condition, indicating that experts in the IN condition changed their ratings to a lesser degree towards the majority opinion than experts in the EX condition. Looking at the median C_1, none of the tests showed a significant difference between the EX and IN condition. With regard to the standard

deviation and the interquartile range of C_1, both tests showed significant differences between conditions: the variation of C_1 across components was greater in the EX condition than in the IN condition. Using C_2 similar results were obtained. The only notable difference is that with regard to the mean C_2 both the bootstrapped t-test and the Mann-Whitney test showed a significant difference between the two conditions.

Table 7.1
Results bootstrapped t-test and Mann-Whitney test for independent samples.

Statistic[a]	Mean		Median		p-value	
	EX condition	IN condition	EX condition	IN condition	bootstrapped t-test	Mann-Whitney test
Mean C_1	0.24	0.17	0.25	0.15	0.06	0.04
Median C_1	0.09	0.04	0.00	0.00	0.26	0.45
SD C_1	0.44	0.37	0.46	0.40	< 0.01	< 0.01
IQR C_1	0.70	0.40	0.75	0.40	< 0.01	< 0.01
Mean C_2	2.24	1.66	2.27	1.49	0.04	0.02
Median C_2	0.84	0.39	0.00	0.00	0.25	0.44
SD C_2	3.84	3.25	4.00	3.51	< 0.01	< 0.01
IQR C_2	6.07	3.47	6.75	3.50	< 0.01	< 0.01

[a] Calculated across components.

Overall, the results indicate that feedback of expert's own initial ratings may have had an effect on the degree to which experts conformed their ratings to the majority opinion. However, because results are not unequivocal, the null-hypothesis which stated that there is no difference between the two conditions regarding the degree to which experts changed their ratings towards the majority opinion, cannot be rejected.

Increase in the level of agreement

Table 7.2 shows the level of agreement among experts in the EX and IN condition based on SA_r and LK_r. In both conditions the level of agreement as expressed by SA_r showed a significant increase from round 1 to round 2. Nonetheless, based on SA_r no significant difference between conditions was found regarding the increase in the level of agreement. Using LK_r a statistically insignificant increase in the level of agreement was found among experts in both conditions. Consequently, no significant difference between conditions was found regarding the increase in the level of agreement. Based on these results the null-hypothesis cannot be rejected in favour of H_3.

Table 7.2
SA_r and LK_r across all experts and components in the EX and IN condition.

Index	IN condition	EX condition	Difference
SA_1	0.50	0.49	0.00
SA_2	0.56	0.57	–0.01
SA_{2-1}	0.06 (0.02, 0.10)[a]	0.07 (0.02, 0.13)[a]	–0.01 (–0.08, 0.06)[a]
LK_1	0.08	0.08	0.01
LK_2	0.11	0.11	–0.01
LK_{2-1}	0.03 (–0.01, 0.07)[a]	0.04 (–0.002, 0.08)[a]	–0.01 (–0.06, 0.04)[a]

[a] 95% bias-corrected and accelerated confidence intervals based on 10.000 bootstrap samples.

7.4 Discussion

This Delphi experiment aimed to provide insight into the effects of feeding back experts' own initial ratings on three outcome measures: (1) the percentage of questionnaire items on which experts changed their opinion, (2) the degree to which experts changed their ratings towards the majority opinion, and (3) the increase in the level of agreement among experts. In this experiment experts who did not receive their own initial ratings (EX condition) changed their opinion relatively more often than experts who did receive their own initial ratings (IN condition). Regarding the degree to which experts changed their ratings towards the majority opinion, no clear evidence of a difference between conditions was found. Overall, results suggested that experts in the EX condition changed their ratings to a greater degree towards the majority opinion than experts in the IN condition. No difference between conditions was found with regard to the increase in the level of agreement among experts.

The difference between conditions regarding the percentage of opinion changes may have been mediated by reinforced egocentric discounting in the IN condition (Bolger & Wright, 2011). Another explanation for at least part of the difference in the percentage of opinion changes may be random error taking place in the EX condition. Experts in this condition who wanted to give a component the same rating as in the previous round may not have remembered well. This would also explain why the difference in the percentage of opinion changes was not translated into a difference on the second and third outcome measure. The two conformity indices suggested that the degree to which experts changed their ratings towards the majority opinion was rather limited in both conditions. Although differences between conditions were observed, only some of these were statistically significant. In any case, they were not reflected in the generally limited increase in the level of agreement among experts.

Two new conformity indices were developed which both provided insight into the degree to which experts changed their ratings towards the majority opinion. Of course, as the indices have not yet been extensively tested or used in practice, their measurement validity and other properties require further scrutinizing. In the context of Delphi studies the word 'conformity' may suggest that experts adopt a herd mentality by merely changing their ratings towards the majority opinion (Woudenberg, 1991). Regarding the two conformity indices introduced in the current study: they do not provide any insight into whether experts adopted a herd mentality or genuinely changed their ratings towards the majority opinion based on all the information that was fed back.

This experiment has several strengths and weaknesses. The use of two newly developed conformity indices for estimating the degree to which experts changed their ratings towards the majority opinion may be considered to be a weakness. Therefore, differences between conditions were tested using two different location and dispersion statistics as well as two different statistical tests. This made it possible to verify results as well as to provide nuance, strengthening the results of the experiment. Furthermore, as this experiment was conducted within a real-world Delphi study in which actual experts gave their opinion on a topic relevant to them, the ecological validity is probably better than that of artificial laboratory Delphi experiments which have been criticized as being inappropriate (Rowe & Wright, 1999; Rowe et al., 1991). Nonetheless, this real-world Delphi study had a specific design, aim, and panel of experts. Therefore, caution is advised when generalizing the results of the current experiment to other settings.

Back in 1991 Rowe et al. (1991) requested more research into the functioning of the Delphi method, such as the effects of different types of controlled opinion feedback on various Delphi features. Since then some relevant experiments have been conducted (Bolger et al., 2011; Campbell et al., 1999; Gowan & McNichols, 1993; Meijering & Tobi, 2016; Rowe & Wright, 1996; Rowe et al., 2005). Nonetheless, this seems to be the first Delphi experiment that investigated the effect of feeding back experts' initial ratings on opinion change and the level of agreement. More Delphi experiments, preferably conducted within different types of real-world Delphi studies, are needed to confirm or contradict the results found in the current study. The two conformity indices that were proposed will be useful in this regard, although they need to be tested more extensively. For example in a simulation study in which the number of experts, items, points on the rating scale, and the level of opinion change towards the majority opinion are systematically varied (like in Birko et al., 2015 and Meijering et al., 2013).

7.5 Conclusion

Several researchers recommended to include experts' own initial ratings in controlled opinion feedback (Boulkedid et al., 2011; Keeney et al., 2006; Murphy et al., 1998). Until now, empirical evidence for this recommendation was lacking. Based on the results of the current experiment feeding back experts' own initial ratings indeed seems to be justified. Although feedback of initial ratings reduced the percentage of questionnaire items on which experts changed their opinion, this was ultimately not reflected in the level of agreement obtained. Furthermore, by providing experts with their own initial ratings, random errors in the form of unintentional opinion changes may be reduced, improving the reliability of experts' ratings.

Chapter 8

Discussion and conclusions

8.1 Introduction

This thesis revolves around the Delphi method and its potential for developing rankings. The Delphi method is named after the Greek Delphic Oracle, an ancient shrine that according to myth enabled the Greeks to consult the god Apollo about various important decisions. Like the Oracle, the Delphi method supports people in making and justifying decisions. However, instead of consulting the god Apollo, the Delphi method is used to collect data from a group of experts. These experts are independently questioned about a particular topic in several subsequent rounds. After each round, the study moderator provides experts with controlled opinion feedback: a summary of the findings from the previous round. The idea behind this procedure is that it allows experts to improve their judgements and achieve a certain level of agreement on the topic.

Rankings consist of objects that have been ordered based on their performance on a ranking attribute. Because of their apparent simplicity they are often used by individuals, organizations, and governments to determine for example where to live or which processes to improve. Remarkably, the methodology of rankings generally receives little attention. This is worrisome, because within the ranking development process many important and often subjective decisions need to be made that may severely influence ranking results. For example, the decision on how to define the ranking attribute determines how it is measured and thus how good or bad each object performs on the attribute.

In the Introduction of this thesis I explained how the Delphi method may support the development of rankings. First, the method may be used to directly obtain a ranking by asking experts to evaluate objects (e.g. landscape architecture research domains) according to a ranking attribute (e.g. usefulness for landscape architecture practice). Second, the method may be used to allow experts to define and operationalize a complex ranking attribute (e.g. urban sustainability). This could then be used to select specific indicators by which the performance of objects on the ranking attribute may be measured.

Surprisingly, it appeared that the potential of the Delphi method for developing rankings had hardly been explored. Moreover, it turned out that the Delphi method itself had some unresolved methodological issues. For example, there are no evidence-based guidelines on how to measure agreement among experts or what information to include in the summary that is fed back to experts after every round. As a result, agreement is measured in various ways and feedback takes on many different forms. To successfully use the Delphi method for developing rankings, these issues required further investigation. Therefore, in this thesis I set out to find an answer on the following overall research question:

> *What are the methodological challenges and opportunities of the Delphi method for developing rankings?*

This final chapter gives an overview of the main findings of this thesis in light of the overall research question. This overview is followed by a discussion on the limitations of the thesis as well as its contribution to science and society. Finally, several directions for future research are given.

8.2 Overview of main findings

To examine the opportunities of the Delphi method for developing rankings, I conducted two Delphi studies. Chapter 3 reported on a Delphi study in which an international sample of landscape architecture experts was invited to evaluate various research domains in three subsequent rounds according to: (1) their importance for landscape architecture research and (2) their usefulness for landscape architecture practice. Based on the results, two rankings of landscape architecture research domains were built. In both rankings the domains 'human dimensions of planning and design' and 'built environments and infrastructure' came out on top. Experts thus considered research into these two domains desirable from both an academic and practice-oriented view. The Delphi study thereby seemed to be the first to provide insights into which research domains should form the core of a future landscape architecture research agenda.

Chapter 5 and 6 explored the opportunities of the Delphi method for supporting the development of urban sustainability rankings that are based on an indicator system. Chapter 5 reported on a method that I developed to critically examine the methodological characteristics of indicator-based city rankings. This method distinguishes several phases in the ranking development process and facilitates the identification of methodological issues within each phase. The method was applied on six existing rankings that aimed to measure the sustainability of European cities. In doing so, various methodological weaknesses were identified. Most remarkably, none of the rankings provided a clear definition of the ranking attribute. This made it hard to determine whether appropriate indicators were selected. Furthermore, the techniques used for normalizing, weighting, and aggregating indicators were hardly substantiated. This is problematic, because the techniques used may severely influence ranking results (Floridi et al., 2011).

Chapter 6 reported on a Delphi study that was conducted in response to some of the methodological weaknesses identified. European urban sustainability experts were invited to evaluate various components (e.g. waste water treatment, safety, business climate) in three subsequent rounds according to their relevance for defining and measuring urban sustainability. Based on the results of the final round, the following seven components were identified as most relevant: air quality, governance, energy consumption, non-car transportation infrastructure, green spaces, inequality, and CO_2 emissions. Furthermore,

weights of these components could be established that reflect their relative contribution to measuring urban sustainability. Thus, by applying the Delphi method it was possible to define and operationalize the ranking attribute urban sustainability into several weighted components that may guide the future development of indicator-based urban sustainability rankings.

In conducting the two Delphi studies some challenges were encountered. For both studies a large heterogeneous sample of experts needed to be assembled. To do so, selection criteria and search strategies were developed for different types of experts (e.g. landscape architects from academia and from professional practice). Still, some types of experts were hard to find. Additionally, non-response and drop-out caused some types to be underrepresented in the studies. In both Delphi studies it was decided to design a first round questionnaire that included a predefined list of items that could be presented to the experts (i.e. landscape architecture research domains in the first Delphi study and components of urban sustainability in the second Delphi study). Drawing up these lists proved to be complicated as the number of items needed to be limited, despite the great diversity of items available. In the final round of both Delphi studies the level of agreement among experts, regarding the most important and useful research domains as well as the most relevant components of urban sustainability, remained rather low. None of the items was regarded as most important, useful, or relevant by a majority of experts.

Several studies were conducted to acquire insight into the methodological challenges of the Delphi method itself. In chapter 2 the measurement of agreement was examined. Based on literature (among others: Banerjee, 1999; Dijkstra & van Eijnatten, 2009; Hubert, 1977), various existing consensus (i.e. DeMoivre index), agreement (e.g. Light's kappa), and association indices (e.g. Cronbach's alpha) were identified that are suitable for estimating the level of agreement among experts. Additionally, I developed a so called Strict Agreement index that expresses the number of expert pairs that agreed on the rating of an item as a proportion of the total number of expert pairs. Unlike the other agreement indices included in the study, this index does not correct for chance agreement. By means of computer simulations, data were generated according to various Delphi scenarios. All scenarios consisted of three Delphi rounds, but differed from each other with regard to the number of experts that participated, the number of items rated, the distribution of item ratings, and the degree to which experts changed their initial ratings towards the mean ratings across rounds. Next, the indices were used to estimate the level of agreement in the generated data. In doing so, it was shown that in the same data the indices suggested different levels of agreement within Delphi rounds. Moreover, the indices suggested different levels of change in agreement across rounds. The study thus made apparent that the choice for a particular index determines the level of agreement that is finally obtained.

Chapter 4 and 7 reported on two experiments into the provision of controlled opinion feedback. Both experiments aimed to find out what effect different types of feedback have on various Delphi outcome measures. Feedback may consist of summary statistics (e.g. the median rating and interquartile range of each item) and rationales (i.e. summaries explaining why experts rated each item as 'very important' or 'very useful'). Additionally, feedback may include experts' own initial ratings. In the first experiment, which was conducted within the landscape architecture Delphi study, experts either received feedback consisting of summary statistics and rationales, or feedback consisting of rationales only. The results of the experiment suggested that feeding back summary statistics and rationales may increase drop-out of experts, whereas the provision of solely rationales may decrease the level of agreement among experts.

In the second experiment, which was conducted within the urban sustainability Delphi study, experts received feedback that either included or excluded their own initial ratings. To properly analyse the data from the experiment, I developed two so called conformity indices that provided insight into the degree to which experts changed their ratings in the direction of the majority opinion. The results of the experiment suggested that feeding back experts' initial ratings increased the reliability of their subsequent ratings. Furthermore, results suggested that feeding back initial ratings reduced the degree to which experts changed their ratings in the direction of the majority opinion. With regard to the level of agreement that was obtained, no statistically significant difference was found between experts who did and those who did not receive their own initial ratings.

To conclude, this thesis showed that the Delphi method offers opportunities for developing rankings. First, the method proved to be useful for building two rankings of landscape architecture research domains. Second, the method made it possible to define and operationalize the complex ranking attribute urban sustainability into several weighted components.

This thesis also showed that using the Delphi method for developing rankings does not come without challenges. The first challenge relates to assembling a heterogeneous sample of experts. Users of the Delphi method need to develop selection criteria and search strategies by which sufficient numbers of different expert types may be found. For some expert types a potentially low response rate and high drop-out rates may be expected. More effort should then be made to find these expert types and to engage them in the study.

A second challenge involves drawing up a predefined list of items included in the first round questionnaire. Based on a possibly wide variety of available items, users of the Delphi method need to draw up a limited list consisting of about 20 to 30 items. This should be carefully done, because it largely determines the final results of the Delphi study (i.e. the

objects that end up in the resulting ranking or the components of the ranking attribute that are identified as most relevant).

The third challenge concerns the provision of controlled opinion feedback. The two experiments reported in this thesis showed that the information included in feedback may influence important Delphi outcome measures. Users of the Delphi method thus need to carefully consider which types of information to feed back (i.e. summary statistics, rationales, experts' own initial ratings). Feeding back summary statistics in addition to rationales may increase drop-out rates, whereas providing solely rationales may decrease the level of agreement among experts. Both effects are clearly undesirable. Feeding back experts' own initial ratings seems to be justified as it may increase the reliability of experts' ratings and does not seem to have an effect on the level of agreement obtained.

A final challenge lies in measuring the level of agreement among experts. Which index is chosen, determines the level of agreement that is finally obtained. Users of the Delphi method are thus advised to consider whether they want to measure consensus, agreement, or association. In addition, they ought to report the value of the chosen index within every Delphi round to provide insight into how the level of agreement developed across rounds. Depending on which index was chosen, users of the Delphi method also need to be aware that the level of agreement in the final round may remain rather limited.

8.3 Limitations

With regard to answering the overall research question, this thesis has several limitations. The challenges and opportunities regarding the actual application of the Delphi method for developing rankings were explored by means of two Delphi studies. Both studies were designed in a similar and specific way. In the first and second Delphi round experts had to rate a number of mostly predefined items using an ordinal rating scale, while in the third round experts had to select a limited number of items. Feedback was incorporated in the second and third Delphi questionnaire and included rather short summaries of experts' rationales, sometimes supplemented with a table of summary statistics and experts' own initial ratings. Of course, there are other ways of designing Delphi studies that may or may not be better suited for developing rankings. Alternative design may also give rise to different challenges and opportunities of the Delphi method for developing rankings.

Two experiments into the provision of controlled opinion feedback were conducted. Although both experiments gave valuable insights into the effects of different types of feedback on various Delphi outcome measures, general guidelines on how to best provide feedback in Delphi studies cannot yet be established. Both experiments were conducted within real-world Delphi studies that, as explained above, had a specific design.

Furthermore, both experiments were conducted within the context of specific topics, using specific samples of experts. Perhaps that in Delphi studies with different designs, topics, and expert samples, different results will be obtained.

In this thesis three new indices were introduced. The two conformity indices were useful for finding out whether feedback of experts' own initial ratings had an effect on the degree to which experts changed their ratings in the direction of the majority opinion. Little is known, however, about how the indices behave within the context of Delphi studies. To a lesser extent this also applies to the Strict Agreement index. This index was used in both Delphi studies and in both Delphi experiments to measure the level of agreement among experts. Although the behaviour of the index was examined in the simulation study, relatively little is known about how the index behaves in real-world Delphi studies. Results based on the three indices should therefore be interpreted with some caution.

8.4 Scientific contribution

This thesis provides new knowledge on how the Delphi method may be used to obtain a ranking of objects on a ranking attribute. With regard to this application, a ranking-type Delphi method was proposed by Dickson et al. (1984) and Schmidt (1997). As explained in the introduction of this thesis, the ranking-type Delphi method requires experts to actually rank objects according to some attribute. This could be regarded as problematic, because it forces experts to make a specific distinction between objects that they may not actually perceive. In chapter 3 of this thesis a different approach was taken. In the first two rounds experts were asked to rate various research domains according to their importance for landscape architecture research and their usefulness for landscape architecture practice, using a 5-point rating scale. Consequently, experts gave many domains rather similar high ratings, which indicates that they did not perceive clear differences between most domains. In the third and final round experts were therefore asked to select the three most important and the three most useful domains. It was shown that based on this approach two rankings of landscape architecture research domains could be built that are actually meaningful for the field of landscape architecture. As a relatively young academic discipline, landscape architecture still struggles with defining its intellectual core (van den Brink & Bruns, 2014). In this sense, the two rankings of landscape architecture research domains may contribute to the maturation of landscape architecture as an academic discipline.

This thesis also provides knowledge on how the Delphi method may be used to obtain a definition and operationalization of a complex ranking attribute. When it comes to developing rankings based on an indicator-system, the consultation of experts has been recommended (Morse & Fraser, 2005; Singh et al., 2009). Chapter 5 of this thesis showed that developers of urban sustainability rankings indeed often consulted experts, although

not by means of the Delphi method. Research into the use of the Delphi method for developing indicator-based rankings also seemed to be lacking. This thesis addressed this knowledge gap by using the Delphi method to define and operationalize the ranking attribute urban sustainability. In the Delphi study experts were asked to rate various components of urban sustainability in two subsequent rounds. In the third round experts were asked to select the five most relevant components and to distribute 100 points across the selected components. It was shown that based on this approach the most relevant components of urban sustainability could be identified and weighted. The Delphi study may thereby help to resolve the ongoing debate about the definition and measurement of urban sustainability.

By conducting a simulation study into the measurement of agreement and two experiments into the provision of controlled opinion feedback, this thesis also contributes to the limited scientific knowledge about the functioning of the Delphi method. Quite some time ago Rowe et al. (1991) already called for more research into the Delphi method. Although the methodological issues of the Delphi method have since then been frequently discussed (see among others: Hasson & Keeney, 2011; Hung et al., 2008), the number of empirical studies into these issues remained limited. The reasons why this is so, can only be guessed. Perhaps that Delphi researchers rather apply the method than do research into its methodological issues or maybe they lack the necessary knowledge to properly set-up a Delphi experiment or simulation study. Anyhow, it is hoped that this thesis may serve as an inspiration for Delphi researchers to conduct more empirical studies into the Delphi method and develop evidence-based guidelines.

The measurement of agreement in Delphi studies has been recognized as a methodological issue for some time now (e.g. Powell, 2003). Nonetheless, the simulation study as reported in this thesis seems to be the first that actually examined and compared the behaviour of various agreement indices within the context of Delphi studies. This thesis thereby provides new knowledge on how the choice for a particular index may influence the level of agreement that is finally obtained and reported. The provision of controlled opinion feedback has been studied more extensively (e.g. Bolger et al., 2011; Gowan & McNichols, 1993; Rowe & Wright, 1996). Still, no evidence-based guidelines exist that help users of the Delphi method to decide whether or not to feed back summary statistics in addition to rationales or whether or not to feed back experts' own initial ratings. The two experiments reported in this thesis provide new knowledge on how these decisions influence several important Delphi outcome measures, such as the degree to which experts change their opinion and achieve agreement. Additionally, this thesis shows how to conduct these kind of experiments within real-world Delphi studies. This is important, because most Delphi experiments had an artificial set-up in the sense that university students or staff participated as experts and were asked their opinion on rather trivial topics. These so called laboratory

Delphi experiments have been heavily criticised as their results are difficult to generalize to real-world settings (Rowe & Wright, 1999).

During the course of this PhD project I developed an agreement index and two conformity indices that will be useful for future substantive Delphi studies and Delphi experiments. The Strict Agreement index provides insight into the level of agreement among experts within rounds and into the change in the level of agreement across rounds. Although the index does not correct for chance agreement, which could be regarded as a disadvantage, it is easily interpretable. Moreover, correcting for chance agreement may be less relevant within the context of Delphi studies in which experts are invited to give their deliberate opinion. The Delphi method has been criticized for inducing experts to simply conform their ratings to the majority opinion (Bolger & Wright, 2011). In this regard, the two developed conformity indices may be used to acquire insight into the degree to which experts in Delphi studies actually changed their ratings in the direction of the majority opinion.

8.5 Contribution to society

In today's society rankings are everywhere and impact, to a greater or lesser extent, the decisions of ordinary individuals, organizations, and governments. It is therefore important that rankings are well developed in the sense that they reflect the actual performance of objects on the ranking attribute as good as possible. This thesis shows how the Delphi method may be used for developing rankings. It is hoped that this knowledge will stimulate future developers of rankings to actually make use of the Delphi method, for example to properly define and operationalize the ranking attribute of interest. As such, this thesis may ultimately contribute to improving the methodological soundness of rankings and thereby the decisions that are based on these rankings.

To effectively apply the Delphi method, however, insight is needed in how to deal with its various methodological issues. In this thesis two of those issues, the measurement of agreement among experts and the provision of controlled opinion feedback, were studied. In doing so, this thesis may contribute to improving the functioning of the Delphi method and thereby its usefulness for developing rankings. Aside from developing rankings, the Delphi method is often applied to solve complex societal problems related to for example the environment and health (Linstone & Turoff, 1975). This thesis may thus improve the use of the Delphi method in these and other societal domains as well.

This thesis also made clear that during the process of developing an indicator-based city ranking, various methodological choices need to be made that may severely influence ranking results. Moreover, it showed that urban sustainability rankings have several methodological weaknesses. End-users of urban sustainability rankings and rankings in

general are thus advised to look beyond the final ranking results and make an effort to find out how the results were obtained, particularly by examining methodological background documents. The method that I developed for examining the methodological characteristics of indicator-based city rankings will be useful here as it enables end-users of urban sustainability rankings to identify methodological weaknesses. The method may also be useful for identifying methodological weaknesses in indicator-based rankings of different objects (e.g. universities instead of cities), although some adjustments will need to be made. Most importantly, a part of the method that specifically focuses on the selection of cities will need to be broadened so that it may include other types of objects.

The results of the two real-world Delphi studies reported in this thesis also have societal value. The rankings of landscape architecture research domains may form the basis of a future landscape architecture research agenda that meets the needs of professional practice. Such an agenda could bridge the current gap between landscape architects from academia and those from professional practice. The operationalization of the ranking attribute urban sustainability in various weighted components may improve the way in which the sustainability of cities is defined, measured, and monitored. Ultimately, this may help to make our cities more sustainable.

8.6 Future research

This thesis represents an important step in uncovering the methodological challenges and opportunities of the Delphi method for developing rankings. However, more studies need to be conducted in which the Delphi method is used to either define and operationalize a ranking attribute or to directly rank objects on a ranking attribute. In doing so, it is worthwhile to try out different Delphi designs. For example, a variation on the ranking-type Delphi method in which experts in the third phase do not rank, but rate objects. By using different designs it will be possible to find out how to best design Delphi studies for the purpose of developing rankings.

More research is also needed with regard to the methodological issues of the Delphi method. When it comes to the measurement of agreement, it will be valuable to find out how the indices studied in this thesis behave in other types of Delphi scenarios. For example, scenarios in which not an ordinal, but an interval or ratio scale is used. Moreover, other indices may exist or may have been developed that require further investigation by means of a new Delphi simulation study. Those who would like to do such a study should also consider to simulate data according to more realistic scenarios. The simulation study reported in this thesis assumed for example that within a scenario all experts either had a low, moderate, or strong tendency to conform their initial ratings to the majority opinion. This was done, because it was unknown what proportion of experts in Delphi studies have a low, moderate,

or strong tendency to conform. The two conformity indices developed during this PhD project could be applied in various real-world Delphi studies to acquire insight into experts' tendency to conform. This knowledge could then be used to more realistically model the degree to which experts in Delphi studies change their ratings towards the majority opinion.

With regard to future research into the provision of controlled opinion feedback, it is crucial to replicate the two experiments reported in this thesis, albeit in a different context. In this way the results of the two experiments may be supported or contested. Additionally, future experiments should not only try to manipulate the type of information that is fed back to experts, but also *the way* a particular type of information is presented to experts. For example, the rationales that were fed back in the landscape architecture Delphi study were rather short and it will be interesting to find out whether more elaborate feedback of rationales would have more influence on the drop-out rate and the level of agreement among experts. By conducting these kinds of experiments within various real-world Delphi studies, enough evidence may eventually be acquired for drawing up clear guidelines about the provision of controlled opinion feedback.

References

Summary

About the author

Completed training and supervision plan

References

Almeida, C., Braveman, P., Gold, M. R., Szwarcwald, C. L., Ribeiro, J. M., Miglionico, A., ... Viacava, F. (2001). Methodological concerns and recommendations on policy consequences of the World Health Report 2000. *The Lancet*, 357(9269), 1692-1697.

Ameen, R. F. M., Mourshed, M., & Li, H. (2015). A critical review of environmental assessment tools for sustainable urban design. *Environmental Impact Assessment Review*, 55, 110-125.

Arcadis. (2015). *Sustainable Cities Index 2015*. Retrieved from http://www.sustainablecitiesindex.com/

Banerjee, M. (1999). Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics-Revue Canadienne De Statistique*, 27(1), 3-23.

Banwell, C., Hinde, S., Dixon, J., & Sibthorpe, B. (2005). Reflections on expert consensus: A case study of the social trends contributing to obesity. *The European Journal of Public Health*, 15, 564-568.

Bardecki, M. J. (1984). Participants' response to the Delphi method: An attitudinal perspective. *Technological Forecasting and Social Change*, 25, 281-292.

Bederman, S. S., McIsaac, W. J., Coyte, P. C., Kreder, H. J., Mahomed, N. N., & Wright, J. G. (2010). Referral practices for spinal surgery are poorly predicted by clinical guidelines and opinions of primary care physicians. *Medical Care*, 48(9), 852-858.

Besecke, A., & Herkommer, B. (2007). *Schönste Stadt - erfolgreichste Stadt - lebendigste Stadt, Sinn und Unsinn von Städterankings*. Berlin: Universitätsverlag der Technischen Universität Berlin.

Best, R. J. (1974). An experiment in Delphi estimation in marketing decision making. *Journal of Marketing Research*, 11(4), 448-452.

Birko, S., Dove, E. S., & Özdemir, V. (2015). 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*, 10(8), e0135162.

Boje, D. M., & Murnighan, J. K. (1982). Group confidence pressures in iterative decisions. *Management Science*, 28(10), 1187-1196.

Bolger, F., Stranieri, A., Wright, G., & Yearwood, J. (2011). Does the Delphi process lead to increased accuracy in group-based judgmental forecasts or does it simply induce consensus amongst judgmental forecasters? [The Delphi technique: Past, present, and future prospects]. *Technological Forecasting and Social Change*, 78(9), 1671-1680.

Bolger, F., & Wright, G. (2011). Improving the Delphi process: Lessons from social psychological research. [The Delphi technique: Past, present, and future prospects]. *Technological Forecasting and Social Change*, 78(9), 1500-1513.

Boulkedid, R., Abdoul, H., Loustau, M., Sibony, O., & Alberti, C. (2011). Using and reporting the Delphi method for selecting healthcare quality indicators: A systematic review. *PLoS ONE*, 6(6), e20476.

Brancheau, J. C., Janz, B. D., & Wetherbe, J. C. (1996). Key issues in information systems management: 1994-95 SIM Delphi results. *MIS Quarterly*, 20(2), 225-242.

Broomfield, D., & Humphris, G. M. (2001). Using the Delphi technique to identify the cancer education requirements of general practitioners. *Medical Education*, 35(10), 928-937.

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

Brown, R. D., & Corry, R. C. (2011). Evidence-based landscape architecture: The maturing of a profession. *Landscape and Urban Planning*, 100(4), 327-329.

Bruns, D., Ortacesme, V., Stiles, R., de Vries, J., Holden, R., & Jørgensen, K. (2010). *Tuning landscape architecture education in Europe (version 26).* Retrieved from ECLAS website http://www.eclas.org/public/ECLAS%20Guidance%20on%20Education.pdf

Buela-Casal, G., Gutiérrez-Martínez, O., Bermúdez-Sánchez, M., & Vadillo-Muñoz, O. (2007). Comparative study of international academic rankings of universities. *Scientometrics*, 71(3), 349-365.

Bunting, S. W. (2010). Assessing the stakeholder Delphi for facilitating interactive participation and consensus building for sustainable aquaculture development. *Society & Natural Resources*, 23(8), 758-775.

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, 46(5), 423-429.

Cambridge Dictionaries Online. (2012, May 22). Retrieved from http://dictionary.cambridge.org/dictionary/british/consensus?q=consensus

Campbell, S. M., Hann, M., Roland, M. O., Quayle, J. A., & Shekelle, P. G. (1999). The effect of panel membership and feedback on ratings in a two-round Delphi survey: Results of a randomized controlled trial. *Medical Care*, 37(9), 964-968.

Chapman, K., & Pike, L. E. (1992). Sources of City Rankings. *Behavioral & Social Sciences Librarian*, 11(1), 1-11.

Charron, N., Dijkstra, L., & Lapuente, V. (2015). Mapping the regional divide in Europe: A measure for assessing quality of government in 206 European regions. *Social Indicators Research*, 122(2), 315-346.

Chen, Z. (2013). *The role of research in landscape architecture practice. (doctoral thesis)*, Virginia Polytechnic Institute and State University, Blacksburg.

Climate Alliance. (2012, November 29). Climate Star 2012 – The award for successful local authorities in the climate alliance. Retrieved from http://www.klimabuendnis.org/666.html

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37-46.

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, 88(2), 322-328.

Cortina, J. M. (1993). What is coefficient alpha - An examination of theory and applications. *Journal of Applied Psychology*, 78(1), 98-104.

Coste, J., Fermanian, J., & Venot, A. (1995). Methodological and statistical problems in the construction of composite measurement scales: A survey of six medical and epidemiological journals. *Statistics in Medicine*, 14(4), 331-345.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334.

Cronbach, L. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, 64(3), 391-418.

Cushman & Wakefield. (2010, November 29). European Cities Monitor '10. Retrieved from http://www.europeancitiesmonitor.eu/

Dalkey, N. C., Brown, B. B., & Cochran, S. (1969). *The Delphi method: An experimental study of group opinion.* Santa Monica: The Rand Corporation.

Dalkey, N. C., & Helmer, O. (1963). An experimental application of the Delphi method to the use of experts. *Management Science*, 9(3), 458-467.

Deming, E., & Swaffield, S. (2011). *Landscape architectural research: Inquiry, strategy, design*. New York: Wiley.

Dempsey, N., Bramley, G., Power, S., & Brown, C. (2011). The social dimension of sustainable development: Defining urban social sustainability. *Sustainable Development*, 19(5), 289-300.

Di Zio, S., & Pacinelli, A. (2011). Opinion convergence in location: A spatial version of the Delphi method. *Technological Forecasting and Social Change*, 78(9), 1565-1578.

Diamond, I. R., Grant, R. C., Feldman, B. M., Pencharz, P. B., Ling, S. C., Moore, A. M., & Wales, P. W. (2014). Defining consensus: A systematic review recommends methodologic criteria for reporting of Delphi studies. *Journal of Clinical Epidemiology*, 67(4), 401-409.

Dickson, G. W., Leitheiser, R. L., Wetherbe, J. C., & Nechis, M. (1984). Key information systems issues for the 1980's. *MIS Quarterly*, 135-159.

Dijkstra, L., & van Eijnatten, F. M. (2009). Agreement and consensus in a Q-mode research design: An empirical comparison of measures, and an application. *Quality & Quantity*, 43(5), 757-771.

Eberman, L. E., & Cleary, M. A. (2011). Development of a heat-illness screening instrument using the Delphi panel technique. *Journal of Athletic Training*, 46(2), 176-184.

Ecken, P., Gnatzy, T., & von der Gracht, H. A. (2011). Desirability bias in foresight: Consequences for decision quality based on Delphi results. [The Delphi technique: Past, present, and future prospects]. *Technological Forecasting and Social Change*, 78(9), 1654-1670.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. London: Chapman & Hall.

Espeland, W. N., & Sauder, M. (2007). Rankings and reactivity: How public measures recreate social worlds. *American Journal of Sociology*, 113(1), 1-40.

European Commission. (2010a). *Making our cities attractive and sustainable. How the EU contributes to improving the urban environment*. Luxembourg: Publications Office of the European Union.

European Commission. (2010b). Europe 2020: A strategy for smart, sustainable and inclusive growth. Retrieved from http://ec.europa.eu/europe2020/index_en.htm.

European Commission. (2011). *Cities of tomorrow - Challenges, visions, ways forward.* Luxembourg: Publications Office of the European Union.

European Commission. (2015). *Will your city be the European Green Capital in 2018?* Luxembourg: Publications Office of the European Union.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low Kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6), 543-549.

Ferguson, N. D., Davis, A. M., Slutsky, A. S., & Stewart, T. E. (2005). Development of a clinical definition for acute respiratory distress syndrome using the Delphi technique. *Journal of Critical Care*, 20(2), 147-154.

Ferri, C. P., Prince, M., Brayne, C., Brodaty, H., Fratiglioni, L., Ganguli, M., ... & Jorm, A. (2005). Global prevalence of dementia: A Delphi consensus study. *Lancet*, 366(9503), 2112-2117.

Field, A. P. (2013). *Discovering statistics using IBM SPSS statistics (fourth edition).* London: Sage.

Field, A. P. (2005). Intraclass Correlation. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science*. New York: John Wiley & Sons Ltd.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378-382.

Fleiss, J. L., & Cuzick, J. (1979). The reliability of dichotomous judgments: Unequal numbers of judges per subject. *Applied Psychological Measurement*, 3(4), 537-542.

Floridi, M., Pagni, S., Falorni, S., & Luzzati, T. (2011). An exercise in composite indicators construction: Assessing the sustainability of Italian regions. *Ecological Economics*, 70(8), 1440-1447.

Foundation of Environmental Education. (2013, November 29). ECO XXI. Retrieved from http://www.eco-xxi.nl/en

Frewer, L. J., Fischer, A. R. H., Wentholt, M. T. A., Marvin, H. J. P., Ooms, B. W., Coles, D., & Rowe, G. (2011). The use of Delphi methodology in agrifood policy development: Some lessons learned. [The Delphi technique: Past, present, and future prospects]. *Technological Forecasting and Social Change*, 78(9), 1514-1525.

Gamer, M., Lemon, J., Fellows, I., & Puspendra, S. (2010). Irr: Various coefficients of interrater reliability and agreement (Version 0.83), R Foundation for Statistical Computing.

Giffinger, R., Fertner, C., Kramar, H., & Meijers, E. (2007a). *City-ranking of European medium-sized cities*. Retrieved from http://www.smart-cities.eu/download/ city_ranking_final.pdf

Giffinger, R., Fertner, C., Kramar, H., Pichler-Milanovic, N., & Meijers, E. (2007b). European smart cities. Retrieved from http://www.smart-cities.eu/index2.html

Giles, M. W., & Garand, J. C. (2007). Ranking political science journals: Reputational and citational approaches. *PS: Political Science & Politics*, 40(04), 741-751.

Giovannini, E., Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., & Hoffman, A. (2008). *Handbook on constructing composite indicators: Methodology and user guide.* Paris: Organisation for Economic Cooperation and Development (OECD).

Gnatzy, T., Warth, J., von der Gracht, H., & Darkow, I. L. (2011). Validating an innovative real-time Delphi approach - A methodological comparison between real-time and conventional Delphi studies. [The Delphi technique: Past, present, and future prospects]. *Technological Forecasting and Social Change*, 78(9), 1681-1694.

Gobster, P. H. (2014). (Text) mining the LANDscape: Themes and trends over 40 years of Landscape and Urban Planning. *Landscape and Urban Planning*, 126(0), 21-30.

Gobster, P. H., Nassauer, J. I., & Nadenicek, D. J. (2010). Landscape Journal and scholarship in landscape architecture: The next 25 years. *Landscape Journal*, 29(1), 52-70.

Goluchowicz, K., & Blind, K. (2011). Identification of future fields of standardisation: An explorative application of the Delphi methodology. [The Delphi technique: Past, present, and future prospects]. *Technological Forecasting and Social Change*, 78(9), 1526-1541.

Goodland, R. (1999). The biophysical basis of environmental sustainability. In J. C. J. M. van den Bergh (Ed.), *Handbook of environmental and resource economics*. Cheltenham: Edward Elgar Publishing.

Gowan, J. A., & McNichols, C. W. (1993). The effects of alternative forms of knowledge representation on decision-making consensus. *International Journal of Man-Machine Studies*, 38(3), 489-507.

Grabow, B. (2006). *Städterankings - Strategische Entscheidungshilfe statt Siegerwettbewerb Brennpunkt Stadt, Lebens- und Wirtschaftsraum, gebaute Umwelt, politische Einheit, Festschrift für Heinrich Mäding zum 65. Geburtstag.* Berlin: Deutsches Institut für Urbanistik.

Graham, B., Regehr, G., & Wright, J. G. (2003). Delphi as a method to establish consensus for diagnostic criteria. *Journal of Clinical Epidemiology*, 56(12), 1150-1156.

Gray, D. E. (2004). Doing research in the real world. London: Sage Publications.

Gupta, U. G., & Clarke, R. E. (1996). Theory and applications of the Delphi technique: A bibliography (1975–1994). *Technological Forecasting and Social Change*, 53(2), 185-211.

Ham, S. A., Levin, S., Zlot, A. I., Andrews, R. R., & Miles, R. (2004). Ranking of cities according to public health criteria: Pitfalls and opportunities. *American Journal of Public Health*, 94(4), 546-549.

Hargie, O., & Tourish, D. (2000). *Handbook of Communication Audits for Organisations*. London: Routledge.

Hassan, A. M., & Lee, H. (2015). The paradox of the sustainable city: Definitions and examples. *Environment, Development and Sustainability*, 17(6), 1267-1285.

Hasson, F., & Keeney, S. (2011). Enhancing rigour in the Delphi technique research. [The Delphi technique: Past, present, and future prospects]. *Technological Forecasting and Social Change*, 78(9), 1695-1704.

Hendriksen, B., Peereboom, E. C., Ballantine, J., Wielaard, N., & Jansen, S. (2010). *City typology as the basis for policy.* Retrieved from https://www.kpmg.de/docs/city_Typology.pdf

Huang, L., Wu, J., & Yan, L. (2015). Defining and measuring urban sustainability: A review of indicators. *Landscape ecology*, 30(7), 1175-1193.

Huang, S. L., Wong, J. H., & Chen, T. C. (1998). A framework of indicator system for measuring Taipei's urban sustainability. *Landscape and Urban Planning*, 42(1), 15-27.

Hubert, L. (1977). Kappa revisited. *Psychological Bulletin*, 84(2), 289-297.

Hung, H. L., Altschuld, J. W., & Lee, Y. F. (2008). Methodological and conceptual issues confronting a cross-country Delphi study of educational program evaluation. *Evaluation and Program Planning*, 31(2), 191-198.

Hussler, C., Muller, P., & Rondé, P. (2011). Is diversity in Delphi panelist groups useful? Evidence from a French forecasting exercise on the future of nuclear energy. [The Delphi technique: Past, present, and future prospects]. *Technological Forecasting and Social Change*, 78(9), 1642-1653.

Jacobs, R., Goddard, M., & Smith, P. C. (2005). How robust are hospital ranks based on composite performance measures? *Medical Care*, 43(12), 1177-1184.

Jones, L. V. (1971). The nature of measurement. In R. L. Thorndike (Ed.), *Educational Measurement (Second ed.)*. Washington, D.C.: American Council on Education.

Jung-Erceg, P., Pandza, K., Armbruster, H., & Dreher, C. (2007). Absorptive capacity in European manufacturing: A Delphi study. *Industrial Management & Data Systems*, 107(1-2), 37-51.

Kahn, M. E. (2006). *Green cities: Urban growth and the environment.* Washington, D.C.: Brookings institution Press.

Keeney, S., Hasson, F., & McKenna, H. P. (2006). Consulting the oracle: Ten lessons from using the Delphi technique in nursing research. *Journal of Advanced Nursing*, 53(2), 205-212.

Keeney, S., Hasson, F., & McKenna, H. P. (2001). A critical review of the Delphi technique as a research methodology for nursing. *International Journal of Nursing Studies*, 38(2), 195-200.

Kern, K. (2008). *Three types of knowledge transfer and learning: Best practice transfer, benchmarking, and certification*. Paper presented at the Nordic Political Science Association Conference, Norway.

Kraemer, H. C. (1980). Extension of the kappa-coefficient. *Biometrics*, 36(2), 207-216.

Kuik, O. J., & Gilbert, A. J. (1999). Indicators of sustainable development. In J. C. J. M. van den Bergh (Ed.), *Handbook of environmental and resource economics.* Cheltenham: Edgar Elgar Publishing.

LaGro, J. A. (1999). Research capacity: A matter of semantics? *Landscape Journal*, 18(2), 179-186.

Landeta, J. (2006). Current validity of the Delphi method in social sciences. *Technological Forecasting and Social Change*, 73(5), 467-482.

Landeta, J., & Barrutia, J. (2011). People consultation to construct the future: A Delphi application. *International Journal of Forecasting*, 27(1), 134-151.

Landeta, J., Barrutia, J., & Lertxundi, A. (2011). Hybrid Delphi: A methodology to facilitate contribution from experts in professional contexts. [The Delphi technique: Past, present, and future prospects]. *Technological Forecasting and Social Change*, 78(9), 1629-1641.

Landis, J. R., & Koch, G. G. (1977). A one-way components of variance model for categorical data. *Biometrics*, 33(4), 671-679.

Lange, R. (2010). Benchmarking, Rankings und Ratings. In D. Simon, A. Knie, & S. Hornbostel (Eds.), *Handbuch Wissenschaftspolitik*. VS Verlag für Sozialwissenschaften.

Lenzholzer, S., Duchhart, I., & Koh, J. (2013). 'Research through designing' in landscape architecture. *Landscape and Urban Planning*, 113(0), 120-127.

Liao, H., Zeng, A., Xiao, R., Ren, Z. M., Chen, D. B., & Zhang, Y. C. (2014). Ranking Reputation and Quality in Online Rating Systems. *PLoS ONE*, 9(5), e97146.

Light, R. J. (1971). Measures of response agreement for qualitative data - Some generalizations and alternatives. *Psychological Bulletin*, 76(5), 365-377.

Lindgren, B. W. (1976*). Statistical Theory*. London: Collier Macmillan Publishers.

Linstone, H. A., & Turoff, M. (1975). *The Delphi method: Techniques and applications*. Massachusetts: Addison-Wesley.

Linstone, H. A., & Turoff, M. (2011). Delphi: A brief look backward and forward. *Technological Forecasting and Social Change*, 78(9), 1712-1719.

Lohuis, A. M., van Vuuren, M., & Bohlmeijer, E. (2013). Context-specific definitions of organizational concepts: Defining 'team effectiveness' with use of the Delphi technique. *Journal of Management & Organization*, 19(6), 706-720.

Lorr, M. J. (2012). Defining urban sustainability in the context of North American cities. *Nature and Culture*, 7(1), 16-30.

Lun, G., Holzer, D., Tappeiner, G., & Tappeiner, U. (2006). The stability of rankings derived from composite indicators: Analysis of the "IL Sole 24 Ore" quality of life report. *Social Indicators Research*, 77(2), 307-331.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is Stochastically larger than the other. *The Annals of Mathematical Statistics*, 18, 50-60.

Maclaren, V. W. (1996). Urban sustainability reporting. *Journal of the American Planning Association*, 62(2), 184-202.

Marchais-Roubelat, A., & Roubelat, F. (2011). The Delphi method as a ritual: Inquiring the Delphic Oracle. *Technological Forecasting and Social Change*, 78(9), 1491-1499.

Maretzke, S. (2006). Regionale Rankings - ein geeignetes Instrument für eine vergleichende Bewertung regionaler Lebensverhältnisse? *Informationen zur Raumentwicklung*, 6, 325-335.

Mayer, A. L. (2008). Strengths and weaknesses of common sustainability indices for multidimensional systems. *Environment International*, 34(2), 277-291.

McManus, P. (2012). Measuring Urban Sustainability: The potential and pitfalls of city rankings. *Australian Geographer*, 43(4), 411-424.

Meijering, J. V., Kampen, J. K., & Tobi, H. (2013). Quantifying the development of agreement among experts in Delphi studies. *Technological Forecasting and Social Change*, 80, 1607-1614.

Meijering, J. V., Kern, K., & Tobi, H. (2014). Identifying the methodological characteristics of European green city rankings. *Ecological Indicators*, 43, 132–142.

Meijering, J. V., & Tobi, H. (2016). The effect of controlled opinion feedback on Delphi features: Mixed messages from a real-world Delphi study. *Technological Forecasting and Social Change*, 103, 166-173.

Meijering, J. V., Tobi, H., Brink, A. van den, Morris, F., & Bruns, D. (2015). Exploring research priorities in landscape architecture: an international Delphi study. *Landscape and Urban Planning*, 137, 85-94.

Mercer. (2012). Mercer's 2012 Quality of Living ranking highlights - Global. Retrieved November from http://www.mercer.nl/articles/quality-of-living-survey-report-2011

Michael, F. L., Noor, Z. Z., & Figueroa, M. J. (2014). Review of urban sustainability indicators assessment – Case study between Asian countries. *Habitat International*, 44, 491-500.

Milburn, L. S., & Brown, R. D. (2003). The relationship between research and design in landscape architecture. *Landscape and Urban Planning*, 64(1–2), 47-66.

Milburn, L. S., Brown, R. D., & Paine, C. (2001). "… Research on research": Research attitudes and behaviors of landscape architecture faculty in North America. *Landscape and Urban Planning*, 57(2), 57-67.

Morse, S., & Fraser, E. D. G. (2005). Making 'dirty' nations look clean? The nation state and the problem of selecting and weighting indices as tools for measuring progress towards sustainability. *Geoforum*, 36(5), 625-640.

Murphy, M. K., Black, N. A., Lamping, D. L., McKee, C. M., Sanderson, C. F., Askham, J., & Marteau, T. (1998). Consensus development methods, and their use in clinical guideline development. *Health technology assessment (Winchester, England)*, 2(3), i-iv, 1-88.

Nesshöver, C., Berghöfer, A., & Beck, S. (2007). *Weltranglisten als Beratungsinstrumente für Umweltpolitik: Eine Einschätzung des Environmental Performance Index.* Marburg: Metropolis-Verl.

Nevo, D., & Chan, Y. E. (2007). A Delphi study of knowledge management systems: Scope and requirements. *Information & Management*, 44(6), 583-597.

Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*, 17(8), 857-872.

Ochel, W., & Röhn, O. (2008). Indikatorenbasierte Länderrankings. *Perspektiven der Wirtschaftspolitik*, 9(2), 226-251.

Olewiler, N. (2006). Environmental sustainability for urban areas: The role of natural capital indicators. *Cities*, 23(3), 184-195.

Pare, G., Cameron, A.-F., Poba-Nzaou, P., & Templier, M. (2013). A systematic assessment of rigor in information systems ranking-type Delphi studies. *Information & Management*, 50(5), 207-217.

Parente, R., & Anderson-Parente, J. (2011). A case study of long-term Delphi accuracy. [The Delphi technique: Past, present, and future prospects]. *Technological Forecasting and Social Change*, 78(9), 1705-1711.

Parke, H. W. (1939). *A history of the Delphic oracle*. Oxford: B. Blackwell.

Pätäri, S. (2010). Industry- and company-level factors influencing the development of the forest energy business — Insights from a Delphi Study. *Technological Forecasting and Social Change*, 77(1), 94-109.

Petschow, U., Rosenau, J. N., & Weizsäcker, E. U. (2005). Governance and sustainability: New challenges for states, companies and civil society. Saltaire: Greenleaf Publications.

Powell, C. (2003). The Delphi technique: Myths and realities. *Journal of Advanced Nursing*, 41(4), 376-382.

Powers, M. N., & Walker, J. B. (2009). Twenty-five years of Landscape Journal: An analysis of authorship and article content. *Landscape Journal*, 28(1), 96-110.

Quacquarelli Symonds. (2013, June 17). QS World University Rankings. Retrieved from http://www.topuniversities.com/university-rankings

Qualtrics [Computer software]. (2015). Retrieved from http://www.qualtrics.com

R Core Team [Computer software] (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from: https://www.R-project.org/

Rauhvargers, A. (2011). Global university rankings and their impact. Belgium: European University Association.

Robinson, D., Bylund, J., Coutard, O., Finnveden, G., Hooimeijer, P., Kabisch, S., ... & Riegler, J. (2015). Transition towards sustainable and liveable urban futures: The strategic research and innovation agenda of JPI Urban Europe. Retrieved from http://jpi-urbaneurope.eu/publications-2/

Rowe, G., & Wright, G. (1996). The impact of task characteristics on the performance of structured group forecasting techniques. *International Journal of Forecasting*, 12(1), 73-89.

Rowe, G., & Wright, G. (1999). The Delphi technique as a forecasting tool: Issues and analysis. *International Journal of Forecasting*, 15(4), 353-375.

Rowe, G., Wright, G., & Bolger, F. (1991). Delphi: A reevaluation of research and theory. *Technological Forecasting and Social Change*, 39(3), 235-251.

Rowe, G., Wright, G., & McColl, A. (2005). Judgment change during Delphi-like procedures: The role of majority influence, expertise, and confidence. *Technological Forecasting and Social Change*, 72(4), 377-399.

Rushton, A., & Moore, A. (2010). International identification of research priorities for postgraduate theses in musculoskeletal physiotherapy using a modified Delphi technique. *Manual Therapy*, 15(2), 142-148.

Schmidt, R. C. (1997). Managing Delphi surveys using nonparametric statistical techniques. *Decision Sciences*, 28(3), 763-774.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, 8(4), 350-353.

Schönert. (2003). Städteranking und Imagebildung: Die 20 grössten Städte in Nachrichten- und Wirtschaftsmagazinen. *BAW Monatsbericht*, 2, 1-8.

Schwengler, B., & Binder, J. (2006). Lösungsansatz zum Gewichtungsproblem der Förderindikatoren beim Übergang zu einem gesamtdeutschen Modell. *Raumforschung und Raumordnung*, 64(4), 284-298.

Science Communication Unit (2015) Indicators for sustainable cities (in-depth report 12). Retrieved from http://ec.europa.eu/environment/integration/research/newsalert/pdf/indicators_for_sustainable_cities_IR12_en.pdf

Scott, W. A. (1955). Reliability of Content Analysis. *Public Opinion Quarterly*, 19(3), 321-325.

Seidel-Schulze, A., Grabow, B., & Tobsch, V. (2009). *Lebenszufriedenheit in Europäischen Städten. Auswertung des Urban Audit European Perception Survey*. Berlin: Deutsches Institut für Urbanistik.

Shane, A. M., & Graedel, T. E. (2000). Urban environmental sustainability metrics: A provisional set. *Journal of Environmental Planning and Management*, 43(5), 643-663.

Shen, L. Y., Ochoa, J. J., Shah, M. N., & Zhang, X. (2011). The application of urban sustainability indicators – A comparison between various practices. *Habitat International*, 35(1), 17-29.

Shih, T., & Fan, X. (2008). Comparing response rates from web and mail surveys: A meta-analysis. *Field methods*, 20(3), 249-271.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420-428.

Siegel, S., & Castellan, N. J. (1988). Nonparametric statistics for the behavioral sciences: McGraw-Hill.

Siemens. (2009). European Green City Index. Retrieved from http://www.siemens.com/entry/cc/features/greencityindex_international/all/en/pdf/report_en.pdf

Silverman, D. (2011). Interpreting qualitative data. London: Sage.

Singh, R. K., Murty, H. R., Gupta, S. K., & Dikshit, A. K. (2009). An overview of sustainability assessment methodologies. *Ecological Indicators*, 9(2), 189-212.

Skulmoski, G. J., Hartman, F. T., & Krahn, J. (2007). The Delphi method for graduate research. *Journal of Information Technology Education*, 6, 21.

Steinert, M. (2009). A dissensus based online Delphi approach: An explorative research tool. *Technological Forecasting and Social Change*, 76(3), 291-300.

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, 80(1), 99-103.

Tai, L. (2003). Doctoring the profession. *Landscape Architecture*, 93(11), 64-73.

Tanguay, G. A., Rajaonson, J., Lefebvre, J. F., & Lanoie, P. (2010). Measuring the sustainability of cities: An analysis of the use of local indicators. *Ecological Indicators*, 10(2), 407-418.

Taylor, Z. (2011). 'Lies, damned lies, and statistics' a critical examination of city ranking studies. Toronto, Canada: Metapolis Consulting.

Tobi, H. (2014). Measurement in interdisciplinary research: The contributions of widely-defined measurement and portfolio representations. *Measurement*, 48, 228-231.

Türksever, A. N., & Atalik, G. (2001). Possibilities and limitations for the measurement of the quality of life in urban areas. *Social Indicators Research*, 53(2), 163-187.

Turoff, M. (1970). The design of a policy Delphi. *Technological Forecasting and Social Change*, 2(2), 149-171.

van de Linde, E., & van der Duin, P. (2011). The Delphi method as early warning: Linking global societal trends to future radicalization and terrorism in The Netherlands. [The Delphi technique: Past, present, and future prospects]. *Technological Forecasting and Social Change*, 78(9), 1557-1564.

van den Brink, A. van den, & Bruns, D. (2014). Strategies for enhancing landscape architecture research. Landscape Research, 39(1), 7-20.

Verhagen, A. P., de Vet, H. C. W., de Bie, R. A., Kessels, A. G. H., Boers, M., Bouter, L. M., & Knipschild, P. G. (1998). The Delphi list: A criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. *Journal of Clinical Epidemiology*, 51(12), 1235-1241.

Ward-Thompson, C. (2010). *The art and science of landscape architecture research*. Paper presented at the Landscape Legacy conference, Maastricht, The Netherlands.

Williams, R., & Van Dyke, N. (2008). Reputation and reality: Ranking major disciplines in Australian universities. *Higher Education*, 56(1), 1-28.

Willis, G. B. (2005). *Cognitive interviewing. A tool for improving questionnaire design.* Thousand Oaks: Sage Publications.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158), 209-212.

Wilson, J., Tyedmers, P., & Pelot, R. (2007). Contrasting and comparing sustainable development indicator metrics. *Ecological Indicators*, 7(2), 299-314.

Woudenberg, F. (1991). An evaluation of Delphi. *Technological Forecasting and Social Change*, 40(2), 131-150.

World Commission on Environment and Development. (1987). *Our Common Future* (Brundtland report). Oxford: Oxford University Press.

Uebersax, J. S. (1982). A generalized kappa-coefficient. *Educational and Psychological Measurement*, 42(1), 181-183.

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, 93(1), 1-13.

Yu, Y., & Wen, Z. (2010). Evaluating China's urban environmental sustainability with data envelopment analysis. *Ecological Economics*, 69(9), 1748-1755.

Zill, J. M., Scholl, I., Härter, M., & Dirmaier, J. (2015). Which dimensions of patient-centeredness matter? - Results of a web-based expert Delphi survey. *PLoS ONE*, 10(11), e0141978.

Summary

In today's society rankings are everywhere. There are rankings of the most prominent universities, the most sustainable cities, and the happiest countries to name only a few. Rankings are often used by individuals, organizations, and governments to make important decisions. Nonetheless, the methodology of rankings generally receives little attention. This is worrisome, because how a ranking was developed may severely influence ranking results.

The development of rankings may be supported by the Delphi method: a structured group communication process in which data are collected from experts in several subsequent rounds with the aim of allowing these experts to achieve agreement on a particular topic. The Delphi method may be used to directly obtain a ranking of objects (e.g. research domains in landscape architecture) on a ranking attribute (e.g. usefulness for landscape architecture practice). Alternatively, the method may be used to obtain a definition and operationalization of a complex ranking attribute (e.g. urban sustainability) on the basis of which an indicator-based ranking may be developed.

Remarkably, it seemed that the potential of the Delphi method for developing rankings had hardly been explored. Moreover, it appeared that the Delphi method itself had some unresolved methodological issues that required further investigation. Therefore, this PhD thesis set-out to find an answer on the following general research question

> *What are the methodological challenges and opportunities of the Delphi method for developing rankings?*

Within the context of this general research question six studies were conducted. Half of the studies focused on two major methodological challenges of the Delphi method itself: the measurement of agreement and the provision of controlled opinion feedback (i.e. the summary of findings that is sent to the experts after every round). The other half aimed to provide insights into challenges and opportunities regarding the actual application of the Delphi method for developing rankings.

In chapter 2 the methodological challenge of measuring agreement among experts in Delphi studies was examined. First, a theoretical distinction was made between indices that measure consensus, agreement, or association. For each type, one or more existing indices were sought. Additionally, I developed a so called Strict Agreement index. The behaviour of all indices was examined by applying them on datasets that were simulated according to various Delphi scenarios. It was shown that within the same data the indices suggested different levels of agreement, and also, different levels of change of agreement across Delphi rounds. Whereas some indices generally showed a limited increase in agreement across rounds (e.g. Light's kappa), other indices showed a substantial increase in agreement

(e.g. Cronbach's alpha). This implies that users of the Delphi method need to clearly justify how they intend to measure agreement among experts as it may severely influence the level of agreement that is finally obtained and reported.

Chapter 3 described a study in which I actually applied the Delphi method to obtain a ranking of objects on a ranking attribute. In this study an international sample of landscape architecture experts was invited to evaluate various research domains in three subsequent rounds according to: (1) their importance for landscape architecture research and (2) their usefulness for landscape architecture practice. Based on experts' subsequent evaluations two rankings of landscape architecture research domains could be built. In both rankings the domains 'built environments and infrastructure' and 'human dimensions of planning and design' came out on top, indicating that these domains may form the core of a future landscape architecture research agenda. Challenges in applying the Delphi method were also encountered. Putting together a heterogeneous sample of landscape architecture experts was difficult as some types of experts were hard to find. Additionally, because of differences in response and drop-out rates some types were underrepresented in the study. Drawing up a wide-ranging, but limited list of research domains that could be presented to the experts in the first Delphi questionnaire also proved to be challenging. In the field of landscape architecture numerous different domains exist, making it difficult to develop a parsimonious list that reflected the great diversity of domains available. Finally, the level of agreement among experts (as measured by the Strict Agreement index) regarding the most important and useful domains remained rather low.

Chapter 4 reported on an experiment that was included in the landscape architecture Delphi study. This experiment examined the methodological challenge of providing controlled opinion feedback. Different types of information may be fed back to experts and it is unknown what effects different types have on various Delphi outcome measures. In this experiment the effects of two controlled opinion feedback conditions were investigated. In one condition feedback consisted of solely rationales (i.e. a summary of experts' explanations on why they evaluated each domain as important or useful). In the other condition feedback consisted of both rationales and summary statistics (i.e. the median and interquartile range of each domain as well as the percentage of experts that evaluated each domain as 'very important' or 'very useful'). Results suggested that feeding back both summary statistics and rationales may increase drop-out of experts, whereas the provision of solely rationales may decrease the level of agreement among experts (as measured by the Strict Agreement index and Light's kappa). Users of the Delphi method are thus advised to carefully consider whether or not to feed back summary statistics in addition to rationales.

Chapter 5 and 6 explored the opportunities of the Delphi method for developing rankings that are based on an indicator system. Chapter 5 reported on a method that I developed for identifying the methodological characteristics of indicator-based city rankings. This method

was applied on six existing urban sustainability rankings. It was shown that all these rankings had methodological weaknesses. Most remarkably, none of the rankings provided a clear definition of the ranking attribute, which made it hard to establish whether appropriate indicators were selected. Furthermore, the rankings hardly substantiated their decisions regarding the techniques that were used to normalize, weight, and aggregate the selected indicators into composite indicator values and rank numbers.

Following these results, chapter 6 described a study in which the Delphi method was applied to define and operationalize the ranking attribute urban sustainability. A sample of European urban sustainability experts was invited to evaluate various components (e.g. air quality, inequality, business climate) in three subsequent rounds according to their relevance for defining and measuring urban sustainability. Based on the results of the final round seven components could be identified as most relevant (air quality, governance, energy consumption, non-car transportation infrastructure, green spaces, inequality, and CO_2 emissions). Additionally, weights of these components could be established that may reflect their relative contribution to measuring urban sustainability. Challenges in applying the Delphi method, similar to those described in chapter 3, were also encountered. Assembling a heterogeneous sample of urban sustainability experts was difficult as some types of experts were hard to find, resulting in a limited representation of those types in the study. Drawing up a parsimonious list of urban sustainability components proved to be challenging, considering the great diversity of components available. Finally, the level of agreement among experts (as measured by the Strict Agreement index) regarding the most relevant components remained rather low.

Chapter 7 reported on an experiment that was included in the urban sustainability Delphi study. Like in chapter 4, this experiment examined the methodological challenge of providing controlled opinion feedback. This time, however, feedback either included or excluded experts' own ratings from the previous round. Results showed that experts who did not receive their own initial ratings changed their opinion relatively more often than experts who did receive their own initial ratings. This may be because experts who did not receive their own initial ratings made more random errors when they wanted to give a component the same rating as in the previous round. Results based on two conformity indices that I developed also suggested that experts who did not receive their own initial ratings changed their ratings to a greater degree towards the majority opinion than experts who did receive their own initial ratings. With regard to the level of agreement that was obtained, no statistically significant difference was found between experts who did and those who did not receive their own initial ratings. Overall, the experiment suggested that feeding back initial ratings is justified as it may improve the reliability of experts' subsequent ratings and does not seem to influence the level of agreement obtained.

Overall, this thesis showed how the Delphi method may be used to: (1) obtain a ranking of objects on a ranking attribute and (2) obtain a definition and operationalization of a complex ranking attribute. It was also shown that these applications of the Delphi method do not come without challenges. First, selection criteria and search strategies need to be developed by which sufficient numbers of different types of experts may be found. Second, for the first Delphi questionnaire a parsimonious list of items (e.g. ranking objects or components of a ranking attribute) needs to be drawn up that largely covers the potentially great diversity of existing items. Third, careful considerations need to be made about which types of information (i.e. summary statistics, rationales, experts' own initial ratings) to feed back to experts after every round as this may influence various Delphi outcome measures. Fourth, the choice for a particular consensus, agreement, or association index needs to be carefully made and justified as it determines the level of agreement among experts that is obtained. Finally, this thesis showed that by means of a simulation study and two real-world Delphi experiments new knowledge about the functioning of the Delphi method could be acquired. More of these studies are needed to establish evidence-based guidelines and to uncover the full potential of the Delphi method for developing rankings.

About the author

Jurian Meijering was born on August 18, 1981 in the city of Ede (the Netherlands). In 2007 he graduated cum laude at the University of Twente with a master degree in communication science. Shortly thereafter he started to work as a project manager within the commercial market research company Blauw Research. At the end of 2009 Jurian went back to academia where he started to work as a lecturer within the Research Methodology group of Wageningen University. As a lecturer Jurian taught several courses to bachelor and master students with various nationalities and from various study programs. He also obtained his University Teaching Qualification.

In December 2010 Jurian was awarded a PhD scholarship with the Wageningen School of Social Sciences (WASS). In September 2011 he became a PhD candidate for five years in which he also remained a lecturer for 20% of his time. As a PhD candidate Jurian conducted several studies into the Delphi method and the application of this method to the development of rankings. Additionally, he worked some time abroad at the German Leibniz institute for Research on Society and Space (IRS), became an active member of the Interuniversity graduate school of Psychometrics and Sociometrics (IOPS), and completed various research related courses. In 2015 Jurian won the WASS best paper award for his paper about the methodological characteristics of European green city rankings (chapter 5 of this thesis). Jurian presented his work at various national and international conferences, such as the European Congress of Methodology and the European Survey Research Association conference.

Jurian Meijering
Wageningen School of Social Sciences (WASS)
Completed Training and Supervision Plan

Name of the learning activity	Department/Institute	Year	ECTS*
A) Project related competences			
Writing the PhD proposal	ECS (RME)	2011	3
Governance for Sustainable Cities (ENP-36806)	ENP	2012	3
Quantitative Data Analysis: Multivariate Techniques (YRM-60306)	ECS (RME)	2012	6
Qualitative Data Analysis: Procedures and Strategies (YRM-60806)	ECS (RME)	2011	3
B) General research related competences			
Introduction course	WASS	2011	1
Techniques for Writing and Presenting a Scientific Paper	WGS	2012	1.2
Information Literacy Including Endnote Introduction	WGS	2013	0.6
What is Psychometrics	IOPS	2015	2
Advising on Research Methods	IOPS	2015	3
Meta-analysis	IOPS	2016	1
C) Career related competences/personal development			
Visiting a Research Institute Abroad	IRS	2012	6
Competence Assessment	WGS	2016	1.5
Mobilising your – Scientific – Network	WGS	2012	1
Project & Time Management	WGS	2012	1.5
Career Orientation	WGS	2016	1.5

Jurian Meijering
Wageningen School of Social Sciences (WASS)
Completed Training and Supervision Plan (continued)

Name of the learning activity	Department/Institute	Year	ECTS*
C) Career related competences/personal development			
'Quantifying the development of agreement among experts in Delphi studies'	European Survey Research Association (ESRA) conference, Ljubljana, Slovenia	2013	1
'Using the Delphi method in an interdisciplinary setting: Opportunities and implications'	European Congress of Methodology (ECM), Utrecht, Netherlands	2014	1
'Ranglijsten van Europese groene steden: De methodologische kenmerken in kaart gebracht'	Workshop on invitation of the Nederlandstalig Platform voor Survey Onderzoek (NPSO), The Hague, Netherlands	2014	1
'Identifying the methodological characteristics of European green city rankings'	Workshop on invitation of the Joint Research Centre of the European Commission, Turin, Italy	2014	1
'Exploring research priorities in landscape architecture: An international Delphi study'	European Council of Landscape Architecture Schools (ECLAS) conference, Tartu, Estonia	2015	1
'Probing the power of Apollo: Methodological challenges and opportunities of the Delphi method for developing rankings'	IOPS Summer Conference, Enschede, Netherlands	2016	1
Total			**41.3**

*One credit according to ECTS is on average equivalent to 28 hours of study load.

Abbreviations
ECS stands for Educational Competence Studies group
RME stands for Research Methodology group
ENP stands for Environmental Policy group
WGS stands for Wageningen Graduate Schools
IOPS stands for Interuniversitaire Onderzoekschool voor Psychometrie en Sociometrie
IRS stands for Institut für Raumbezogene Sozialforschung

Printed by: Digiforce B.V. || Uitgeverij Boxpress

Cover design: Jurian Meijering
Photo front cover obtained from: iStock.com/Marcus Lindstrom
Photo back cover obtained from: iStock.com/Gargolas