Metrics, Indicators and Analytics to Support Government Excellence Programme: The Case of Dubai Government Website Excellence Model (WEM)

Hazza Khalfan Matar Hadday Alnuaimi

Doctor of Business Administration

Aston University

June 2019

Metrics, Indicators and Analytics to Support Government Excellence Programme: The Case of Dubai Government Website Excellence Model (WEM)

Hazza Alnuaimi

Doctor of Business Administration

2019

Summary

This research is focused on the construction of composite indicators: a complex process involving various steps that have significant impact on the results. One of the main problems in constructing composite indicators is its reliance on multiple subjective judgments (Cherchye et al., 2008). This was clearly demonstrated in the case of Website Excellence Model (WEM) scores, whose main purpose is to assess and compare the performance of Dubai Government departments' website. Many subjective judgments were being made by different parties in each of the three main stages of the WEM process: pre-assessment, assessment and post-assessment stage. This level of subjectivity led to a problem where many departments end up being unsatisfied with the overall scores and the general process of deriving the results.

This research indicates that at each stage of the WEM process, the reliability, validity and fairness of the results were affected. To construct a more accurate, flexible, equitable and transparent WEM scoring methodology, we proposed the use of geometric data envelopment analysis model (G-DEA) along with some general guidelines to be followed during different stages of the process. G-DEA methodology combines positive characteristics of geometric aggregation, Analytical Hierarchy Process (AHP) and DEA. Geometric aggregation makes improvements on two different levels. First, it is better suited for constructing WEM scores than the "standard" additive aggregation, for much the same reasons as for why the switch from additive to geometric aggregation took place for Human Development Index back in 2010. Second, it allows for DEA-like models to be easily extended and applied to a composite indicator irrespective of how complex its hierarchy structure may be. The elements of AHP and DEA contribute through their own well-known properties, such as the reduction of decision bias (AHP and DEA) and an equitable evaluation of departments relative to the observed best practices (DEA).

In short, this thesis proposes the use of G-DEA model and discusses the most relevant theoretical and practical aspects and features of that method when applying it to WEM scores. G-DEA methodology is well suited for the WEM scoring framework but there are certainly many other applications, relating to the construction of composite indicators that could benefit from the same methodology. Overall, this study aims to provide both practitioners and academics in the field of composite indicators with a clear application focus on using G-DEA to assess website performance, penetrating the area which so far has never been used in the context of composite indictors. In addition, this study clearly illustrates how G-DEA can combine many good qualities of different well-known techniques for constructing composite indicators.

Keywords: Composite Indicators, Data Envelopment Analysis (DEA), weighting, aggregation, performance measurement

Dedication

This thesis is dedicated to my dearest sons Khalfan and Rakan

Acknowledgments

My doctorate study has been one of the greatest learning experiences I have embarked on a few years back and possibly the most challenging. Along this journey, I have gained a lot of knowledge in a field that was totally novel to me, managed to balance between the different demands of work and academia, getting married and being blessed with two loving boys. My doctorate days are ones I will cherish forever and I am deeply appreciative of so many people and institutions that have been integral to this research and without whom this work would have never been possible.

Firstly, I would like to express my gratitude to The Executive Council of Dubai Government for encouraging me to pursue my doctorate study and providing the funds and time to do so. I am specifically thankful to His Excellency Abdullah Al Basti and His Excellency Abdullah Al Shaibani for their exceptional encouragement and support during this journey. My sincere thanks also goes to Dubai Smart Government and particularly Ms. Yara Kakish for supplying all the information needed and the data that this research has relied upon.

A big recognition and appreciation goes to my supervisors Dr. Ali Emrouznejad and Dr. Ozren Despic whom have been of invaluable support since the early days of my studies. My profound gratitude goes to my main supervisor, Dr. Ozren, who's input and guidance have shaped the final form that the thesis has taken. His continuous trust, patience and support have helped me accomplish this research for which I am extremely grateful.

I am especially grateful to my mother Amna and brother Rashid, who have been unwavering in their encouragement, love and support.

Last, but never least, I thank the three most important people in my life: my dearest wife Shumous whom I'm greatly indebted to for her dedication, support and belief over the past few years, and my two little sons Khalfan and Rakan whom have given me purpose in life. I am forever grateful.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction and Background of the Study

1.1. Introduction

Many organizations around the world, particularly governments and international regulatory bodies, develop, customize, and use what are known as composite indicators to understand and measure relative progress. These are mathematically calculated figures that are used to assess, and comparatively rate their own performance with respect to a range of activities, such as the effectiveness of policy, economic and other forms of development, or even quality of customer service (Cherchye et al., 2008). Composite indicator (CI) comprise multiple data, weighted and aggregated into a single value, which should ideally be meaningful, objective, and useful for comparing the overall performance of the units assessed on a specific concept. This is best understood by way of a conventional example, the United Nation's Human Development Index (HDI), with which most are familiar. The HDI is a CI that can be used to rank countries based on their aggregate performance relative to their education, income, and life expectancy. These broader categories further categorize other measurable factors like literacy rate, access to healthcare, and GDP, among numerous other criteria. In addition to governments, international media and the development sector also make use of such CIs to provide the wider public with an understanding of relative global social progress.

While CIs are understood to provide understanding of ideas such as progress, achievement, or under development, in a tiered manner relative to objective scales of measurement, their methodological construction has known drawbacks. Cherchye et al. (2008) show how a number of "subjective judgements" inform the making of composite indicators, and therefore undermine the objective understanding that they are designed and aim to represent. These subjective judgements include decisions about what data should be included or excluded, and therefore what data qualifies as a meaningful sub-indicator to larger indicators. Another problem is that it is also frequently not clear what weights will be assigned to each sub-indicator in relation to the others included. These drawbacks are often accounted for by using

a methodology based on data envelopment analysis (DEA). These methodologies can help with understanding how composite indicators are constructed whilst they are being constructed, offering opportunities for deriving insights about the most promising directions for the assessed unit to improve their performance. This process is made further robust through the inclusion of uncertainty and sensitivity analysis which help to identify how subjective assessments are made and how errors in data occur (Cherchye et al., 2008).

Drawing significantly from Cherchye et al. (2008), this thesis is a case study examining an initiative within the Dubai government to improve and measure the excellence of its performance in terms of its use of information communication technology. It was to this end that the Dubai Smart Government (DSG) department was established in 2000, including a mission to "deliver world-class smart services and infrastructure to create happiness" (Smartdubai.ae, 2019). Its role is to encourage various city government departments to institutionalize web technology in their operations for better performance and improved customer satisfaction. Currently, DSG uses a composite indicator called the Website Excellence Model (WEM) to assess the quality of government websites, with respect to such factors as ease of use, efficiency, and quality of information, among others. The DSG department is responsible for administering this assessment and results are reported to the Dubai Government Excellence Program (DGEP). WEM is part of a larger initiative known as Government Excellence Model (GEM) with the aim to foster public sector improvement. GEM is an organizational excellence assessment framework managed by the DGEP. The program's main function is to assess performance across Dubai Government departments. It is in this way that the Dubai government has institutionalized the idea of excellence as a standard of performance it needs to achieve internally, for stakeholders, and for the public that is serves and governs. Dubai Government has 32 departments in total but only 19 departments are being assessed and considered in this research due to the availability of data.

1.2. The Research Problem

In order to discuss the central challenges of constructing composite indicators as they apply to the DSG department's work, the next chapter will first introduce the wider context of the WEM scoring approach and its importance within the GEM initiative. Drawing from Cherchye et al. (2008), the main problem in constructing composite indicators is its reliance on multiple subjective judgments. As the authors elaborate, this is a problem of multiple subjective judgments which when improperly controlled engender biases and errors, which can be used to manipulate results (Cherchye et al., 2008). This then leads to inaccurate and possibly prejudiced conclusions. This outcome is particularly problematic as it undermines the trustworthiness of the whole process of composite indicator construction, which is actually meant to provide a meaningful evaluative understanding of a comparative concept.

In reference to WEM (the DSG's CI), there were a number of subjective choices made by the different parties involved throughout the assessment process. This subjectivity led to a problem of unsatisfied Government departments in the overall scores and the general process of deriving the results. The process does not nurture and support a healthy competition; it does not help the departments to understand their scores nor does it allow them to understand how to most effectively improve their performance. These issues have been identified through the post assessment feedback survey of Government departments. The results of the survey were clearly pointing to a high level of dissatisfaction among many departments, which was the main trigger to scrutinise the current methodology and to design a new one that would better serve its purpose.

1.3. Approach and Justification

Cherchye et al (2008) discuss the successful use of a DEA-based methodology to build composite indicators that help to control for subjective judgments. They note that one of the most useful features of this methodology is its ability to "maximize the overall score for each decision-making unit". In this case study, the application of DEA-based construction of WEM

scores will be tested and discussed. In a DEA-based construction, which is able to produce weights and aggregation at once, there are a number of resolutions to multiple subjective judgments that don't require prior information on normalization of indicators or on approved set of weights.

This case study, then, involves examining DSG's role in envisioning, constructing and using WEM composite indicators to measure the performance of government departments relative to their own websites. The analytical approach is focused on comparing DSG's construction of CI within the framework laid out by Cherchye et al. (2008), on the methodological shortcomings found in DSG's application with those identified by the authors, and finally on the use of DEA to control for subjective judgements in WEM. The DEA was selected to be at the core of the new methodology due to its ability to control subjectivity by allowing flexible weights (specifying the weights in the form of ranges) and therefore allowing the departments being assessed to align the weights better to their intrinsic characteristics and motivations. Another important feature and another reason why DEA was selected as the method of choice for deriving WEM scores is related to the ability of its multiplicative version (called geometric DEA) to deal with complex hierarchical structures frequently encountered within the context of composite indicators.

In this way, this case study will perform a comprehensive assessment of websites in government services. This assessment will provide a useful example on which to model a larger processual transition to smart government, specifically, and government excellence in general, across all Dubai government departments. Based on the analyses of the success and shortcomings of WEM, these results can be used to inform additional smart transformation and government excellence efforts. This case study will therefore examine several aspects of WEM operations including the purpose of measuring excellence and the efficiency of smart transformation, while also identifying the shortcomings of the current methodology used to produce WEM scores. Following this, it will also propose solutions for overcoming the

drawbacks of a methodology that is reliant on subjective judgments. Thus, this study will support Dubai's journey to smart government by contributing to the effort of measuring excellence, in general, and measuring efficiency in smart transformation in particular.

1.4. Contribution and Significance of the Study

In transferring relevant results available in the work of Cherchye et al. (2008) to this research case study, different possibilities and limitations of constructing composite indicators affecting WEM scores, and ultimately the overall Government Excellence Model (GEM) scores will be demonstrated. The significance of this research is that it investigates and highlights the importance of how composite indicators are constructed, specifically with respect to the WEM scores. WEM scores are considered essential in transforming how government departments operate and carry out activities and service delivery on the Internet. A study examining the methodology of WEM scoring is particularly important in the context of Dubai, given the city-stated desire to institutionalize excellence. The findings of this study can be used to understand current knowledge and practices used in constructing WEM scores and be used to design improved methodologies and processes, which will support overcoming the problems mentioned in the above section.

The main practical contribution of this research is to penetrate the area which has so far never been applied in the context of composite indictors. This will be carried out through a clear application on the methodology being applied on a specific issue – to assess website performance. In addition, this study shows how the good qualities of different techniques can be combined through the application of DEA-based methodology to derive composite indicators. The research will also contribute to the literature by further raising awareness among academics and practitioners about frequently overlooked and yet rather superior properties of multiplicative aggregation relative to its additive counterpart, especially in the presence of flexible weights.

In general, the study directs the attention to the inherent difficulties of constructing meaningful composite indicators with respect to WEM and its scoring system. This problem is addressed by identifying and implementing the most suitable DEA-based model for constructing WEM scores, that would effectively inhibit most of the existing sources of subjective elements in the construction process. This will ultimately result in WEM scores that are transparent and trustworthy. Both properties are well-known features of DEA-based methodology. Transparency of the results is achieved by clearly illustrating performance targets, efficient peers and possible improvement paths. Data visualisation techniques could further support understanding of a seemingly complex DEA mechanism and its results through a single interactive dashboard designed for each department assessed. Trustworthiness of the scores is based on the ability of DEA-based models to allow flexible weights between lower and upper bounds. The bounds themselves can be specified or derived by the Government and experts and they ensure that the results are aligned with Governments' objectives while the flexibility of the weights within the bounds ensure that the departments have their say in choosing the set of weights that better reflect their own conditions, aspirations and motivations. Hence, the WEM scores derived through DEA-based methodology will yield the results that are equitable from both the Government's and departments' perspectives.

In this thesis, a comprehensive classification of the key government excellence pillars, specifically smart-government transformation will be put forth. This work will be grounded through an investigation of existing models and theories in order to understand how composite indicators are applied efficiently. What follows is a comparative empirical study of the Dubai public sector's use of composite indicators, with a focus on DSG's role administering and implementing WEM scoring models. These findings will help decision makers in the city to avoid problems and overcome obstacles when they construct composite indicators and assess smart-government transformation in specific as well as overall government performance. The outcome of the research will also help in assessing the current levels of excellence and identify

areas of change, which in turn will help government departments improve current performance levels and deliver better services to the public. Aim and Objectives of the Research

Our aim was to design the new assessment process that will address many of the weaknesses of the current process, which led to the following objectives:

1. *To remove various types of decision biases featuring in the existing WEM model*. The most prominent ones in the current practice are overconfidence in judgment accuracy and the use of a uniform measurement scale for assessing all the sub-indicators. Overconfidence in judgment accuracy ignores intrinsic uncertainties about the value of weights attached to sub-indicators while the use of the uniform scale for all the sub-indicators creates an unnecessary loss of information in the process.

2. *To encourage a balanced performance across different criteria for all the departments*. The current WEM model allows departments to achieve high scores even in the presence of very poor scores on some sub-indicators and can appear as better than some other departments whose performance is reasonably good across all the sub-indicators. This property goes directly against the Dubai Governments' strategic objectives and needs to be either removed or significantly weakened.

3. *To clearly demonstrate the logic and the fairness behind the scores derived*. The fairness of the current WEM model has been disputed many times by the assessed departments. While the main objective for the new WEM framework is to create a fair and equitable scoring system, this may increase the complexity of the logic behind the model. Hence, fine tuning the balance between credibility of the scores and ease of comprehension will be important for a successful practical implementation of the new model.

1.5. Thesis Structure and Chapter Outline

This first chapter of the thesis has discussed the research context that forms the background of this case study. While composite indicators are one of the most accepted ways for governments to internally assess performance, their methodological approach is inherently compromised, allowing for a number of subjective judgments to drive the indicator and sub-indicator selection process. This issue taints the objective of providing standardized, measurable, and comparative understandings of the concept under assessment.

The main emphasis of Chapter 2 is to address the context of the research case study (WEM score) and the main problems in the process of its construction. Chapter 2 also provides an overview of Dubai Government Excellence Program (DGEP), Government Excellence Model (GEM), and Dubai Smart Government (DSG). The former provides a comprehensive comparative framework through which to understand how the latter operates. Chapter 2 will therefore provide an overview of the Dubai Government Excellence Program (DGEP) and explain how it has been established and tasked with helping the city achieve its goals of excellence. Here, the thesis will outline how government departments in Dubai are organized around the broader initiative of improving governance and customer service through communication technology. Then, various DGEP awards with particular focus on the GEM awards will be covered, which will lead us to a discussion of the Smart Government transformation indicator and its assessment method, i.e., the WEM scoring method. Dubai Government's shift to smart government will be also highlighted within the larger context of the UAE's move in this direction, and its ambition to be 'one of the best countries in the world' (UEA.Vision2021.ae, 2019). Dubai's particular desire to further improve governance and the ways in which it sees information and communication technologies as central tools for achieving this objective will be discussed. Moreover, it will be argued that if this is a matter of policy for the government, then an understanding of how it constructs WEM is essential for it to have a more trustworthy and result-driven measure of its progress.

The emphasis of Chapter 3 is to describe the complexity in constructing composite indicators, their objectives, steps, applications and their relevancy to the research problem. Chapter 3 also studies the recent body of literature on the construction of composite indicators. The chapter opens with the discussion of the body of literature debating how to measure and use multidimensional phenomena. A number of social scientists, economists, and statisticians agree that the development of composite indicators is one of the most useful ways to combine diverse measures under a common understanding or index (Cherchye et al., 2008). Following this overview, an illustration on how composite indicators are calculated will be given while drawing attention to the detailed methodological care needed in their construction.

The emphasis of Chapter 4 is to introduce the theoretical background behind the proposed methodology and to explain its features and properties. Having provided this background, Chapter 4 proposes and demonstrates how to use some DEA-based models to produce more reliable and more equitable WEM scores. This chapter will also explore the methodological construction of composite indicators in technical detail including presenting the basic DEA approach with its differing properties, and respective advantages and disadvantages. The chapter will close with an introduction to the Geometric Data Envelopment Analysis (G-DEA), an extended version of DEA, which will be applied to the WEM score in the following chapter.

Having outlined the problems with WEM scoring methodology, followed by a discussion of the theoretical and methodological frameworks behind composite indicators, Chapter 5 will show how to improve the construction process of WEM scores, where G-DEA methodology is used at the core of that process and where the main focus is to achieve the aim and the objectives as defined in the current chapter.

The conclusion will reflect on the case study, followed by a discussion of the learning process. Most importantly, this chapter will pay attention to the practical implementation of the selected methodology, highlighting the limitations and the future research. This chapter will also describe the main contribution of the research and importance of the proposed

methodology for constructing composite indicators to decision makers within the central government, practitioners from the government departments and the assessors who conduct the assessment.

Chapter 2: The Dubai Government Excellence Program

2.1. Introduction

This chapter will introduce the Dubai Government Excellence Program (DGEP) and its main assessment framework, the Government Excellence Model (GEM). It will also present the current assessment methods used for Smart Government Transformation and the WEM score, which are discussed in detail, along with a consideration of their limitations.

Federal and local government departments in the United Arab Emirates are undergoing tremendous change due to comprehensive transformations towards smart government. This initiative refers to the use of the latest ICTs (smart application/mobile phones) by government agencies to support government operations, activities and services they provide to their customers, businesses, and other stakeholders. Technological innovation has brought the world into the homes of people and has increased the transfer of information and ideas. Most significantly, it has changed people's expectations and desires. Smart government is considered as one of the latest revolutions dramatically changing the way government departments communicate and interact with their stakeholders through developing innovative processes and providing high quality public service.

The transformation to smart government specifically in Dubai is of significant importance to the UAE Vice-President and Prime Minister and Ruler of Dubai HH Sheikh Mohammed Bin Rashid Al Maktoum. Through his book *My Vision,* he has contended strongly on institutionalising excellence and developing programs for upgrading the performance and services of Dubai's government. Excellence in government refers to the outstanding practices in organizations to achieve and sustain superior levels of performance in order to meet or exceed the expectations of all their stakeholders (Adaep.ae, 2019). Relative to this continuous improvement and sustainability of high performing results, Sheikh Mohammed has stated that, "the best way to maintain excellence is to develop it into a social conduct, so that it becomes an integral part of people's behaviour and psyche" (Al Maktoum, 2006). He also stated that, "We

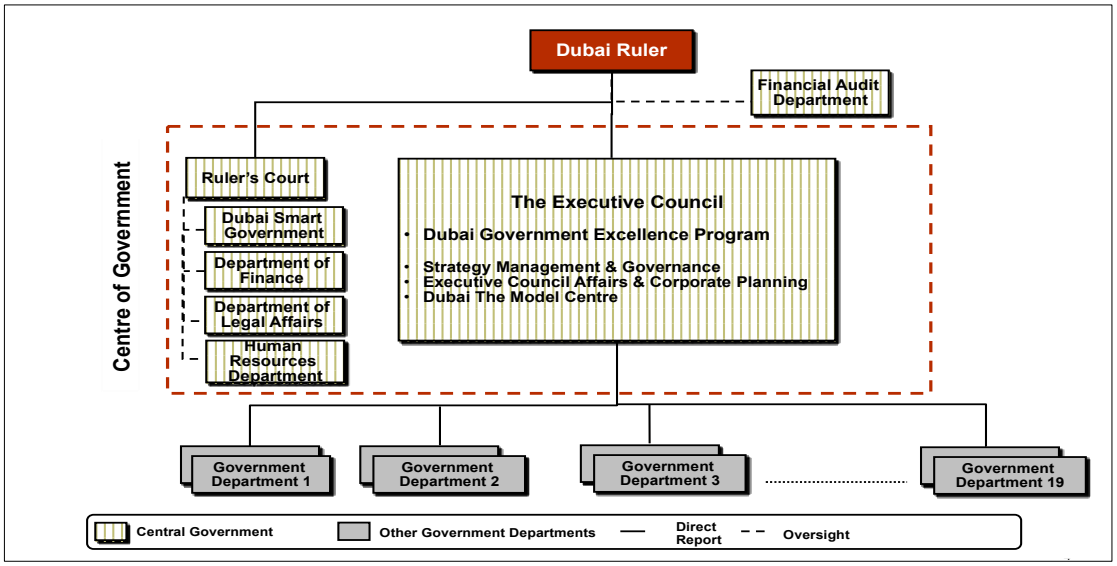
cannot possibly stop the global race for excellence – we must join it". One main aspect of His Highness' vision is to achieve the society's wellbeing and happiness. There is a strong need to transform to smart government to reach every single individual and making it a two-way communication between government and people to ensure that the population's needs are taken care of 24 hours a day, 7 days a week and 365 days a year.

Dubai as a city is aspiring to improve the performance of its government sector. This will be achieved through compliance with the latest developments and technologies in all fields. Enhancement in the Government's capacity to implement modern administrative principles is based on customer satisfaction, resource development, procedures simplification, systems documentation, innovation encouragement and capability development (Dubaiplan2021.ae, 2019). Furthermore, Dubai seeks to build a sound working environment by motivating and supporting government departments to adopt strategies inducing comprehensive development, efficient servicing of the business community, excellent investment conditions, support to the private sector and promotion of free entrepreneurship.

Pursuant to what precedes, HH Sheikh Mohammed Bin Rashid Al Maktoum ordered the establishment of the Dubai Government Excellence Program (DGEP) in 1997 to develop the government sector and improve its performance and services through honouring awards, incentives, motivational working environment, constructive cooperation and positive competition (Al-Maktoum, 1997). Sheikh Mohammed strived to pursue a vision of attaining the highest levels of excellence in government performance in Dubai. It is construed that the focus of the program would not only be on improving the management of government departments, but would also provide rewards to departments, teams and individuals based on their contribution to the improvement of government services. In this way, departments are encouraged to set clearly defined and well communicated objectives, with measurement systems and public reporting of the results. These activities are centrally anchored on their interaction with, and improved satisfaction of, their customers (Al-Maktoum, 2006).

I am currently working within the Dubai Government Excellence Program (DGEP), which is part of The Executive Council (TEC) of Dubai Government - the governing 'board'/cabinet of the Local Government. TEC is part of the Dubai central government department along with five other departments that reside under the Ruler's Court HH Sheikh Mohammed bin Rashid Al Maktoum, as shown in Figure 1. The DGEP's main role is to assess departments' performance of which 19 will be considered in this thesis.

Figure 1: Dubai Government Map



(Adapted from the Dubai Executive Council, 2009)

2.2. DGEP Awards Categories

DGEP offers both honours and financial awards to government departments, divisions and teams as well as to government employees fulfilling a specific set of assessment criteria. DGEP awards are generally classified into three different categories and 22 awards that take into account the diversity of the work nature of government departments, to meet the Dubai Government Excellence Program's objectives and to help improve the performance of governmental sectors to achieve leadership in all fields.

The three categories of DGEP Awards include organizational excellence known as the Government Excellence Model (GEM), employee excellence known as Dubai Medals for Government Excellence and independent awards. These categories are regularly measured

during DGEP annual assessment cycle and consist of 8, 11 and 3 awards respectively. The section below looks in detail at the Government Excellence Model category. This specific category contains an award supporting government smart transformation, which includes the Website Excellence Model (WEM) score as one of its main key performance indicators.

2.3. Government Excellence Model (GEM)

On the 19th of April 2016, His Highness mandated the implementation of the Government Excellence Model (GEM) in Dubai Government as well as the assessment process applied at the UAE federal level. The GEM was previously launched at the federal level on the 7th of March 2015, representing innovative management trend with universally Emirati content.

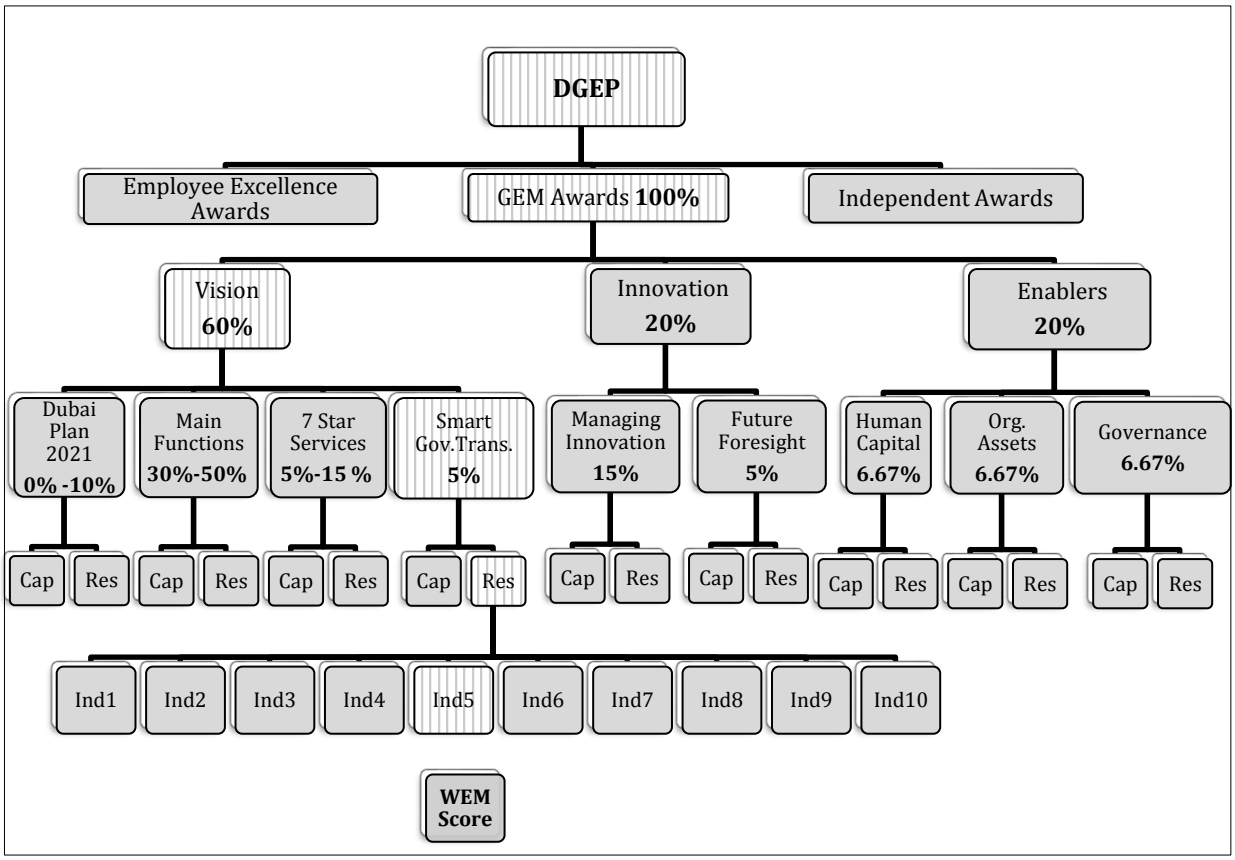
The GEM was developed to serve the ambitious vision of the UAE, which is to be amongst the best countries in the world by 2021 (UAE vision2021, 2010). It is considered a new way of thinking in planning, implementing and developing government services and operations. This new way is based on innovative principles and concepts which have been applied within the UAE Government that have proven their effectiveness in achieving pioneering results.

The GEM has been adopted as a basis for assessing the participating government departments in the Dubai Government Excellence Award and serves as an extension of the development journey launched over 20 years within the Government of Dubai and the UAE Government. The aim is to ensure that all Dubai Government departments reach a maturity level beyond excellence while focusing on performance leadership, innovation and smart transformation.

The GEM was designed in cooperation with key central government departments such as Department of Finance (DoF), Department of Human Resources (DoHR), Financial Audit Department (FAD) and Dubai Smart Government (DSG). This collaboration includes assessing Dubai Government departments in their respective area of specialization and provides results (scores) on specific key indicators measured across the Government.

The structure of the GEM consists of three main pillars: Vision, Innovation and Enablers, as shown in Figure 2. The first pillar, Vision, consists of four main indicators: Dubai Plan 2021, Main Functions, 7 Star Services and Smart Government Transformation.

Figure 2: DGEP awards category



The Smart Government Transformation indicator encompasses the Website Excellence Model (WEM) score as one of its results' sub-indicators as highlighted in Figure 2. These represent the core operations through which government departments work on achieving their vision, strategic objectives and the objectives of the Dubai plan 2021. The second pillar, Innovation, consists of two main indicators: Managing Innovation and Future Foresight. The third pillar, Enablers, consists of three main indicators: Human Capital, Organizational Assets and Governance. In total there are nine main indicators; each of which is further split into two main sections: capabilities sub-indicators and results sub-indicators. The capabilities sub-indicators reflect all the efforts undertaken by the departments to achieve their goals and improve their processes, services, policies and projects. The results sub-indicators reflect the level of

achieved targets by the departments. The GEM has been designed in a way that guarantees a direct link between capabilities and relevant results of the same main indicator. The way in which GEM assessment process works and how it's derived is explained next.

2.4. The Process of Deriving GEM Scores Using the DGEP Approach

The GEM score is derived using the bottom level sub-indicators that belong to capabilities and results. The weight assigned to the capabilities is 30% of the total weight of the main indicator while the remaining weight of 70% is assigned to the results. DGEP did not assign a specific weight for each of the bottom level sub-indicators. During the assessment, the assessor looks at both, the capabilities' and the results', sub-indicators in general without giving a score to each sub-indicator but rather giving an overall score for the capabilities and results using the GEM assessment tool, which is illustrated in Table 1.

Table 1: GEM assessment tool

Capabilities [30%]	Weights	Description*
Effectiveness	60%	Do capabilities meet the needs of all stakeholders and contribute to achieving the strategy? Are capabilities suitable to the entity's nature of work? Do they conform to international best practices?
Efficiency	20%	Are capabilities implemented in ways that ensure optimal utilization of various resources and rational spending?
Learning & Development	20%	Are capabilities improved using creative ideas and innovative methods, based on analysis and learning from performance results and best practices?
Results [70%]	Weights	Description*
Comprehensiveness & Usability	50%	Are all the appropriate indicators to monitor, understand and forecast the performance of the capabilities and level of success in achieving the strategic plan measured?
Achievement of Targets	20%	Are the specified targets sound and ambitious? Were the targets achieved?
Performance Improvement	20%	Is the learning and development process in the entity effective?
Leading Position	10%	Have the results that have been achieved helped Dubai and the UAE in reaching a leading position worldwide?

* Descriptions support the assessors to understand the departments' performance relative to each factor.

The assessment tool for the capabilities sub-indicators focuses on three different factors: effectiveness, efficiency and learning & development. The results sub-indicators are assessed on four factors: comprehensiveness & usability, achievement of targets, performance improvement and leading position. The weights assigned to each factor in the capabilities and results assessment tools are also shown in the table.

It should be noted that the results sub-indicators are measured in two distinct ways. Some sub-indicators are measured by the departments under assessment, which are in turn assessed and scored by one of the assessment teams using the GEM assessment tool. The remaining sub-indicators are assessed and scored by a central government department according to their own assessment tool. The scores of the central government department assessment are provided then to the concerned assessment team for their assessment using the GEM assessment tool. The WEM score is one of the sub-indicators that is assessed and scored by the central government department. As shown in Figure 2, the WEM score, which is the main focus of this thesis, is one of the 10 results' sub-indicators that belong to the Smart Government Transformation main indicator. The full list of those sub-indicators, referred to as Ind1 to Ind10 in Figure 2, is provided in Appendix A.

After each department is assessed on each factor, the weighted sum of scores, using the weights shown in Table 1, is calculated to get the overall score for capabilities section and the overall score for results section. The final score with respect to the corresponding main indicator is then obtained as the weighted sum of the overall capabilities' score and the overall results' scores, using aforementioned weights of 30% and 70%, respectively.

Each government department is assessed by one of four different assessment teams that integrate their assessment findings and scores in one feedback report. The assessment teams include: Subject Matter Experts (SMEs) Team, Smart Government Transformation Team, Managing Innovation Team and Human Capital Team. The team members are experts specialized in government department's work, smart government transformation,

organizational innovation, and human resources, respectively. They possess knowledge and experience in the international best practices in their relevant fields. Each assessment team consists of two to four experts. The number of experts is determined according to the diversity and the nature of the department's work and its size. The assessment of the GEM is distributed among the four assessment teams as shown in Table 2.

Table 2: Distribution of GEM assessment teams over the nine main indicators

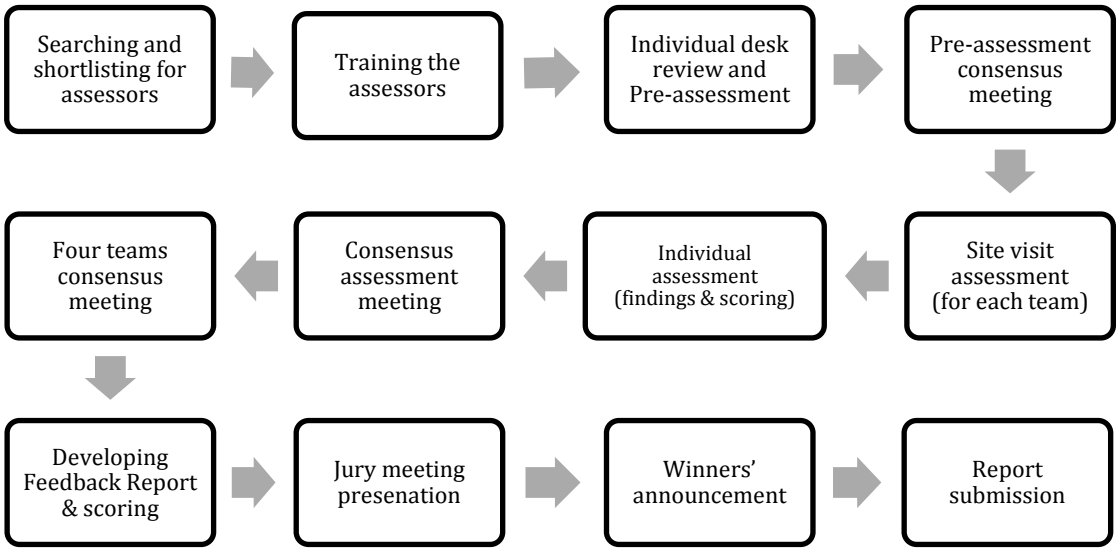
Main Indicator	Weight	Assessment Team			
		SMEs	Smart Government	Managing Innovation	Human Capital
Dubai plan 2021	55%	✓			
Main Functions					
7 Star Services					
Smart Gov. Trans.	5%		✓		
Managing Innovation	15%			✓	
Future Foresight	5%	✓			
Human Capital	6.67%				✓
Org. Assets	6.67%	✓			
Governance	6.67%	✓			
Total Weight	100%	73.33%	5%	15%	6.67%

DGEP assigns weights for the GEM's main pillars (60% for Vision, 20% for Innovation, 20% for Enablers) and for the six of the nine main indicators (5% for Smart Government Transformation, 15% for Managing Innovation, 5% for Future Foresight, and equal weights of 6.67% to Human Capital, Organizational Assets and Governance). The allocation of weights was based on the government vision and leadership direction set for the government. The weights allocated to the first three main indicators under the Vision pillar are shown in ranges (non-fixed weights) due to the fact that different departments require different weights to be assigned to those three main indicators. Their total weight is worth 55% while the remaining 5% is fixed for the smart transformation indicator, as shown in Table 2. The weight of 10% for Dubai Plan 2021 is applicable only for the departments that have been assigned key performance indicators (KPIs) for Dubai Plan 2021. For those that are not assigned any such KPIs, those 10% are transferred to the Main Functions indicator. Similarly, the weight for 7

Star Services indicator is set at 5% for the departments that are less service oriented, 10% for the departments that have medium service orientation and 15% for those that have high service orientation. The difference between 15% and the actual weight assigned for 7 Star Services is then also transferred to the Main Function indicator.

In the allocation of weights, the assessed government departments are not consulted at all. With the weights and aggregation method (weighted sum) already being pre-set by the DGEP, the only missing components are the scores relating to all the results' and capabilities' factors for each of the nine main indicators. As we will see later on, something similar happens when deriving the WEM scores and many improvements that we will make in that context will be also applicable at this higher level where the overall scores for GEM awards are constructed. As for getting the scores relating to all the results' and capabilities' factors for each of the nine main indicators, the four assessment teams, distributed according to Table 2 and using the GEM assessment tool shown in Table 1, are given that task, which in itself is a very long and intensive process, as illustrated in Figure 3.

Figure 3: DGEP assessment process



(Adapted from DGEP)

The assessment process starts with the search for assessors as per the categories mentioned earlier (SMEs Team, Smart Government Transformation Team, Managing Innovation Team, and Human Capital Team) and ends with the presentation of the final scores, winners' announcement and reports submission to the government departments. This long process takes approximately up to five months. The lengthiest and the costliest part of the process starts from training the assessors until the jury meeting presentation. That part of the process is referred to as "assessment cycle".

The search for international assessors is based on their background and expertise in the relevant required categories. Once their resume is reviewed, DGEP shortlists the most suitable assessors for GEM evaluation and classifies them into the four assessment teams as illustrated in Table 2. A comprehensive two-day training module is conducted for all the assessors to ensure that they are aware of GEM's assessment methodology and indicators. Each member of the four teams is tasked with pre-assessing each government department individually and respective to their scope of assessment. This is done through conducting an on-desk assessment initially by reviewing the departments' submission form. Once the individual assessment is completed, each assessment team reviews the individual findings and consensually agrees on the key business factors - the issues that need to be looked into and agreed upon during the subsequent on-site visit assessment.

The next step involves each assessment team conducting on-site visit assessment according to a pre-agreed schedule between the teams. The SMEs Team conducts the visit for 2-4 days while the other three teams conduct the visit for 1 day only. The on-site visit assessment schedule is synchronized to ensure that the Smart Government, Innovation and Human Capital Teams finish their on-site assessment during the 2-4 days which is the site visit assessment duration of the SMEs Team. Once the on-site visit is completed, each member of each of the four teams is tasked to assess the department individually and provide the relevant scores for each department visited. Each assessment team then conducts a consensus assessment meeting

wherein each team member agrees on the main findings of the assessment site visit and on the set of final scores for all the indicators assessed. If three assessors have not reached a consensus and there were huge variations in the scoring, the team leader gets involved in agreeing on the final score with the respective team members.

The next step involves each team developing a final draft of the assessment feedback report, which is then reviewed by the team leader. The team leaders ensure the harmonization of scoring, feedback report format and style among all assessment teams. This report is then presented to the jury members for final review and endorsement. The technical jury meeting is held to discuss the outcome of the assessment for each department with the following objectives:

- Ensure the overall integrity of the assessment process wherein the assessment is held according to the set methodology and assessment tool.

- Ensure that the assessment is taking into consideration the specifics and the key business factors of the government department.

The jury endorses the final scoring list and recommendation for the winners within different categories. In addition, the jury members develop the final jury report, which includes the recommendations on the improvement of the assessment process. The final report is reviewed by a DGEP team for a final check on the feedback consistency among all the participating departments. The final assessment feedback reports are then submitted to the government departments following the DGEP Award Ceremony and winners' announcement.

2.5. Smart Government Transformation (SGT) Indicator

Of particular importance in motivating the efforts of smart government initiatives in Dubai government departments is an organizational excellence award within the GEM known as the "Distinguished Smart Government Department". This award corresponds to the fourth main

indicator, Smart Government Transformation (SGT), which is one of the four main indicators of the Vision pillar, as illustrated in Figure 2. SGT indicator focuses on the daily operation of government departments, the establishment of contacts and the provision of services and information to the customers through the Internet.

Just as all the other main indicators, SGT is divided into two main sections: Capabilities and Results. Under both sections, there are a number of sub-indicators. The results section is spilt into 10 sub-indicators (Ind1 – Ind10) as shown in Figure 2, one of which is the WEM score – the main focus in this study.

Once each department is assessed on all the sub-indicators, the overall SGT score is derived, which is used by the DGEP to award the best department, which then becomes the Distinguished Smart Government Department. This award and the title are given once every two years to ensure that all the departments have sufficient time to enhance and improve their activities, based on the evaluation and feedback provided following the assessment cycle. The assessment of the SGT indicator is conducted in cooperation with the Dubai Smart Government (DSG), as a central government department, which is leading the effort of smart government transformation. Accordingly, DGEP ensures that DSG's efforts are met through the assessment cycle.

DSG participates effectively during DGEP assessment cycles in assessing each department on each of the 10 sub-indicators of the SGT main indicator. These 10 sub-indicators reflect the results of smart transformation of all the government departments in Dubai. One of these sub-indicators (Ind5), known as Website Excellence Model (WEM), serves a largely heterogeneous population that comprises of users with vastly different learning styles and expertise levels. Accordingly, DSG is committed to collaborating with government departments in providing quality websites to individuals and businesses while it also provides DGEP with all the departments' WEM scores. This collaboration is shared in the ongoing journey towards excellence in Dubai Government's smart transformation efforts in alignment with the

directives of His Highness Sheikh Mohammed Bin Rashid Al-Maktoum. The focus of this study is the derivation of the WEM score since this score itself is considered to be a rather complex composite indicator with an already complex hierarchy structure of GEM. WEM score is an aggregate score of many scores obtained by direct evaluation of departments relative to the sub-indicators from the bottom level of the WEM hierarchy. This level of complexity is not seen in any other sub-indicator of the SGT. Using WEM score as the main focus of this study will therefore help us in selecting the most appropriate method and tools for constructing this composite indicator but it will also help us to ultimately generalise this new approach on the other components of SGT in specific as well as the other components of GEM in general.

Having this in mind, it is important to state that the current DSG approach is not founded on the best methodologies available with some apparent limitations. Full details of the current approach with a special emphasis on the most problematic aspects of that approach is the main subject of the next section.

2.6. The Process of Deriving WEM scores using the DSG Approach

As Part of the GEM assessment process, 19 Government departments' websites were evaluated. This website evaluation was the first assessment based on the revised sub-indicators published by Dubai Smart Government Department in 2011, i.e., the Website Excellence Model (WEM). The results of this evaluation represent the baseline for Dubai Government websites for several upcoming biennial evaluations. The Dubai Smart Government introduced the Website Excellence Model to formulate government-wide guidelines to be adopted by Dubai Government departments in their websites. It aims to achieve comprehensive maturity in Government websites in line with international best practices and standards. The purpose of this indicator is to ensure that Dubai Government websites are customer focused, accessible, well-designed and usable, have appropriate content and policies, achieve high levels of website usage and finally lead to high levels of customer satisfaction. These specific requirements for government websites have been articulated in several surveys run by DSG. The results clearly

indicate that most users visit government websites to seek information and to benefit from services provided via the websites. If ease of access, usability and the required content of a particular website are in a poor condition, users might not return to the website again. In conclusion, WEM intends to ensure that government websites efficiently serve customers, effectively retain existing users and acquire new ones. Therefore, DSG started evaluating government websites quality based on their compliance with WEM through its four main sub-indicators: Accessibility, Usability & Design, Content and Policy.

Accessibility: this sub-indicator measures whether the website serves a large heterogeneous population that comprises users with vastly different learning styles and capability levels. Throughout all website development phases starting from concept development implementation, website designers must strategically keep in mind all possible access barriers. This is to create a government website that is inclusive and accessible to the widest possible audience.

Usability & Design: this sub-indicator intends to measure whether the website conveys to all users a single and unified message. It is necessary that a single brand identity is implemented to a certain extent and promoted by the relevant government departments through their respective communication channels. Implementing seamless and usable website design through the use of logos, taglines, colour palettes and uniform templates adds to the user experience when browsing any page of the department's website.

Content: this sub-indicator measures whether the required content is available on the website to meet users' needs and expectations. This is a very important element in ensuring the success of a website and the use of it. Though the control of content to be included in the website is left to the respective government entity, it needs to be current, accurate, relevant and easy to read to ensure future return of the users to the website.

Policy: Since the information published on government websites has legitimate implications and concerns for the department itself and users, this sub-indicator intends to measure whether clear and unambiguous policies are stated for the users accessing the website. Such policies should address issues related to data protection, accessibility and responsibility among others on the websites.

All those indicators intend to provide the necessary controls in achieving two main objectives: high-levels of customer satisfaction and high-levels of website usage. These objectives are considered through three different elements of the overall website evaluation, wherein the WEM score itself represents 35% of the total score and it is related to how the organization views itself. The other two elements represent customer satisfaction (50% of the total score), which looks at how customers view the organization, and website usage (15% of the total score), which addresses the usage statistics of the organization's services. These scores are provided separately to DGEP, which represent the sub-indicators Ind5, Ind6 and Ind7, respectively. The overall website evaluation score is for DSG use only. The three elements with their scores and the split of the WEM score into its four main sub-indicators and their weights are shown in Table 3.

Table 3: Weights Allocation for Website Overall Evaluation

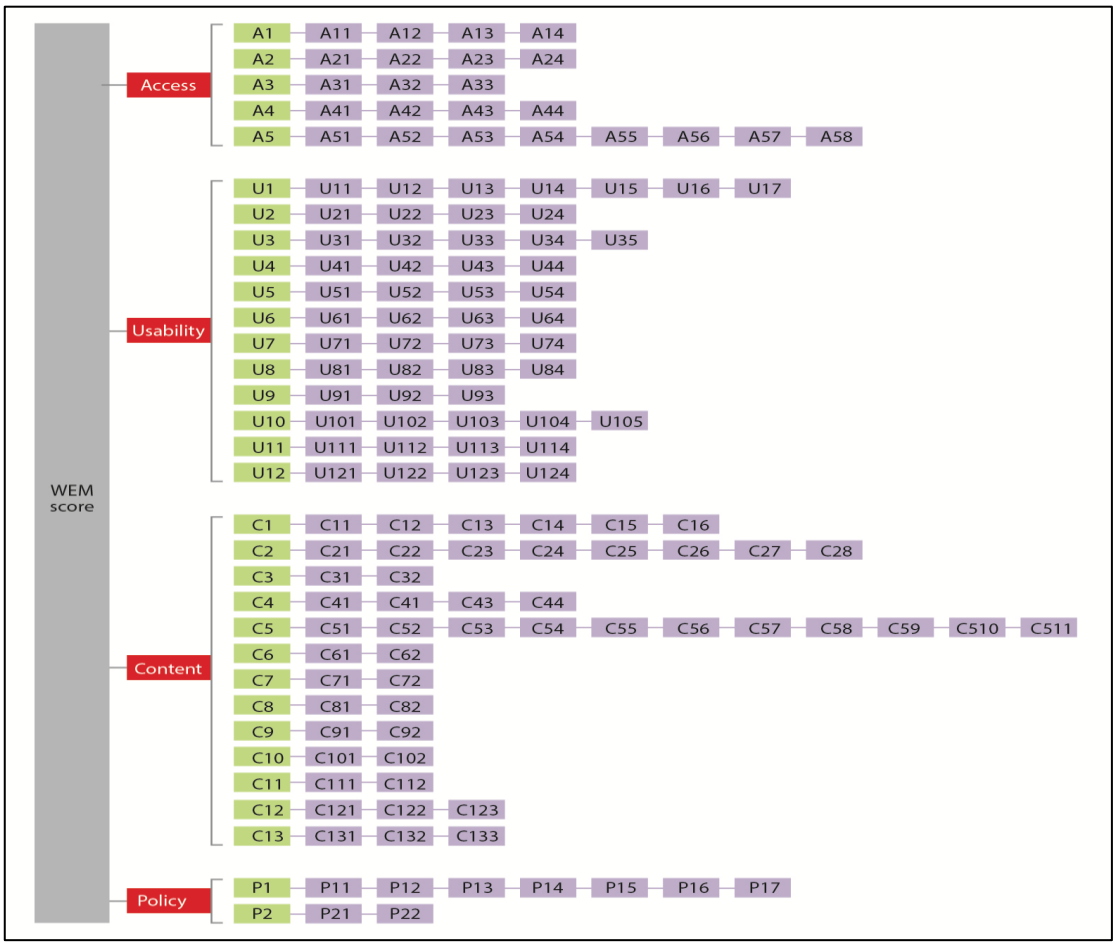
ELEMENTS	WEIGHTS
Customer Satisfaction	50%
Website Excellence Model (WEM Score)	35%
Website Usage	15%
TOTAL WEIGHT	100%

Accessibility	8%
Usability & Design	12%
Content	10%
Policies	5%

The WEM score is a composite indicator that summarizes in a single number (through the weighted sum model) the value derived from the website – this score is calculated for each

government department in Dubai. The complexity of WEM score does not stop with the four main sub-indicators. All of those main sub-indicators are further split into their own sub-indicators, all of which are then also split into their sub-indicators, which finally make the complete decomposition of the WEM score. In other words, DSG constructs WEM scores using a hierarchy structure consisting of 4 levels with the 4 indicators on the 2nd level (described above), the total of 32 indicators on the 3rd level and a massive number of 133 indicators on the 4th level. This is illustrated in Figure 4 where, for the sake of brevity, the indicators on the 3rd and the 4th level are not named. Also, the "Usability & Design" indictor is renamed into "Usability" only, but we will keep in mind that it stands for both: usability and design.

Figure 4: WEM score hierarchy structure



Each sub-indicator score is derived from a set of underlying indicators, which are structured into a hierarchy for ease of the analysis. Using these indicators, DSG conducts the evaluation

on a biennial basis and according to a predefined scoring methodology. The scoring methodology itself is at the heart of our investigation and this is what we are going to focus on in our analysis. The design is fairly straightforward; it is a one-way bottom up "causality stream" between a group of sub-indicators and their respective indicator in the upper level. This is due to each lower level feeding into the upper level.

The methodology adopted for assessing the WEM score is explained in the following sections using Figure 4 as background. DSG follows three main stages in the process of assessing government websites and deriving WEM scores for each, which are pre-assessment, assessment and post assessment stage. DSG's assessment process is very similar to DGEP's assessment process in general wherein it starts with searching for assessors and ends with the awarding the winners and submitting feedback reports to the government departments as illustrated in Figure 3.

The remaining part of this chapter provides full details of the steps taken within each of the three stages of the process: pre-assessment, assessment and post-assessment. These steps are important to understand since it is exactly through those steps where we will first identify problematic aspects of the process and later on look for ways to address those problems by providing alternative methods that would better fit the purpose.

2.6.1 Pre-assessment Stage

This stage involves steps that will set the foundation of the assessment cycle and its objectives. It is important to note that the DSG's role in this stage is not as innocent as it may appear. While the assessors will be expected to score the departments on all the indicators from the bottom level of the WEM hierarchical structure in the assessment stage, it is the DSG who determines the aggregation method as well as the relative importance (weights) of all the indicators in the hierarchical structure. Just like it was the case with the GEM Awards hierarchy, neither the assessors nor the departments being assessed have any control on the weights and aggregation method as both of these are fully set by the DGEP in case of GEM Awards scores and by DSG in

case of WEM scores. The details of these two important aspects of the WEM model are described below.

WEM score weighting method

For simplicity, DSG's experts could have chosen to apply equal weights throughout the WEM score structure. However, after much consideration, and bearing in mind their experience with the government departments, they decided to give extra weight to the components that were deemed to be more important than some others. The decision was that the largest weight is to be given to the indicators reflecting the usability and design of the website. The indicator of using the appropriate content in the website is considered as the second most important element, followed by the indicator of customer accessibility. The least important indicator was the policy for the users accessing the website. Assigning weights for all indicators was a mutual effort agreed upon by all the DSG's internal experts. First, the total weight of 35% allocated to the WEM score as part of the overall website evaluation model (see Table 3) is decomposed into the weights of the four main WEM indicators, which are then normalised and finally rounded. Table 4 shows this decomposition as well as the number of the sub-indicators at the next third level of the WEM hierarchy.

Table 4: Weights, normalised weights and number of sub-indicators for the 4 main WEM indicators.

Indicators	Weights	Weights (normalised & rounded)	Number of sub-indicators
Accessibility	8%	23%	5
Usability & Design	12%	34%	12
Content	10%	29%	13
Policies	5%	14%	2
Totals	35%	100%	32

Notice already here how the original weights for the four main indicators (8%, 12%, 10% and 5%) happen to be like this only because the total sum was required to be 35% as per yet another subjective judgment (the percentage split among Ind5, Ind6 and Ind7). If the total was to be 100% to begin with, we would almost certainly end up with the set of weights that is different from the one obtained through normalisation and rounding. While these may be

minor details with relatively minor effects on the overall scores, it is the principle of the approach, which is problematic. Namely, it is clear that experts' knowledge cannot yield such a high level of precision that would generate crisp weight of 23%, 34%, 29% and 14% for the four main indicators.

For level 3 and level 4 indicators, the weights were assigned following the same approach. However, it needs to be noted that there were some sub-indicators from level 3 that were not assessed during the last assessment cycle due to their apparent inapplicability for the government departments. In those cases, the weight that should have been allocated to an irrelevant sub-indicator is distributed amongst the remaining sub-indicators. Table 5 illustrates this scenario for the level 3 indicators that fall under the Accessibility main indicator. The full list of weights for the level 3 and level 4 sub-indicators, as set by the DSG, can be found in Appendix B and Appendix C.

Table 5: Level 3 weights distribution

Indicator	Sub-indicator		Original Weights	New* Weights
Accessibility	A1	Provide Access to the Website Through an Easy to Remember URL including an Appropriate Representation of the department Name under. Gov.ae domain	4%	4.8%
	A2	Provide a Quick Access to the Website from a Search Engine	4%	4.8%
	A3	Provide Access to the Website with Identical and Consistent Results through a Wide Range of Web Browsers	4%	4.8%
	A4	Provide a Functional Bilingual Website	5%	5.8%
	A5	Provide Appropriate Access to Website Files	2%	2.8%
	A6	Provide Access to the Website for People with Disabilities	4%	-
Total			23%	23%

* The final set of weights after the weight of the un-assessed sub-indicator (A6) is equally distributed across all the remaining sub-indicators.

It was not clear at all how exactly the weights assigned by the DSG were determined, starting from the ones allocated to Ind5, Ind6 and Ind7 all the way down to the allocation of weights to the indicators in levels 3 and 4. After further clarification with their experts, it became clear that a rather subjective process has been apparently followed, which is very similar to the

method knows as the Budget Allocation Process (BAP). This method along with its strengths and weaknesses will be discussed in the next chapter.

WEM score aggregation method

The aggregation approach typically depends on the nature and the context of the case in hand. In the case of DSG assessment, the DSG team is using simple weighted sum model, or what is also known as additive aggregation model. To illustrate potential problems of additive aggregation, let us focus only on Usability & Design and Content as if these two were the only main indicators of WEM score. We will further simplify the matters by assuming that these two indicators are equally weighted and we will look only at three different sets of hypothetical scores representing departments A, B and C. Suppose that department A is performing decently well with respect to both indicators, department B has an excellent Content but very poor Usability & Design while department C has an excellent Usability & Design but very poor Content. Table 6 shows a specific set of scores that may reflect one such a situation. As it can be observed, even though the overall aggregated scores are exactly the same for all three departments, it stands to reason that the performance of the department A should be much more preferred than those exhibited by departments B and C. This was also subsequently confirmed by the DSG team. After all, what is the use a very good content if that content is not usable? Also, what is the use of a very good web-site design and usability if its content is poor?

Table 6: An illustration of problems with the current DSG additive aggregation

Indicator	Department A	Department B	Department C
Content	50%	90%	10%
Usability & design	50%	10%	90%
Overall Score	50%	50%	50%

Although DSG aims to improve the performance of departments with respect to all WEM indicators, they applied an inappropriate aggregation method for aggregating the WEM score. Weighted sum method is a very widely used method in the evaluation process even though it has many limitations. It is apparent from Table 6 that this type of aggregation (additive) does

not give attention to the unbalanced performance of the website indicators, as the overall scores will be the same in the end.

In addition, if we allow some degree of flexibility in the weights assigned to the two indicators, the situation becomes even worse. For example, we could allow departments to choose the weights for the two indicators to be anywhere between 40% and 60% while making sure that the sum of the two is equal to 100%. Such a flexibility may indeed be required so to avoid dealing with an absurd and imaginary level of precision granted to experts' subjective judgments. In this case, department A will still have the overall score of 50% while departments B and C can push up their scores to 58% by attaching more weight to the criterion where they perform better. These examples illustrate why additive aggregation may not be the best fit for the nature of the case study here. More details on the different types of aggregation will be presented in the following Chapter.

2.6.2 Assessment stage

The process in this stage involves the assessors applying the measurement scale set by DSG experts to assess the WEM score indicators. It includes the following steps: the individual assessment, the assessors' consensus meeting and the team leader review. Prior to elaborating on the measurement scale applied by DSG, a brief description of the assessment steps is provided as follows:

- Individual assessment of department websites: Each assessor assesses 19 government department websites individually and assigns scores for each using the WEM score methodology.

- Assessors' consensus meeting: Assessors meet and review their scores and collectively agree on awarding the final score to each department website. If both assessors have not reached a consensus, the team leader gets involved to agree on the final score.

- Team leader review: The team leader reviews the assessors' final scores and does a random check on a few websites to verify the assigned results. If any discrepancy arises, the team leader requests correction from the assessors.

The main issues in the whole assessment stage process are therefore related to assigning scores and reaching consensus on the scores assigned. DSG once again is not an innocent party here since it is DSG who requires the use of the same measurement scale for all WEM's end-indicators (the sub-indicators in level 4 of the hierarchy highlighted in Figure 4). The score for each department relative to each end-indicator is assigned from a 4-point rating scale to determine the compliance of the department with the end-indicator. That is, in case of zero-compliance (guideline not implemented or not available) then a 0 score is assigned to that sub-indicator. For partial compliance, a score of 1 or 2 is assigned, where a score of 1 represents a 33% compliance completion, and a score of 2 represents a 66% compliance completion. For full compliance (100%) a score of 3 is awarded.

DSG experts stated that all 133 end-indicators can be evaluated subjectively using the four-point scale despite the fact that there are many end-indicators that would be more appropriately measured using different scales depending on the nature of the indicator. The reason they opted for this approach is due to the convenience and ease of application across all the end-indicators. Applying the same approach (subjective judgments) across all end-indicators would allow them to avoid a tedious and problematic task of normalising scores. While this may be a valid concern, using the same measurement scale for all the end-indicators does introduce a significant amount of unnecessary friction and imprecision in the process of getting the final set of scores. We will look at these details in chapter 5 but, for now, it will suffice to say that such a uniform scale of measurement is the main cause for the potential loss of information and for a too lengthy process of reaching the consensus. Using more natural scale of measurement for each criterion, which better fits the nature of the criterion measured as well as the subjective nature of the assessors, would be desirable.

2.6.3 Post Assessment stage

The assessors along with the team leader present to the DSG team the assigned scores of the 19 government department websites for approval. DSG submit the final departments' scores to DGEP including the ranking of each department to be incorporated within GEM's final score. These scores will help the GEM's assessors in assessing the smart government transformation indicator and also identify the winner in the category that represents this indicator entitled "distinguished smart government department".

This stage also involves presenting the final results of WEM scores for all departments in a lengthy report titled "Dubai Government Websites Evaluation Report". The detailed report presents each of the four main WEM indicators through three main sections: the respective sub-indicators average scores, the sub-indicators' compliance percentage, and all government departments' scores for that particular main indicator. While the report itself is produced with care and presented clearly, there is hardly any transparency in terms of how different scores were obtained. This has caused a lot of friction and complaints in the past that could have been avoided if the results were to be more transparent and more encouraging for cooperation between different departments.

2.7. Summary

This chapter discussed the Dubai Government Excellence Program (DGEP) and explained its importance in making Dubai a centre of excellence. The different DGEP awards were explained, with particular focus on the GEM awards. This led on to a discussion of the Smart Government transformation indicator and its assessment method, i.e., the WEM scoring method. The three stages of the scoring process methodology were described in details and the main problems were briefly mentioned. Using this information as background, the next chapter will discuss different analytical tools and theoretical background of the methodology.

Chapter 3: The Construction of a Composite Indicator

3.1. Introduction

Over the last several years, there has been some debate over how the measurement of multidimensional phenomena can be used to assess the performance of countries. The multidimensional combination can be obtained by applying methodology known as composite indicators (CIs) (Salzman, 2003; Mazziotta and Pareto, 2013). CIs are extensively used tools to assess and compare countries' performance in multiple fields such as economy, society, environment and technology performance. Recently, composite indicators have been effectively utilized as tools for policy analysis and public communication (Cherchye et al., 2007). CIs are useful in identifying trends and drawing attention to particular issues. They can also be helpful in setting policy priorities and in benchmarking or monitoring performance (Saltelli, 2007).

According to the Organization for Economic Co-operation and Development's (OECD) statistical definition, a CI is formed when individual indicators are compiled into a single index on the basis of an underlying model of a multi-dimensional concept that is being measured. Examples of the well-known CIs include Technological Achievement Index (TAI), Human Development Index (HDI), Environmental Performance Index (EPI), Website Index (WI) and Global Innovation Index (GII). All of the above-mentioned fields are very general by definition and for each area there are various individual indicators and sub-indicators that provide information on how a country or an organisation is performing. The aggregation of all available indicators and sub-indicators in any field leads to the development of composite indicators. The Organization for Economic Co-operation and Development (OECD, 2008) has clearly identified a sequence of ten steps on how to design, develop and construct composite indicators that will support decision makers in improving the quality of the intended outcomes. These steps involve a theoretical framework, selection of the data, imputation of missing data, multivariate analysis, normalization, weighting and aggregation techniques, uncertainty and

sensitivity analysis, back to the data, links to other indicators and finally visualization of the results. The first two steps are interconnected. The choice of the data in the second step depends strictly on the selected indicators and the phenomena defined by the theoretical framework. Moreover, in order to support the reliability of the data and assess its impact on the composite indicators developed, an estimation of missing values should be conducted in the third step. This is followed by a multivariate analysis, which aims to study the overall structure and the nature of the dataset of the composite indicators through different analytical techniques. These analyses will help group the indicators of units being assessed based on their association, similarity and degree of correlation. Normalisation, another analytical step, includes different techniques, which may be applied to the data prior to any mathematical application. This could be due to the data set being of a different type (qualitative or quantitative data) or having different measurement units. Weighting and aggregation are two integrated approaches and represent an integral part in the sequence of developing the composite indicator. Weights are assigned to each indicator based on a number of factors such as statistical models or their influence, importance, expert opinion on the individual indicator. This is followed by the aggregation of all the information in the set of individual indicators into one single number, the composite indicator. Sensitivity analysis tests are applied to test the robustness of the composite indicator and improve transparency and the structure of all previous steps. The application of steps eight, nine and ten provides an extension of the analysis, a link with other similar variables and measurement types and finally leads to a presentation of the overall indicator in an efficient way. To a great degree, the helpfulness and credibility of a CI depends heavily on the fundamental weighting and aggregation schemes. Therefore, the study on data weighting and aggregation has always been an interesting but debatable matter in the field of constructing CIs (Esty et al., 2006). Rigorous investigation has been conducted during the stage of data aggregation regarding the applicability of Multiple Criteria Decision Analysis (MCDA). (See the following references: Hatefi and Torabi, 2010, Despic, 2006, Hajkowicz, 2006 and Zhou et al., 2010).

Moreover, a key problem in applying MCDA aggregation methods to construct a CI is the determination of the weights for the sub-indicators under analysis. There are numerous weighting methods that can be applied in deriving the weights for sub-indicators. OECD (2008) have illustrated the pros and cons of different weighting techniques. Often, results from expert judgment or public opinion poll methods can be used as a basis to derive the weights for sub-indicators (Hope et al., 1992). Typically, equal weights tend to be applied when such information is not readily available. However, it should be noted that one key challenge from such practice is the disagreement exhibited by the units evaluated, as each sees itself having unique features and preference (Lau and Lam, 2002).

This chapter aims to provide some general guidelines for the construction of a CI. In the sections below, further discussions will be conducted on some of the key steps of constructing composite indicator with the main focus on normalisation, weighting and aggregation as they are vital in choosing or building the composite indicator construction process that would be the most fitting to a given application. Though, this chapter is largely about literature review some elements of the case study will be introduced in parallel by using the case study data to illustrate some of the features and differences of the various aggregation techniques that could be used in the CI construction process. In our case, the new applied methodology will have a direct impact on the WEM score construction process.

3.2. Pre-normalisation steps

When creating a composite indicator, there are a number of steps involving numerous options, with subsequent possibilities that will affect the usability, fairness, credibility, robustness and reliability of the results. One of the most important steps are the first four steps where the hierarchical structure of the composite indicators is finalised. The first step requires a sensible theoretical framework to be identified. This framework must clearly define the active phenomenon to be measured, and in subsequent steps link sub-components and fundamental indicators together. For instance, the Growth Competitiveness Index (GCI) which has been

developed by the World Economic Forum is based on three categories: the macroeconomic environment, the quality of public institutions and technology" (Nardo et al., 2005). Another example is the Technology Achievement Index (TAI), which is broken down into four groups of technological capacity: creation of technology, diffusion of recent innovations, diffusion of old innovations and human skills.

These well-connected structures give more clarity on the driving factors that make up the composite indicator but the important questions that need to be addressed are: "Are all the important categories that represent the intended concept included in the structure?" and "Are the categories included sufficiently independent of each other so that there are no significant overlaps?" The importance of these questions and the importance of the first four steps of the composite indicator construction process is nothing less but critical for the successful construction. Depending on the nature and the importance of a composite indicator constructed, these stages can last for a very long time. For very complex indicators, it may be necessary to measure "degree of incompleteness", which represents maximum possible change in the value of composite indicator while keeping the scores of its main categories fixed. Once the full structure is developed, it is then also necessary to check for any redundancy among the criteria since it is possible that by removing one or more categories from the structure, the degree of incompleteness remains unchanged. While all the above makes it clear that those first four steps cannot be lightly handled, they are outside of the scope of our aim and objectives and will therefore not be discussed any further. They are mentioned here briefly only to keep the reader aware of the importance of those steps in the overall construction process. Our focus here is on the following three steps of the process involving normalisation, weighting and aggregation, which when taken together provide a complete analytical machinery for deriving composite indicator scores once the data is obtained by measuring performance of units with respect to all the categories found at the end-level of the hierarchical structure.

3.3. Normalisation

This step aims to make the selected indicators comparable, particularly when the indicators within a specific data set have different units of measurement. This is done through normalising individual indicators in order to render them comparable (OECD, 2008). Normalisation typically precedes any data aggregation and essentially standardizes the diverse measurement units that might appear. There are several normalisation methods such as ranking, z-scores, min-max and distance to reference, which are briefly outlined below. Before these analytical techniques are initiated, we will introduce notation, where y_{jp} is used to denote the value of an indicator *j* for a generic unit *p*. Therefore, in order for the indicator to be normalized, we must change each indicator y_{jp} to a normalised indicator, Y_{jp}.

Ranking is one of the simplest normalisation techniques used to measure unit performance over time (OECD, 2008). In this case the normalised indicator will be: $Y_{jp} = rank\ (y_{jp})$. In the United States, The Information and Communication Index and Healthcare Performance, have utilized this method (Nardo et al., 2005). While this normalisation causes potentially problematic loss of information, it also has its advantages such as not being sensitive to outliers and that it allows the performance of countries/departments to be followed over time in terms of their relative positions.

z-score is another widely used method. Here, the values y_{jp} are standardized into a common scale with a mean of zero and standard deviation of 1. First, the average score $\overline{y}_j$ and standard deviation σ_j is calculated using the scores of all units for indicator *j*. The normalised indicator values Y_{jp} are calculated for each unit *p* using the following expression: $Y_{jp} = \left(y_{jp} - \bar{y}_j\right)/\sigma_j$.

Another normalisation technique commonly used is the min-max normalisation, which changes the value of an indicator so that its normalised value is always within the range between 0 and 1. This is accomplished by calculating the distance from the minimum value,

and dividing it by the range of the indicator values in the sample. Normalised indicator values Y_{jp} are calculated using this expression: $Y_{jp} = \left(y_{jp} - \min_p y_{jp}\right) \Big/ \left(\max_p y_{jp} - \min_p y_{jp}\right)$.

There is another similar technique for normalisation known as the "distance to a reference". This technique gauges the distance to the value of a reference unit *k* instead of the min and max values of the sample (Nardo et al., 2005). This reference unit could be the best performing unit or an external benchmark, or an average group, or the group leader. This technique has been used effectively by the EU Lisbon Agenda, where they have taken the US and Japan as benchmarking units (OECD, 2008). This method allows the expression of the ratio to lie between the real value of an indicator and the reference value. Normalised value is obtained using either $Y_{jp} = y_{jp}/y_{jk}$ or $Y_{jp} = \left(y_{jp} - y_{jk}\right)/y_{jk}$.

It is clear that each of those four singular normalisation processes can produce different results. For example, the first three methods have a varied dependency on the outliers' unit of performance. The first process is unaffected when outliers are in the example, but the composite indicator in the second process is much more affected by an outlier indicator. In the third process, there is a greater dependence on outliers than in the first and second. All these different properties create ambiguous results. Also, it is clear that the max and min values can vary over a period of time. This makes the process too sensitive to very few scores and so the obtained normalized scores can be unreliable.

Many approaches yield the values that change over time and require a recalculation when an outlier is there in order to keep them similar. Note also that in the case of the last normalisation process described, the normalised values depend on extreme values, which are normally used as a reference point. Consequently, the benchmark choice is vital, which frequently leads to unreliable results. Those normalization processes that heavily depend on extreme values may consider adding an additional step of identifying and removing outliers before performing

normalization. Yet, even that process (removing outliers) requires a rather subjective view on what makes an extreme value to be an outlier.

This short description of various normalisation approaches and their various properties is important to understand and relate to the Data Envelopment Analysis (DEA) approach, which is our method of choice. Namely, DEA does not require any normalisation of data due to its unit invariance to measurement units (Cherchye et al., 2007). This may appear as a significant advantage since all those normalisation methods carry some problems with them. The fact is that when a DEA-based model is used to construct a composite indicator score, the normalisation is implicitly carried out since the models come with the normalisation built-in. At the same time, the normalisation carried out by DEA-based models is a very sophisticated one. It is most similar to the "distance to a reference" approach although it does not use a single reference unit but a number of different units that sit on what is known as the best-practice frontier. These details will be discussed in the next chapter but for now it is just important to note that this type of normalisation reduces or completely eliminates problematic features of the "distance to a reference" approach. One of the most important problems avoided is related to the fact that composite indicators are multi-dimensional and so when a single reference unit is used, then there will be units that behave very differently from the reference unit and yet they are directly compared to it. To use a light-hearted explanation, while "distance to a reference" approach may compare apples and oranges with an excuse that they are both fruit, the DEA approach is strict in comparing apples only to apples and oranges only to oranges. The differences between different sorts of apples or different sorts of oranges may not be accounted for by the DEA approach but at least apples are not compared to oranges and the other way around. The bottom line is that when using DEA, the normalisation step can be skipped but we need to be aware that normalisation is nevertheless there. The good news is that the built-in normalisation is a very good one, definitely not worse than some of the best ones out there.

3.4. Weighting

Once the normalisation stage is completed, the next steps are to assign value judgments and measure of importance (weights) to the normalized indicators and find the most suitable approach to aggregate them into a single composite indicator. Weights heavily influence the outcome of a composite indicator and ranking in a benchmarking exercise. Thus, the weights should ideally be selected and agreed upon based on the theoretical framework. Further studies about weighting in the context of constructing composite indicators elaborated the significance of taking into account consensus amongst citizens relative to the importance of each indicator (Hagerty and Land, 2007). The individual indicators may be assigned either equal weights amongst all or they may be given different weights relative to their importance to the composite indicator (Freudenberg, 2003). Thus, the primary decision that must be made on the importance of the indicator is between equal or different weighting, which will directly affect the final scores.

3.4.1 Equal weighting

Many composite indicators depend on an equal weighting technique wherein the same weight is assigned to all the sub-indicators. Thus, this implies that all these sub-indicators are relatively of the same importance in the composite indicator, which could conceal the absence of a statistical or a practical basis. Equal weighting has been typically applied in cases where the causal relationships amongst the individual indicators was not known or in cases where there was lack of agreement of an alternative technique in the assignment of weights, i.e., differential weighting. Applying equal weighting in a composite indicator leads to an unbalanced structure of the composite indicator itself due to the unequal weighting of its dimensions. This results from grouping the sub-indicators into dimensions on a higher-level, which are then further aggregated into the composite indicator. Although the sub-indicators will all have equal weights, the dimensions they're grouped into will have different weights based on the number of indicators grouped. In this case, the dimension with a larger number

of sub-indicators will have a higher weight compared to a dimension with a smaller number of sub-indicators (Maggino and Ruviglioni, 2009).

In adopting equal weighting technique, the choice of assigning weights is seemingly less subjective; and the subjective element arises exclusively in the choice of indicators selected. Thus, equal weighting method remains controversial despite its popularity. It is "obviously convenient but also universally considered to be wrong" (Chowdhury and Squire, 2006).

An alternative to the equal weighting technique is differential weighting, which does not necessarily relate to the selection of different weights but rather looks into the selection of the most appropriate approach to identify the weights amongst the different approaches (Nardo et al., 2005). There are numerous techniques of assigning weights that can be found in the literature. Some of those methods are based on a pure statistical approach while some other ones are participatory in their nature. The list of methods is very long and some of the most prominent ones are: factor analysis (FA), principle component analysis (PCA), unobserved components models (UCM), budget allocation processes (BAP), public opinion (PO), conjoint analysis (CA) and analytical hierarchy processes (AHP). In some cases, weights might be assigned based on statistical methods and, in other cases, they may be used to either reward or punish specific indicators, depending on expert opinion, to reflect policy priorities or other factors. The first three methods in the list above belong to the class of statistical weighting methods. The remaining four are normally categorised as participatory weighting methods because they require subjective judgments to be made by various stakeholders.

3.4.2 Statistical Weighting Methods

Principle component analysis (PCA) and factor analysis (FA) are statistical approaches with the aim of reductionism (Greco et al., 2018). The idea is that PCA and FA are based on the statistical dimensions of the data, which describe the highest possible variation in indicators, using the smallest possible number of principle components. The weights for each principle component are calculated by considering their proportions in explaining the variance in the

data set. Therefore, the larger the data variance explained by the component, the greater is the weight assigned to it. The original data in the PCA approach can be defined by a series of equations that represent the number of indicators. These equations are fundamentally linear transformations of the original data which are constructed in a form that leads to the representation of the maximum variance of the original variables by the first equation, the second highest variance by the second equation and so on. Through the FA approach differs from the PCA, the outcome is quite similar. In FA, the original data depends essentially on underlying common and specific factors which explains the resulting variance in the original data set. Liu, et al. (2009) conducted study employing factor analysis to identify an instrument to measure the service quality of website portals. The results indicated that the instrument is a four-factor model that includes adequacy of information, appearance, usability, and privacy and security. There are, however, some drawbacks to such an approach. Srinivasan, (1994) states that it is difficult to interpret the obtained linear combination of indicators such as in the case of the human wellbeing index. He argues that the correlations amongst the indicators do not necessarily reflect their influence on wellbeing. It has been pointed out that a multidimensional approach is needed to overcome this drawback since important dimensions of the human wellbeing index are not strongly correlated (Somarriba and Pena, 2009).

Another statistical method for deriving weights is the unobserved components model (UCM), which is based on the idea sub-indicators depend on an unobserved variable and an error term. For this method to be used in practice, it is necessary to have a set of sub-indicators all measuring an unknown phenomenon represented by an unobserved variable. Accordingly, estimating the unknown variable will highlight the relationship between the composite indicator and its respective sub-indicators. The derived weights will be set to minimise the error in the composite indicator. The method is in many ways like regression analysis with the main difference being that the dependent variable in the UCM is unknown. This method has been applied in the construction of governance indicators (Kaufmann et al., 2004). The main

disadvantages of using UCM are the dependence of results on the availability of sufficient data and that it does not work well with highly correlated sub-indicators.

3.4.3 Participatory Weighting Methods

Budget Allocation Process (BAP)

BAP is a commonly used weight assigning approach. It is based on the experts' opinion and experience on the subject analysed, wherein each expert is asked to allocate an assigned "budget" to each individual indicator according to their perceived importance (Mascherini and Hoskins, 2008). This is how WEM scores are calculated. The average score for every indicator is assigned, as it is a respective weight. The final WEM composite indicator is calculated using the additive aggregation method. In terms of applying BAP, in the case of the TAI environmental sector in 1991, 400 experts from different fields had to allocate a budget to several environmental indicators to analyse an air pollution issue (Moldan et al., 1997). The main advantages of BAP approach lie in its relatively straightforward nature and short duration. A drawback of this method is that it might require taking into consideration unrelated views from experts in different fields. However, expert opinion can be a very useful tool in the weight selection for composite indicators. This is especially the case, considering that applying expert opinion will likely increase the legitimacy of the composite and the perception of decision-makers and public opinion. Another drawback of BAP is that the weights generated may not measure the importance of individual indicator even. The focus might be instead on assigning weight based on specific preference, or urgency to mitigate a certain issue. The OECD argues that budget allocation is optimal for a maximum of 10-12 indicators. If too many indicators are involved, this method can induce serious stress in the experts who are asked to allocate the budget. Moreover, when a CI is attained from a large number of sub-indicators, then this could make the allocation of the points by the experts rather difficult.

Public Opinion (PO)

An alternative approach to BAP is to determine the weights for indicators within a composite indicator by consulting with the general public. Public opinion polls have been widely used for many purposes over the years, including the assignment of weights by the public as they are typically easy to conduct and inexpensive (Parker, 1991).

Parker has discussed the application of public opinion due to the concern about the Environmental Index (EI). The main advantages identified for this approach is that it deals with issues on the public agenda and it let all the stakeholders express their preferences in the specific topic and creates an agreement for policy action. One of the main disadvantages in applying the public opinion approach is that it could result in inconsistencies when dealing with a large number of indicators, as it is difficult to ask the public to allocate points to number of individual indicators than to express a degree of concern (e.g. happy or not happy, great or small) about a particular issue. Another drawback, public opinion might imply a measurement of concern, where people are expressing an individual concern rather than a public opinion.

Conjoint Analysis (CA)

This technique utilizes a combination of statistical analysis and opinions of experts. It has been effectively used in marketing (McDaniel and Gates, 1998) and in consumer research (Green and Srinivasan, 1978). In the context of CIs, Ulengin et al. (2001) used this approach to measure the urban quality of life in Istanbul. Conjoint Analysis is a decomposition multi-criteria approach. The idea is to evaluate a set of alternative scenarios, each of them composed by a certain set of sub-indicators. Experts are asked to choose their preference among these scenarios (sets of sub-indicators). This preference is then decomposed with respect to the components (sub-indicators) based on the evaluations. Weights represent how the preference is changed when changing a sub-indicator (trade-offs), which is the approach's main advantage. Thus, the matter of compensability and its possible desirability arises. Overall, it is

a complex method of estimating weights, it needs a large sample of experts and a well-defined framework of the questions.

Analytical Hierarchy Processes (AHP)

Another widely used technique designed to solve a complex problem in the field of multi-criteria decision-making in hierarchical structures is the Analytical Hierarchy Processes (AHP) (Saaty, 1987). It allows decision-makers to identify their preferences using natural language comparisons, which are themselves translated onto a numeral scale between 1-9 as illustrated in Table 7. The numerical values of comparisons made between each possible pair of sub-indicators are then used to compute the relative importance of all the indicators with respect to their parent indicator.

Table 7: Nine-point scale for pairwise comparison

Importance description	Numerical Value
Equal importance: two activities contribute equally to the object	1
Moderate importance: slightly favours one over another	3
Essential or strong importance: strongly favours one over another	5
Demonstrated importance: dominance shown in practice	7
Extreme importance: favouring of highest possible order of affirmation	9
Intermediate values: when compromise is needed	2, 4, 6, 8

AHP is used as a weighting technique in several composite indicators such as the Technology Achievement Index (TAI), Environmental Index (EI) and Human Development Index (HDI). For example, in the case of HDI in Iran, AHP was applied to rank five factors including income, culture, healthcare, knowledge and civil rights (Paktinat and Danaei, 2014). Experts were asked to make judgments and give a crisp value about the relative importance of each pair of five indicators and rank them based on the AHP technique using additive aggregation in calculating the final composite indicator. Another example of the use of AHP in a hierarchy structure is the e-readiness Index (Gupta et al., 2007), which is based on 6 broad indicators. Each of these indicators is broken down into several sub-indicators. The problem of such methodology is that there is a possibility of inconsistency between experts' opinion, which

could lead to unreliable results. Entani, et al., (2001) dealt with interval comparison matrix that contains the decision makers' uncertainty judgements for four output indicators and obtained the interval importance values instead of crisp using interval AHP. Emrouznejad and Marra (2017) analysed the publication of 8441 of the application of AHP in several sectors such as education, health, energy, computer science, supply chain and ecology over the period of 40 years. They highlighted the advantages and drawbacks of using AHP for decision-making. AHP can be combined with other techniques such as mathematical programming and Data Envelopment Analysis (DEA) due to its simplicity and flexibility of application. Moreover, Emrouznejad and Marra (2017) highlighted how the integration of AHP with other methods has helped to overcome the shortcomings of individual approaches. Ho (2008) also highlighted the integrated AHP's advantages and its applications from 1997 to 2006. Some problematic properties of AHP have been pointed out such as using crisp values to express the decision maker's opinion in comparison to alternative approaches wherein the former can be uncertain in reality (Wang and Chen, 2007).

3.4.4 Data envelopment analysis (DEA)

In the literature relating to the construction of composite indicators, DEA is sometimes listed separately from the benefit of the doubt (BOD) approach even though BOD is nothing else but a DEA specifically tailored for the construction of composite indicators. We do not distinguish one from the other here simply because there are many different DEA models and hence many ways to tailor the corresponding BOD approach. Also, in the literature on CIs, DEA is frequently categorised as one of the statistical weighting methods. However, even though a pure DEA approach does not require any subjective judgments to be made, its practical value is significantly increased in the domain of CIs if we allow subjective judgments to be incorporated in the form of weight restrictions. Due to all the above and since DEA-based methodologies are the main focus of our interest for WEM score construction, we are discussing DEA (and BOD), separately from all the other weighting methods discussed above.

All the weighting methods discussed in previous sections have a common feature of assigning the same indicator weights for all the units under assessment. (UCM is an exception but only in those cases where not all the units have data on all individual indicators). Having the same indicator weights for all the units gives an important advantage of being able to compare all the units on a common ground. At the same time this advantage is possibly one of the biggest disadvantages in the presence of unit-specific characteristics, which can make different units see and treat the same indicator very differently. Using the same indicator weights for all the units simply ignores any such differences, making it also very difficult to examine true reasons for any poor performance of the units assessed (Shen et al., 2013).

As opposed to all the methods discussed so far, DEA allows different weights to be used for the same indicator by different units. This is also one of the most important characteristics of DEA (BOD) when used for constructing CIs. In general, DEA is a well-known non-parametric method for the estimation of production frontiers based on linear programming. This method has first been introduced by Charnes, et al. (1978) as a tool to measure relative efficiency between different units sharing homogeneous activities. For each unit, the efficiency is measured as the ratio of the weighted sum of outputs to the weighted sum of inputs relative to the maximum value of this ratio observed among all the units. The weights for each of the outputs and inputs might differ from one unit to another and are selected to optimize efficiency of the unit assessed. The best performing units able to achieve an efficiency score of 1 are assumed to form the best-practice production frontier, which are essentially used as benchmarks to all the other units. Theoretical features of DEA are explained in detail in the following chapter while the main focus in the remaining part of this section is on its properties as a tool for assigning weights to indicators and its use in various applications.

DEA has received significant attention in the area of construction of CIs. There are several interesting features of DEA in comparison to the other techniques in developing CIs. Firstly, it offers a new way of aggregating multiple indicators without having prior knowledge of their

weights. Secondly, each unit gets its own best possible weight for each of its indicators. DEA accordingly assesses the relative performance of a specific unit using endogenous weights while considering the performance of all the other units. In this way, the major issues for each unit under assessment can be identified. At the same time, the units assessed cannot complain about the unfair weighting due to the DEA's positioning of the unit in the best way possible. The weights are assigned in a way to optimize the overall score of the unit assessed and so any other weighting techniques would generate a lower score. The poor performing unit based on the most favourable set of weights cannot raise concern due to inappropriate evaluation process. The most favourable weighting arrangement for unit performance can be determined from the unit data themselves.

The idea behind such approach is that a good relative performance in a specific indicator for a unit assessed essentially means that the unit considers that indicator as relatively important. On the contrary, a unit assigns less importance to those indicators that it apparently performs weak in relative to the other units in the set (Rogge et al., 2006). These weaknesses represent the dimensions that a unit should further develop in order to achieve a better score. In the context of constructing CIs, this means that any lack of agreement on the appropriate weighting scheme and the uncertainty surrounding the process can be overcome by one such data-oriented weighting technique.

With reference to the aforementioned strengths of DEA in the construction of CI, the DEA (or BOD) approach has been extensively investigated in various studies. Gaaloul and Khalfallah (2014) reassessed the Digital Access Index (DAI), by allowing the weights to be associated with all indicators involved unlike those weights (fixed for all) initially proposed by the International Telecommunication Union. There are many studies on the construction of CIs using the BOD approach, such as: Storrie and Bjurek (2000) who used this method to measure unemployment. Cherchye (2001) adopted this technique to build indicators of macro-economic performance. Mahlberg and Obersteiner (2001) re-assessed the Human

Development Index (HDI) using BOD. Cherchye and Kuosmanen (2002) assessed the durable development. Cherchye et al. (2004) evaluated the phenomenon of social inclusion through the application of the BOD approach. Fare and Grosskopf (2004) studied the environmental performance index. Despotis (2005) looked into the Human Development Index using DEA. Ramanathan (2006) assessed the macro-economic performance index. Zhou et al. (2007) applied this approach to the Sustainable Energy Index. Cherchye et al. (2008) also used this method to re-measure the Technological Achievement Index (TAI).

Some studies were also making further advances in theory relating to the use of DEA for constructing CIs. Shen et al. (2011) suggest a generalized multiple layer DEA (MLDEA) model to present the layered hierarchy structure of inputs and outputs through the incorporation of different possible types of weight restrictions in each category across the various layers. Shen et al. (2013) highlighted a key drawback of the DEA-based models wherein it treats all the indicators equally since all the indicators have to be within a single layer. They have argued that treating all indicators within the same layer equally may lead to a weak discriminating power and unrealistic allocation of weights. Subsequently, Shen et al. (2013) successfully applied MLDEA approach for the construction of the road safety performance index. In MLDEA method, the weights in a particular category of a layer can be interpreted as the importance level of the corresponding indicator. Thus, the decision makers' value judgments can be incorporated through weight restrictions in a specific category.

From the perspective of achieving the objectives relating to the WEM scoring methodology, this extension of DEA to deal with multiple layers is very attractive since WEM score itself is based on a rather massive hierarchical structure as presented in Chapter 2. Still, before settling on any specific methods, there are more important issues that need to be carefully considered such as the issue of aggregation, which is discussed in the following section.

As for the use of DEA as one of the weighting methods, there are still some important considerations that need to be addressed such as those relating to incorporating value

judgments into the models. Yet, since these are closely related to the mathematical formulation of DEA, they will be discussed in the next chapter when we introduce all the technical details behind DEA and its multiplicative variant known as geometric DEA (G-DEA).

3.5. Aggregation

Composite indicator (CI) has been applied as a measurement tool to monitor performance, conduct benchmark comparisons, policy analysis and decision making in a variety of fields. Yet, CI is just a mathematical aggregation of a set of individual indicators that assess multi-dimensional issues that have no common units of measurement (Nardo et al. 2005).

In CI construction, the selection of an appropriate aggregation approach is an integral step, which has gained much attention in the literature. A key aspect that affects the choice of the aggregation method used is the type of indicators selected. On one hand, sub-indicators within a CI are considered 'substitutable' if a deficit in one of the indicators can be compensated by a surplus in another indicator. Thus, a low value in a particular indicator can be offset by a high value in another indicator. On the other hand, sub-indicators within a CI are considered 'non-substitutable' if the compensatory feature is not permitted. An example of this case is, a low value in "hospital beds per 1,000 inhabitants' cannot be offset by a high value in "medical doctors per 1,000 inhabitants" and vice versa. Therefore, the aggregation approach is defined as 'compensatory' or 'non-compensatory' based on whether compensability is allowed amongst the sub-indicators (Casadio Tarabusi and Guarini, 2013).

The compensatory approach involves applying additive methods such as the arithmetic mean. In the cases of partially compensatory or non-compensatory approaches (in further text referred to simply as non-compensatory approaches), non-linear methods such as the geometric mean or the multi-criteria analysis are applied (Mazziotta and Pareto, 2013).

Additive aggregation is not an attractive approach for constructing WEM score due to its undesirable feature of implying full compensability. Geometric aggregation, on the other hand,

is better suited since it is only partially compensatory. It is also suitable when sub-indicators are conveyed in diverse ratio-scales (OECD, 2008). By using geometric aggregation, units that have no low scoring sub-indicators will be in advantage over the units that may have a mixed level of scores ranging from very low to very high. This property is very much along the line of our second objective, where we wanted to encourage a more balanced performance across different criteria for all the departments.

The main factors that affect the choice of the aggregation method selected, are the objective of the work under analysis and the type of users (researchers or the general public). Typically, an aggregation method is either defined as 'simple' or 'complex'. An aggregation method is considered as 'simple' when an easily understandable mathematical function is applied. An aggregation method is considered as 'complex' when sophisticated models or multivariate approaches are applied. Additive aggregation (weighted sum or weighted arithmetic mean) is the most well-known representative of a simple compensatory approach while multiplicative aggregation (weighted product or weighted geometric mean) is the most well-known representative of a simple non-compensatory approach. Multiplicative aggregation is perhaps not as simple as additive aggregation but in certain applications the attractiveness of its non-compensatory nature is much stronger than the unattractiveness of its slightly more complex mathematical formulation. The change in how HDI is computed that took place in 2010 is a good example of this: the weighted sum was replaced by the weighted product of the three main factors (life expectancy, education index and gross national product). The main reason for this change was to reduce the level of substitutability between the three main factors, which was deemed as unreasonably high when weighted sum was used. Closely related to this is the investigation by Ebert and Welsch (2004) on the differences between the arithmetic mean and geometric mean when constructing environmental indices. One of their major finding was that indices in the form of an arithmetic mean are not meaningful because the variables do not satisfy the property of interval-scale unit comparability. With respect to the construction of

WEM score, high level of substitutability induced by additive aggregating goes directly against the government's objective to support its departments in improving their performance across all the dimensions. As this is an important issue for WEM score, we will soon take a deeper look into multiplicative aggregation and how exactly it differs from additive aggregation, especially in the presence of flexible weighting schemes such as the ones provided by DEA.

DEA itself has been extensively applied in the construction of CIs (Zhou et al. 2008). There are two forms of using the DEA in aggregating the CIs. One form involves the traditional method of DEA in constructing an aggregated index, which first distinguishes inputs and outputs among a set of sub-indicators. Examples of such application include the Welfare Achievement and Improvement Index (Zaim et al. 2001), the Economic Wellbeing Index (Murias et al. 2006), the Environmental Performance Index (Zhou et al. 2007b), and the University Quality Assessment Index (Murias et al. 2008).The second form involves transforming the sub-indicators into similar type of variables first, whether benefit or cost type variables, and then aggregating them into a CI using DEA-like models. This application has gained attention in the past decade and several researches have been conducted on this line, which include Despotis (2005a, b), Zhou et al. (2007a) and Cherchye et al. (2007a, 2007b, 2008). The MLDEA approach proposed by Shen et al. (2013) is of particular interest here due to its ability to represent multiple levels of hierarchy in a single aggregation formula as well as allowing for flexible weighting scheme, which are all desirable features for the WEM score constructions. However, that approach is essentially relying on additive aggregation, whose non-compensatory properties become even more pronounced in the presence of flexible weights. The problems relating to this issue were briefly addressed next to Table 6 in section 2.6.1, where the current aggregation method for WEM score was presented. In the next section we will explore this issue further.

3.5.1 Aggregation: Additive vs. Multiplicative

To elaborate further on the above, this section will look into the case study data to clearly illustrate the differences between the additive and multiplicative aggregation methods. Let us

first compare results between additive and multiplicative models when applied on the WEM scores data for 19 departments. Figures 5 and 6 show the WEM scores using additive and multiplicative models, respectively. The scores are normalised so that the maximum score is equal to one and the 19 departments are ordered from the largest to the lowest score. The distributions show a slight drop in scores when applying the additive model while the multiplicative model displays a significant drop in scores after the top three departments.

Figure 5: Distribution of WEM scores using additive aggregation model

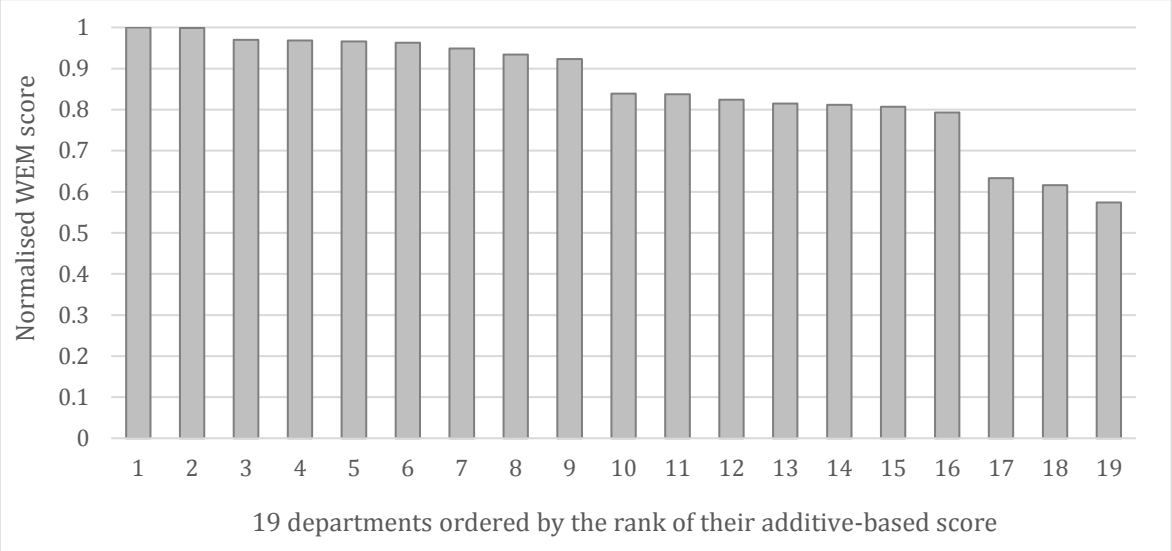
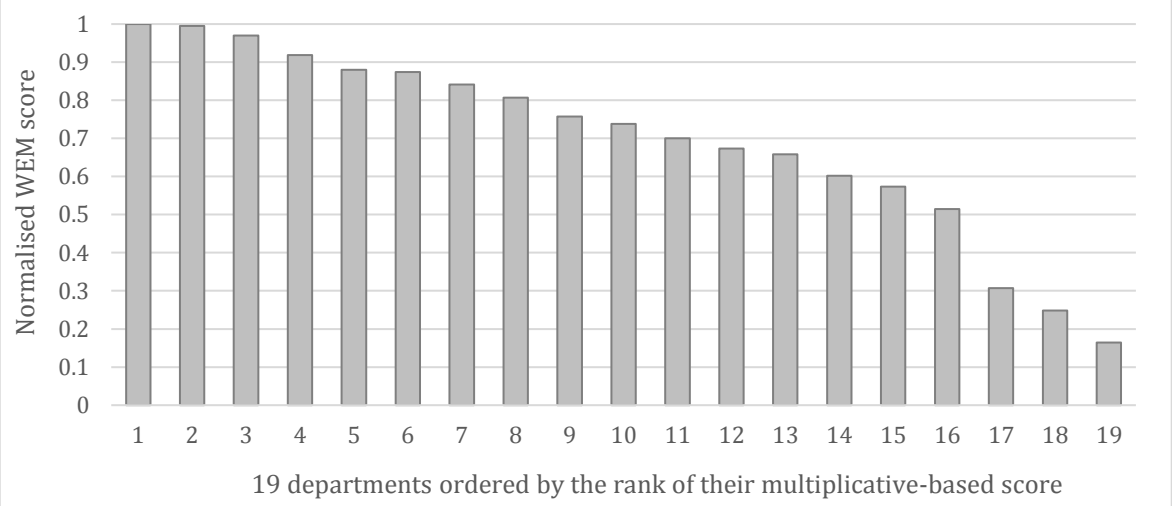
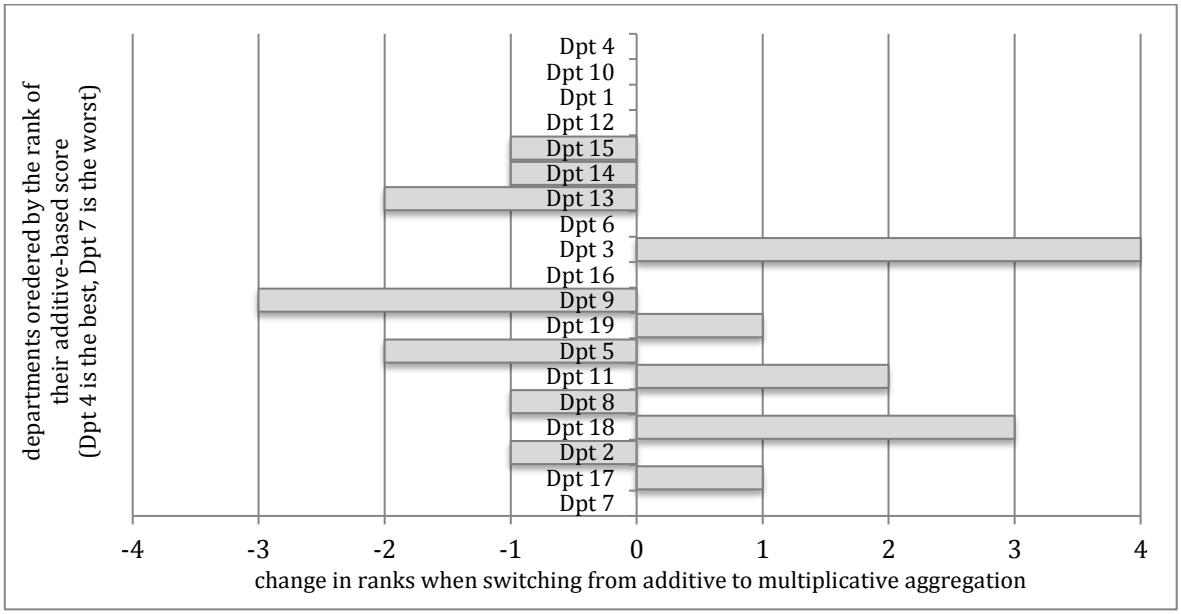


Figure 6: Distribution of WEM scores using multiplicative aggregation model



If we look at the differences in ranks across the two approaches, as illustrated in Figure 7, at one extreme we have a gain of 4 places using the geometric WEM score (Department 3), a fall of 7 departments between 1-3 places (departments 15, 14, 13, 9, 5, 8, and 2) while 7 departments maintained the same rank. The average absolute change in ranks is found to be 1.16 rank positions.

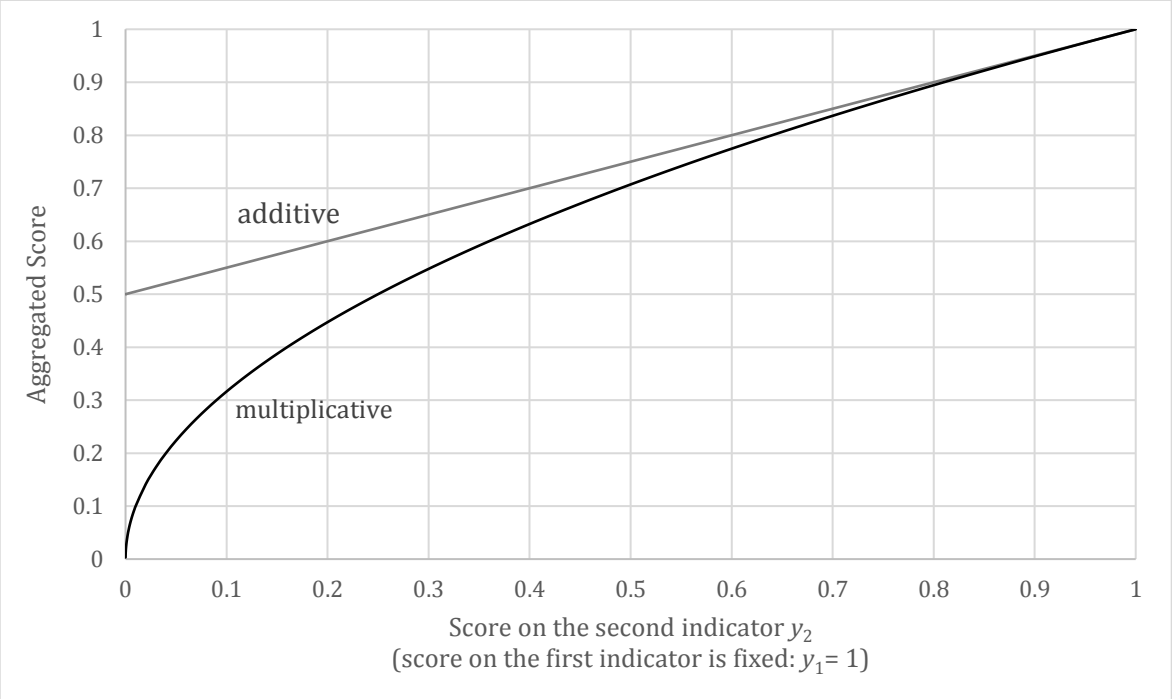
Figure 7: Difference in ranks between additive and geometric WEM scores approaches.



Change in the rankings of the departments due to different aggregation method used and especially a significant jump of department 3 from the 9th to the 5th position, confirms that the decision on the type of aggregation is an important one to consider for the construction of WEM scores. It is necessary then to look at these differences closely. Differences between additive and multiplicative aggregation are rather well researched and only the main points are covered here. Yet, there is hardly any research that investigate the differences between the two aggregations in the presence of flexible weights (such as the ones allowed by DEA). Since flexibility of weight is also a desirable feature for constructing WEM score, we will pay special attention to that aspect.

Let's start with two indicators (y_1, y_2) that have equal weights and positive scores within the range of zero to one. If all the departments have the same total of the two scores $(y_1 + y_2)$, then their overall score and ranking would be the same under additive aggregation. Under the multiplicative aggregation, the score $(y_1 \times y_2)$ will not necessarily be the same and one might rank higher than the other. The geometric score is maximized when $y_1 = y_2$. For example, if we consider the case where $y_1 + y_2 = 1$, then the individual scores of (0.5, 0.5) give the highest geometric mean winning over the combination (0.6, 0.4) or (0.7, 0.3) or any other unequal distribution of scores. This is due to the well-known inequality between the two means, which states that for any given positive scores, their geometric mean will always be less than or equal to their arithmetic mean (equality occurs when the scores are equal). Referring to the above example, the arithmetic mean is always the same if the total score is 1. However, the geometric score is maximized when the indicators are equal and if they diverged away from equality, the geometric score declines. This is illustrated in Figure 8 where one of the scores is fixed to 1.

Figure 8: Comparing additive and multiplicative aggregations using crisp weights.



It is worth noting how the differences in total score are increasingly more pronounced the more the two scores differ from each other. When the score on the second indicator is above 0.5 (i.e., above 50% of the score on the first indicator), the differences in the aggregated score are rather small and are never bigger than 6% in relative terms. On the other hand, when the score on the second indicator is below 0.3 (i.e., below 30% of the score on the first indicator), the difference in the aggregated score is greater than 15% and that difference increases very quickly reaching about 25% for the score of 0.2, and 45% for the score of 0.1.

Let us now look at this inequality from a different perspective: on one hand, a department could benefit from a wide spread of scores and rate higher under the additive approach but with such scores it would have an undesirable impact under the multiplicative approach. On the other hand, a department will do well under the multiplicative approach if it had more balanced and less scattered scores. This discussion assumed equal weights. However, it also applies to the case of unequal weights. Assuming we have non-negative weights b_j which add up to unity, then the general arithmetic-geometric mean inequality states that

1) $$\sum b_j y_j \geq \prod y_j^{b_j},$$

where the left-hand side of the equation is the weighted arithmetic mean and the right-hand side is the weighted geometric mean of sub-indicators y_j. Equality would appear again only when all y_j are the same. Just as before, if two departments have the same additive score and rank, the one with smaller lever of variations among individual indicators will tend to have a higher geometric mean and rank compared to the department that has larger variations among its individual indicators. Conversely, for the two departments that have the same geometric score and rank, the one with equal indicators will have the lowest additive score compared to the department that has indicators diverging from each other.

It is of special interest to see and compare the means when departments are allowed some flexibility in choosing the weights for individual indicators. For this purpose, we will compare aggregated scores of two indicators whose total score $(y_1 + y_2)$ adds up to 1. We present five

different scenarios starting from the one where the weights are fixed at 50% for each indicator and in increments of ± 10% moving towards the scenario where the weights can be chosen from 50% ± 40% interval. Figures 9 and 10 illustrate the effects to aggregated scores.

Figure 9: Additive aggregation of two scores using different levels of weight flexibility.

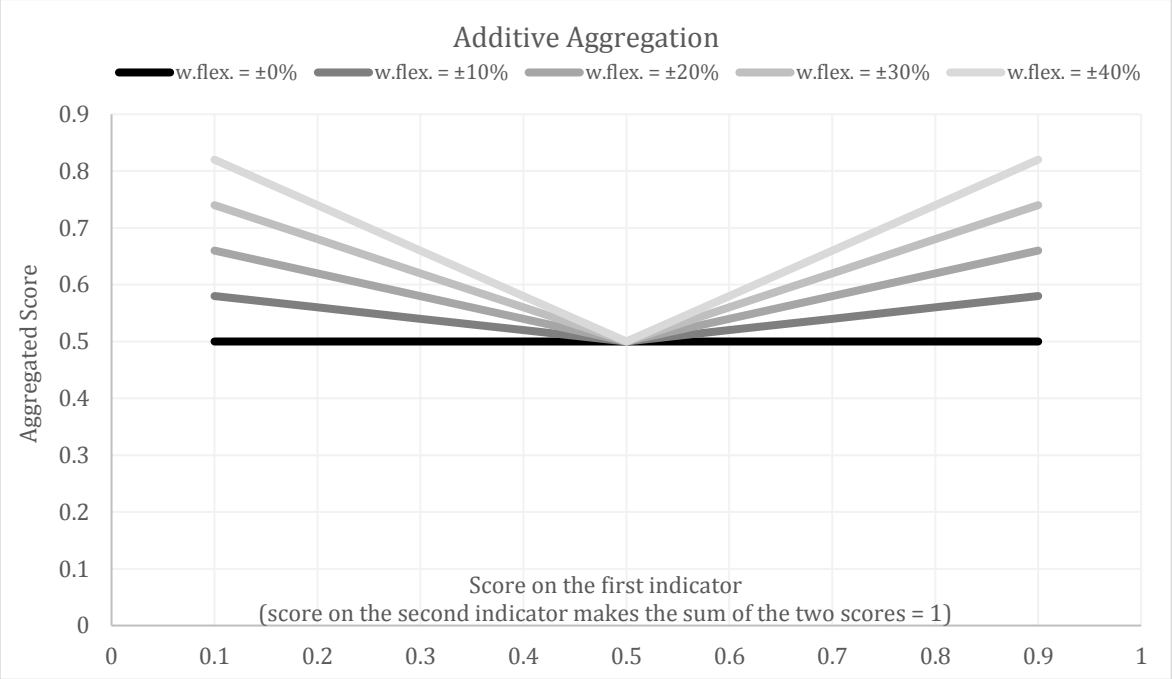
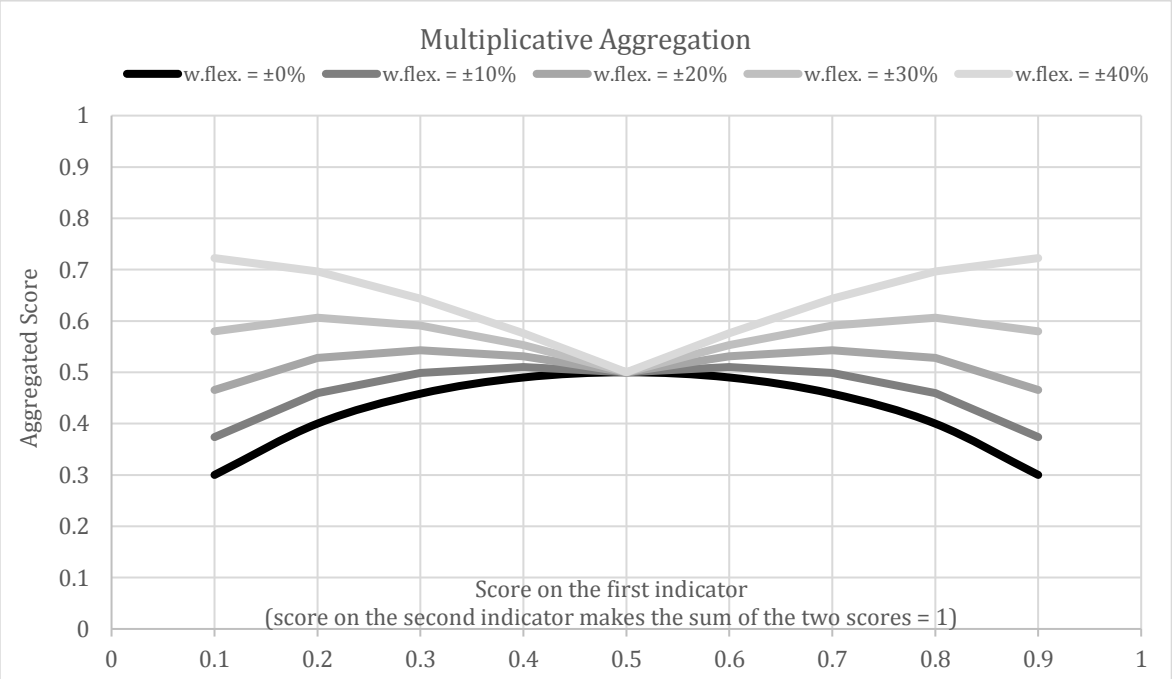


Figure 10: Multiplicative aggregation of two scores using different levels of weight flexibility.



To produce the scores in both figures, it was assumed that departments will use the weighting scheme (when given any flexibility on weights) that will maximise their aggregated score. As expected, the multiplicative aggregation clearly exhibits a much lower compensatory power than additive aggregation. Compensatory nature of the additive model is on the other hand even further amplified by flexibility of weights.

The main reason to consider flexible weights is due to uncertainties that naturally exists in subjective evaluation of weights by the experts and/or any other stakeholders. Allowing for the weights to be specified in ranges rather than as crisp values will affect the time needed for the stakeholders to reach consensus on weights: in general, the wider the allowed ranges, the less time is needed to reach the consensus. Yet, as illustrated by Figures 9, wide ranges for weights under additive aggregation lead to some rather eccentric aggregated scores. So, when faced with uncertainties in the evaluation of the weights of the criteria, additive aggregation propagates this uncertainty fully to the "extreme units" allowing them to be scored much better than the "more balanced" units. The more extreme the unit is, the higher overall score will be assigned to it under additive model, irrespective of how small the uncertainties are in evaluating the weights. Multiplicative model, on the other hand, does not allow for such a one-sided treatment unless the uncertainty in the subjective evaluation of weights becomes very extreme as illustrated in Figure 10 by the line corresponding to the flexibility of ±40%, which is allowing the weights to be anywhere between 10% and 90%. Aggregated scores under multiplicative aggregation stay at a very reasonable level as long as the flexibility on weights does not exceed ±20%. Combining practical considerations concerning difficulties in accurately specifying the weights and theoretical considerations concerning the abnormality of aggregated scores under very flexible ranges, the multiplicative model allowing for about ±10% to ±20% flexibility for weights seems to be the best of both worlds.

In summary, under the additive aggregation and due to its compensatory power feature, poor performance in some indicators could be compensated to a greater extent by a good

performance elsewhere. This drawback is overcome under the geometric aggregation wherein consistent performance is rewarded to a greater extent across the different indicators. In the final chapter of the thesis we will implement this idea to improve WEM scoring methodology, which will be based on a multiplicative model allowing for a small level of flexibility on weights. This will create conditions where more attention will be given to a balanced performance and where excelling in a few indicators will not lead to high scores and rankings.

3.6. Normalisation, Weighting and Aggregation for some well-known CIs

In the literature, several composite indicators were used as samples for the application of different existing approaches of weight and aggregation. However, these indices are just an example of a varied and growing list of composite indicators developed and utilised globally. Table 8 gives a brief summary of several indices mentioned in the literature, specifying the normalisation, weighting and aggregation methods used to form them.

Table 8: Composite Indicators Summary

Composite Indicators	Normalisation	Weighting	Aggregation
Website Excellence Model	-	BAP	Additive
Web Index	Z-score	Equal weights and expert opinion	Additive
Technology Achievement Index	Min/max value, logarithm	Equal weights	Additive
Environmental Performance Index	Min/target value	AHP	Additive
Internal Market Index	Z-score	PCA and BAP	Additive
Human Development Index	Min/target value, logarithm, reference	Equal weights	Geometric
Global Innovation Index	Min/max value	Equal weights	Additive
Quality of Life Index	Min/target value	Expert opinion	Additive

The table shows all combinations of normalization, weighting and aggregation techniques. One can notice two major points by looking at the characteristics of the present composite indicators in the literature. First, because of its perceived theoretical complication, and the fact that it is utilised in productive efficiency measurement, despite its numerous advantages, DEA is hardly used as a weighting method. Second, geometric mean (multiplicative aggregation method) is also hardly used, while weighted arithmetic mean (an additive aggregation method) is definitely the most frequent type of aggregation. Even though multiplicative aggregation is harder to understand than additive aggregation, it has many advantages: it allows for an easy way to deal with complex hierarchical structures (as we will see in Chapter 4) and it does not allow for full compensability between individual indicators.

3.7. Robustness and sensitivity

As previously stated, a number of problems can turn up when going through the process of selection weighting and aggregating indicators into what can effectively be considered a composite indicator. The outcome and related unit rankings therefore largely depend on the selected approach. This is why sensitivity tests should be followed to study the effect of various decisions made in the construction process, such as: counting or discounting indicators, the alteration of weights and the use of various normalization methods. A wide variety of statistical tests can smooth the process of the sensitivity analysis so that the composite indicator is robust and not deeply dependent on the choice of normalization, weighting approaches, or levels of aggregation of sub-indicators. As we saw above, when constructing a composite indicator, there are an assortment of problems that can be found with respect to the selection, normalization, weighting, aggregating of indicators into a composite.

The outcomes and rankings of each unit on the composite may mainly depend on the decisions taken. As such, an important factor is the use of a sensitivity analysis to discover the strength of rankings to the inclusion and exclusion of certain indicators, changes in any weighting

system, using different aggregation methods, and setting different decision rules to construct the composite (Freudenberg, 2003).

3.8. Presentation and visualization of the results

Presenting composite indicators to decision makers is a significant step. Visualizing the composite indicators results requires specific consideration as it could influence both the relevance and interpretability of the results by the users. Due to the complex nature of composite indicators, the concerned stakeholders, whether them being the general public or policymakers, tend to ignore reading the methodological notes or keynotes stated. Thus, their understanding and interpretation of the results is essentially based on the messages conveyed through the different means of displaying the results.

Moreover, it is essential that the composite indicators clearly depict the picture to the users rapidly and accurately. Visual models used for presenting composite indicators tend to flag warning signals for decision makers and usually highlight the critical issues that require policy interventions, which ensure corrective actions and continuous improvement. It has been stated that, "if arguments are not put into figures, the voice of science will never be heard by practical men" (OECD, 2008). There are various means for displaying the results of the composite indicators. These could include but not limited to the following: tables, graphs, charts, dashboards, sophisticated figures such as the four-quadrant model applied for sustainability index etc. Presenting the results in a table format might be considered as a comprehensive approach, as applied in WEM scores to display the results; however, the drawback is that it may be too detailed and not visually appealing to the user.

3.9. Composite Indicators and Business Excellence Models

As this whole chapter essentially provides a literature review on composite indicators, it is not out of place to make here a quick but an important overview of the works within the area of business excellence models, which may be relevant to the issues investigated in this thesis.

As explained in Chapter 2, WEM score is just one component of the Government Excellence Model, which in turn is one of the three categories of DGEP Awards. As such, DGEP in its broadest sense belongs to the class of government excellence models. In fact, DGEP is one of the members of GEM Council representing the Government Excellence Model in the UAE and the Middle East and North Africa region. Through its GEM Council membership, DGEP can benchmark its Excellence model and Award process with those of other countries and regions including Singapore, USA, China, Europe, Japan, India, Central/South America and Malaysia. DGEP's participation within the GEM also helps guide its future development plans and enables it to share its experiences with the global network. For this reason, it is important to briefly address the literature relevant to business excellence models (BEM).

The use of BEM has become popular near the end of 20th century and along with the Total Quality Management (TQM), it has been the most popular approach to enhance organisational performance and management capabilities over the last 25 years (Dahlgaard, 2013). While the literature on BEM is vast and appears in over 30 scientific journals, most of that body of literature focuses on traditional approaches and existing excellence framework such as EFQM (European Foundation for Quality Management) or on selecting relevant criteria for different business sectors. A very small percentage of that literature is focused on the analytical mechanism behind the models used. Rather, simple weighted averages are taken for granted or, if discussed at all, then they are discussed in the context of distribution of weights (i.e., budget allocation process) or whether to use any scoring system at all.

Those relatively sporadic papers that do focus on the analytical mechanisms are mainly concerned about the weight structure, such as Kanji (1998), Eskildsen et al. (2001) or Dahlgaard (2013), and hardly ever about the functional form through which the weights are aggregated. Still, there are some exceptions, such as Tavana et al. (2011), where the authors develop a benchmarking framework by combining EFQM excellence model with various multi-criteria decision analysis tools with the main purpose to deal with the problem of subjectivity

of weights. Namely, they propose calibrating the subjective weights with the objective weights determined through the entropy concept (the idea that the most important criteria are those that have the greatest discriminating power between the units assessed). Another example is provided by Tomaževič et al. (2016). They consider a specific application in police services and they suggest that the 5 enablers and 4 results criteria, typically used by the EFQM, could be used as inputs and outputs in a single DEA model. These two studies are rare examples of a deeper investigation of the analytical mechanism employed by BEM that appear in BEM oriented journals. The former is published in "Benchmarking: An International Journal" while the latter is in "Total Quality Management & Business Excellence". For a more serious treatment of the analytical mechanisms behind the BEM, one needs to search journals outside of typical BEM domain, such as Applied Soft Computing (see Hosseini Ezzabadi et al., 2015) or Expert Systems with Applications (see Moreno-Rodríguez et al., 2013)

All the above points to the fact that there is a considerable gap between typical BEM publications, which do not pay much attention to the analytical mechanisms behind the excellence model and almost makes no notice of the recent developments in that area that are present in what could be categorised as decision science and performance measurement body of literature. Even a deeper analysis of the websites of GEM members (EFQM, 2019) does not provide any information on how their Business Excellence Models address the issue of weighting and aggregation. However, through DGEP's membership within the GEM, it has been deduced that all the members follow a similar approach in selecting the weights and aggregating the final scores. That is, and as mentioned in Chapter 2, the weights have been selected following subjective process which is similar to the Budget Allocation Process (BAP) and the departments' scores have been aggregated using simple additive aggregation method. This research therefore represents a significant step, that essentially comes from the BEM practitioners' side, to bridge this gap and brings a much-deserved attention to the analytical

mechanism behind the excellence models directly to GEM Council and that global network of its members.

3.10. Summary

Many research papers, many reports and even some books have been written on the construction of composite indicators. This is not surprising considering the complexity of the process itself on one side and the popularity of composite indicators in policy making as well as in popular press.

In this chapter we have looked into the main issues relating to the construction process. A special attention was given to the parts of the process that are directly related to the stated objectives of this study. For this reason, we took a deeper look into normalisation, weighting and aggregation steps. We have presented some of the most popular approaches for each of these steps, discussed their advantages and disadvantages and briefly reflected on their usage in practical applications. We have also directly considered how some of these approaches fit the aim of this study, which is to design the new assessment process that will address many of the weaknesses of the current process in constructing the WEM score.

Through the discussion and analysis conducted in this chapter, we have identified some of the most attractive tools and features for the WEM scoring methodology, which are: DEA-based models, multiplicative aggregation and flexible weighting scheme that does not allow the ranges given to weights to be too wide. In the next chapter, after presenting theoretical aspects of DEA, we will design a DEA-based model that can accommodate all those desirable features.

Chapter 4: DEA-Based Composite Indicator

4.1. Introduction

One common way of measuring the performance of an organization is to compare their results against the results of similar organizations. This comparison can be done using a relative measurement of efficiency - best practice function (Luna et al. 2013). Data Envelopment Analysis (DEA) is one of the recent tools developed for modern management science, which can be utilized to measure performance, derive weights and construct an aggregated composite indicator. In this chapter we will take a look at several different DEA based models that have been used in the construction of composite indicators and one new one that we are going to use in this research.

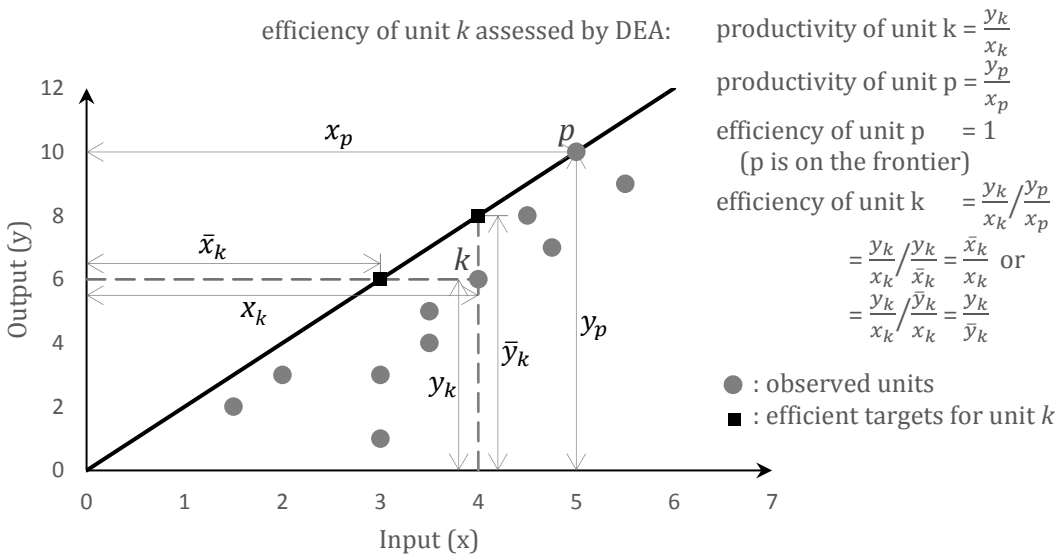
4.2. DEA theoretical background

The theory behind DEA was first introduced by Farrell (1957) to address the issues faced by many organizations in productivity improvement. He proposed an activity analysis approach that could adequately overcome these issues by measuring productive efficiency. Two decades later, Charnes, Copper and Rhodes (1978), formally introduced DEA to address the need for suitable procedures to assess the relative efficiencies of multi-input multi-output production units. DEA is an analytical tool designed to evaluate the comparative efficiency of homogeneous organizations known as Decision Making Units (DMUs) such as banks, hospitals, schools, business firms and water companies.

DEA can identify the best practice frontier with a simple restriction that all DMUs lie on or below the efficiency frontier. At the same time, it can identify possible efficiency improvements that may help DMUs achieve their potential. Figure 11 illustrates the idea, showing two different ways for the inefficient unit *k* to improve and become efficient. Charnes et al. (1978) developed a DEA that utilizes linear programming to attain an efficient frontier, consisting of

efficient DMUs in a sample. This analysis also assesses efficiency according to the distance between the frontier and the unit measured as illustrated for the case of unit *k* in Figure 11.

Figure 11: An illustration of an assessment by Data Envelopment Analysis



An equivalent mathematical way to describe this way of calculating efficiency is by seeing the efficiency as the ratio between the productivity level achieved by the unit assessed and the maximum possible level of productivity that could be achieved based on the observed practices. This can be illustrated through the formulas shown below, which defines the relative efficiency e_k for a generic unit *k*:

2) $$e_k = \frac{f_k}{f_{max}}$$

In the formula above, f_k is the productivity level of unit *k* and f_{max} is the maximum possible productivity level (for a given set of assumptions). The efficiency score of unit *k*, e_k, is simply the ratio between these two values. In classical DEA models, f_{max} is taken to be the productivity level that corresponds to one of the points on the efficiency frontier, such as point corresponding to unit *p* in Figure 11.

The productivity level f is usually also expressed in the form of a ratio. For the special case of single input and single output the function f_k take this form:

3) $$f_k = \frac{Output_k}{Input_k}$$

In a more general setting with many variables affecting the productivity, all the variables are categorised either into inputs or into outputs. Decreasing any input or increasing any output is in general assumed to lead to a more efficient unit. Many classical DEA models generalize function f for the case of multiple inputs and multiple outputs in the form shown below:

4) $$f_k = \frac{\sum_j b_j y_{jk}}{\sum_i a_i x_{ik}}$$

In the formula, y_{jk} represents *j-th* output of unit k and x_{ik} represent *i-th* inputs for unit k. The parameters a_i and b_j are the weights assigned to each input and output respectively. It is assumed that inputs x_{ik} and outputs y_{jk} are non-negative and each unit has at least one positive input and output value.

DEA optimally assigns non-negative weights, a_i and b_j, by using mathematical programming technique in such a way to maximize productivity f_k while ensuring through a set of suitable constraints that under the same weights productivity of all the units stays below one. This will essentially calculate the efficiency score of unit k as defined in (2). Basically, the unit(s) that achieve the maximum productivity level under the selected weights will correspond to the binding constraint(s) in the set of constraints requiring that no other unit exceeds the efficiency score of 1. Such unit(s) will be effectively preventing the unit k to get even higher efficiency score.

Deriving the efficiency measure e_k of unit k in the manner described above is the fundamental idea behind classical DEA. It's important to note that the weights a_i and b_j are determined for each unit k under assessment without any need to specify them in advance. Ultimately, this allows the selected weights by unit k to maximise its efficiency score.

The actual calculation of the optimal weights and the efficiency score for any unit k can be done by formulating and solving a suitable optimisation problem. The function f takes the same

form as in (4) to express the productivity level of each observed unit including unit k. The complete formulation of an optimization model that takes care of finding the optimal weights and the efficiency score of unit k is shown in equation (5).

5)
$$\begin{aligned} e_k = \quad & \max \frac{\sum_j b_j y_{jk}}{\sum_i a_i x_{ik}} \\ & \textit{subject to} \quad \frac{\sum_j b_j y_{jp}}{\sum_i a_i x_{ip}} \leq 1, \quad \forall p \\ & \qquad\qquad a_i, b_j \geq 0, \quad \forall i \; and \; \forall j \end{aligned}$$

The model above can be easily converted into a linear programming model but before we do that, it is instructive to note that formulation of function f in (4) is just one possible generalization of (3) for the case of multiple inputs and multiple outputs. Different generalizations of (3) are possible and one of them, which is based on weighted product rather than on weighted sum of inputs and outputs will lead us towards geometric DEA (G-DEA) model, as defined in Despic (2013). We will explore G-DEA model in more details in section 4.3. of this Chapter.

The transformation of the model in (5) into a linear programming model, as developed by Charnes and Cooper (1962), can be achieved by selecting a representative solution for which $\sum_i a_i x_{ik} = 1$. Once this equality is substituted into (5), the following model is obtained:

6)
$$\begin{aligned} e_k = \quad & \max \sum_j b_j y_{jk} \\ & \textit{subject to} \quad \sum_i a_i x_{ik} = 1 \\ & \qquad\qquad \sum_j b_j y_{jp} - \sum_i a_i x_{ip} \leq 0, \quad \forall p \\ & \qquad\qquad a_i, b_j \geq 0, \quad \forall i \; and \; \forall j \end{aligned}$$

The above linear programming model is known in the literature as input-oriented value-based constant-returns-to-scale DEA Model. There are other alternative formulations, such as output-oriented instead of input-oriented, envelopment instead of value-based (the dual of the above model) and variable-returns-to-scale instead of constant-returns-to-scale model. In the context of constructing a composite indicator, such as the one for WEM score, we will be

dealing only with outputs and assume a dummy input for all the units, rendering the issue of returns-to-scale irrelevant and beyond the scope of this study. The output-oriented model is obtained by linearizing the inverted form of (5), where the inverted ratio for unit k is minimised while ensuring that inverted ratios of all the units are restricted to be greater than or equal to 1. The transition to linear form is then achieved by selecting a representative solution for which $\sum_j b_j y_{jk} = 1$. Finally, the envelopment model, which is the dual of the value-based model, is useful for directly finding efficient peers and efficient targets for an inefficient unit *k*. This form will be useful for us in the implementation stage of the model but for now our focus remains on the value-based model whether input or output oriented.

There are a couple of major differences that needs some attention when using DEA in the context of composite indicators. One is related to restricting the weights, as Cherchye et al. (2008) suggested, in a much stronger way (to be greater than a certain value) than simply requiring them to be non-negative as in equation (6). Cherchye et al. (2008) has argued that complete freedom of weights has a few disadvantages. If the units being assessed were given a lot of freedom, they will tend to ignore many inputs and many outputs by assigning a weight of zero; whilst focusing on a single input and single output that they are performing best in. This is much greater flexibility than what we would like to allow in the context of the composite indicators.

The other thing that differentiates the use of DEA in the context of composite indicators is that categorising indicators into inputs and outputs, as indicated in equations (3)-(6) might not be convenient. It is frequently the case that we do not have any indicators that could be inputs (those that are of a minimizing nature: the less the better). Normally, all indicators used in the construction of a composite indicator, convey properties that require maximizing performance (the more the better) for units under assessment. This is certainly the case for our data set, where all we have are output-like indicators (i.e., the more the better).

These two issues are the two modifications to the above model that essentially create the model developed by Melyn and Moesen (1991) and later named by Cherchye et al. (2004) as the Benefit of the Doubt (BOD) Approach.

4.2.1 The Benefit of the doubt (BOD) Approach

The BOD model is really nothing else but classical DEA model adjusted to fit the context of composite indicators. As already mentioned in the end of the previous section the main modifications are based on removing inputs and restricting weights in some way. Removal of inputs can be seen equivalent to all the units having a single input equal to one, also known as dummy input. When no inputs are taken into account and when weighted sum of outputs is taken as a generalisation of a single output for the case of multiple outputs, then expression (2) can be re-written for the case of composite indicators as shown in equation (7):

7) $$e_k = \frac{\sum_j b_j y_{jk}}{\sum_j b_j y_{jp*}}$$

where $p*$ is the unit in the sample that maximizes the weighted sum $\sum_j b_j y_{jp}$ using the weights selected by unit k. A linear programming model that follows from equation (7) is shown in equation (8):

8) $$\begin{aligned} &\max \ \sum_j b_j y_{jk} \\ &\textit{subject to} \quad \sum_j b_j y_{jp} \leq 1, \quad \forall p \\ &\qquad\qquad\quad b_j \geq 0, \ \ \forall j \end{aligned}$$

This model is equivalent to the formulation in (6), where $\sum_i a_i x_{ip} = 1$ is taken to be true for all units, which will have to be true given that $\sum_i a_i x_{ik} = 1$ is true and that a dummy input (equal to 1) is assumed for all the units. The model in (8), just as the DEA model in (6), allows flexibility on weights and therefore it allows different units to emphasise their different strengths. Since each unit obtains the best conceivable score related only to the existing strengths of its peers, this weighting technique is at the minimum strictness level possible. This evades the consequence of complaints about unfair weighting by a single unit. This inherited advantage from DEA, can be considered a greatly desired asset in the public arena during debate on these

issues. Yet, the equation (8) when used in the context of composite indicators normally gets some additional constrains that relate to stricter restrictions imposed on the weights. These restrictions normally reflect some minimum requirements selected by the evaluating authorities. The restricted weights hence act as a compromise and ensure that the results are equitable from both the company's (in our case, the government's) and participant's (in our case, the department's) perspectives. Equation (8) is maximizing the score$\sum_j b_j y_{jk}$ which is equal to what we called efficiency score e_k in (6). In the context of composite indicators, we will rename this score and call it performance score, while in the context of our case study we will be referring to it simply as WEM score. Equation (8) along with any additional weight restrictions added to the model will then generate the set of optimal weights as well as the resulting performance score of unit k.

The performance score, just like the efficiency score in DEA, will have to be between 0 and 1, and it articulates the performance of the k-thunit relative to the best performers found among all the units. The specific best performers (peers) identified by unit k can be seen as role models with similar features as the assessed unit but with a better performance. The best performers are some of the real units in the sample and therefore the level of performance exhibited by them should be practically attainable by the unit assessed. Therefore, it is convenient to identify the best performers for a given unit that can be valuable for the decision-makers. It gives them examples for better practice to be considered to increase performance.

Furthermore, another advantage of the BOD approach, also inherited from DEA, is that no normalization of the indicator values is required. The model easily amends the weights in order to maximize the final score, thus the indicators' measurement units are not relevant. This is due to the feature of DEA known as "units invariance", which makes the normalization stage redundant (Cherchye et al., 2008). Cherchye et al (2008) also emphasized that, "normalization obscures the original purpose of the indicator [… as] one is no longer summarizing the original data, but re-scaled scores." This property helps us to evade the unnecessary complications that

result from the normalization procedure. It is also allowing the assessors to use the most natural scale of measurement for each indicator, which improves the reliability of the scores assigned to each unit on each individual indicator.

The model shown in equation (8) is the one that allows a very flexible distribution of weights, essentially allowing units to place an insignificant or zero weight to some or even most of the indicators. (In DEA we talked about inputs and outputs but within the context of composite indicators and in the absence of any inputs we will refer to the output valuesy_{jk} as indicator values). For instance, when a specific unit is the best among all units for a single individual indicator, then that unit will always obtain the maximum possible performance score of 1. This will happen even though it may be performing poorly among all the other individual indicators. Due to this, the model in (8) may eventually find many units to be the best performers. This is what is known in DEA as a "weak discriminatory power". This is especially pronounced in the cases where the number of individual indicators is much larger in comparison to the number of units measured. This is the main reason why Cherchye et al. (2008) recommend the incorporation of weight limitations into the model as a solution to this sort of a problem.

4.2.2 Weight restrictions (WR)

Thanassoulis, (2001) stated that weight restrictions are chosen to maximize the efficiency rating of a particular unit subject to the restriction that the unit should be positive, therefore the WR is an additional constraint added to the DEA model which will be computed separately for each unit generating an optimal weight which will vary from unit to unit. DEA permit each unit to select any weight that wants for its input and output following the restriction that no weight should be negative, and the ratio of the result should not exceed 1 (Hadad, 2003). Three approaches to estimate parameters of weight restrictions in DEA identified by Thanassoulis (2001) are as follows:

- Assurance regions of type I (ARI): This approach is named due to Thompson et al. (1986,1990) based on using the available information and expert opinion and acting as a direct restriction on weight, restriction link only input or only output weight.
- Assurance regions type II (ARI): This approach is also based on expert opinion restricting virtual inputs and outputs and can have equal set of input-output weights.
- Absolute weights restrictions: This approach is used to keep all inputs-outputs in the analysis and prevent them from been ignored.

Weight restrictions are usually elicited through expert opinions or monetary considerations and subsequently added as additional constraints to the linear programming model in equation (8). The former approach translates the perceived relative importance of input and output factors into the relation between the corresponding weights. The latter requires that higher weights be placed on more expensive resources and outputs with higher prices (Podinovski, 2004). When weight restrictions are added to any DEA model, such as the one in (6), the resulting efficiency score can only be worse and never better than the efficiency score obtained without any weight restrictions. This means that weight restrictions will also potentially reduce the number of efficient units. Since a poor discriminatory power of DEA is frequently listed as one of its most prominent disadvantages in practical applications, then using weight restrictions will be a much-needed improvement whenever the use of weight restrictions can be justified. Due to the weight restrictions, and since the weights allocated by the model are dependent on the units of measurement, the absence of a normalization stage in the BOD approach causes a problem. That is why it would be too difficult to convert opinions such as "indicator I_1 is twice as important as indicator I_2" into inequalities involving the relevant weights. There are two major ways to deal with this problem:

- normalising the data, which leads to the drawbacks specified in Chapter 3;
- expressing the weights restrictions in terms of "pie-shares" constraints.

Pie-share, S_{jk}, is an idiom in the BOD vocabulary and it designates the share of the performance score e_k made up by a particular individual indicator, as shown in equation (9)

9) $$S_{jk} = b_j y_{jk}$$

Obviously, the sum of all the pie-shares for a unit k matches the performance score for that unit:

10) $$e_k = \sum_j S_{jk}$$

Pie-shares percentage contribution, s_{jk}, to the performance score is calculated as

11) $$s_{jk} = \frac{b_j y_{jk}}{\sum_j b_j y_{jk}}$$

Restrictions to the percentage of pie-shares are now more aligned with the way experts form their judgments and opinions. The simplest way to formulate weight restrictions in this setting is to ask a group of experts for their judgments on the allowable ranges for each pie-share percentage contribution. Those judgments are essentially defining upper and lower limits, l_j and u_j, on the pie-share percentages, and so the constraints will be written in this form:

12) $$l_j \leq s_{jk} \leq u_j$$

Adding these types of restrictions to the model in (8) has been successful in challenging the above-mentioned drawbacks. For example, Vierstraete (2012) suggests a BOD version of the Human Development Index to measure the effectiveness of countries capitalizing their resource to develop the health and education of their populations.

In the next section, we shall propose the Geometric DEA approach to constructing composite indicators, after which we will illustrate the workings of the model and demonstrate its benefits in comparison to the previous classical DEA models. In addition, weight restrictions for G-DEA will also be elaborated.

4.3. Geometric DEA (G-DEA)

G-DEA effectively tackles the problems in constructing composite indicators identified in Chapter 2, which we were able only to partially address with all the models mentioned so far. As already mentioned in section 4.2, G-DEA is just another way of generalizing the model

shown in equation (3). We will start the description of the G-DEA from there and by first looking at some of the theoretical aspects of this method in comparison to the classical DEA.

4.3.1 Theoretical overview of G-DEA in comparison with classical DEA

The G-DEA method was originally introduced by Despic (2012) to assess technical efficiency and it is a variation of the DEA method. Theoretical development of this topic is quite recent and G-DEA has not been used in the construction of composite indicators before.

To illustrate the G-DEA model and to make it easier to compare it with the classical DEA model, we will first reformulate the classical DEA model in (5) using its maxmin formulation:

13) $$e_k = \max_{\substack{\sum a_i=1 \\ \sum b_j=1 \\ a_i,b_j\geq 0}} \min_{p} \frac{\sum_i a_i \frac{x_{ip}}{x_{ik}}}{\sum_j b_j \frac{y_{jp}}{y_{jk}}}$$

Note that even though that we are using the same notations for the weights, a_i, b_j, as in equation (5), those weights are not identical to the weights in equation (5). In equation (13), the weights are unitless while in equation (5) the weights have units. For equation (13) to be fully equivalent to equation (5) the weights, a_i, b_j need to be non-negative and must add up to 1, as indicated under the max function in (13).

Both models, (5) and (13) yield the same efficiency scores as proven by Despic (2004). In contrast to the classical DEA formulation in (5), the DEA model in (13) features relative input values (x_{ip}/x_{ik}) and relative output values (y_{jp}/y_{jk}) instead of inputs and outputs. When these ratios are oriented in a "the larger, the better" form for unit k, then we can talk about relative input strength ($RIS_{jk} = x_{ip}/x_{ik}$) and relative output strength ($ROS_{jk} = y_{jk}/y_{jp}$) of unit k in reference to unit p. The greater these relative strengths are, the better is the performance of the unit k relative to unit p. The classical DEA model, as formulated in (13), can then be described as the product between the weighted arithmetic mean of RIS's and weighted harmonic mean of ROS's of unit k. The G-DEA model directly follows from (13), where instead of using weighted arithmetic mean for RIS's and weighted harmonic mean for ROS's, the

weighted geometric mean is used to aggregate both, RIS's and ROS's. This is illustrated in equation (14):

14) $$e_k = \max_{\substack{\sum a_i=1 \\ \sum b_j=1 \\ a_i,b_j\geq 0}} \min_p \prod_j \left(\frac{y_{jk}}{y_{jp}}\right)^{b_j} \prod_i \left(\frac{x_{ip}}{xik}\right)^{a_i}$$

The min operator in front of these products simply ensures that unit p selected for comparing unit k with is one of the best performers among all the units. The max operator does what it normally does in all DEA models, which is to find the optimal set of weights that will maximize the score of unit k.

Both maxmin models, (13) and (14), can be transformed into their corresponding linear programming form, as shown in Despic (2013). Model (13) translates into:

15) $$\begin{aligned} &\max \quad \omega_k \\ &\textit{subject to} \quad \textstyle\sum_j \beta_j \left(\frac{y_{jp}}{y_{jk}}\right) - \sum_i a_i \left(\frac{x_{ip}}{x_{ik}}\right) \leq 0, \quad \forall p \\ &\qquad\qquad \textstyle\sum_i a_i = 1 \\ &\qquad\qquad \textstyle\sum_j \beta_j = \omega_k \\ &\qquad\qquad a_i, \beta_j \geq 0, \ \ \forall i \ and \ \forall j \end{aligned}$$

In the process of transforming (13) to (15), ω_k was introduced as a new variable, representing efficiency score of unit k, while β_j's were used instead of b_j's and where $\beta_j = b_j \omega_k$. The model (15) corresponds to an input-oriented model. Applying minmax, instead of maxmin, operator to the inverse of the objective function in (13) and subsequently transforming that model into its linear programming form will yield an output-oriented version. This model is shown below.

16) $$\begin{aligned} &\min \quad \omega_k \\ &\textit{subject to} \quad \textstyle\sum_j b_j \left(\frac{y_{jp}}{y_{jk}}\right) - \sum_i \alpha_i \left(\frac{x_{ip}}{x_{ik}}\right) \leq 0, \quad \forall p \\ &\qquad\qquad \textstyle\sum_i \alpha_i = \omega_k \\ &\qquad\qquad \textstyle\sum_j b_j = 1 \\ &\qquad\qquad \alpha_i, b_j \geq 0, \ \ \forall i \ and \ \forall j \end{aligned}$$

The above model is in many ways similar to the input-oriented model but with some important differences. First, ω_k represents the inverse of the efficiency score of unit k. Also, the weights b_j's remained unaffected while what happened to b_j's in the input-oriented model, it now happened to a_i's in this model; they are substituted by α_i's, where $\alpha_i = a_i\omega_k$.

When G-DEA model, as shown in (14), is converted into its linear programming form, then the following model is obtained:

17)
$$\max \quad \theta_k$$
$$\textit{subject to} \quad \sum_j b_j \ln\left(\frac{y_{jp}}{y_{jk}}\right) - \sum_i a_i \ln\left(\frac{x_{ip}}{x_{ik}}\right) + \theta_k \leq 0, \quad \forall p$$
$$\sum_i a_i = 1$$
$$\sum_j b_j = 1$$
$$a_i, b_j \geq 0, \quad \forall i \; and \; \forall j$$

When transforming (14) to (17) there is no need to make any substitutions for weights so that both set of weights, a_i's and b_j's still add up to 1. The efficiency score e_k of unit k in model (17) will be equal to $exp(\theta_k)$. There are many interesting observations that could be made when comparing the model in (17) with the models in (15) and (16). We will limit ourselves to observe only those details that are relevant for this study, namely to those details that make an important difference within the context of composite indicators.

Note that the G-DEA model in (17) has no natural orientation and that both, RIS's and ROS's, are treated in a similar way. This is best seen in a slightly rearranged form of the model (17):

18)
$$\min \quad \varphi_k$$
$$\textit{subject to} \quad \sum_j b_j \ln\left(\frac{y_{jk}}{y_{jp}}\right) + \sum_i a_i \ln\left(\frac{x_{ip}}{x_{ik}}\right) + \varphi_k \geq 0, \quad \forall p$$
$$\sum_i a_i = 1$$
$$\sum_j b_j = 1$$
$$a_i, b_j \geq 0, \quad \forall i \; and \; \forall j$$

Model (18) is better aligned with model (14) in a sense that both RIS's and ROS's are written in the same, the more the better, orientation. Variable φ_k represents the total improvement

needed by unit k to become efficient although, strictly speaking, it is a natural logarithm of that improvement. From the first constraint in (18), it is clear that any improvement in efficiency required by unit k can be arbitrarily split into two parts. One part could be used to improve inputs and the other one to improve outputs of unit k. In other words, if φ_k is split into φ_k^a and φ_k^b in such a way so that $\varphi_k^a + \varphi_k^b = \varphi_k$, then φ_k^a can be used to improve RIS's and φ_k^b to improve ROS's of unit k. This essentially means that each input of unit k would be divided by $exp(\varphi_k^a)$ while each output would be multiplied by $exp(\varphi_k^b)$. Input and output levels obtained in this way are the efficient targets for unit k.

The most important insight here is that G-DEA model treats inputs and outputs in essentially the same way. If we were to invert all the inputs in the original data set (making them output-like in a sense of the orientation), then there would be no difference in how two groups, relating to what we called RIS's and ROS's, are treated by the G-DEA model. At the same time, the efficiency score would be the same as the one obtained using the original input values. No such equivalency in treatment of inputs and outputs exists within the classical DEA model shown in (15) and (16). All this means that G-DEA model is much easier to generalise for the cases when dealing more than two groups of criteria or if dealing with nested criteria spreading on several different hierarchical levels. Being able to seamlessly encapsulate all the criteria, regardless of the complexity of their hierarchical structure, within a single optimisation model is an important advantage of G-DEA over classical DEA models. This will be better illustrated in the next section but for now let us just observe one more important difference between G-DEA and DEA models.

The weights, a_i's, b_j's, α_i's and β_j's in the above models are all unitless. Those weights are in fact identical to pie-shares in BOD models and so the process of incorporating expert judgments and opinions would follow a similar procedure as described in section 4.2.2. Still, there are some important differences between a_i's and b_j's on one side and α_i's and β_j's on the other side. Weights a_i's and b_j's, in all the models above add up to 1, which means that, in

addition of representing pie-shares, they also represent pie-share percentages. The most typical expert judgments and opinions therefore directly translate into simple restrictions on those weights or on their relationships. Weights α_i's and β_j's, on the other hand, add up to either the inverse efficiency score $1/e_k$ of unit k or to its efficiency score e_k, respectively. Converting α_i's and β_j's into pie-share percentages can be done by multiplying each α_i or by dividing each β_j by the efficiency score e_k of unit k. The problem with this conversion is that different units have different efficiency scores meaning that any restrictions imposed on pie-share percentages relating to α_i's and β_j's are going to be specific to the unit being assessed. In other words, if we add restrictions on α_i's or β_j's, then different units will be assessed using different feasible region. This can spoil the strict equity criteria in the relative evaluation of units, which is one of the key advantages claimed by DEA.

The easiest way to avoid the above problem is to avoid setting restrictions on β_j's in the input-oriented model (15) and on α_i's in the output-oriented model (16). While this solution preserves strict equity criteria, it certainly imposes additional limitations in modelling when incorporating experts' judgments and opinions. Yet, the G-DEA model, presented in (17) and (18), does not contain any α_i's and β_j's, which means that this specific limitation never shows up if G-DEA model is used instead of a classical DEA model.

4.3.2 G-DEA model in its BOD form

In section 4.2.1, BOD model was presented as a DEA model adjusted to fit the context of composite indicators. The BOD model itself was presented in equation (8), which follows directly from the DEA model in (6) when all the units are assumed to have a single dummy input equal to 1. If the same assumption of a dummy input is introduced in the G-DEA model in (17) or (18), a similar transformation will occur where there will be no inputs featuring in the model. The model obtained in such a way can be seen as G-DEA model in its BOD form. However, as mentioned in the previous section, the G-DEA model treats inputs and outputs in the same way, which means that in its BOD form we could easily have two, three, or as many

as we want groups of outputs featuring in its main constraints. Equation (19) illustrates G-DEA model in its BOD form when all the outputs are split into three separate groups.

19) $\max \quad \theta_k$

$$\textit{subject to} \quad \sum_i a_i \ln\left(\frac{y_{ip}}{y_{ik}}\right) + \sum_j b_j \ln\left(\frac{y_{jp}}{y_{jk}}\right) + \sum_r c_r \ln\left(\frac{y_{rp}}{y_{rk}}\right) + \theta_k \leq 0, \quad \forall p$$

$$\sum_i a_i = 1$$

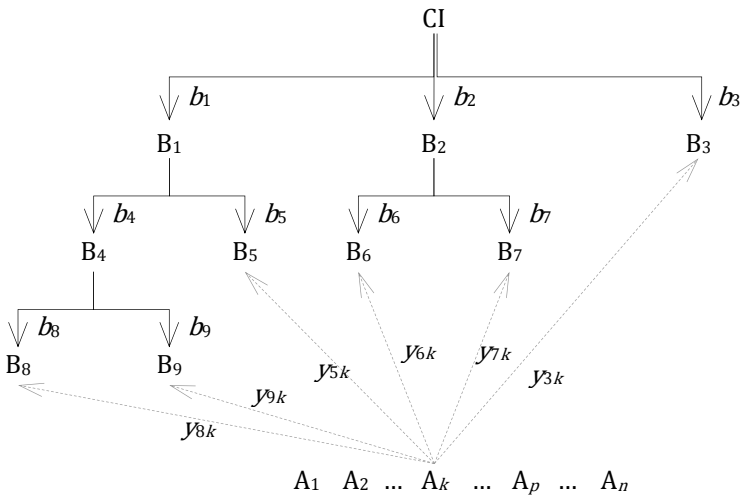
$$\sum_j b_j = 1$$

$$\sum_r c_r = 1$$

$$a_i, b_j, c_r \geq 0, \quad \forall i, \forall j \; and \; \forall r$$

Variables y_{ip}, y_{jp} and y_{rp} are all outputs – they are just being split into three separate groups. Note that the model assumes that all three groups are of equal importance. If that is not the case, then that can be adjusted by setting the appropriate values for $\sum a_i$, $\sum b_j$and $\sum c_r$. The above flexibility also means that we can easily adjust the BOD form of the G-DEA model so that it encapsulates all the outputs relating to the criteria from different levels of a hierarchy in a single model. Complex hierarchies are commonly seen in the context of CIs and this is certainly true for the WEM score. To illustrate the G-DEA model, a simple example of a complex hierarchical structure for composite indicator CI is shown in Figure 12:

Figure 12: An illustration of a complex hierarchical structure for a composite indicator



(Source: Despic, 2013)

The above hierarchy shows how the composite indicator CI is made out of three main sub-indicators B_1, B_2 and B_3. B_1 and B_2, depend on their own sub-indicators: B_4 and B_5 aggregate into B_1 while B_6 and B_7 aggregate into B_2. Finally, B_4 itself is an aggregate of its sub-indicators B_8 and B_9. The indicators that do not split further into their own sub-indicators, also known as end-indicators, are B_3, B_5, B_6, B_7, B_8 and B_9. Any unit k being assessed with respect to CI is directly measured or in some other way evaluated with respect to end-indicators only. Those quantities are represented by the values $y_{3k}, y_{5k}, y_{6k}, y_{7k}, y_{8k}$ and y_{9k} in Figure 12. To address the above example of a hierarchy structure and any subjective judgments that might be elicited by the experts/decision makers, we can extend the G-DEA model in (14). Assuming that all the indicators are of maximising type, the efficiency formulation of G-DEA model is expressed as follows:

$$20) \qquad e_k = \max_{b_j \in B} \min_{p} \left(\left(\left(\frac{y_{8k}}{y_{8p}} \right)^{b_8} \left(\frac{y_{9k}}{y_{9p}} \right)^{b_9} \right)^{b_4} \left(\frac{y_{5k}}{y_{5p}} \right)^{b_5} \right)^{b_1} \left(\left(\frac{y_{6k}}{y_{6p}} \right)^{b_6} \left(\frac{y_{7k}}{y_{7p}} \right)^{b_7} \right)^{b_2} \left(\frac{y_{3k}}{y_{3p}} \right)^{b_3}$$

If any of the indicator scores are input-like oriented, i.e., the less the better, then they will need to be inverted before using them in the model.

We can note that the hierarchy structure in Figure 12 is clearly reflected in the above equation, which includes the condition $b_j \in B$, which represents non-negativity restrictions on all the weights ($b_j \geq 0$) and their normalisation ensuring that the weights of each group of sub-indicators of the same criterion add up to one. In this example, we have: $b_8 + b_9 = 1, b_4 + b_5 = 1, b_6 + b_7 = 1, b_1 + b_2 + b_3 = 1$.

In a similar way, any hierarchical structure can be represented by a suitable G-DEA model, such as the one in (20). Any such model can be then converted into its linear form by substituting the product of local weights b_j's by their corresponding global weight w_j's and then using the same transformation as applied to (14) to reach (17). As for transforming local weights b_j's into global weight w_j's any model such as the one in (20) can be "flattened" by using the

weighted product of the relative strengths of the end-indicators only. For example, when the equation (20) is flattened we obtain the following model:

$$21) \qquad e_k = \max_{w_j \in W} \min_{p} \left(\frac{y_{8k}}{y_{8p}}\right)^{w_8} \left(\frac{y_{9k}}{y_{9p}}\right)^{w_9} \left(\frac{y_{5k}}{y_{5p}}\right)^{w_5} \left(\frac{y_{6k}}{y_{6p}}\right)^{w_6} \left(\frac{y_{7k}}{y_{7p}}\right)^{w_7} \left(\frac{y_{3k}}{y_{3p}}\right)^{w_3}$$

Clearly $w_8 = b_1 b_4 b_8$, $w_9 = b_1 b_4 b_9$, $w_5 = b_1 b_5$, $w_6 = b_2 b_6$, $w_7 = b_2 b_7$ and $w_3 = b_3$. The global weights w_j exists for end-indicators only and due to the conditions imposed on local weights b_j to add up to 1, it will always be the case that the sum of all the global weights is also equal to 1. In other words, if we let the index *j* to go over the end-indicators only, then we have $\sum w_j = 1$. In the above example, taking into account the conditions specified by $b_j \in B$, the following is true:

$$\begin{aligned}
\sum w_j &= b_1 b_4 b_8 + b_1 b_4 b_9 + b_1 b_5 + b_2 b_6 + b_2 b_7 + b_3 \\
&= b_1 b_4 (b_8 + b_9) + b_1 b_5 + b_2 (b_6 + b_7) + b_3 \\
&= b_1 b_4 + b_1 b_5 + b_2 + b_3 \\
&= b_1 (b_4 + b_5) + b_2 + b_3 \\
&= b_1 + b_2 + b_3 \\
&= 1
\end{aligned}$$

In the case of a large numbers of end indicators, such as the WEM score with 133 end-indicators, or in general for any number of end-indicators and any hierarchy, the corresponding G-DEA model can be flattened into the following model:

$$22) \qquad e_k = \max_{\substack{\sum w_j = 1 \\ w_j \geq 0}} \min_{p} \prod_{j \in \{end-indicators\}} \left(\frac{y_{jk}}{y_{jp}}\right)^{w_j}$$

Note that while the local weights $b_j$represent relative importance of the criteria locally with respect to their parent indicator only, the global weights $w_j$represent relative importance of the end-criteria with respect to the main composite indicator. To be able to use this model in practice, it is now necessary to understand how any weight restrictions imposed onto local weights b_j can be transformed into weight restrictions imposed onto global weights w_j.

4.3.3 G-DEA Weights restriction

The easiest way to explain the transformation of weight restrictions imposed on b_j's into the weight restrictions on w_j's is through an illustrative example. Suppose, that for the hierarchy in Figure 12, the following relationship was elicited from experts: $b_3 \geq 2b_2$. To transform this relationship into the equivalent restriction involving global weights, we can notice that $b_3 = w_3$ while b_2 must be equal to $w_6 + w_7$ since $w_6 + w_7 = b_2b_6 + b_2b_7 = b_2(b_6 + b_7) = b_2$. So, the relationship represented by $b_3 \geq 2b_2$ can be substituted by $w_3 \geq 2(w_6 + w_7)$.

In a similar way any relationship that may be specified by experts addressing local weights b_j's can be transformed into their corresponding relationships among the global weights w_j's. While it is theoretically possible for the experts to state their judgments and opinions relating to the end-criteria and their global weights directly, in practice it is much easier to form an opinion relating to local weights. After all, one of the main purposes for developing a hierarchy of criteria for a composite indicator is exactly that: to make the process of subjective judgments manageable and easier to apply in practice.

If experts, on the other hand, simply want to set a lower or an upper limit on a specific local weight, the transformation of such a restriction follows similar reasoning as in the case of the relationship between the local weights. For example, if it is required that $b_1 \geq 10\%$ then this translates into $w_8 + w_9 + w_5 \geq 10\%$ since $w_8 + w_9 + w_5 = b_1b_4b_8 + b_1b_4b_9 + b_1b_5 = b_1b_4 + b_1b_5 = b_1$. In short, the relationship between b_j's and w_j's is relatively straightforward: any local weight b_j of a specific criterion anywhere in the hierarchy is equal to the sum of the global weights w_j of the end-criteria that aggregate into that specific criterion.

From all the above, it is clear that G-DEA can address complex hierarchies with ease and in a manner that is similar to the way multiplicative AHP is applied. Yet, as opposed to multiplicative AHP, there is no need to search for the weights that best fit experts' opinions expressed through pairwise comparisons. Rather, the whole process of eliciting experts'

opinions is now much more flexible. Even without any experts' opinions, the G-DEA model will still work, in which case it will behave much like a classical DEA model where no restrictions on weights are imposed. Still, in the context of composite indicators, some limitations on weights are almost necessity and as long as they come in the form of pairwise comparisons or as upper or lower limits on relative importance of local weights, the G-DEA model can readily accept all such requirements and apply them in the same form for all the units evaluated.

One of the most important things to note here is that G-DEA can accommodate any level of details relating to the relative importance of weights. This means that in practical applications, the effort of eliciting any preferences relating to weights should be equivalent to the perceived relative importance of the criteria in each group. For example, in Figure 12, the group of the criteria from the second level of the hierarchy (sub-criteria of CI: B_1, B_2 and B_3) may be much more important than the group of the criteria from the fourth level (sub-criteria of B_4: B_8 and B_9). If that is the case, then much more effort should be spent in eliciting experts' opinions and judgements in relation to the first group than in relation to the second group. In general, for more complex composite indicators such as WEM score, the range of techniques used to elicit restrictions on weights can go from the simplest approaches such as budget allocation process (BAP) or not imposing any restrictions at all up to the most demanding processes such as completing a full pairwise matrix for a group of criteria as done in a classical AHP way.

Finally, when using the most demanding processes to elicit preferences on weights, such as pairwise comparisons, it needs to be noted that G-DEA offers some extra flexibility even within that process itself. Namely, experts need not to be forced to strictly agree on any specific value when performing pairwise comparisons. In particular, if the process seems to be too expensive or too time consuming to reach consensus, then it is possible to use an interval of values, which all experts may agree with. That interval can be used to set lower and/or upper bounds on the relationship between the local weights in question and then subsequently transformed into the weight restrictions on the corresponding set of global weights and such inserted directly into

the G-DEA model. Another form of flexibility exists even when the experts are pushed to agree on a crisp value for all entries in a pairwise matrix. Namely, the classical AHP procedure does not have to be followed in full and it can be cut short after the entries are verified to satisfy consistency criteria. After that, instead of deriving the best-fit weights for the criteria in the pairwise matrix, the inconsistencies in the entries (if any exists) can be transformed into lower and upper bounds on the weights so that any entry in the pairwise matrix is fully consistent with the ranges defined by those lower and upper bounds.

This particular transformation of uncertainty manifested through the inconsistency among the values in the pairwise matrix into the uncertainties specified by lower and upper bounds on individual weights can be achieved in many ways. One such approach was suggested by Salo and Hämäläinen (1995).

Implementation of the G-DEA model as shown in equation (22) as well as implementation of some of those techniques for eliciting weight restrictions will be further illustrated through the WEM score case study, which is the main topic of the next chapter.

4.4. Summary

DEA is certainly a very promising tool for constructing composite indicators. Its main advantage is that it allows for flexibility of weights and therefore allows the units being assessed to align the weights better to their intrinsic characteristics and motivations. Restriction on weights are for this reason one of the most critical elements of DEA model when used in the context of composite indicators. They simply must be there to ensure that aims and the objectives of the assessor (central government in the case of WEM score) are respected but small degree of flexibility is nevertheless needed to account for differences in the way different units operate.

Other important feature we needed to be concerned with was the one relating to the ability of DEA to deal with complex hierarchies frequently encountered within the context of composite

indicators. Finally, the last but not the least important feature was the nature of aggregation. In Chapter 3 we saw how the additive aggregation, especially in the presence of flexible weights can easily produce abnormal aggregate scores. For this reason, it was necessary to set up a DEA model which is based on multiplicative aggregation. All these concerns can be successfully addressed by the geometric DEA-model, which when applied to the context of CIs becomes something like geometric version of the BOD model.

Chapter 5: Data Analysis and Findings

5.1. Introduction

The methodology currently being applied by the DSG has its limitations. Therefore, a better alternative methodology has to be used for the assessment of website performance from different perspectives such as usability of the results, transparency, fairness, equity, credibility, robustness and reliability. This chapter discusses in detail several methods that provide the DSG with the necessary tools for the construction and practical implementation of a sound WEM methodology. The case study conducted on assessing government website performance is presented, explained and analysed in the sections below. The new proposed methodology known as the geometric data envelopment analysis (G-DEA) will be applied to the Website Excellence Model (WEM) for assessing website performance.

G-DEA was designed as an improvement to the existing composite indicator methodology. This improvement focuses on the choices made in terms of weighting and aggregation methods used, their rationale, and the varying results different methods produce. Through these alterations, departments were compared in terms of their web service attributes, including access, usability, content and policy. In addition, the departments that are doing better on all four attributes and in each attribute individually will be presented. Through this process, the aim is to show the possibility of deriving WEM scores with greater confidence and more meaningful, objective and nuanced end results.

The proposed G-DEA methodology comes as a result of analysing together with DSG experts how different methodologies and aggregation methods progressively improve the process of deriving composite indicators starting from additive, then moving to multiplicative, and finally to the G-DEA type of aggregation. In the sections below, different aspects of the new methodology are examined and compared showing their similarities, differences, advantages and disadvantages. Also, specific examples in ranking departments' performance are shown in order to compare how G-DEA produces results that are easier to agree with by the departments

as well as by the assessor relative to other aggregation and weighting methods that may produce irrational results due to various problems stemming mainly from difficult subjective decisions about the value of weights and an inappropriate aggregation method.

Moreover, the chapter outlines the way in which the weights calculated using the G-DEA methodology produce useable and credible results, while at the same time, demonstrate greater transparency and reliability in the calculation process. In this way, it can be argued that the proposed G-DEA, with its underlying geometric aggregation, offers a more elegant and effective approach to deal with multiple-layer structures. The case study illustrates all of the aforementioned advantages of using G-DEA through its practical implementation for the case of the Website Excellence Model (WEM).

5.2. Deriving WEM scores using the G-DEA approach

An important capability of G-DEA is that it can maintain the hierarchical structure of indicators when that is desirable in an assessment. G-DEA can help in overcoming the shortfalls of the previous method used by the DSG department for deriving WEM scores (as discussed in Chapter 2) as it is an easily applicable and functional tool for creating composite indicators. This point will be illustrated below when it is explained how the G-DEA version of WEM operates. To show the different steps involved, the G-DEA model will be applied to evaluate the website performance of the 19 government departments in Dubai (as mentioned in Chapter 1) that have been assessed previously using the DSG approach. In the effort to gain appreciation of the new method by the DSG experts, we needed to compare the two approaches using the same starting point. Namely, we needed to start from the same data used in constructing DSG's original WEM scores so to be able to make better comparison of the results and gain a better understanding of the reasons for some of the key differences in the set of the final scores.

Applying G-DEA to address the WEM problem will enable us to better depict the advantages of a new G-DEA approach. The conducted analysis involves three main stages, which are pre-

assessment, assessment and post assessment. These stages will be explained in detail below to provide a better understanding on how the methodology can be applied in this context.

5.3. Pre-assessment Stage

As noted earlier in Chapter 2, the development of WEM scores involves the use of subjective judgements entailing uncertainties throughout the process of constructing the composite indicator. The subjectivity and uncertainty in this stage arises from the assigning of weights to the indicators and the selection of the aggregation method. These choices significantly influence the departments' WEM score and should be taken into consideration prior to conducting the assessment. Therefore, the analysis in this section will focus on improving the resulting uncertainties mentioned in the above process.

5.3.1 Uncertainty due to weighting techniques

Due to the subjective nature of judgement followed by the DSG in assigning weights (reflecting the experts' opinions and their experiences in dealing with government departments), the issue of uncertainty and reliability in weights setting surfaced in this case (see Section 2.6.1). Such approach is causing a lot of friction amongst Government departments when the results are announced, thus bringing into the question the accuracy and trustworthiness of the results.

To overcome this issue and to derive more reliable weights leading to higher confidence in the end-results, the weights of all indicators and sub-indicators were elicited through a pairwise comparison using a standard nine-point numerical scale of the AHP framework. AHP has the advantage of eliciting complex and subjective judgements of different experts in a common platform. The reliability of weights comes from the process of the pairwise comparison which allows flexibility to the experts to express their opinion and continuously compare each pair of indicators until reaching the consistency level. Most importantly, by using AHP style of eliciting weights, the degree of subjectivity is significantly reduced compared to the BAP approach used by the DSG experts to derive original WEM scores. In addition, AHP's in-built method of

calculating the inconsistency index can further help to ensure consistency of these judgements, that they are provided with sufficient care and that any error due to negligence is eliminated. The descriptions of the numerical values of the scale have been presented in Chapter 3.

Following a consensual discussion with two experts from DSG, the hierarchical structure constructed was used to draw pairwise comparison matrices using the nine-point scale for the standard AHP technique. This step involves eliciting the pairwise comparison values, which includes calculating the corresponding priority vector (local weights) of the indicators. Examples of the pairwise matrices for the four main indicators (level 2) are shown in Table 9.

Table 9: Pairwise comparison for the indicators in level 2 of the WEM score hierarchy

WEM	Access	Usability	Content	Policy	Local Weights
Access	1	1/3	1/2	4	18%
Usability	3	1	2	6	47%
Content	2	1/2	1	5	29%
Policy	1/4	1/6	1/5	1	6%

The experts initially carried out a comparison by providing crisp relative importance values between each pair of indicators. Based on their opinions, AHP matrices were created for each parent indicator in the WEM hierarchy. With respect to the four main indicators from level 2, the experts clearly indicated that the highest relative importance was given to the indicator reflecting the usability of the website. Usability was given an importance of 3, 2 and 6 times greater than accessibility, content and policy, respectively (see Table 9). The second most important indicator is content, which reflects an importance of 2 and 5 times greater than accessibility and policy, respectively. The policy indicator which indicated the policies stated for the users accessing the website was perceived to be of least importance.

Preference ordering of the four main indicators was in full agreement with the ordering obtained by the original BAP approach. Relative importance (local weights) of these four main indicators in the original approach were presented in Table 4 in Chapter 2 and these were 23%,

34%, 29% and 14% for accessibility, usability, content and policy, respectively. When comparing these values with the ones obtained in Table 9, some significant differences can be observed, especially with respect to the local weight attached to policy indicator for which the local weight dropped from the original 14% down to 6%. DSG experts however agreed that the AHP approach yielded local weights that are more realistic and better reflect their opinion about the relative importance of the four main indicators. The original weights were deemed to be skewed in favour of policy indicator and one of the reasons for this was due to the fact that the BAP weights were allocated to the four indicators using the total of 35% rather than 100% (see Table 4). This basically meant that changing the originally allocated BAP weights by just 1% would be equivalent to the change in almost 3% when the weights are normalised to add up to 100%. At the same time the 5% (out of 35%) allocated towards policy indicator in the original BAP approach was perceived as a small enough percentage and that there was no need to push it down any further.

AHP matrices were constructed for all the parent indicators in the WEM hierarchy but the detailed analysis was given only to the four main indicators shown in Table 9. Once DSG experts agreed that AHP style of eliciting weights generate the weights better aligned with their experts' opinions than using the BAP approach, they were happy to apply AHP for all the parent indicators in the hierarchy. Appendix D shows all the pairwise matrices constructed by the experts while all the calculated local weights are provided in Appendix E and F.

Notice that while we are using AHP methodology to get the experts to think in terms of pairwise comparisons rather than directly allocating the weights using BAP approach, we have no need to follow AHP approach any further than to check on the consistency of the pairwise matrices. The ultimate purpose of starting with the AHP method is only to increase the confidence level in the subjective judgments on weight. Once we are ready to use G-DEA model, the pairwise evaluations in AHP matrices are directly used to form the weight restrictions, which are added to the model. This is done through a code written in MATLAB, where upper and lower bounds

on weights are obtained using the minimum and maximum ratios between two indicators that could be derived from any not fully consistent pairwise matrix. (Fully consistent matrices would yield the same lower and upper bounds resulting in no flexibility on the weights in the G-DEA model). To illustrate this idea using the matrix in Table 9, let us focus on the ratio between usability and policy. In their direct comparison, that ratio is 6. Yet, we could derive this ratio by going indirectly through all the other indicators in the matrix. So, going through accessibility indicator, the ratio between usability and accessibility is 3 while the ratio between accessibility and policy is 4. Hence the ratio between usability and policy must be 12. Doing the same process using content indicator as an intermediate factor, we obtain the value of 10 for the ratio between usability and policy. Due to certain level of inconsistency in specifying the pairwise comparisons (which is acceptable based on the AHP inconsistency index of 0.02), we could argue that the ratio between usability and policy could be anywhere between the lowest value of 6 and the largest value of 12. These values are then immediately transformed into the corresponding lower and upper bound on this ratio within the G-DEA model. In Chapter 4 we explained how these can be easily transformed into the restriction on global weights and added directly to the model.

Across many models developed for eliciting weights, pairwise comparison matrices provide a framework that elicits the preferences from decision makers and have been used in various applications such as education, engineering, government, industry, management, manufacturing, personal, political, social, and sports (William, 2008). However, due to the subjective nature of human judgements as well as the complexity and uncertainty experienced in decision making in the real world, it is sometimes unrealistic to expect and difficult to agree on specific crisp values for some pairwise comparison judgements. An easier and better alternative is to provide interval judgements for some or all of the judgements in a pairwise comparison matrix. Again, just as in the case of inconsistent matrices with crisp values, the

matrices with interval judgments can be directly utilised within the G-DEA model to derive the corresponding weight restrictions.

To incorporate the relative importance in the form of ranges, we have first calculated all the local weights (Appendix E and F) for all the pairwise matrices (Appendix D) and then calculated global weights for all the indicators in the WEM hierarchy. Then, the indicators were rank ordered by their global weights (see Appendix G) so to find out which indicators are the most important (largest global weight) and hence account for the most significant portion of the final WEM score. The first 12 most important indicators were singled out and selected for a more thorough re-evaluation of their pairwise matrices but this time allowing for range values to be used instead of crisp values. The main reason behind using ranges instead of crisp value is to allow experts to be more confident about their subjective judgments. The 12 indicators singled out for re-evaluation are presented in Table 10.

Table 10: List of the most important indicators and their global weights top to bottom

Indicators/Sub-indicators	Global weights (Highest to Lowest)
Usability	0.47
Content	0.29
Access	0.18
U12	0.103
A3	0.098
U1	0.080
U4	0.080
U122	0.065
U7	0.064
C5	0.063
A31	0.062
Policy	0.06

The first 12 indicators from the list were re-evaluated for the following reasons:

1) All the remaining sub-indicators had a maximum of 5% of the global weight, which meant that it was a reasonable cut-off point following the top 12th indicator.

2) In reviewing the percentage differences amongst the list of indicators from top to bottom, a significant relative distance between the 12th and the 13th indicator was apparent.

It is of no surprise that the four main indicators in the second level matter very much, so it would be most appropriate to engage the experts into an AHP-like exercise where full consensus does not have to be reached. Allowing this type of subjective judgments would most likely result in significant savings of both time and effort in the future. An illustration of level 2 re-evaluation of pairwise values is shown in Tables 11 and 12. In our example, this step involved asking the experts to provide pairwise values with a high degree of confidence even if they have to use intervals instead of crisp numbers. The experts were given the flexibility to leave the crisp values drawn initially as long as they were very confident on their selection. In the process, the experts were re-evaluating only the entries at the upper triangle of the matrix. Clearly the values in the lower triangle of the matrix, being inverse of the values in the upper triangle, have the upper and lower values switched between the two matrices.

Table 11: Level 2 Lower bound pairwise matrices

WEM-Lower	Access	Usability	Content	Policy
Access	1	1/3	1/3	3
Usability	3	1	3	7
Content	3	1/3	1	5
Policy	1/3	1/7	1/5	1

Table 12: Level 2 Upper bound pairwise matrices

WEM-Upper	Access	Usability	Content	Policy
Access	1	1/1.5	1	5
Usability	1.5	1	5	9
Content	1	1/5	1	7
Policy	1/5	1/9	1/7	1

Using this approach, where we slowly moved from BAP to interval-based pairwise matrices was necessary to gain better understanding and appreciation by the DSG experts for the G-DEA methodology. In general, this is very significant in the context of an argument on the relative

importance of the indicators, since all the interested parties (DSG experts, assessors and Government departments) will then be able to react to the weights and to suggest modifications whenever they may see appropriate. Following the application of G-DEA methodology in assigning weights, DSG experts realised the significance of this approach and how the resulting model can indeed incorporate their subjective judgments with much greater confidence in comparison to their original approach.

5.3.2 Uncertainty due to aggregation techniques

The DSG have applied a generic and unified measurement scale across all end-indicators for ease of calculation and to avoid the tedious task of normalisation when aggregating the scores. However, the DSG could have applied different measurement scales that cater to the distinctive nature of the indicators without the concern of normalising the data by applying the G-DEA approach. This is due to the DEA feature of "units invariance", which makes the normalisation stage redundant as mentioned in Chapter 4. Unfortunately, despite the fact that there is a clear possibility to use more appropriate measurement scale better suited to the intrinsic nature of the measured indicators, for the verification and comparison purposes we had to apply the new methodology using the same scores as assigned by the assessor within the original WEM score framework. This proved to be a rather complex problem, especially since this rather arbitrary scale 0, 1, 2 and 3, as used in the original framework, contained the values of 0, which is not a suitable value to use for multiplicative type of aggregation.

As described in Chapter 2, the WEM score is a composite indicator that summarises a weighted average in a single number. WEM consists of 4 levels, with 4 indicators on the 2nd level, 32 indicators on the 3rd level and 133 indicators on the 4th level. The analysis in this section examines different aggregation approaches so to provide a smooth departure from the original WEM scoring methodology by first moving to additive aggregation but based on the newly derived weights from AHP matrices, followed by multiplicative aggregation using the same AHP-based weight, and finally G-DEA that uses AHP comparisons directly in the form of weight

restrictions rather than the weights that may be derived through AHP. Through this process of deriving the scores using different approaches, the resulting WEM scores are evaluated and compared and any large discrepancies are analysed in detail. This was necessary for the DSG expert to better understand where and why the differences in scores occur.

To compare DSG original score with new methods, local weights derived from AHP matrices were used for additive and multiplicative aggregation. Although the DSG applied weighted average sum known as additive aggregation, it differs from the additive WEM score (A-WEM) we derived due to the fact that we were using new set of local weights. Recall from the discussion in Chapter 3 that an undesirable feature of the A-WEM is its compensability nature of aggregation, wherein the poor performance of some indicators can be compensated by the sufficiently high values of the other indicators. Thus, A-WEM will not entirely reflect the desired performance of a department across all of its individual indicators.

The shortcomings of the A-WEM approach can be partially overcome by using multiplicative aggregation for the WEM score (M-WEM). As explained in Chapter 3, multiplicative aggregation is a simple method that involves a less compensatory approach than the additive aggregation method in which departments with low scores in some indicators would be better represented with an A-WEM aggregation. That is, an increase in an indicator value would have higher marginal utility on the composite indicator if the indicator value is low.

To demonstrate the difference between A-WEM and M-WEM methods, a comparison between different departments was conducted to show how these approaches are calculated. The former method is essentially a calculation of the weighted average of a set of values, and, predictably, when there is a significant difference between some numbers this becomes evened out. In a M-WEM aggregation, however, large differences between numbers, particularly values of zero, become very significant, distorting the final values. Thus, we cannot really use this type of scoring if an indicator has a zero value. The decision to use multiplicative aggregation for WEM scores means that zero score cannot be used when scoring the departments on the

criteria unless some indicators are seen as critical and if it makes sense not to allow the departments who "fail" on this critical element to even compete for the reward. This was not a problem in the case of DSG original WEM scoring methodology as they were able to derive the scores through weighted average. Although it is likely that a particular department might, for example, not carry out a specific activity such as content development for the website and hence have no rating on that indicator. In such an instance, one would be dealing with a department that is different from the others, and, some would argue, that such practice does not constitute a departmental website. One would have to assess such departments' websites separately.

To resolve the problem of zero scores when M-WEM is used, a sensitivity analysis of M-WEM aggregation with a series of different scores close to a zero value was conducted to show how this is a practical solution that minimise distortions relative to original values. In all the other cases, whatever the lowest score is to be selected (such as 1 on scale from 1 to 100), the Dubai Government, DSG and all the departments need to be comfortable and agreeable with the level of penalty and level of detrimental effect such a low score may have on the overall score. Choosing the lowest score for geometric aggregation is not the matter of being correct or being scientific; it is the matter of comfort and the thing to be agreed on by all the interested party, where the main issue to consider is how much an unbalanced performance is penalized versus a more balanced performance. In the following sub-section, we will investigate what kind of suggestion for the DSG and the departments would be the least painful to accept and get used to, i.e., the one that would be the least different from what they may intuitively expect due to their habit (and, indeed, due to the habit of all of us) to think in terms of additive aggregation. This fundamental difference between A-WEM and M-WEM aggregation methods will later on help us better appreciate the differences in scores obtained using G-DEA approach.

The problem of replacing zero scores

DSG experts assign a score of zero to the indicator that is not implemented or not available (as shown in Chapter 2. When using geometric weighted means (multiplicative aggregation), zeros in the data set are problematic. Indeed, one of the main multiplicative method requirements is that all values need to be free of zeros and negative values. To understand this requirement, let us consider the following five data sets: (100, 100, 100, 4), (100, 100, 100, 3), (100, 100, 100, 2), (100, 100, 100, 1) and (100, 100, 100, 0). When applying an additive aggregation for each data set, we will get the following results: 76.00, 75.75, 75.50, 75.25 and 75.00, respectively. As expected, there is no significant difference between all data sets despite the presence of a zero value in the last data set. In a case where multiplicative aggregation for the same data set is applied, the following results are generated: 44.72, 41.62, 37.61, 31.62 and 0, respectively. Therefore, in spite of the fact that the geometric mean and the arithmetic mean both decrease, the former does so in greater increments, and most importantly jumps down to zero once one of the values is zero. Therefore, we can see that a single zero value overpowers the other indicators, to the point where they do not matter at all. The geometric mean will always be zero in this case, whether or not the data set contained large numbers.

To overcome such challenges when dealing with M-WEM, we first need to understand the reasoning behind the DSG's allocation of a value of zero to the end-indicators. To approach this question, it is important to understand the intended representation of the end-indicators. In this case study all the zero values have been revised and a sensitivity analysis has been performed as shown in Table 13. This is to check whether the zero values can successfully be replaced with an alternative small value such as 0.00001, 0.001, 0.01, 0.1. These different values were arbitrarily selected in the analysis to see what happens to the scores and rankings under the different approaches when replacing the zero value. This in turn will allow the selection of the "safest choice" for the replacement of zero. In this way, we can obtain M-WEM score that can be more readily compared to A-WEM scores.

Table 13: Sensitivity analysis of replacing zero values to calculate the M-WEM scores.

	"0.1"		"0.01"		"0.001"		"0.00001"	
	Score	Rank	Score	Rank	Score	Rank	Score	Rank
Dept. 1	97%	3	97%	3	96%	3	94%	3
Dept. 2	48%	18	25%	18	13%	18	3%	18
Dept. 3	90%	7	88%	5	86%	5	81%	5
Dept. 4	100%	1	100%	1	100%	2	98%	2
Dept. 5	71%	15	57%	15	46%	15	30%	15
Dept. 6	89%	8	81%	8	73%	8	59%	9
Dept. 7	41%	19	16%	19	7%	19	1%	19
Dept. 8	69%	16	51%	16	38%	16	21%	16
Dept. 9	74%	12	60%	14	49%	14	31%	14
Dept. 10	99%	2	99%	2	100%	1	100%	1
Dept. 11	74%	13	67%	12	61%	12	50%	12
Dept. 12	94%	4	92%	4	89%	4	83%	4
Dept. 13	88%	9	76%	9	65%	10	48%	13
Dept. 14	91%	6	84%	7	78%	7	65%	7
Dept. 15	93%	5	87%	6	82%	6	72%	6
Dept. 16	77%	10	74%	10	70%	9	64%	8
Dept. 17	50%	17	31%	17	19%	17	7%	17
Dept. 18	71%	14	66%	13	61%	13	51%	11
Dept. 19	76%	11	70%	11	65%	11	54%	10

The sensitivity analysis was conducted as follows:

- The value of 0.01 represents 1% of the normalised score that is considered very close to 0 from a practical perspective. 1% is probably a good idea if the score scale from 0 to 100 which is considered as a good scale to use for whatever is being assessed. This is typical for the applications where the scores are expressed in percentages since that is essentially the same as using the scale from 0 to 100. In that case, using 1% instead of 0, means that the worst score is 1 and the best score is 100. WEM framework was traditionally using percentages between 0 and 100 (albeit only four different values: 0%, 33%, 66% and 100%) and it was very natural to test the effects of using 1 (1%) as the worst score instead of 0. As we see, this did have some effects and some departments' rankings went down (such as departments 9, 5 and 13) while some departments' rankings of course went up (such as departments 3, 11 and 18). However, the changes in rankings were not extreme so it looks that this kind of replacement of 0% by 1% could be reasonably well accepted by

all the assessors and all the assessed. However, to better understand the changes, fluctuations and perturbations, we have investigated also what happens if 0% is substituted by 10%, by 0.1% and 0.001% (completely arbitrarily chosen on both sides from 1% just to see the differences and to compare the changes

- It was not advisable to replace the 0 score by the values 0.001 and 0.00001 as high fluctuation was observed in both the scores and rankings. It is apparent that the smaller the values used to replace 0, the lower the overall scores across most of the departments except department 10 wherein the scores have shown minor improvement as has its ranking. [This particular change for department 10 clearly indicates that its scores over all indicators are better balances than the scores of the department 4. Both departments were the top two departments under all scoring methodology and department 4 just slightly better than department 10 when additive aggregation is used]. Moreover, if we observe the ranking of all departments where 0.001 and 0.00001 were used to replace 0 score, there was an improvement in ranking for departments 16, 18 and 19, although their respective overall scores have decreased which is out of the norm in comparison to the other departments. Also, a big change in ranking is apparent for department 13 in comparison to other departments.

- The value of 0.1 represents 10% of the normalised score which is not considered very close to 0 from a practical perspective. The selection of 10% may be a good idea if a natural scale of measurement is from 0 to 10, in which case 10% means that the worst score is 1 and the best score is 10 – many real word application use this kind of scale anyway even if they do not have problem of using geometric aggregation. It is also apparent from the above table, that the overall scores have increased, especially the low scores in comparison to the scores that were initially high when using the other alternative values (0.01, 0.001 and 0.00001). For example, the scores for departments 2 and 7 increased (almost doubled), but the

ranking did not change and remained the lowest (18 and 19, respectively). Accordingly, we can deduce that 0.1 might not be the most appropriate value to replace the 0 score.

In summary, the sensitivity analysis is done to see which value when used instead of 0 has the smallest potential to be rejected by the people involved, i.e., the value which will create the least amount of psychological friction when the results are seen. Therefore, 1% could be considered as the most natural choice to use (given the fact that the scores used for WEM were in percentages) and it also looks that the resulting "disturbances" are not too intensive. While 10% creates even fewer intensive "disturbances" but there is no big difference between disturbances created by 1% and 10%. So, it may actually be difficult to decide between 1% and 10% - they both create low disturbances. However, 1% is more natural choice given the scale used, while 10% creates less disturbances and the lowest overall scores generated (such as 41%, 48% and 50%) are perhaps psychologically easier to accept than the ones produced when using 1% instead of 0% (the lowest score here are 16%, 25% and 31%). Using 0.1% or 0.001% values instead of 0%, on the other hand, creates larger disturbances and also creates some scores that are possibly "unwanted" by the assessed departments, such as the overall scores of 1%, 3% and 7% in case of using 0.001% as the replacement for 0% or the overall scores of 7%, 13% and 19% in case of using 0.1% as the replacement for 0%.

Collectively with the DSG experts, it has been agreed that the above choice indeed resolves the issue. In terms of percentages, we can see from a practical purpose that a value of 1% is very close to 0%. Thus, it can be used without having a significant impact on the overall result (as we can see from the above table that the absolute values of the scores are not as important as their relative values). Such a minor change in the calculation process shows that the DSG must develop a new measurement scale for evaluating end-indicators, free of zeros and negative values. With such a modified data set and by using the M-WEM model, the issues faced using zero values will no longer arise. For the purposes of comparing the scores using different aggregation and original scores used by the assessor in the original WEM framework, M-WEM

scores were calculated using 0.01 instead of 0 scores. Table 14 shows a comparison among all departments scores and ranking using A-WEM and M-WEM approaches.

Table 14: A-WEM vs. M-WEM scores and rank

Department	A-WEM Scores	M-WEM Scores	A-WEM Rank	M-WEM Rank
Department 1	97%	97%	3	3
Department 2	63%	25%	17	18
Department 3	92%	88%	9	5
Department 4	100%	100%	1	1
Department 5	81%	57%	13	15
Department 6	93%	81%	8	8
Department 7	57%	16%	19	19
Department 8	81%	51%	15	16
Department 9	84%	60%	11	14
Department 10	100%	99%	2	2
Department 11	81%	67%	14	12
Department 12	97%	92%	4	4
Department 13	95%	76%	7	9
Department 14	96%	84%	6	7
Department 15	97%	87%	5	6
Department 16	84%	74%	10	10
Department 17	62%	31%	18	17
Department 18	79%	66%	16	13
Department 19	82%	70%	12	11

The next section compares the scores derived from the A-WEM and M-WEM approaches for different departments.

Comparison between A-WEM and M-WEM scores

Further investigations were conducted to better understand what gave rise to the differences between the scores. We can clearly notice a reduction in all the departments' scores when moving from the A-WEM approach to the M-WEM as shown in Table 14. The scores will always decrease when moving from A-WEM to M-WEM since the weighted geometric mean is always smaller than the weighted arithmetic mean (as mentioned in Chapter 3). The intensity of the reduction is mainly dependent on the degree of unbalance in the sub-indicators' scores. If the scores of the sub-indicators were all equal, then there would be no reduction at all when moving from one approach to another. The larger the variations amongst the scores being aggregated, the larger the reduction will be in the aggregate scores when moving from A-WEM

to M-WEM. This observation does not apply to the ranking order wherein some departments have experienced an improvement in ranking; some have had a reduction whilst a few ranked the same. To ease the concerns of DSG experts, it was necessary to dig deep into the sources of any significant difference in scores and/or rankings.

A comparison was carried out between departments, 5 and 18 which experienced significantly different drops in scores. Under A-WEM, they both have about 80% score. Yet, when moving from A-WEM to G-WEM, department 5's score was reduced by 34% and department 18's score by 23%. This has also caused different directions in change of their ranking under the different scoring approaches. Under A-WEM score, departments 5 and 18 ranked 13th and 16th respectively while under M-WEM score, department 5 falls by 2 whilst department 18 rises by 3 places, thus ranking 15th and 13th respectively. If we look closely at their scores for the four main indicators, then we can see (Table 15) that the main source of differences is the accessibility indicator, where the weighted reduction for this indicator is 9.8% for department 5 versus a mere 0.8% of weighted reduction for department 18.

Table 15: Comparison between departments 5 and 18 at level 2 scores: A-WEM vs. M-WEM.

Department 18						
Indicators	A-WEM weights	A-WEM Scores	M-WEM Scores	M-WEM weights	Reduction	Weighted Reduction
Access	**0.178**	**93.8%**	**89.5%**	**0.175**	**4.6%**	**0.8%**
Usability	0.474	60.7%	50.2%	0.476	17.2%	8.2%
Content	0.288	75.0%	50.1%	0.290	33.2%	9.6%
Policy	0.060	77.4%	48.9%	0.059	36.7%	2.2%
Level 1 Score		**71.7%**	**55.5%**		**22.6%**	
Department 5						
Indicators	A-WEM weights	A-WEM Scores	M-WEM Scores	M-WEM weights	Reduction	Weighted Reduction
Access	**0.178**	**74.0%**	**33.3%**	**0.175**	**55.1%**	**9.8%**
Usability	0.475	70.9%	51.7%	0.476	27.1%	12.8%
Content	0.288	75.0%	52.2%	0.290	30.4%	8.8%
Policy	0.060	88.5%	58.8%	0.059	33.6%	2.0%
Level 1 Score		**73.7%**	**48.3%**		**34.4%**	

Since the accessibility indicator was the main source of the differences in scores for departments 5 and 18, further investigation into the accessibility's sub-indicators in level 3 has to be conducted. We can infer from Table 16 that the main source of the differences relates to the sub-indicator A3. The following was observed: a weighted reduction of 38.8% representing a significant reduction for department 5 versus a weighted reduction of only 0.4% for department 18.

Table 16: Comparison between departments 5 and 18 at level 3 scores: A-WEM vs. M-WEM

Department 18						
Indicators	A-WEM weights	A-WEM Scores	M-WEM Scores	M-WEM weights	Reduction	Weighted Reduction
A1	0.117	100.0%	100.0%	0.113	0.0%	0.0%
A2	0.173	100.0%	100.0%	0.171	0.0%	0.0%
A3	0.551	**96.5%**	**95.8%**	0.557	0.6%	**0.4%**
A4	0.104	95.8%	95.1%	0.104	0.8%	0.1%
A5	0.056	31.9%	22.5%	0.055	29.5%	1.6%
L2 Score		**93.8%**	**89.5%**		**4.6%**	
Department 5						
Indicators	A-WEM weights	A-WEM Scores	M-WEM Scores	M-WEM weights	Reduction	Weighted Reduction
A1	0.117	79.0%	38.0%	0.113	51.9%	6.0%
A2	0.173	100.0%	100.0%	0.171	0.0%	0.0%
A3	0.551	**63.3%**	**18.8%**	0.557	70.3%	**38.8%**
A4	0.104	100.0%	100.0%	0.104	0.0%	0.0%
A5	0.056	40.3%	33.0%	0.055	18.0%	1.0%
L2 Score		**74.0%**	**33.3%**		**55.1%**	

To identify why the sub-indicator A3 is the main source of difference in scores, a more detailed analysis into its respective level 4 end-indicators was conducted. Department 18's scores for the sub-indicators A31, A32, A33 are 3, 3 and 2 respectively under the DSG approach, which translate into 100%, 100% and 67%, respectively when normalised. The department's scores generate very similar weighted averages for both A-WEM and M-WEM case: 96.5% vs. 95.8% respectively and the overall reduction is only 0.6%.

On the other hand, department 5's scores for the same sub-indicators are very unbalanced. The scores for sub-indicators A31, A32, A33 are 3, 0 and 0, respectively under the DSG approach,

and have been normalised as 100%, 0%, 0% for A-WEM and as 100%, 1% and 1% for M-WEM. In this case, we can clearly notice the significant differences in the weighted average between A-WEM and M-WEM; 63.6% for A-WEM and only 18.8% for M-WEM, which represents a reduction of 70.3%. Thus, we can deduce that this is the main cause for the significant change in ranking between department 18 and department 5 when moving from A-WEM to M-WEM.

Table 17: Comparison between departments 5 and 18 at level 4 scores: A-WEM vs. M-WEM

Department 18						
Indicators	A-WEM weights	A-WEM Scores	M-WEM Scores	M-WEM weights	Reduction	Weighted Reduction
A31	0.633	100.0%	100.0%	0.637	N/A	N/A
A32	0.260	100.0%	100.0%	0.258	N/A	N/A
A33	0.106	66.7%	66.7%	0.105	N/A	N/A
L3 Score		**96.5%**	**95.8%**		**0.6%**	
Department 5						
Indicators	A-WEM weights	A-WEM Scores	M-WEM Scores	M-WEM weights	Reduction	Weighted Reduction
A31	0.633	100.0%	100.0%	0.637	N/A	N/A
A32	0.260	0.0%	1.0%	0.258	N/A	N/A
A33	0.106	0.0%	1.0%	0.105	N/A	N/A
L3 Score		**63.3%**	**18.8%**		**70.3%**	

A quick look at Appendix C will reveal that those indicators A31, A32 and A33 correspond to the compatibility of the web site with different browsers. The question that now needed to be answered by the DSG experts was "Which of the two aggregation methods better reflects the reality?" and thus more fitting to be used as a more appropriate aggregation method for DSGs' practical context and case study. DSG experts were asked to select a score (out of 100) for department 5 as an example based on their objective in assessing government department websites:

- What score (out of 100) for the overall accessibility/compatibility should be assigned to the website, which is perfectly compatible with Internet Explorer but not at all with Chrome, Safari and Firefox?

As mentioned in Chapter 2, the DSG clearly stated that their objective in assessing government department websites is to improve performance relative to all WEM indicators and not having scores where one indicator scored very high and the other indicators scored very low. Despite the fact that compatibility with Internet Explorer (A31 indicator) was assigned significantly higher weight, the DSG experts were more comfortable with the overall A3 score of 18.8% representing the multiplicative aggregation approach (M-WEM) which was more fitting for their practical purposes rather than the score of 63.6% representing the additive aggregation approach (A-WEM) which was more in line with their current approach.

In summary, due to the compensatory power feature of the additive aggregation approach, poor performance in some indicators could be compensated to a greater extent by good performance elsewhere. This drawback is overcome under the multiplicative aggregation wherein consistent performance is rewarded to a greater extent across the different indicators. Having said this, excelling in a few indicators will not necessarily imply a higher score and ranking.

In subsequent sections, it will be shown how G-DEA can be used to improve WEM scores, in addition to its advantages over the previously mentioned approaches. Prior to applying the G-DEA model to WEM scores, it is worth noting that the original WEM methodology applied has several drawbacks; one of which is using compliance levels of 0%, 33%, 67% and 100% and these were subsequently translated into scores of 0, 1, 2 and 3, respectively. Such practice has been applied to all the previous aggregation approaches. However, with the proposed G-DEA approach there is no need for this conversion and any percentage compliance can be taken into the model as a raw score, except of course, for the score of 0%, which would be replaced by 1% (0.01). This essentially means that there is no difference made between those who scored 0% and those who scored 1% as used in the M-WEM. Accordingly, when running G-DEA using raw scores (0, 1, 2 or 3 where 3 is the highest score among all the raw scores), then 0.01 (which simply meant 1% in the previous setting) will be equivalent to 0.03 when those raw scores are

used. Thus, all zeros will be substituted with value of 0.03 so that the original scale 0, 1, 2 and 3 is converted into 0.03, 1, 2 and 3 scale.

Accordingly, we derived overall scores for each department using the G-DEA Model. The code for G-DEA was run in MATLAB. The hierarchy structure was built based on the case in hand. For the next step, it was necessary to provide the evaluation of all departments with respect to all the end-indicators of the AHP value tree. Once both steps (constraints produced through the evaluation of the set of pairwise comparison matrices and values of departments with respect to the end indicators) have been successfully completed, we run the DEA model in MATLAB to calculate the final efficiency scores of all departments together with the weights used by each department using the G-DEA methodology.

However, as mentioned in the previous chapter, the weighting method in G-DEA can be adapted using different approaches (weights restrictions), starting from very strict to complete freedom. One approach is to derive the individual indicator weights using crisp values as in the case of A-WEM and M-WEM. G-DEA will be applied and compared into two different applications of WEM: G-DEA wherein experts agree on preferences (EAP) and G-DEA wherein experts disagree on preferences (EDP). The scores and ranks for G-DEA (EAP) and G-DEA (EDP) are presented in Table 18.

In G-DEA (EAP), the experts use pairwise matrices to create crisp values for the weights and in G-DEA (EDP), the experts use the same pairwise matrices to create interval values for the weights to allow ranges rather than crisp values. The above two applications were run in MATLAB using two different codes but essentially using one model (G-DEA). The only difference is that G-DEA (EAP) generates intervals weights based on the internal inconsistency of the matrices initially produced, while G-DEA (EDP) allows flexibility for the experts to agree on a range rather than crisp values.

The G-DEA (EAP) approach has some practical issues in the allocation of the weights; nevertheless, it allows keeping the hierarchical structure of the data into account. In this case study, WEM score has a relatively high number of indicators and the data structure is fairly complex; thus, it is not easy for the experts to give exact opinions (i.e. crisp values) for the weights. It is also clear that many of the indicators' weights within the same category have the same value (i.e. equal weighting) (See appendix B or C). This can be interpreted as a "safe choice" from the experts, as it is common practice to give equal weights when in doubt about the relative importance of the two alternatives. This downside of G-DEA (EAP) can be avoided with the use of the G-DEA (EDP) approach that allows experts to give judgements in the form of ranges rather than crisp values, and that easily incorporates the hierarchical structure of the data. The following example for department 18 will clarify this case further.

Table 18: G-DEA scores comparison derived from using crisp and interval values

Department	G-DEA (EAP)	Rank	G-DEA (EDP)	Rank
Department 1	95.81%	3	100.00%	1
Department 2	37.94%	18	45.09%	18
Department 3	93.69%	5	97.93%	8
Department 4	99.84%	2	100.00%	1
Department 5	72.97%	14	83.12%	15
Department 6	86.85%	10	93.91%	11
Department 7	24.91%	19	31.50%	19
Department 8	65.28%	16	81.73%	16
Department 9	70.30%	15	88.14%	14
Department 10	100.00%	1	100.00%	1
Department 11	84.44%	12	92.55%	12
Department 12	94.48%	4	100.00%	1
Department 13	80.69%	13	96.98%	9
Department 14	90.14%	8	96.71%	10
Department 15	92.68%	7	100.00%	1
Department 16	93.15%	6	100.00%	1
Department 17	46.13%	17	61.14%	17
Department 18	89.01%	9	100.00%	1
Department 19	86.23%	11	91.16%	13

We can observe that department 18's score reduced significantly when moving from A-WEM to G-DEA (EDP). It was ranked 16 under A-WEM. Yet, its overall score went all the way to 100%

under G-DEA (EDP) becoming one of the 7 departments with this highest score. The main reason for the difference in scores is due to the subjectivity in the experts' judgements and the agreement on crisp values using pairwise matrices, where typically it is unrealistic and infeasible to obtain exact judgements as in the case of A-WEM, M-WEM and G-DEA (EAP) as shown in Table 19.

In contrast, the G-DEA (EDP) approach allows flexibility through containing experts' uncertain judgements within interval importance values for each indicator rather than the selection of crisp values. Moreover, G-DEA (EDP) selects the most optimal weight for each indicator within the interval importance values obtained by the experts for department 18. Such a flexible approach coupled with the fact that department 18's scores across individual indicators are better balanced than other departments' scores, allowed department 18 to be one of the seven top performing departments with a score of 100%.

In comparing the different approaches, A-WEM & G-DEA (EDP), the most significant changes can be observed in the elicited weights for the usability and content indicators. Under the A-WEM approach, the weights for the usability and the content indicators are 47% and 29% respectively; while the G-DEA (EDP) generated weights are 70% for the usability and 14% for the content. It is clear that a significant change in weights occurred, representing an increase of 68% in the usability and a decrease of 49% in the content when moving from A-WEM to G-DEA (EDP). This significant difference might create frictions and disagreements when the results are announced as the accuracy, reliability and validity of weights might be undermined.

Table 19: Department 18 level 2 weights and ranks under different scoring methods

Indicators	A-WEM	M-WEM	G-DEA (EAP)	G-DEA (EDP)	Min G-DEA (EDP)	Max G-DEA (EDP)
Access	18%	18%	17%	14%	5%	35%
Usability	47%	48%	52%	70%	49%	77%
Content	29%	29%	26%	14%	11%	27%
Policy	6%	6%	4%	2%	2%	5%
Rank	16	13	9	Between 1-7		

Once the pre-assessment was concluded, we moved on to the next phase of the analysis, i.e., the assessment stage. This stage is explained below with the use of examples.

5.4. Assessment stage

The process of deriving WEM scores during the assessment stage involves the use of subjective judgement steps entailing uncertainties, additional efforts and time consumption on agreeing on the final results. The above uncertainties arise from the choice of the measurement scale seen most appropriate by the assessors during the following steps: the individual assessors' assessment, the collective assessors' consensus meeting and finally the team leader review as discussed in Chapter 2. All these assessment steps may influence the departments' WEM score and should be taken into consideration when conducting the assessment. Therefore, the analysis in this section will focus on reducing the resulting uncertainties and improving the departments' WEM score.

5.4.1 Uncertainty in measurement scale

DSG experts have evaluated 133 end-indicators subjectively using a categorical scale of 0, 1, 2 and 3, despite the fact that there were many end-indicators that perhaps could have been measured using different scales depending on the nature of the indicator. Table 22 shows an example of the end-indicators A21, A51, and C54, that could have been measured using one or more of the following scales: interval, binary and ratio.

The end-indicator A21 can be assessed using a binary scale of 0 or 1 or a scale of 0, 1 and 2 instead of 0, 1, 2 and 3. In assessing this indicator, the assessor has been given keywords representing the departments' scope of work to access the website from a search engine, e.g., the keyword "accident" should lead to the police department's website. In one scenario, when using a scale of 0 or 1, it is possible for a score of 1 to be given if the website is accessible using any of the keywords provided and a score of 0 otherwise. In another case, when using a scale of 0,1 and 2, it could be argued that 0 and 2 would reflect the same as the above for accessibility

to the website or not. However, assigning a score of 1 as a middle ground would be in the case where it is possible to access the website but in the situation where the link is not found on the first page of the search engine.

In the case of end-indicator A51, it reflects the availability of five different attributes: name, description, size, format and date. This indicator could be assessed using a scale of 0, 1, 2, 3, 4 and 5 rather than 0, 1, 2 and 3 as what is currently being applied in WEM. In the case where none of the 5 attributes are available, a score of 0 will be given. The remaining scores 1, 2, 3, 4 and 5 would represent the availability of 1 or 2 or 3 or 4 or 5 attributes, respectively. Therefore, scales of 0, 1, 2 and 3 for such an indicator will render departments incomparable.

Moreover, the end-indicator C54 (Service description: a brief description about the service) which comes under the C5 indicator (Provide Sufficient Information about Government department services and eServices) can be assessed using a ratio scale instead of 0, 1, 2 and 3. In such a case, one could inquire about the percentage completion of the brief description about the service.

Table 20: WEM end-indicators and their description

End-indicator	Description
A21	Keyword 1 English: Provides a Quick Access to Website from a Search Engine
A51	File 1: File attributes are available (name, description, size, format, date)
C54	Service description: a brief description about the service

Thus, using only a scale of 0, 1, 2 and 3 is adding a superficial level of uncertainty into the model where in fact experts can be more accurate and reliable on the assessment outcomes by applying more precise measurement scales. Thus, a generic scale is not suitable for all indicators equally. The main problem resulting from this is the probable inaccuracy in the assessment and the extra time needed for assessors to agree on the scores. In fact, each of the 133 end-indicators deserves their own most appropriate scale in order for the assessors to feel comfortable when assigning the scores and collectively agree on the scores instantly. This will

save time, effort and cost for the whole assessment stage. Therefore, there are some improvements that the DSG could have applied to the assessment process to add clarity, accuracy and avoid the current conflicts arsing amongst assessors when assigning scores due to the vague and inappropriate measurement scales.

Moreover, it is worth mentioning that with such categorical scales, it is difficult to follow compliance completion increases over time. For example, a department website may have increased its compliance from 35% to 60% during the two years between evaluations; however, the score will remain a 1. This exclusion of transformation information will affect the validity of the results and will not show a proper representation of the changes that have taken place over time. Thus, the process of selecting a measurement scale should be based on the nature of the indicator and not enforcing all the indicators on a unified measurement scale.

5.5. Post assessment stage

WEM scores should be able to communicate an overall picture of the departments' scores to the targeted audience (decision-makers and practitioners) in a timely and accurate manner. DSG have followed an exhaustive approach in displaying the WEM scores, which may be too detailed, not visually appealing and missing important information such as the indicators' weights. The current WEM score report consists of three main sections for each level 2 indicator, which are: the respective sub-indicators' average scores, the sub-indicators' compliance percentage and government departments' scores for that particular indicator reflected in three different figures. Such structure is not ideal and obscures critical information, i.e., the indicators' weights are not reflected in the report for decision makers to act upon.

The proposed visualisation of the scores includes an interactive figure that communicates the most important information. This information would show the department's overall score, scores obtained on each of the four main indicators with their G-DEA (EDP) optimal weight within the ranges of lower and upper values in comparison to the rest of the departments in

one single dashboard. If the assigned weight to any of the four indicators were to be manually adjusted (high or low), it will be easy for the departments to see that their new scores will always be less than their G-DEA (EDP) optimal score. This information could be communicated in a concise, clear, accurate and appealing manner through a well-designed web-based dashboard.

One aspect of the third objective of this study was to finely tune the balance between credibility of the scores derived on one end and ease of comprehensions and their utilisation by the departments assessed on the other end. The initial vision was that this could be done through a well-designed web-based dashboard. However, due to a number of practical concerns relating to the switch from the current methodology to the one we propose here, this objective felt out of scope of this research at this point of time and it was left to be part of the future work. Difficulties of practical implementation were underestimated initially and only after we went through the process of getting the experts to appreciate the value of the new methodology, it was clear that implementing the new methodology requires this to be done in stages. More details about the observed difficulties in implementing G-DEA methodology is given in the final chapter of the thesis.

5.6. Summary

One of the main lessons in the effort to construct the scoring methodology, which would be the best fitting for the specific context of the problem, was that the road towards acceptance and implementation of a new methodology is never as smooth and as easy as theoretical aspects of that methodology may suggest. In this chapter we saw many obstacles and difficulties that needed to be overcome so to keep a healthy balance between simplicity of the process and the accuracy and validity of the model. For this reason, we ended up with something what could be seen as a mixed bag of tools, where at different points we borrow from different methodologies depending on what makes the better fit. While G-DEA remains at the heart of the whole process, the whole process is really a mix and match approach. For example, after

the whole validation process with DSG experts was completed, it was clear that the process of weight elicitation does not have to be a uniform one. Rather, for different levels of hierarchy and for different level of importance of the criteria, different methods are better suited ranging from simple BAP approach (for the least important criteria or where we have many sub-criteria) through AHP pairwise comparison approach (for all the other criteria) and finally ending up with AHP pairwise comparison requiring high degree of confidence and hence allowing interval values (for the most important criteria).

Chapter 6: Conclusion, Contribution and Future Research

6.1. Research conclusions

This research has shown that the construction of a composite indicator is a complex process involving various steps that have significant impact on the results. One of the main problems in constructing composite indicators is its reliance on multiple subjective judgments (Cherchye et al., 2008). With regards to the construction of WEM scores (the DSG's CI), there were several subjective judgments made by different parties involved in different stages of the construction process: the pre-assessment, assessment and post assessment stage. This subjectivity led to a problem of unsatisfied Government departments in the overall scores and the general process of deriving the results. As demonstrated in this thesis, derivation of WEM scores is an intricate process with a complex hierarchy structure, which is used to provide the final scores and award the best department(s) accordingly.

The current WEM scoring methodology has many problems. This research has only focused on the most important ones which have a direct implication on the departments' scores and the way they are presented to them. This has been deduced from the outcomes of the post-assessment survey highlighting Government departments concerns and dissatisfaction. This research indicates that at each of the three stages of the construction process adopted in the current approach, the reliability, validity and fairness of the results were affected.

In the first pre-assessment stage, the following issues arise:

- Weights are determined using a questionable and highly subjective approach. Yet, the level of precision is incredibly high, meaning that this hard-to-believe knowledge is assumed to exist and is incorporated into the model.
- Aggregation method allows for full compensation among indicators and hence does not encourage a more balanced development and progress of the departments.

In the assessment stage, the main problem is the enforcement of a four-points scale [0, 1, 2, 3]

to be used by the assessors to assess all indicators even though many indicators would fit more naturally to a different measurement scale. Due to the use of this scale, some information is lost where it could/should have been preserved. At the same time the scoring process by the assessors is unnecessarily made more difficult than it could have been and as such may have had a further negative impact on the accuracy of the scores recorded.

In the final, post-assessment stage of the process, the main problem is the presentation of the results. Current practices do not foster healthy competition nor encourage learning. The results are largely not trusted by the departments assessed and are one of the main causes for arguments and dissatisfaction in the whole assessment process.

To address the problem at the three different stages, we have developed specific objectives that will support in overcoming these problems and designing the new assessment process of WEM. In the process of achieving the first objective a significant effort was directed towards removing various types of decision biases featuring in the existing WEM model. To that end, the proposed model, using pairwise comparison process and using lower and upper bounds on the weights instead of crisp weight values, will substantially reduce the problems relating to overconfidence in judgment accuracy. It is clear however that this new approach will notably increase the amount of time required to create the pairwise entries and to reach the consensus on those values. Yet, two mitigating circumstances will make it possible to put this approach into practice. The first one is that this kind of time-consuming exercise needs to be done only once and the results can be used for several assessment cycles or until there is a need to change either the structure of WEM or to change priorities of different indicators due to change in Government's policies. The second one is that the proposed model is flexible enough to encapsulate different specifications about the weights at different places of the WEM structure and still run all the specifications through a single model. Namely, in case of WEM score, it is feasible to assume that the most time-consuming and the most rigorous procedures, relating to agreeing on each pairwise comparison value, will need to be performed only for a small

subset of the most important indicators. Elicitation of weights for the indicators of smaller importance or of minor impact to the overall WEM score may be performed using less rigorous procedures such as budget allocation approach. In short, the flexibility of the proposed model allows for the mix of different weight elicitation techniques to be used, which will in the end all be presented through the set of G-DEA weight restrictions, as illustrated in Section 4.3.3., and the weights will be optimised within a single model for each department. At the same time, reducing the bias relating to overconfidence in judgment accuracy in this particular way has a parallel effect on creating a more equitable scoring system since all the intrinsic uncertainties about weight values were left for individual departments to exploit to their advantage so that the final set of weights as chosen by their corresponding G-DEA model will better fit their own distinct characteristics and motivations.

Another significant source of bias in the existing WEM model is the use of uniform measurement scale for all the end-indicators. This issue was present in the existing WEM scoring system "for the sake of simplicity". Yet, that "simplicity" only existed on the side of the authorities who proposed using such a scale and it was never questioned in the context of its use by the assessors and in the context of the accuracy of the measurement. Using uniform scale of measurement was in part also needed so to avoid the issue of normalised the data, which would be necessary for the existing additive form of aggregation, but this was only a secondary issue. The most problematic aspect of this enforced "simplicity" was the unnecessary loss of information and practical difficulty faced by the assessors when using the given scale for the indicators that are not easily measured on such a scale. These problems were discussed in detail in Section 5.4.1. The proposed G-DEA model alleviate all these problems due to its unit-invariance property inherited from the DEA model, which in practical terms means that the assessors may now use any scale of measurement they feel most comfortable with and to use different scales for different indicators. Hence, this new form of true "simplicity" prevents any loss of information or enforced inaccuracies to be present in the

end-indicators' scores assigned by the assessors.

In process of achieving the second objective, extensive effort and analysis was made to encourage a balanced performance across different criteria for all the departments. This is done through examining different aggregation approaches discussed in detail in section 5.3.2. The current WEM model allows departments to achieve high scores even in the presence of very poor scores on some sub-indicators and can appear as better than some other departments whose performance is reasonably good across all the sub-indicators. In this regard, the proposed model using geometric aggregation will reduce the level of substitutability; the departments cannot anymore linearly compensate low score in one dimension by high score in another dimension as in the case of usability and content presented in section 2.6.1 or in the case of department 5 presented in section 5.3.2. Under geometric aggregation, the departments will be much more prone to improve on the indicators they are performing poor at since these improvements will generate much greater increase in the overall score than if they were to focus on improving on the indicators where they already perform well. Geometric aggregation penalizes poor performance in any dimension: the poorer the department's scores on a single dimension, the stronger the decline in the results. This is exactly why the geometric aggregation model is supporting the Dubai Governments' strategic objectives of encouraging a balanced performances and improvements across all the indicators.

In the process of achieving the last objective, the proposed G-DEA model with its property of allowing each department to select their own set of weights for a given set of lower and upper bounds will create a fair and equitable scoring system. G-DEA model demonstrates fairness, which in the current WEM model has been disputed many times by the assessed departments. While the main objective for the new WEM framework is to create a fair and equitable scoring system, this may increase the complexity of the logic behind the model. Hence, fine tuning the balance between credibility of the scores and ease of comprehension will be important for a

successful practical implementation of the new model.

In summary, the main outcomes achieved in this research by applying the G-DEA model can be considered as recommendations for all concerned stakeholders (the central government and the government departments) listed below:

- Conducting a pairwise comparison which has been found extremely useful because it calibrates and reduce the level of subjectivity of the experts. Also, it is not necessary to reach the consensus on the pairwise values since the ranges can be also incorporated into the model.
- Allowing the departments to perform in their best possible strength by giving them some degree of freedom to choose their weights in the form of ranges rather than through fixed weights. These ranges are derived either through any internal inconsistencies within standard pairwise matrices or through on-purpose specified ranges by the experts in consultation with the departments
- Allowing the assessors to choose the most appropriate scale of measurement and not restricted to one for all indicators. The only restriction is that zeros should not feature in a selected scale, which is hardly limiting since many judgment scales are arbitrary anyway.
- Applying geometric aggregation which encourages more balanced performance for all departments. This recommendation does depend on the actual strategy of the central government and so may not be applicable for all the strategies.

6.2. Research contribution

To construct a more accurate, flexible, equitable and transparent WEM scoring methodology, we proposed the G-DEA methodology with some general guidelines to be followed during the assessment stages. The accuracy of the model is based on the fact that various decision biases are reduced or eliminated from the model, such as the ones relating to overconfidence in subjective judgment or an enforced scale of measurement.

The flexibility of the model is multi-dimensional. First, G-DEA allows the representation of any however complex hierarchy structure and still be able to derive the set of final scores through a single model for each assessed department. Another form of flexibility of the model is that it allows for different weight elicitation techniques to be used for indicators of different complexity and different importance. The less important or the more complex indicators (those having many sub-indicators) are, the simpler weight elicitation method should be used such as equal weighting or budget allocation process. With an increasing importance of the indicators, the more complex weight elicitation techniques could be used, such as pairwise comparisons with crisp value in the matrices or, in case of the most important indicators, pairwise comparisons where the entries in the matrices are specified in ranges, which will in turn reduce the time required to reach consensus on the pairwise entries and at the same time allow greater flexibility for the assessed departments to select their optimal weights from within the specified range. Finally, the third type of greater flexibility in the process comes from allowing the assessors to choose the most natural measurement scales and not stick with a single scale for all indicators. This benefit comes directly from DEA weighting and aggregation methods, which eliminate the need to normalise the data. The only reduction in this flexibility is not allowing the assessors to give zero scores, which for the applications where judgment scales are used is not of any significance since the scales are arbitrary anyway.

The model also provides fairness and equity to all assessment stakeholders (the central government and the government departments). From the departments' perspective, the model is equitable because the departments can select the weights within given ranges to show themselves in the best possible light and to align the set of weights that is best suited to their own individual circumstances and aspirations. The model is also equitable from the central government perspective for two reasons. First, the central government can provide their point of view on the importance of criteria through the ranges of weights (the upper and lower limit of each weight). Second, through the use of geometric aggregation, they can better direct the

effort of the departments towards a more balanced performance. When balanced performance across different criteria is desired, then geometric aggregation produces more sensible overall results than additive aggregation, as illustrated in the example in Chapter 2 and Chapter 3 and further elaborated in the research findings in Chapter 5. It is important to obtain compliance with what is expected: the results Dubai government can get from geometric aggregation is well aligned with their expectations, while additive aggregation takes us outside this realm producing some scores that are unexpected and irrational given the underlying set of scores.

Transparency and simplicity can be provided to all the assessed departments using data visualisation, which will support us in explaining relatively complex models in one interactive dashboard. While this component of the model was not given due attention in this research, it is worth noting that such an interactive dashboard could feature not only the scores and its dependence on the scores of sub-indicators but also all the elements that could be normally obtained from any DEA-based analysis such as performance targets and efficient peers.

The research essentially contributes in making a complex methodology like G-DEA on creating CI for WEM very clear and providing a template on how it can be implemented. Such methodology can be employed in the future for many other indicators related to Government excellence and possibly outside Government excellence. To the best of our knowledge, this is the first implementation of the G-DEA methodology and based on our findings, it promises to bring significant improvements relative to the current methodology and along all those important attributes of accuracy, fairness, equity and transparency. Given these qualities, the proposed methodology should be a good candidate to explore within the context of any other types of composite indicators, especially very complex ones such as those relating to sustainability, which usually involve many indicators that come from different scales of measurements.

Applying the G-DEA methodology in practice combines many positive characteristics of different methodologies. One is the multiplicative aggregation, which has been recognised by

the Human Development Index (HDI) as being better than additive aggregation. The second one is the use of pairwise comparisons for reducing decision bias while not forcing the experts to reach a full consensus on any pairwise entries. The third one is the benefit of DEA-like methodology, which has been emphasised by the BOD approach in the recent handbook of the construction of composite indicators. In fact, our proposed G-DEA model for composite indicators could be seen as an enriched multiplicative version of the BOD model.

In addition to all the above contributions, this research makes further contributions to the literature and knowledge through investigating and analysing the differences between the additive and multiplicative aggregation methods and comparing their results. This analysis will contribute in raising the awareness among the researchers and practitioners equally about the weakened position of the additive aggregation, which is especially pronounced in the presence of flexible weights as illustrated in Figures 9 and 10 in Chapter 3. Last but not the least, the contributions made here make this thesis to be a pioneering work in bridging the gap between the body of the literature on business excellence models and the body of the literature on decision analysis and composite indicators. While there is some effort to bridge this gap from the decision analysis side, to the best of the author's knowledge this work is a rare attempt to bridge this gap from the BEM practitioner's side.

The WEM score was a perfect trial for applying the new method, so we can address any issues that had not been previously observed. It is not too costly to make a mistake in WEM because it is not as prestigious and important as GEM and if everything turns out to be fine, then it is perfect to apply it to GEM.

6.3. Limitations and recommendations for future research

Although the proposed methodology offers many advantages and improvements over the currently used methodology, it also has some limitations. Its main drawback is its apparent complexity. From a practical perspective, applying such a complex method will require a lot of

effort and time to make the results easy to comprehend by all the parties involved in the assessment. One way to overcome such limitation is to introduce the methodology in stages rather than at once. For example, the first step will be introducing flexible weights, which will give the government departments the freedom to choose from the range of weights rather than having crisp weights whilst keeping the current system as it is right now. The following step would be changing the aggregation from additive to multiplicative. This step can be justified to the departments relatively easily by explaining the difference between both approaches and the reason for such change from additive to multiplicative, which is mainly due to the need of obtaining a balanced performance. We can build on the 9-year experience of the Human Development Index (HDI) when they switched from additive to multiplicative aggregation. These steps will eventually support us to introduce the G-DEA methodology.

Another limitation relates to those cases where it is very natural to assign zero scores. Such limitation will affect the ease with which the proposed methodology can be implemented. However, we will support the assessors to come with a replacement of zero such as the one drawn from our sensitivity analysis where we forced to replace zeros for no other reason but to be able to compare the results obtained by different methodologies. For example, when it is natural for the assessors to use zero score like in the case of a binary scale of measurement using only 0 and 1, the most likely scenario will be to use 1 and 100 [instead of 0 and 1].

An additional generic weakness of this study relates to the limited time available to conduct the research. It was not possible to have a full coverage of all aspects relative to this research. These include the psychological aspect that deals with human perception and the e-government aspect that has similar studies to the case study in this research. The psychological aspect mentioned above specifically would address the social behavioural perspective of the assessors. Though all assessors have been trained on the assessment tools mentioned in chapter 2, we cannot ignore the possibility of them being influenced by self-interest, background, experience and surrounding environment which in turn could impact their

rational choices. Further research in this area could provide insight on how scores have been assigned to specific indicators which would impact the results.

Moreover, similar studies in the e-government aspect specifically measuring website performance was not fully covered in this research. The research focused solely on WEM score wherein an extensive amount of research was conducted to align with the main framework of Dubai Government; the Government Excellence Model. This research would have been stronger had we been able to study in more depth other similar experiences. A significant amount of time during this study was spent on identifying all the problems related to the analytical mechanisms of the existing methodology and on finding an adequate cure by implementing new analytical tools and process, ultimately culminating in the proposed methodology with G-DEA at its core.

An interesting area that can be considered for future study is measuring the performance of the government departments over time using the Malmquist Index. By doing this, we could actually see how much the departments improved and in which direction they improved, thus gaining valuable insights in their performance. However, even though we have several of past results, unfortunately the assessment was always different using different indicators and kept changing. We have now developed a stable approach with the new Government Excellence Model that yields reliable results. The application of the Malmquist Index can further consolidate our findings and provide new insights into departmental performance.

6.4. Summary

This research has been applied on a practical case study in order to develop a novel way of assessing government departments website performance. The findings can be applied to the Government Excellence Model and beyond. The significance of this study is therefore both theoretical and practical. It is hoped that future research will develop the findings put forward here in order to further enhance our knowledge of construction composite indicators.

References

Adaep.ae. Award cycle [online]. Available from: https://www.adaep.ae/en/Pages/default.aspx. [Accessed June.2019].

Al Maktoum, M.B.R. (1997). About Government Excellence Program [online]. Available from: https://dgep.gov.ae/en/about-us [Accessed December 2013].

Al Maktoum, M.B.R. (2006). My vision challenges in the race for Excellence. Dubai: Motivate publishing.

Casadio Tarabusi., E. and Guarini G. (2013). An Unbalance Adjustment Method for Development Indicators. Social Indicators Research, 112(1), 19-45.

Charnes, A. and Cooper, W.W. (1962). Programming with Linear Fractional Functional. Naval Research Logistics Quarterly, 9(3/4), 181-185.

Charnes, A., Cooper, W.W. and Rhodes, E. (1978). Measuring the efficiency of decision-making units. European Journal of Operational Research, 2(6), 429-444.

Cherchye, L. (2001). Using data envelopment analysis to assess macroeconomic policy performance. Applied Economics, 33(3), 407-416.

Cherchye, L. & Kuosmanen, T. (2002). Benchmarking sustainable development: A synthetic meta-index approach. EconWPA Working Papers.

Cherchye, L., W. Moesen and T. Van Puyenbroeck. (2004). Legitimately Diverse, yet Comparable: On Synthesizing Social Inclusion Performance in the EU. Journal of Common Market Studies, 42(5), 919-955.

Cherchye, L., Moesen, W., Rogge, N. and Van Puyenbroeck, T. (2007a). An Introduction to 'Benefit of the Doubt' Composite Indicators. Social Indicators Research, 82(1), 111-145.

Cherchye, L., Lovell, C. K., Moesen, W. and Van Puyenbroeck, T. (2007b). One market, one number? A composite indicator assessment of eu internal market dynamics. European Economic Review, 51(3), 749-779.

Cherchye, L., Moesen, W., Rogge, N., Van Puyenbroeck, T., Saisana, M., Saltelli, A., Liska, R. and Tarantola, S. (2008). Creating composite indicators with DEA and Robustness analysis: The case of the Technology Achievement Index. Journal of the Operational Research Society, 59(2), 239-251.

Chowdhury, S. and Squire, L. (2006). Setting weights for aggregate indices: An application to the commitment to development index and human development index. The Journal of Development Studies, 42(5), 761-771.

Dahlgaard, J.J., Chen, C.K., Jang, J.Y., Banegas, L.A. and Dahlgaard-Park, S.M. (2013) Business excellence models: limitations, reflections and further development, Total Quality Management & Business Excellence, 24:5-6, 519-538.

Despić, O. (2004). DEA-R relative efficiency model: some theoretical and practical considerations. Working Paper RP0435, Aston Business School, UK.

Despić, O. (2012). Geometric DEA Models and their Properties. In Charles, V.& Kumar, M. (eds.) Data Envelopment Analysis and Its Applications to Management. Newcastle Upon Tyne: Cambridge Scholars Publishing, 29-50.

Despić, O. (2013). Some properties of geometric DEA models. Croatian Operational Research Review, 4, 1-18.

Despić, O. and Prasanta, D. (2006). Flexible Multiplicative Analytic Hierarchy Model. Aston Business School Research Papers. Birmingham: Aston University.

Despotis, D.K. (2005a). Measuring human development via data envelopment analysis: the case of Asia and the Pacific. Omega 33(5), 385-390.

Despotis, D.K. (2005b). A Reassessment of the Human Development Index Via Data Envelopment Analysis. Journal of Operational Research Society, 56(8), 969-980.

DubaiPlan2021.ae. Dubai Strategic Plan 2015 [online]. Available from: https://www.dubaiplan2021.ae/dsp-2015-2/ [Accessed June 2019].

Ebert, U. and Welsch, H. (2004). Meaningful environmental indices: A social choice approach. Journal of Environmental Economics and Management, 47(2), 270-283.

EFQM. Global Excellence Council [online]. Available from: http://www.efqm.org/index.php/community/global-excellence-council/ [Accessed May 2019].

Emrouznejad, A., and Marra, M. (2017). The state of the art development of AHP (1979–2017): A literature review with a social network analysis. International Journal of Production Research, 55(22), 6653-6675.

Entani, T., Ichihashi, H. and Tanaka, H. (2001). Optimistic Priority Weights with an Interval Comparison Matrix. In Terano, T. et al. (eds.) JSAI 2001 Workshops, LNAI 2253. Berlin Heidelberg: Springer-Berlin, 344-348.

Eskildsen, J.K., Kristensen. K. and Juhl, H.J., 2001. The criterion weights of the EFQM excellence model. International Journal of Quality & Reliability Management. 18(8), 783-795.

Esty, D.C., Levy, M.A., Srebotnjak, T., de Sherbinin, A., Kim, C.H. and Anderson, B. (2006). Pilot 2006 Environmental Performance Index. New Haven: Yale Center for Environmental Law & Policy.

Färe, R. and Grosskopf, S. (2004). Modelling undesirable factors in efficiency evaluation: Comment. European Journal of Operational Research 157(1), 242-245.

Farrell, M.J. (1957). The measurement of productive efficiency. Journal of the Royal Statistical Society, 120(3), 253-290.

Freudenberg, M. (2003). Composite Indicators of Country Performance: A Critical Assessment, OECD Science, Technology and Industry Working Papers 2003/16.

Gaaloul, H. and Khalfallah, S. (2014). Application of the "benefit-of-the-doubt" approach for the construction of a digital access indicator: A revaluation of the "digital access index". Social Indicators Research, 118(1), 45-56.

Greco, S., Ishizaka, A., Tasiou, M., and Torrisi, G. (2018). On the methodological framework of composite indices: A review of the issues of weighting, aggregation, and robustness. Social Indicators Research, 141(1), 61-94.

Green P.E., and Srinivasan V. (1978). Conjoint analysis in consumer research: issues and outlook, Journal of Consumer Research 5(2), 103-123.

Gupta, M.P., Bhattacharya, J. and Agarwal, A. (2007). Evaluating e-government. e-governance: case-studies, s.l.: CSI SIG on e-Governance.

Hagerty, M.R. and Land, K.C. (2007). Constructing Summary Indices of Quality of Life: A Model for the Effect of Heterogeneous Importance Weights, Sociological Methods & Research, 35(4), 455-496.

Hajkowicz, S. (2006). Multi-attributed environmental index construction. Ecological Economics, 57(1), 122-139.

Hatefi, S.M. and Torabi, S.A. (2010). A common weight MCDA–DEA approach to construct composite indicators. Ecological Economics, 70(1), 114-120.

Hope, C., Parker, J. and Peake, S. (1992). A pilot environmental index for the UK in the 1980s. Energy Policy 20(4), 335–343.

Ho, W. (2008). Integrated analytic hierarchy process and its applications - A literature review. European Journal of operational research, 186(1), 211-228.

Hosseini Ezzabadi, J., Dehghani Saryazdi, M. and Mostafaeipour, A. (2015). Implementing Fuzzy Logic and AHP into the EFQM model for performance improvement: A case study, Applied Soft Computing Journal, 36, 165-176.

Kanji, G.K. (1998). Measurement of business excellence, Total Quality Management, 9(7), 633-643

Kaufmann, D,. Kraay, A,. and Mastruzzi, M., (2004). Governance Matters III: Governance Indicators for 1996, 1998, 2000, and 2002. Washington, DC: World Bank. © World Bank. https://openknowledge.worldbank.org/handle/10986/17136

Lau, K.N. and Lam, P.Y. (2002). Economic freedom ranking of 161 countries in year 2000: a minimum disagreement approach. Journal of the Operational Research Society 53(6), 664-671.

Liu, C. T., Du, T. C., and Tsai, H. H. (2009). A study of the service quality of general portals. Information & Management, 46(1), 52-56.

Luna, D. E., Gil-Garcia, J. R., Luna-Reyes, L. F., Sandoval-Almazan, R., and Duarte-Valle, A. (2013). Improving the performance assessment of government web portals: A proposal using data envelopment analysis (DEA). Information Polity, 18(2), 169-187.

Mahlberg, B. and Obersteiner, M. (2001). Remeasuring the HDI by Data Envelopment Analysis. IIASA Interim Report. IIASA, Laxenburg, Austria: IR-01-069.

Maggino, F. and Ruviglioni, E. (2009). Obtaining weights: from objective to subjective approaches in view of more participative methods in the construction of composite indicators, Seminar on New Techniques and Technologies for Statistics (NTTS) – EUROSTAT, Brussels.

Mascherini, M. and Hoskins, B. (2008). Retrieving expert opinion on weights for the Active Citizenship Composite Indicator, European Commission – Institute for the protection and security of the citizen – EUR JRC46303 EN.

Mazziotta, M. and Pareto, A. (2013). Methods for constructing composite indices: one for all or all for one? RivistaItaliana di Economia Demografia e Statistica, LXVII(2), 67-80.

McDaniel, C. and Gates R. (1998), Contemporary Marketing Research. West Publishing, Cincinnati, OH.

Melyn W. and Moesen W. (1991). Towards a synthetic indicator of macroeconomic performance: Unequal weighting when limited information is available. K.U.Leuven, Centrum voor Economische Studiën; Leuven.

Mishra, S.K. (2007). A comparative study of various inclusive indices and the index constructed by the principal components analysis. Available at SSRN: https://ssrn.com/abstract=990831

Moldan, B., Billharz, S. and Matravers, R. (1997). Sustainability Indicators: Report of the Project on Indicators of Sustainable Development, SCOPE 58. Chichester and New York: John Wiley & Sons.

Moreno-Rodríguez, J.M., Cabrerizo, F.J., Pérez, I.J. and Martínez, M.A. (2013). A consensus support model based on linguistic information for the initial-self assessment of the EFQM in health care organizations, Expert Systems with Applications 40(8), 2792-2798.

Murias, P., de Miguel, J.C. and Rodriguez, D. (2008). A composite indicator for university quality assessment: The case of Spanish higher education system. Social Indicators Research, 89(1), 129-146.

Murias, P., Martinez, F. and de Miguel, C. (2006). An economic wellbeing index for the Spanish provinces: A data envelopment analysis approach. Social Indicators Research, 77(3) 395-417.

Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A. and Giovannini, E. (2005). Handbook on constructing composite indicators: methodology and user guide. OECD Statistics Working Paper 2005/3. Available from: http://www.olis.oecd.org/olis/2005doc.nsf/LinkTo/std-doc(2005)3.

Nicoletti, G., Scarpetta, S. and Boylaud, O. (1999). Summary indicators of product market regulation with an extension to employment protection legislation, OECD ECO Working Paper No. 226. Available at SSRN: http://dx.doi.org/10.2139/ssrn.201668

OECD (2005). Handbook on Constructing Composite Indicators: Methodology and User Guide. Available from: http://dx.doi.org/10.1787/533411815016

OECD (2008). Handbook on Constructing Composite Indicators: Methodology and User Guide. Available from: http://www.oecd.org/sdd/42495745.pdf

Paktinat, M. and Danaei, A. (2014). An application of fuzzy AHP for ranking human resources development indices. Management Science Letters, 4(5), p.993-996.

Parker, J. (1991). Environmental reporting and environmental indices. Ph.D. Thesis. Cambridge, United Kingdom.

Podinovski, V. (2004) Production trade-offs and weight restrictions in data envelopment analysis, Journal of the Operational Research Society, 55:12, 1311-1322.

Ramanathan, R. (2006). Data envelopment analysis for weight derivation and aggregation in the analytic hierarchy process. Computers and Operations Research, 33 (5), 1289-1307.

Rogge, N., Cherchye, L., Moesen, W. and Van Puyenbroeck, T. (2006). 'Benefit of the doubt' composite indicators. Paper presented at the European Conference on Quality in Survey Statistics.

Saaty, R.W. (1987). The analytic hierarchy process: What it is and how it is used. Mathematical Modelling, 9(3-5), 161-176.

Salo A.A. and Hämäläinen, R.P. (1995). Preference programming through approximate ratio comparisons. European Journal of Operational Research, 82(3), 458-475.

Saltelli, A. (2007). Composite Indicators between Analysis and Advocacy. Social Indicators Research, 81(1), 65-77.

Salzman, Julia and Andrew Sharpe (2003), Methodological Choices Encountered in the Construction of Composite Indices of Economic and Social Well-Being, Paper presented to the Canadian Economics Association conference, 31 May 2003. Available: http://www.csls.ca/events/cea2003/salzman-typol-cea2003.pdf

Shen, Y., Hermans, E., Ruan, D., Wets, G., Brijs, T. and Vanhoof, K. (2011). A generalized multiple layer data envelopment analysis model for hierarchical structure assessment: A case study in road safety performance evaluation. Expert Systems with Applications, 38(12), 15262-15272.

Shen, Y., Hermans, E., Brijs, T. and Wets, G. (2013). Data envelopment analysis for composite indicators: A multiple layer model. Social Indicators Research, 114(2), 739-756.

Smartdubai.ae. (2019). About Us. [online] Available at: https://www.smartdubai.ae/about-us [Accessed June 2019].

Somarriba, N., and Pena, B. (2009). Synthetic indicators of quality of life in Europe. Social Indicators Research, 94(1), 115–133.

Srinivasan, T. N. (1994). Human development: A new paradigm or reinvention of the wheel? American Economic Review, 84(2), 238-243.

Storrie, D. and H. Bjurek. (2000). Benchmarking European Labour Market Performance with Efficiency Frontier Techniques. Technical Report, CELMS Discussion papers, Goteborg University.

Tavana M., Yazdi, A.K., Shiri,M. and Rappaport, J. (2011). An EFQM-Rembrandt excellence model based on the theory of displaced ideal, Benchmarking: An International Journal, 18(5), 644-667.

Tomaževič, N., Seljak, J. and Aristovnik A. (2016). TQM in public administration organisations: an application of data envelopment analysis in the police service, Total Quality Management & Business Excellence, 27(11-12), 1396-1412

UAE.Vision2021. (2019). UAE vision [online]. Available from: https://www.vision2021.ae/en [Accessed June 2019].

Ülengin, B., Ülengin, F. and Güvenç, Ü., (2001). A multidimensional approach to urban quality of life: the case of Istanbul, European Journal of Operational Research, 130(2), 361-374.

Vierstraete, V. (2012). Efficiency in human development: A data envelopment analysis. The European Journal of Comparative Economics, 9(3), 425-443.

Wang, T. C. and Chen, Y.H. (2007). Applying Consistent Fuzzy Preference Relations to Partnership Selection. Omega, 35(4), 384–388.

Zaim, O., Färe, R. and Grosskopf, S. (2001). An Economic Approach to Achievement and Improvement Indexes. Social Indicators Research, 56(1), 91-118.

Zhou, P., Ang, B.W. and Poh, K.L. (2007a). A mathematical programming approach to constructing composite indicators. Ecological Economics, 62(2), 291-297.

Zhou, P., Ang, B.W. and Poh, K.L. (2007b). A non-radial DEA approach to measuring environmental performance. European Journal of Operational Research, 178(1), 1-9.

Zhou P., Ang B.W. and Poh K.L. (2008). Measuring environmental performance under different environmental DEA technologies. Energy Economics, 30(1), 1-14.

Zhou, P., Ang, B.W. and Zhou, D.Q. (2010) Weighting and Aggregation in Composite Indicator Construction: A Multiplicative Optimization Approach. Social Indicators Research, 96(1), 169–181.

Appendix A: Results' sub-indicators scores of Smart Government Transformation Indicator measured by the Dubai Smart Government (DsG)

In1 - Electronic/smart maturity index

In2 - The score of completing and enabling electronic services on smart devices

In3 - The score of adopting electronic/smart government services

In4 - The score of completing the electronic/smart transition for internal processes

In5 - Website Excellence Model (WEM) score

In6 - The score of government department website customer satisfaction

In7 - The score of government department website usage

In8 - The score of current and new services that utilize smart technologies such as Internet of Things, sensors, smart glasses and wearable devices

In9 - The score of aligning the budget allocated to smart and electronic transition with the strategy of the Government of Dubai

In10 - The score of utilizing common services and systems provided by the Smart Government of Dubai to government departments such as Government Resources Planning Systems

Appendix B: WEM: weights allocation for levels 2 and 3

Indicator	Sub-indicator		Weight
Accessibility (23%)	A1	Provide Access Through an Easy to Remember URL; gov.ae domain	4.8%
	A2	Provide a Quick Access to the Website from a Search Engine	4.8%
	A3	Provide Identical and Consistent Results through different browsers	4.8%
	A4	Provide a Functional Bilingual Website	5.8%
	A5	Provide Appropriate Access to Website Files	2.8%
Total			**23%**
Indicator	**Sub-indicator**		**Weight**
Usability & Design (34%)	UD1	Provide a Clearly Defined Website Header and Footer	4.1%
	UD2	Provide a Clear and Readable Entity & Dubai Government Logos	2.1%
	UD3	Provide a Functional Link to the Official Portal of Dubai Government	2.1%
	UD4	Provide a Well Designed Customer Focused Homepage	4.1%
	UD5	Provide a Functional Homepage Link Available Across all Web Pages	2.1%
	UD6	Provide a Well Structured and Effective Sitemap	3.1%
	UD7	Provide an Effective and Efficient Search Functionality	3.6%
	UD8	Provide a Logically Organized and Easy to Navigate Website	2.1%
	UD9	Provide a Proper and Easy to Use Navigation Facility	2.1%
	Website links		2.1%
	UD10	Use an Appropriate Design for Website Links	
		Provide Active Internal and External Links	
		Provide Clear and Meaningful Links on the Website	
	Website Forms		2.1%
	UD11	Provide Simple and Easy to Use Forms	
		Provide Functioning and Properly Working Forms	
		Provide Proper and Easy to Understand Guidelines for Completing the Online Forms	
	Website Design		4.2%
	UD12	Provide a Consistent Font Style Across the Website Pages	
		Provide a Consistent Format Throughout The Website	
		Provide Well Designed Website Page Titles	
Total			**34%**
Indicator	**Sub-indicator**		**Weight**
Content (29%)	C1	Provide Information about the Entity in “About Us” Section	3.5%
	C2	Provide Entity Contact Information in "Contact Us" Page	3.5%
	C3	Provide a Facility to Submit Feedback on the Website	2.0%
	C4	Provide Effective and Efficient FAQ page on the website	3.0%
	C5	Provide Sufficient Information about Entity Services & eServices	4.0%
	C6	Provide Accurate Website Copyright Information	1.5%
	C7	Provide a Proper "Site Maintained By" Message	1.5%
	C8	Provide a Functional Link to eJob	1.5%
	C9	Provide a Functional Link to eSuggest	1.5%
	C10	Provide a Functional Link to eComplain	1.5%
	C11	Provide a Functional Link to Ask Dubai	1.5%
	C12	Define/Use Proper and Meaningful Metadata on Almost Every Page	3.0%
	C13	Provide Accurate Dates on the Website Pages	1.5%
Total			**29%**
Indicator	**Sub-indicators**		**Weight**
Polices (14%)	P1	Provide Information Regarding Protection and Handling of Privacy	7.5%
	P2	Provide Information Regarding the Website Terms and Conditions	6.5%
Total			**14%**

Appendix C: WEM: weights allocation for end-indicators (level 4)

A1		Weight
A11	Short and easy to remember.	10%
A12	Clear and unequivocal in referring to the entity name or its abbreviation.	20%
A13	Under (UAE) top-level domain, for Arabic website.	20%
A14	Under gov.ae top-level domain, for English website.	50%
A2		**Weight**
A21	Keyword 1 English	25%
A22	Keyword 2 English	25%
A23	Keyword 1 Arabic	25%
A24	Keyword 2 Arabic	25%
A3		**Weight**
A31	The website is compatible with I.E	40%
A32	The website is compatible with Chrome	40%
A33	The website is compatible with Safari or Firefox	20%
A4		**Weight**
A41	Bilingual link is available at a consistent location at the page header	40%
A42	Bilingual link is clear and recognizable	20%
A43	English bilingual link directs the user to the same page in the other language (test at least 5 Pages)	20%
A44	Arabic bilingual link directs the user to the same page in the other language (test at least 5 Pages)	20%
A5		**Weight**
A51	File 1: File attributes are available (name, description, size, format, date)	20%
A52	File 1: File format (HTML or others) in case of PDF or other format a link to download software is available.	5%
A53	File 2: File attributes are available (name, description, size, format, date)	20%
A54	File 2: File format (HTML or others) in case of PDF or other format a link to download software is available.	5%
A55	File 3: File attributes are available (name, description, size, format, date)	20%
A56	File 3: File format (HTML or others) in case of PDF or other format a link to download software is available.	5%
A57	File 4: File attributes are available (name, description, size, format, date)	20%
A58	File 4: File format (HTML or others) in case of PDF or other format a link to download software is available.	5%
U1		**Weight**
U11	Header: clearly defined and separated from the rest of the content.	15%
U12	Header: consistently used throughout the entire website (on all the pages of the website) (check at least 10 pages)	15%
U13	Government of Dubai official logo on the left side	20%
U14	The entity's official logo and name on the right side at the top of the page header	20%
U15	Logos are on a white strip, no distracting elements in the middle or top	10%
U16	Footer: clearly defined and separated from the rest of the content	10%
U17	Footer: consistently used throughout the entire website (on all the pages of the website) (check at least 10 pages)	10%
U2		**Weight**
U21	The entire logo should be clickable so that the user should not guess which part is clickable	40%
U22	Clicking on the entity logo directs the user to the homepage of the corresponding language (Arabic to Arabic homepage)	25%

U23	Clicking on the entity logo directs the user to the homepage of the corresponding language (English to English homepage)	25%
U24	Logos are clear (good quality)	10%
U3		**Weight**
U31	"Dubai.ae" (for English website)	20%
U32	" Dubai.ae" (for Arabic website)	20%
U33	Both Logos are available at a consistent location in the page header throughout the website	30%
U34	Link to http://www.dubai.ae either English or Arabic website depends on the user's language	20%
U35	Both links are clear and readable	10%
U4		**Weight**
U41	The entity's services and eServices are presented on the homepage and ensure easy & quick access for the users	30%
U42	The services or eServices are highlighted on the homepage in a proper categorization (e.g. Customer segment, service categories, etc)	30%
U43	Overall design and layout of the homepage (please add any comments, e.g. horizontal and vertical scrolling, images)	20%
U44	Overall content of the homepage (please add any comments, for example irrelevant content)	20%
U5		**Weight**
U51	Homepage link is available at a consistent location in the page header	30%
U52	Homepage link is clear and recognizable	10%
U53	Homepage link directs the user to the home page no matter where they are on the website through a single click	30%
U54	Homepage link directs the user to the correct language home page (English/English, Arabic/Arabic)	30%
U6		**Weight**
U61	Sitemap is available at a consistent location (all pages) either in header or footer	25%
U62	Sitemap is bilingual	25%
U63	Sitemap structure is well organized	25%
U64	Sitemap links function properly	25%
U7		**Weight**
U71	Search tool is available at a consistent location at the page header	20%
U72	Search tool is functioning properly with basic keywords search (try at least 4 English/Arabic)	30%
U73	Search results are language consistent ** if search in English then results should be displayed in English and vice versa	25%
U74	Search result page is properly organized	25%
U8		**Weight**
U81	The website navigation menu is different from the rest of content (can be easily recognized)	20%
U82	The website navigation menu is at a consistent location on every page of the website	20%
U83	The website navigation menu titles are short and descriptive	20%
U84	Website content is organized logically (for example, providing an eService link only under "About Us" will not make any sense to the user)	40%
U9		**Weight**
U91	Navigation facility is available at a consistent location throughout the website pages	30%
U92	Navigation facility in a consistent style across all the website pages	30%
U93	Navigation facility is working properly	40%
U10		**Weight**
U101	Different colors for visited and non-visited destination links	15%
U102	Links are clear and descriptive	15%

U103	Overall broken links on the websites (internally and externally) (please suggest scoring based on the tool selected)	40%
U104	External links open in a new page	15%
U105	External links open in a related language, if the external links are not available in the language the user should be notified	15%
U11		**Weight**
U111	Mandatory fields are marked and data format is available	30%
U112	Instructions on completing online forms are available	20%
U113	Confirmation screen is provided, upon submitting the form, along with a reference number for follow up purposes in case needed. The confirmation screen may also contain the contact number or email, which should be used with this reference number for an enquiry or to obtain any clarification.	30%
U114	Provide a notification about time/date of the request completion	20%
U12		**Weight**
U121	Same font is used across the website, to a certain extent (10 pages)	30%
U122	Format of the website is consistent among website pages e.g. colors, scrolling (refer to guidelines for examples) (10 pages)	60%
U123	Page titles are short and descriptive(10 pages)	5%
U124	Page titles available in related language (10 pages)	5%
C1		**Weight**
C11	"About Us" link is available at a consistent location in the page header	10%
C12	The entity's vision statement is available in "About Us" page	20%
C13	The entity's mission statement is available in "About Us" page	20%
C14	The entity's mandate is available in "About Us" page	20%
C15	The entity's objectives are available in "About Us" page	20%
C16	General contact information: include general entity contact information with a link to contact us for further contact information	10%
C2		**Weight**
C21	Contact us link is available at a consistent location in the page footer	10%
C22	The physical mailing address of the entity's head office and branches/service centers	20%
C23	The street address of the entity's head office and branches/service centers with location maps (it is important to ensure that the location maps are available in both languages, Arabic & English)	20%
C24	The entity's and branches/service centers' telephone number(s). The telephone number should include area code	10%
C25	The entity's and branches/service centers' fax number(s). The fax number should include area code	10%
C26	The entity's and/or branches/service centers' e-mail address(es)	10%
C27	A point of contact within the entity that is responsible for user enquiries (does not necessarily have to be an individual name; it can also be an email address labelled as customer service@... or questions@...., etc …). Government entities are encouraged to provide an autoreply email for users informing them for example with a response time and follow up details (please refer to the guideline for UD.16 for further details)	10%
C28	The entity's and branches/service centers' hours of operation for over the counter and telephone based interactions, in case the entity directly deals with the public	10%
C3		**Weight**
C31	A feedback form is available on the website (either in "Contact Us" page or separately) at a consistent location	40%
C32	Feedback form is available in both languages (English & Arabic)	60%
C4		**Weight**
C41	FAQ link is available at a consistent location in the website header or footer (FAQ should either be relevant to the website or services or eServices)	20%
C42	FAQs are available in both languages	30%

C43	FAQs Page: The page is properly organized (organization of questions, grouping, etc.)	35%
C44	A facility to ask a new question should be available in case the user request is not fulfilled	15%
C5		**Weight**
C51	Service catalogue or service information is available on the website at a consistent location and easily accessible	10%
C52	The list or catalogue of services includes all entity services (check for at least 8 services)	8%
C53	Service name: the name of the service should be self-explanatory; the user should not need to read the service description unless he/she needs more information	10%
C54	Service description: a brief description about the service	10%
C56	Service requirements: details of the requirements needed for this service (e.g. documents)	10%
C57	Service procedures: list of steps needed for this service, the steps should be clear and available in sequence	10%
C58	Service forms: if the services require form(s) to be filled, an option should be available to download	8%
C59	Service expected completion time: the expected average time to complete this service	9%
C510	Service fees: the fees for this service (in some cases different fees might be required depending on certain conditions)	9%
C511	Service centres: physical locations to access the service	9%
C6		**Weight**
C61	Provided with the appropriate corresponding year followed by the entity name	50%
C62	Available at a consistent location throughout the website in the website footer	50%
C7		**Weight**
C71	Include the entity name as “This site is maintained by the [Entity Name]”	50%
C72	Be available at a consistent location throughout the website in the website footer	50%
C8		**Weight**
C81	eJob link is available in careers or vacancies page	40%
C82	eJob link is functioning properly (links to eJob in related language)	60%
C9		**Weight**
C91	Navigation facility is available at a consistent location throughout the website pages	40%
C92	Navigation facility in a consistent style across all the website pages	60%
C10		**Weight**
C101	eComplain link is available in a proper location	40%
C102	eComplain link is functioning properly (links to eComplain in related language)	60%
C11		**Weight**
C111	Ask Dubai link is available in a proper location	40%
C112	Ask Dubai link is functioning properly (links to Ask Dubai in related language)	60%
C12		**Weight**
C121	Appropriate structure exists on all pages to accommodate meta data	40%
C122	Title, description and keywords are populated in all pages	35%
C123	Minimum 3 keywords exist in all pages	25%
C13		**Weight**
C131	Site last modified or updated date is available at a consistent location in the website footer	30%
C132	Site last modified or updated date reflects the latest update on the site (check news page or frequently updated pages)	40%
C133	The same date format is used across the website (check at least 5 pages)	30%
P1		**Weight**
P11	Websites policies(can be security or privacy policy) are available at a consistent location of the website footer	10%

P12	Collection & use of information: the website should address what user information is collected, and how this information is used and shared by the entity	15%
P13	IP addresses & cookies: the website should address if users' IP addresses are collected and how they are used. If cookies are used, the policy must address the purpose of using them.	15%
P14	Protection of information: the website should address to whom and/or what entities users' information will be available and the policy for sharing the information with third party(ies), if any.	15%
P15	Security of information: the website should address what measures are taken to preserve the security of users' information. Mandatory use of a secure and encrypted method for transmission of personal data or financial transactions over the internet.	15%
P16	Disputes: steps a person should take if they have reasonable doubt that their privacy is being compromised.	15%
P17	Third party website: if the entity has links to other website(s), the website should address all concerns and issues related to these links, such as responsibility, accuracy of information, security, liability of information, etc.	15%
P2		**Weight**
P21	Link available throughout the website at a consistent location in the website footer	40%
P22	Terms and condition information is appropriate (e.g. usage of content, registration termination, etc.)	60%

Appendix D: Pairwise matrices for WEM indicators & sub-indicators

WEM score	Access	Usability	Content	Policy
Access	1	1/3	1/2	4
Usability	3	1	2	6
Content	2	1/2	1	5
Policy	1/4	1/6	1/5	1

Access	A_1	A_2	A_3	A_4	A_5
$\mathbf{A_1}$	1	1/2	1/5	1	3
$\mathbf{A_2}$	2	1	1/5	2	3
$\mathbf{A_3}$	5	5	1	5	7
$\mathbf{A_4}$	1	1/2	1/5	1	2
$\mathbf{A_5}$	1/3	1/3	1/7	1/2	1

Usability	U_1	U_2	U_3	U_4	U_5	U_6	U_7	U_8	U_9	U_{10}	U_{11}	U_{12}
$\mathbf{U_1}$	1	5	5	1	5	3	2	5	5	5	5	1/2
$\mathbf{U_2}$	1/5	1	1	1/5	1	1/3	1/5	1	1	1	1	1/5
$\mathbf{U_3}$	1/5	1	1	1/5	1	1/3	1/5	1	1	1	1	1/5
$\mathbf{U_4}$	1	5	5	1	5	3	2	5	5	5	5	1/2
$\mathbf{U_5}$	1/5	1	1	1/5	1	1/3	1/5	1	1	1	1	1/5
$\mathbf{U_6}$	1/3	3	3	1/3	3	1	1/2	3	3	3	3	1/3
$\mathbf{U_7}$	1/2	5	5	1/2	5	2	1	5	5	5	5	1/3
$\mathbf{U_8}$	1/5	1	1	1/5	1	1/3	1/5	1	1	1	1	1/5
$\mathbf{U_9}$	1/5	1	1	1/5	1	1/3	1/5	1	1	1	1	1/5
$\mathbf{U_{10}}$	1/5	1	1	1/5	1	1/3	1/5	1	1	1	1	1/5
$\mathbf{U_{11}}$	1/5	1	1	1/5	1	1/3	1/5	1	1	1	1	1/5
$\mathbf{U_{12}}$	2	5	5	2	5	3	3	5	5	5	5	1

Content	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	U_{11}	U_{12}	C_{12}	C_{13}
$\mathbf{C_1}$	1	1	5	2	1/2	7	7	7	7	7	7	2	7
$\mathbf{C_2}$	1	1	5	2	1/2	7	7	7	7	7	7	2	7
$\mathbf{C_3}$	1/5	1/5	1	1/5	1/7	2	2	2	2	2	2	1/5	7
$\mathbf{C_4}$	1/2	1/2	5	1	1/3	5	5	5	5	5	5	1	5
$\mathbf{C_5}$	2	2	7	3	1	7	7	7	7	7	7	3	7
$\mathbf{C_6}$	1/7	1/7	1/2	1/5	1/7	1	1	1	1	1	1	1/5	1
$\mathbf{C_7}$	1/7	1/7	1/2	1/5	1/7	1	1	1	1	1	1	1/5	1
$\mathbf{C_8}$	1/7	1/7	1/2	1/5	1/7	1	1	1	1	1	1	1/5	1
$\mathbf{C_9}$	1/7	1/7	1/2	1/5	1/7	1	1	1	1	1	1	1/5	1
$\mathbf{C_{10}}$	1/7	1/7	1/2	1/5	1/7	1	1	1	1	1	1	1/5	1
$\mathbf{C_{11}}$	1/7	1/7	1/2	1/5	1/7	1	1	1	1	1	1	1/5	1
$\mathbf{C_{12}}$	1/2	1/2	5	1	1/3	5	5	5	5	5	5	1	5
$\mathbf{C_{13}}$	1/7	1/7	1/7	1/5	1/7	1	1	1	1	1	1	1/5	1

Policy	P_1	P_2
$\mathbf{P_1}$	1	3
$\mathbf{P_2}$	1/3	1

$\mathbf{A_1}$	$\mathbf{A_{11}}$	$\mathbf{A_{12}}$	$\mathbf{A_{13}}$	$\mathbf{A_{14}}$
$\mathbf{A_{11}}$	1	1/3	1/3	1/5
$\mathbf{A_{12}}$	3	1	1/2	1/4
$\mathbf{A_{13}}$	3	2	1	1/4
$\mathbf{A_{14}}$	5	4	4	1

$\mathbf{A_2}$	$\mathbf{A_{21}}$	$\mathbf{A_{22}}$	$\mathbf{A_{23}}$	$\mathbf{A_{24}}$
A_{21}	1	1	1	1
A_{21}	1	1	1	1
A_{21}	1	1	1	1
A_{21}	1	1	1	1

$\mathbf{A_3}$	$\mathbf{A_{31}}$	$\mathbf{A_{32}}$	$\mathbf{A_{33}}$
A_{31}	1	3	5
A_{32}	1/3	1	3
A_{33}	1/5	1/3	1

$\mathbf{A_4}$	A_{41}	A_{42}	A_{43}	A_{44}
$\mathbf{A_{41}}$	1	5	5	5
$\mathbf{A_{42}}$	1/5	1	1	1
$\mathbf{A_{43}}$	1/5	1	1	1
$\mathbf{A_{44}}$	1/5	1	1	1

$\mathbf{A_5}$	A_{51}	A_{52}	A_{53}	A_{54}	A_{55}	A_{56}	A_{57}	A_{58}
A_{51}	1	5	1	5	1	5	1	5
A_{52}	1/5	1	1/5	1	1/5	1	1/5	1
A_{53}		5	1	5	1	5	1	5
A_{54}	1/5	1	1/5	1	1/5	1	1/5	1
A_{55}	1	5	1	5	1	5	1	5
A_{56}	1/5	1	1/5	1	1/5	1	1/5	1
A_{57}	1	5	1	5	1	5	1	5
A_{58}	1/5	1	1/5	1	1/5	1	1/5	1

U_1	U_{11}	U_{12}	U_{13}	U_{14}	U_{15}	U_{16}	U_{17}
U_{11}	1	1	1/2	1/2	3	3	3
U_{12}	1	1	1/2	1/2	3	3	3
U_{13}	2	2	1	1	5	5	5
U_{14}	2	2	1	1	5	5	5
U_{15}	1/3	1/3	1/5	1/5	1	1	1
U_{16}	1/3	1/3	1/5	1/5	1	1	1
U_{17}	1/3	1/3	1/5	1/5	1	1	1

U_2	U_{21}	U_{22}	U_{23}	U_{24}
U_{21}	1	5	5	7
U_{22}	1/5	1	1	5
U_{23}	1/5	1	1	5
U_{24}	1/7	1/5	1/5	1

U_3	U_{31}	U_{32}	U_{33}	U_{34}	U_{35}
U_{31}	1	1	1/5	1	5
U_{32}	1	1	1/5	1	5
U_{33}	5	5	1	5	7
U_{34}	1	1	1/5	1.00	5
U_{35}	1/5	1/5	1/7	1/5	1

U_4	U_{41}	U_{42}	U_{43}	U_{44}
U_{41}	1	1	3	5
U_{42}	1	1	3	3
U_{43}	1/3	1/3	1	1
U_{44}	1/5	1/3	1	1

U_5	U_{51}	U_{52}	U_{53}	U_{54}
U_{51}	1	5	1	1
U_{52}	1/5	1	1/5	1/5
U_{53}	1	5	1	1
U_{54}	1	5	1	1

U_6	U_{61}	U_{62}	U_{63}	U_{64}
U_{61}	1	1	1	1
U_{62}	1	1	1	1
U_{63}	1	1	1	1
U_{64}	1	1	1	1

U_7	U_{71}	U_{72}	U_{73}	U_{74}
U_{71}	1	1/5	1/3	1/3
U_{72}	5	1	3	3
U_{73}	3	1/3	1	1
U_{74}	3	1/3	1	1

U_8	U_{81}	U_{82}	U_{83}	U_{84}
U_{81}	1	1	1	1/7
U_{82}	1	1	1	1/7
U_{83}	1	1	1	1/7
U_{84}	7	7	7	1

U_9	U_{91}	U_{92}	U_{93}
U_{91}	1	1	1/3
U_{92}	1	1	1/3
U_{93}	3	3	1

U_{10}	U_{101}	U_{102}	U_{103}	U_{104}	U_{105}
U_{101}	1	1	1/7	1	1
U_{102}	1	1	1/7	1	1
U_{103}	7	7	1	7	7
U_{104}	1	1	1/7	1	1
U_{105}	1	1	1/7	1	1

U_{11}	U_{111}	U_{112}	U_{113}	U_{114}
U_{111}	1	3	1	3
U_{112}	1/3	1	1/3	1
U_{113}	1	3	1	3
U_{114}	1/3	1	1/3	1

U_{12}	U_{121}	U_{122}	U_{123}	U_{124}
U_{121}	1.00	1/5	7	7
U_{122}	5	1	9	9
U_{123}	1/7	1/9	1	1
U_{124}	1/7	1/9	1	1

C_1	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}
C_{11}	1	1/3	1/3	1/3	1/3	1
C_{12}	3	1	1	1	1	3
C_{13}	3	1	1	1	1	3
C_{14}	3	1	1	1	1	3
C_{15}	3	1	1	1	1	3
C_{16}	1	1/3	1/3	1/3	1/3	1

C_2	C_{21}	C_{22}	C_{23}	C_{24}	C_{25}	C_{26}	C_{27}	C_{28}
C_{21}	1	1/3	1/3	1	1	1	1	1
C_{22}	3	1	1	3	3	3	3	3
C_{23}	3	1	1	3	3	3	3	3
C_{24}	1	1/3	1/3	1	1	1	1	1
C_{25}	1	1/3	1/3	1	1	1	1	1
C_{26}	1	1/3	1/3	1	1	1	1	1
C_{27}	1	1/3	1/3	1	1	1	1	1
C_{28}	1	1/3	1/3	1	1	1	1	1

C_3	C_{31}	C_{32}
C_{31}	1	1/7
C_{32}	7	1

C_4	C_{41}	C_{42}	C_{43}	C_{44}
C_{41}	1	1/3	1/4	2
C_{42}	3	1	1/3	5
C_{41}	4	3	1	7
C_{42}	1/2	1/5	1/7	1

C_5	C_{51}	C_{52}	C_{53}	C_{54}	C_{55}	C_{56}	C_{57}	C_{58}	C_{59}	C_{510}
C_{51}	1	1	1	1	1	1	1	1	1	1
C_{52}	1	1	1	1	1	1	1	1	1	1
C_{53}	1	1	1	1	1	1	1	1	1	1
C_{54}	1	1	1	1	1	1	1	1	1	1
C_{55}	1	1	1	1	1	1	1	1	1	1
C_{56}	1	1	1	1	1	1	1	1	1	1
C_{57}	1	1	1	1	1	1	1	1	1	1
C_{58}	1	1	1	1	1	1	1	1	1	1
C_{59}	1	1	1	1	1	1	1	1	1	1
C_{510}	1	1	1	1	1	1	1	1	1	1
C_{511}	1	1	1	1	1	1	1	1	1	1

C_6	C_{61}	C_{62}
C_{61}	1	1
C_{62}	1	1

C_7	C_{71}	C_{72}	C_8	C_{81}	C_{82}	C_9	C_{91}	C_{92}	C_{10}	C_{101}	C_{102}	C_{11}	C_{111}	C_{112}
C_{71}	1	1	C_{81}	1	1/2	C_{91}	1	1/2	C_{101}	1	1/2	C_{111}	1	1/2
C_{72}	1	1	C_{82}	2	1	C_{92}	2	1	C_{102}	2	1	C_{112}	2	1

C_{12}	C_{121}	C_{122}	C_{123}	C_{13}	C_{131}	C_{132}	C_{133}
C_{121}	1	3	5	C_{131}	1	1/3	1
C_{122}	1/3	1	3	C_{132}	3	1	3
C_{123}	1/5	1/3	1	C_{133}	1	1/3	1

P_1	P_{11}	P_{12}	P_{13}	P_{14}	P_{15}	P_{16}	P_{17}
P_{11}	1	1/2	1/2	1/2	1/2	1/2	1/2
P_{12}	2	1	1	1	1	1	1
P_{13}	2	1	1	1	1	1	1
P_{14}	2	1	1	1	1	1	1
P_{15}	2	1	1	1	1	1	1
P_{16}	2	1	1	1	1	1	1
P_{17}	2	1	1	1	1	1	1

P_2	P_{21}	P_{22}
P_{21}	1	1/2
P_{22}	2	1

Appendix E: Local weights and inconsistency index for level 2 & 3

Local Weights		ICI		Local Weights	ICI		Local Weights	ICI		Local Weights	ICI		Local Weights	ICI
b_{Access}	0.18	0.02	b_{A_1}	0.12	0.02	b_{U_1}	0.17	0.01	b_{C_1}	0.16	0.01	b_{P_1}	0.00
$b_{\text{Usability}}$	0.47		b_{A_2}	0.17		b_{U_2}	0.03		b_{C_2}	0.16		b_{P_2}	
b_{Content}	0.29		b_{A_3}	0.55		b_{U_3}	0.03		b_{C_3}	0.05			
b_{Policy}	0.06		b_{A_4}	0.10		b_{U_4}	0.17		b_{C_4}	0.11			
			b_{A_5}	0.06		b_{U_5}	0.03		b_{C_5}	0.22			
						b_{U_6}	0.08		b_{C_6}	0.02			
						b_{U_7}	0.13		b_{C_7}	0.02			
						b_{U_8}	0.03		b_{C_8}	0.02			
						b_{U_9}	0.03		b_{C_9}	0.02			
						$b_{U_{10}}$	0.03		$b_{C_{10}}$	0.02			
						$b_{U_{11}}$	0.03		$b_{C_{11}}$	0.02			
						$b_{U_{12}}$	0.22		$b_{C_{12}}$	0.11			
									$b_{C_{13}}$	0.05			

Appendix F: Local weights and inconsistency index for level 4

Local Weights		ICI	Local Weights		ICI	Local Weights		ICI	Local Weights		ICI
$\boldsymbol{b_{A11}}$	0.08	0.06	$\boldsymbol{b_{U11}}$	0.15	0.00	b_{C11}	0.07	0.09	b_{P11}	0.08	0.00
$\boldsymbol{b_{A12}}$	0.15		$\boldsymbol{b_{U12}}$	0.15		b_{C12}	0.21		b_{P12}	0.15	
$\boldsymbol{b_{A13}}$	0.21		$\boldsymbol{b_{U13}}$	0.27		b_{C13}	0.21		b_{P13}	0.15	
$\boldsymbol{b_{A14}}$	0.56		$\boldsymbol{b_{U14}}$	0.27		b_{C14}	0.21		b_{P14}	0.15	
$\boldsymbol{b_{A21}}$	0.25	0.00	$\boldsymbol{b_{U15}}$	0.05		b_{C15}	0.21		b_{P15}	0.15	
$\boldsymbol{b_{A22}}$	0.25		$\boldsymbol{b_{U16}}$	0.05		b_{C16}	0.07		b_{P16}	0.15	
$\boldsymbol{b_{A23}}$	0.25		$\boldsymbol{b_{U17}}$	0.05		b_{C21}	0.08	0.00	b_{P17}	0.15	
$\boldsymbol{b_{A24}}$	0.25		b_{U21}	0.61	0.08	b_{C22}	0.25		b_{P21}	0.33	0.00
$\boldsymbol{b_{A31}}$	0.63	0.03	b_{U22}	0.17		b_{C23}	0.25		b_{P22}	0.67	
$\boldsymbol{b_{A32}}$	0.26		b_{U23}	0.17		b_{C24}	0.08				
$\boldsymbol{b_{A33}}$	0.11		b_{U24}	0.05		b_{C25}	0.08				
$\boldsymbol{b_{A41}}$	0.63	0.00	b_{U31}	0.14	0.07	b_{C26}	0.08				
$\boldsymbol{b_{A42}}$	0.13		b_{U32}	0.14		b_{C27}	0.08				
$\boldsymbol{b_{A43}}$	0.13		b_{U33}	0.54		b_{C28}	0.08				
$\boldsymbol{b_{A44}}$	0.13		b_{U34}	0.13		b_{C31}	0.13	0.00			
$\boldsymbol{b_{A51}}$	0.21	0.00	b_{U35}	0.05		b_{C32}	0.88				
$\boldsymbol{b_{A52}}$	0.04		b_{U41}	0.41	0.009	b_{C31}	0.13	0.00			
$\boldsymbol{b_{A53}}$	0.21		b_{U42}	0.36		b_{C41}	0.12	0.03			
$\boldsymbol{b_{A54}}$	0.04		b_{U43}	0.12		b_{C42}	0.27				
$\boldsymbol{b_{A55}}$	0.21		b_{U44}	0.11		b_{C43}	0.54				
$\boldsymbol{b_{A56}}$	0.04		b_{U51}	0.31	0.00	b_{C44}	0.06				
$\boldsymbol{b_{A57}}$	0.21		b_{U52}	0.06		b_{C51}	0.09	0.00			
$\boldsymbol{b_{A58}}$	0.04		b_{U53}	0.31		b_{C52}	0.09				
			b_{U54}	0.31		b_{C53}	0.09				
			b_{U61}	0.25	0.00	b_{C54}	0.09				
			b_{U62}	0.25		b_{C55}	0.09				
			b_{U63}	0.25		b_{C56}	0.09				
			b_{U64}	0.25		b_{C57}	0.09				
			b_{U71}	0.08	0.01	b_{C58}	0.09				
			b_{U72}	0.52		b_{C59}	0.09				
			b_{U73}	0.20		b_{C510}	0.09				
			b_{U74}	0.20		b_{C511}	0.09				
			b_{U81}	0.10	0.00	b_{C61}	0.50	0.00			
			b_{U82}	0.10		b_{C62}	0.50				
			b_{U83}	0.10		b_{C71}	0.50	0.00			
			b_{U84}	0.70		b_{C72}	0.50				
			b_{U91}	0.20	0.00	b_{C81}	0.33	0.00			
			b_{U92}	0.20		b_{C82}	0.67				
			b_{U93}	0.60		b_{C91}	0.33	0.00			
			b_{U101}	0.09	0.00	b_{C92}	0.67				
			b_{U102}	0.09		b_{C101}	0.33	0.00			
			b_{U103}	0.64		b_{C102}	0.67				
			b_{U104}	0.09		b_{C111}	0.33	0.00			
			b_{U105}	0.09		b_{C112}	0.67				
			b_{U111}	0.38	0.00	b_{C121}	0.63	0.00			
			b_{U112}	0.13		b_{C122}	0.26				
			b_{U113}	0.38		b_{C123}	0.11				
			b_{U114}	0.13		b_{C131}	0.20	0.00			
			b_{U121}	0.27	0.09	b_{C132}	0.60				
			b_{U122}	0.62		b_{C133}	0.20				
			b_{U123}	0.05							
			b_{U124}	0.05							

Appendix G: WEM's indicators rank ordered by their global weights

Indicators	Values (High to Low)	Indicators	Values (High to Low)	Indicators	Values (High to Low)	Indicators	Values (High to Low)
Usability	0.47	C32	0.012	P17	0.006	U91	0.003
Content	0.29	U11	0.012	C51	0.006	U92	0.003
Access	0.18	U12	0.012	C52	0.006	P11	0.003
U12	0.103	C22	0.012	C53	0.006	C131	0.003
A3	0.098	C23	0.012	C54	0.006	C133	0.003
U1	0.080	A14	0.012	C55	0.006	U22	0.003
U4	0.080	A41	0.012	C56	0.006	U23	0.003
U122	0.065	U84	0.011	C57	0.006	A42	0.002
U7	0.064	A33	0.010	C58	0.006	A43	0.002
C5	0.063	C12	0.010	C59	0.006	A44	0.002
A31	0.062	C13	0.010	C510	0.006	C81	0.002
Policy	0.06	C14	0.010	C511	0.006	C91	0.002
C1	0.047	C15	0.010	U111	0.006	C101	0.002
C2	0.047	A5	0.010	U113	0.006	C111	0.002
U6	0.039	U61	0.010	U123	0.005	U31	0.002
P1	0.037	U62	0.010	U124	0.005	U32	0.002
U72	0.033	U63	0.010	U71	0.005	A51	0.002
U41	0.033	U64	0.010	U51	0.005	A53	0.002
C4	0.031	U103	0.010	U53	0.005	A55	0.002
C12	0.031	U43	0.010	U54	0.005	A57	0.002
A2	0.031	U21	0.009	C82	0.004	U34	0.002
U42	0.029	U93	0.009	C92	0.004	C44	0.002
U121	0.028	C42	0.009	C102	0.004	U112	0.002
A32	0.026	U44	0.009	C112	0.004	U114	0.002
P2	0.022	C132	0.008	A13	0.004	C31	0.002
U13	0.022	U33	0.008	U15	0.004	A11	0.002
U14	0.022	C122	0.008	U16	0.004	U81	0.002
A1	0.021	A21	0.008	U17	0.004	U82	0.002
C121	0.020	A22	0.008	C21	0.004	U83	0.002
A4	0.019	A23	0.008	C24	0.004	U101	0.001
C43	0.017	A24	0.008	C25	0.004	U102	0.001
U2	0.015	P21	0.007	C26	0.004	U104	0.001
U3	0.015	C6	0.007	C27	0.004	U105	0.001
U5	0.015	C7	0.007	C28	0.004	U52	0.001
U8	0.015	C8	0.007	C41	0.004	U24	0.001
U9	0.015	C9	0.007	C11	0.003	U35	0.001
U10	0.015	C10	0.007	C16	0.003	A52	0.0004
U11	0.015	C11	0.007	C61	0.003	A54	0.0004
P22	0.015	P12	0.006	C62	0.003	A56	0.0004
C13	0.014	P13	0.006	C71	0.003	A58	0.0004
C3	0.014	P14	0.006	C72	0.003		
U73	0.013	P15	0.006	C123	0.003		