Supplier sustainability assessment validation and evaluation models

Scholte, T.

Award date:
2019

Technische Universiteit Eindhoven University of Technology

Department of Industrial Engineering & Innovation Sciences
Operations, Planning, Accounting and Control (OPAC) group

SUPPLIER SUSTAINABILITY ASSESSMENT VALIDATION AND EVALUATION MODELS

T. Scholte

In partial fulfilment of the requirements for the degree of
Master of Science in Operations Management and Logistics

Supervisors:
Dr. T. Tan (TU/e)
Prof. Dr. E. Demerouti (TU/e)
Dr. S.S. Dabadghao (TU/e)
M. Baren (Philips)
M.C. McNeill (Philips)

March 24, 2019

Keywords: Supplier Sustainability, Sustainability Assessment, Sustainability Scoring, Supplier Classification, Questionnaire Validation, Self-Assessment, Expected Correlations, Inconsistency Model, TOPSIS

Abstract

Auditing remains a popular monitoring approach for supplier sustainability assessment, whilst critics increase due to conflict of interest and supplier corruption, among others. It is suggested that focal firms should maximize collaborations with suppliers instead, although time and financial resources are lacking. Making use of self-assessment, with peer assessments, is most efficient to select suppliers for collaborations. The gap in literature found is the lack of truthfulness or accuracy of this approach. This thesis bridges this gap by validating both assessments based on connected questions and stating the harmful inconsistency of suppliers. This knowledge serves as initial insight in truthfulness and feedback to increase the consistency of answering. Next to that, the decision making process of Philips to select suppliers for collaborations is revised and aligned with both risk and improvement objectives. It is suggested that the simple TOPSIS method structures and improves the current process, as well as allowing adjustability and using the maximum number of resources available to Philips. Altogether, by the use of increasing accuracy of data, less shifted results should be present, and correct decisions can be made regarding the selection of suppliers for collaboration.

Executive Summary

Introduction

In 2004, Philips started assessing their suppliers on sustainability performance, by the help of third-party assessors. Since Philips believes that audits do not fully uncover the real sustainability levels, because audits are 'violation-based' methods, Philips worked towards a more constructive method in which honesty, transparency and continuous improvement are most important. In 2016, the audits stopped and the "Beyond Auditing" program started(Philips, 2017). Instead of "only trying to pass", suppliers are encouraged to be honest and improve continuously, regardless of their situation. The program thus focuses on collaboration, transparency, commitment and meeting agreed targets.

In order to retrieve accurate information and high credibility for sustainability assessment, the supplier's information source should be unbiased, accurate and truthful. However, sustainability is not one single concept that fits all, so this might differ per industry or per company. The goal of this research is thus to establish a more structured supplier sustainability evaluation model, although sustainability has different concepts among companies and/or industries. This starts with describing methods to validate the information source (i.e. supplier questionnaire data) and retrieving useful sustainability indicators for assessment. Based on this information, suppliers are evaluated and subsequent actions to improve their sustainability performance follow. In this way, this research contributes to the literature in terms of describing more standard steps from (accurate and unbiased) information source to supplier evaluation, and contributes to the business in terms of implementation of this supplier evaluation model within the SSP program and its supply chain. Next to that, initial insights in the suppliers' truthfulness are gained that can be used to increase the transparency and truthfulness of suppliers.

Literature Research

The first research question relates to supplier sustainability assessment approaches known in the field and concludes on what option is the most suitable for focal firms. Whilst auditing remains one of the most popular approaches, many critics relate to conflicts of interest, supplier corruption, auditing fatigue, and the pass/fail mentality. These critics align with the reasons of Philips to start the "Beyond Auditing" program. Literature suggests that focal firms should instead collaborate as much as possible with their first-tier suppliers, so that transparency, openness, sustainability performance, and shared value increase. However, most focal firms lack time and financial resources to collaborate with their (thousands of) suppliers, which limits the adoption of these collaborations. On the other hand, efficient use of monitoring approaches can select a subset of the most risky and/or low performing suppliers for these collaborations instead. Literature suggests that self-assessments, with the use of peer assessments, are the most efficient approach to use.

Data

By the use of Excel VBA all information regarding the assessment workbooks of the Beyond Auditing program since 2016 is gathered and consolidated into one dataset. This dataset possess information of 662 assessment workbooks from 272 unique suppliers in scope of the SSP program, with 2269 fields per assessment workbook. The information collected per assessment workbook and within scope of this research are the workbook/supplier characteristics, dashboards, self-assessment questionnaire (SAQ) answers and supporting evidence document (ED) availability. Besides that, this dataset is extended with spend per supplier and predicted relative improvements as provided by a fellow researcher within the same broader research project. To make accurate comparisons between the current situation and research models, some assessment workbooks are omitted (for the supplier evaluation model only) due to a lack of spend data.

When exploring the data, and more specifically the sustainability levels of suppliers, some interesting results are found that influences the choices made within this research. For example, it is found that suppliers improve over time within the SSP program, based on their sustainability level as calculated by Philips. One of the most important findings is the difference between suppliers that are visited on site by Philips (SSIP suppliers, which are expected to be in need of more time spend by Philips due to high risk) and the suppliers that are not (DIY suppliers, which are expected to be mature enough). SSIP suppliers are found to improve more than DIY suppliers, from one (yearly) sequence to another. This might indicate the effectiveness of visiting suppliers on site and collaboratively working together on improvement actions, which is supported by an average improvement of sustainability score when comparing it before and after the on-site visit.

Models

The second research question relates to validating the supplier's self-assessment questionnaire (SAQ) and supporting evidence documents (ED), so that initial insights in truthfulness are gained and more accurate and unbiased results are in place. When comparing the document availability of a subset of connect questions the focus is set on investigating the false positives, i.e. when the suppliers claims to possess a document in the SAQ whilst not being able to provide it in the ED, which are assumed to be the most harmful. These harmful inconsistencies could indicate window-dressing or social desirability, as well as unawareness or misinterpretation. By aggregating these outcomes per connected question, a harmful inconsistency (HIC) score is generated on question or supplier level.

The third research question relates to revising the decision making process of the supplier evaluation model (selecting suppliers for SSIP or DIY), which is currently unstructured and only based on risk criteria. However, the SSP program's objective is to improve as much as possible, which is thus not in line with this model. Two heuristic models are proposed as revision this

decision making process, which both take risk and improvement into account when selection suppliers for SSIP (on-site collaborations) and DIY. Next to that, the maximum number of resources available to Philips are taken into account as well, so that the group of SSIP suppliers never exceeds the time and financial resources. The risk-based heuristic model first ranks the group of suppliers into risk levels (high, medium, low) and then selects the suppliers from high to low risk, according to predicted improvement (again from high to low). The distance-based heuristic model makes use of the Technique for Order of Preference by Similarity to Ideal Situation (TOPSIS) method, which calculates the relative (Euclidian) distance of one supplier to the best-case supplier, and ranks accordingly. This is based on the objectives and accessory weights.

Results

To conclude on the inconsistency model, an average HIC-score of 23% is found which corresponds to less of a quarter of the connected questions that were not provided by suppliers on average. On the other hands, suppliers that answer consistently score both low and high (ED) scores, which means that not only the high scoring suppliers answer consistently. Furthermore, it is seen that suppliers learn over time (in terms of consistency) and this can even be increased by collaborating and visiting the supplier on-site. However, these findings cannot be statistically supported, although the SSP program is just a few years in place.

Besides that, it is suggested to use the distance-based (TOPSIS) heuristic as revision of the supplier evaluation model, because of its simplicity and adjustability. However, within the SSP program the use of either heuristic model already improves the current situation in terms of both risk and improvement potential, of which the latter was not use beforehand. Currently the weights assigned to the objectives are based on expert-opinion, although statistical testing would improve this model even further towards mathematical model for decision support. On the other hand, multiple other mathematical models might be of great value besides the TOPSIS model. Since it is mainly aimed to show the potential of revised decision making models, improvements remain possible.

Conclusion

In conclusion, with the knowledge gained from all three research questions, the implementation of both the inconsistency model (resulting in the HIC-score) and revised decision making process for supplier evaluation models (distance-based TOPSIS heuristic) sounds promising in terms of working towards accurate and unbiased data, as well as structuring the supplier evaluation model for the SSP program. The HIC-score give initial insights in truthfulness and can serve as feedback to increase accuracy, transparency and truthfulness. The distance-based TOPSIS heuristic allows adjustability of objectives and weights, and serves as first step towards decision support within the SSP program. Altogether, this research contributes to increasing accuracy and truthfulness of questionnaire answering, as well as standardizing the decision making process in place.

Preface

This is the last page I write in this thesis, so that I can take some time to look back on the incredible journey that has led me here. More than seven year ago I started studying at the Eindhoven University of Technology, whilst only two years later figuring out that I started at the wrong place and study. My lifelong dream of becoming a famous architect made contact with the real world and shifted towards the study that many already advised me to do: Industrial Engineering. This switch was scary at first, but whilst making new friends again and becoming an active member of the study association Industria, I felt the right decision was made. Years of studying in Paviljoen, attending drinks in the Villa, actively participating in Industria activities, and having the best time of my study life will always be in my mind.

Although writing my master thesis feels like the finishing touch of my student life, I can honestly say that it has been anything but an easy and restful time. Although the whole project is such an interesting topic to be part of, and I could easily spend multiple months or years on further researching the future directions of the SSP program of Philips, I am enjoying the fact that I am almost halfway this page before finishing and I can spend the rest of the page as thank-you-note .

At first, I want to thank all my supervisors for enabling me to write my thesis and trying to improve any aspect of it. Tarkan, thank you for all the time spent as my mentor over the last two years, giving me the opportunity to work on such an interesting project within the field of sustainability, and your humour during all meetings. Evangelia and Shaunak, thank you for your insights and feedback during the meetings that improved my research after each discussion. Thank you all for your knowledge and expertise, I really appreciate all feedback you shared with me.

Next to that, I want to thank all direct colleagues at Philips for the enjoyable time I spent while conducting my thesis. Marco, thank you for your inspiring conversations and vision of the whole research project. Dylan, thank you for your expertise and time spent on supervising me, I look forward to continue the good times we have. I want to thank all my direct colleagues and fellow researchers for making the office a fun place to return to every day, this really helped me through.

And last, but not least, I want to thank all my close friends and family for their support and laughter during my student life. My friends, you were always there for me and made me enjoy all time shared. My parents and brothers, thank you for all time shared with me during the weekends and holidays, as well as your support and faith in me. And the biggest thank you for my girlfriend, your unconditional love and presence in my life keeps me up and running every day!

Tom Scholte

Table of Contents

List of Figures

List of Tables

1. Introduction

In 2004, Koninklijke Philips N.V. (hereafter: Philips) started assessing their suppliers on their sustainability performance, by the help of third-party assessors. Since Philips believes that audits do not fully uncover the real sustainability levels, because audits are 'violation-based' methods, Philips worked towards a more constructive method in which honesty, transparency and continuous improvement are most important. In 2016, the (third-party) audits stopped and the "Beyond Auditing" program started (Philips, 2017). Instead of "only trying to pass the audit", suppliers are encouraged to be honest and improve continuously, regardless of their situation. The program thus focuses on collaboration, transparency, commitment and meeting agreed targets. This currently implemented approach is a reactive approach, which reacts to the results of the conducted assessments. One of the ultimate goals of the broader research project, "Supplier Sustainability Improvement", is to create a predictive approach. A shift is thus needed from reactive to predictive, so that Philips can act earlier, faster, and more efficiently. The first research project took the first step towards that future state, by creating a predictive model with 10 to 13 questions that have equal accuracy to the approximately 700 questions. The next step is to make sure that supplier rate themselves as fair as possible, so that sustainability levels are accurate. Smouter (2018) assumed that the final score resembled the actual sustainability, without looking at supporting evidence, which influenced its prediction accuracy and bias.

In the relatively new field of research in sustainability assessment, practices vary significantly in terms of method, weighting, and aggregating of sustainability indicators (Dobrovolskienė et al., 2017; Singh, Murty, Gupta, & Dikshit, 2008). These indicators of sustainability are often a composite value or consist of multiple values (e.g. environment, society, and economy). In order to retrieve accurate information and high credibility for sustainability assessment, the supplier's information source should be unbiased, accurate and truthful (Gualandris, Klassen, Vachon, & Kalchschmidt, 2015). In other words, before suppliers are assessed and subsequent actions follow, there should be clear and useful indicators, which are based on accurate and unbiased data.

The goal of this research is thus to establish a more standard supplier sustainability evaluation model, although sustainability has different meanings among companies and/or industries (Hahn & Scheermesser, 2006). This starts with describing methods to validate the information source (i.e. supplier data) and retrieving useful sustainability indicators for assessment. Based on this sustainability assessment, suppliers should be then evaluated and subsequent actions to improve their sustainability should follow from each evaluation. In this way, this research contributes to the literature in terms of describing more standard steps from information source to supplier evaluation, and contributes to the business in terms of implementation possibilities of this supplier evaluation model within the own organization and its supply chain.

This research is part of a broader research project, namely the "Supplier Sustainability Improvement" project as submitted to TKI Dinalog by the consortium of Eindhoven University of Technology, Philips Electronic NL BV, Fairphone and ELEVATE (Philips, 2017). The aim of the broader project is to create a predictive approach addressing the sustainability improvement of the suppliers, to develop decision support for supplier selection and structural improvement, and to implement sustainability within the own organization and its supply chain. But before these predictions can be made, supplier information and sustainability scores should be accurate and unbiased, and there should be a standard and structured model on how to evaluate suppliers, which influences the cooperation with the supplier and the focus on sustainability improvement actions.

1.1 Outline

The Cross-Industry Process for Data Mining (CRISP-DM) model of Chapman et al. (2000) is used as guidance within this research. This model divides projects into six phases, although iterative steps can be taken since it is rather a continuous cycle of steps to take, and the project might not end after this research. It is suggested that new insights might trigger future research, which seems appropriate to this research. The model divides projects in the following phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment.

The Business Understanding phase relates to chapter 2 and 4 within this research. In chapter 2 the scope of the SSP program is stated and chapter 4 relates to the literature research to supplier sustainability assessment approaches and scoring methods in the field. The gap in the current literature on supplier sustainability assessment and scoring can best be stated in twofold: no further insights in truthfulness of suppliers are known besides (first-, second-, third-party) auditing, and the step from supplier assessment data to decision making and taking action seems missing. With the knowledge gained, RQ1 can be answered. After that, in chapter 5, the Data Understanding and Data Preparation phases are explained. With the use of Excel VBA and programming language R this Data Understanding results in two datasets within the scope of RQ2 and RQ3. After that, in chapter 6 models are explained that relate to RQ2 (the validation of questionnaires, i.e. the inconsistency model) and RQ3 (the revised supplier evaluation model). Both relate to the Modelling phase and serve as a start of answering the research questions, although results are needed to support them. The results are presented and discussed in chapter 7, after which chapter 8 concludes on all research questions. Here it is mainly explained that with the use of the inconsistency model questionnaires can be validated and aggregated to a harmful inconsistency score (HIC-score) on supplier and/or question level. Next to that, attention is paid to the use and adjustability of the revised supplier evaluation model and improvement of the current situation. After that, the Evaluation phase has ended and both RQ2 and RQ3 can thus be answered too. In the following chapters the implementation within Philips (chapter 9), limitations of this research (chapter 10), and future research directions (chapter 11) are presented and elaborated on.

2. Scope

Since 2004, Philips has conducted thousands of sustainability audits at their suppliers that were done by a third party auditor. Nowadays, Philips does not hold interest in this data anymore. The first reason is unreliability, since the guidelines of those audits, the codes of conduct, changed. The comparison of the audit outcomes can thus not be done due to different content and focal points over time. The second reason is inaccuracy, because Philips discovered that audits cannot fully uncover the real supplier's sustainability level. By conducting audits suppliers only care about passing the audit, since that would mean a continuing relationship with their buyer. Because of the unreliable nature of this audit data (R. M. Locke, Qin, & Brause, 2007), this is left out of scope.

Since the introduction of the 'Beyond Auditing' approach within Philips in 2016, suppliers are yearly asked to conduct (1) self-assessment questionnaires (SAQs), (2) provide evidence documents (ED), and (3) are potentially visited for a site assessment (SA) (Philips, 2017). This SAQ (1) consists of approximately 700 questions, depending on answers given that could be followed up by more questions. This waterfall structure is designed such that more detailed questions open up when a supplier claims to possess certain attributes of the subject. The SAQ is divided into seven sections, of which the first section is an introduction and request for commitment of the supplier, and the second section requests general information only. The other five sections correspond to five topics of sustainability, according to the Responsible Business Alliance (RBA, formerly IECC) code, international standards and Philips' site questionnaires.

Filling in the SAQ (1) then results in a Supplier Dashboard score card with different scores, which can be seen in Figure 1 below. The SAQ results in 45 unique fields that all relate to one certain topic and element (maturity level). Each question of the SAQ relates to one topic and one element, and is assigned a specific weight to calculate the weighted scores (i.e. five topics scores, nine element scores, and final score). This is quite similar to ED (2) provided, which all relate to one of the fields again. Each ED (e.g. pictures, files, etc.) is check manually within Philips for its availability and quality, and receives a score of 0, 0.5, or 1 per ED. Within this initial process multiple scorecards exist over time. In short, at first, the supplier fills in the SAQ (1) and this results in a scorecard with a final score, namely the SAQ score. After that, the supplier submits the ED (2), of approximately 200 questions, which result in a second scorecard with a weighted score, namely the ED (validation) score. The latter dashboard is used to compare and assess suppliers on.

Dashboard

Topics	Weighted Section Scores	Policy	Procedures	Implementation	Management Responsibility	Communication	Risk control	Target Setting &Tracking	Corrective action approach	Supplier management
Quality	99%	100%	100%	100%	100%	99%	100%	95%	100%	100%
Environment	82%	100%	72%	87%	100%	82%	65%	81%	83%	83%
Health and Safety	78%	27%	63%	73%	83%	90%	71%	98%	100%	100%
Business Ethics	40%	100%	65%	46%	12%	76%	23%	3%	30%	59%
Human Capital	74%	100%	68%	58%	65%	84%	81%	74%	100%	82%
Weighted average	75%	87%	74%	73%	72%	87%	68%	69%	83%	85%

Figure 1: Example of the supplier dashboard

When the Final score is known and communicated to the supplier, suppliers are evaluated and allocated in different categories, namely Best in Class (BiC), Do It Yourself (DIY), SSIP (Supplier Sustainability Improvement Plan), or No Zero Tolerance (NZT). This allocation and its characteristics can be seen in Figure 2 below. Currently, only the classifications SSIP and DIY are used. DIY suppliers are seen as mature enough to make sufficient improvement based on a development plan provided by Philips. On the other hand, SSIP suppliers show low maturity and receive extra attention by Philips, which consists of a site assessment (3) and discussion with the supplier to agree upon improvements to be made. With a site assessment Philips actively takes part in the evidence gathering process on the site of the supplier, which also results in yet another score, often slightly higher than the validated SAQ score. It is expected that this is caused by the insecure nature of the suppliers who do not feel fully comfortable yet to share all evidence, although this is (partially) valid evidence.

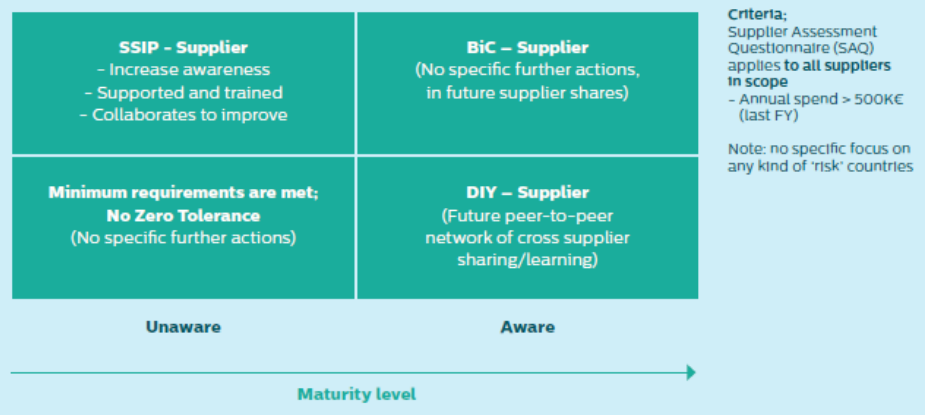

Figure 2: Supplier evaluation model

This Supplier Sustainability Performance program is executed every year, but once the SAQ is filled in, the supplier only have to update changes to the SAQ, so that not all questions need to be answered again. In short, the SAQ (1) is thus validated by ED (2), which can be validated again by a site assessment (3). For this reason, the validated ED is seen as more objective than the SAQ. When the different scores are analysed, the average ED is 25% lower than the average SAQ score.

In the earlier research the quality topic was already left out of scope because it is transferred to the supplier quality department, and the final score was corrected for this fact. The SAQ questions belonging to the quality chapter were still used as input for the model, since these can be important for predicting sustainability. Besides that, the ED was left out of scope for the model since the prediction should be made based on easy-to-retrieve information. Within this research the evidence is in scope, to get insights in the consistency of answering, plus to bridge the gap between SAQ and ED data. Next to that, this ED data also included (potential) zero tolerances. At last, the decision making behind the supplier evaluation concept (Figure 2) is revised.

3. Research design

This research is the second master project in line and builds upon the previous research as carried out by Smouter (2018). That research implemented feature selection on a real dataset of supplier sustainability data of Philips in order to identify a subset of variables that can predict the supplier's sustainability level, with the use of a random forest. It was found that a subset of questions, 10 to 13 instead of approximately 700, already predicts as accurately as using all (approximately) 700 questions. In this proactive way, supplier sustainability levels can be indicated and most improvement areas can possibly already be found beforehand. Unfortunately, a subset of only general information questions did not yield satisfactory prediction results. One of the suggestions made for future research was the bridge the gap between self-assessment and supporting evidence so that suppliers rate themselves as truthful as possible.

Since multiple projects, with different approaches, are done in parallel regarding the broader research project, this research is mainly focussed on validation of supplier information and taking the first steps towards revising the current evaluation model. So, the objectives of this research are to create a methodology on how to validate the SAQ and ED, and to revise the decision making of supplier evaluation model. By bridging the gap between SAQ and ED, the actual situation is best represented and scores might be less shifted when making decisions based on these scores (Distelhorst, Hainmueller, & Locke, 2016; Gualandris et al., 2015; Yamin, Parker, Xi, & Stanley, 2017). Next to that, by structuring the current evaluation model with appropriate criteria, this results in a more standard and usable supplier evaluation model (Foerstl, Reuter, Hartmann, & Blome, 2010). By the use of accurate and unbiased data, correct decisions can actually be made in the supplier evaluation model.

3.1 Research questions

The following research questions arise from the research objectives, and relate to phase 2.2 and 2.3 of the "Supplier Sustainability Improvement" project proposal as submitted to TKI Dinalog:

- **RQ1: What approach can be used to assess suppliers and establish actual sustainability scores?**

This research question contributes to finding the best approach on how to assess suppliers' sustainability and how to establish sustainability scores that best reflect the actual situation. Currently, differences in the literature exist between scoring suppliers' sustainability in one or multiple values, i.e. in one composite value or multiple values (Foerstl et al., 2010; Wilhelm, Blome, Bhakoo, & Paulraj, 2016). Next to that, methods to weight and aggregate sustainability indicators are investigated. The deliverable resulting from this RQ1 is thus an approach how to assess suppliers, an overview of methods how to weight and aggregate these assessed sustainability indicators, and what variables should be used for decision making.

- **RQ2: What method can be used to validate the suppliers' (SAQ and evidence) data?**

This research question contributes to creating a method to bridge the gap between self-assessment (SAQ) and supporting evidence (ED), and validating how inconsistent suppliers assess themselves. Currently, there is no use of feedback on question level, so inaccurate scores might be present. These inaccurate scores might lead to shifted results (Distelhorst et al., 2016; Gualandris et al., 2015; Yamin et al., 2017). Insights are given into the current situation of inconsistent data, i.e. the supplier's data is compared to each other and investigated to find out whether deviations in scores and/or answers deviate within the process and over time. The deliverable resulting from this RQ2 is thus a method how to validate the supplier's data in order to increase consistency among suppliers and to gain initial insights in the truthfulness of suppliers.

- **RQ3: How should suppliers be classified and what characteristics are appropriate to use?**

This research question contributes to structuring the currently used evaluation model and extend it with new objectives and the maximum number of resources available at the focal firm. This is firstly done by setting up the supplier evaluation model qualitatively, and secondly evaluating the model with quantitative data. This is done by finding mainstream mechanisms to connect theory and practice (e.g. RQ1 and RQ2) (Tachizawa & Wong, 2015; Zhou & Xu, 2018) and comparing the results with the currently used model. The deliverable resulting from this RQ3 is thus a revised decision making method of the supplier evaluation model that states the best method to use.

4. Literature research

In 1987, when the Brundtland report "Our common future" was released, the consequences of our economic behaviour were stated and it was suggested that change was needed in business activities; namely a change towards a (more) dynamic perspective (World Commission on Environment and Development (WCED), 1987). Besides the economic behaviour, there should also be an understanding about, for example, human rights, labour, environment, and anti-corruption (United Nations Global Compact (UNGC), 2018). Instead of approaching them as traditional stand-alone issues, these should be integrated and interrelated to serve as dimensions of a higher-order construct (Carter & Jennings, 2002, 2004). An important limitation of this operationalization is the absence of an economic perspective (Carter & Rogers, 2008).

Although all definitions of sustainability slightly vary throughout the years (Ahi & Searcy, 2013), most definitions have an overlapping core perspective, namely the triple bottom line (TBL) concept developed by Elkington (1998). This theoretical framework of sustainability integrated the available (stand-alone) literature, and resulted in multiple criteria: the concept namely considers and balances economic, environmental, and social (or in terms of the three pillars: profit, planet, and people) goals and suggest that there exists a core area in which not only environmental and social gains are found, but also long-term economic benefits and thus competitive advantage (Dyllick & Hockerts, 2002). Despite the number of pillars (A. Van Weele & Van Tubergen, 2017), it is argued that only considering TBL aspects is insufficient from a practical point of view (Wu, Liao, Tseng, & Chiu, 2016; Wu, Zhu, Tseng, Lim, & Xue, 2018). It is stated that by defining co-benefits, i.e. generating win-win situations for reaching individual and universal goals (Kwan & Hashim, 2016) more clearly, companies do not have to rely on the TBL aspects only when developing and selecting the appropriate sustainability indicators.

Taking the present economics, globalization, outsourcing to developing countries, market changes, and demand uncertainty into account, supply chains are getting more important (Andersen & Skjoett-Larsen, 2009; Varma, Wadhwa, & Deshmukh, 2006) and it is necessary for organizations to start interacting within their supply chain. Focal firm, i.e. buying firms, are nowadays expected to assure social and ecological sound production within their supply chain by external stakeholders (Foerstl et al., 2010), and irresponsible supplier behaviour is projected to the focal firm. This could cause adverse publicity, reputational damage, and costly legal obligations (Amaeshi, Osuji, & Nnodim, 2008; Carter & Jennings, 2004; Hoejmose, Roehrich, & Grosvold, 2014). In the last decade, the transition from economic-based collaborations to environmental and societal relationships was already visible in supplier selection, although it is suggested that most environmental and societal reviews happen after selection (Ladd & Badurdeen, 2010).

To create so-called 'shared value' in the supply chain requires responsible purchasing by creating transparency, traceability, collaboration, and implementation of sustainability (Lintukangas, Hallikas, & Kähkönen, 2015). Companies can be ranked in different stages of maturity on this integration of sustainability in their value chain. Corporate Social Responsibility (CSR) relates to this responsibility to integrate sustainability into the company and its value chain. CSR thus relates to integrating sustainability, in short, and different models exist which focus on the steps to make and achieve this integration, while other models rather focus on external practices or stages of maturity. More and more MNCs influence the CSR policies of their suppliers by integrating and monitoring sustainability indicators in the supplier selection process to thoughtfully select based on sustainability performance (Matthyssens & Faes, 2013). Different adoption models and stages exist in the literature, which all identify multiple stages that companies go through when developing a sense of corporate responsibility as they move along the learning curve (Nidumolu, Prahalad, & Rangaswami, 2009; Van Tulder, Van Tilburg, Francken, & Da Rosa, 2013; A. Van Weele & Van Tubergen, 2017; Zadek, 2004; Zimmerli, Holzinger, & Richter, 2007).

In order to understand the impact of the supply chains, supplier assessment thus seems important to performance, reputation, and sustainability. Besides that, companies can easily seek appropriate action to minimize the impact, which can be done effectively by setting up regular cycles of assessment (GRI, 2014). In many cases, this assessment is done by sending out questionnaires to suppliers to retrieve the information required to assess. The problem may arise that suppliers are lacking knowledge or tools, but this can be resolved by additional training or specific industry standard questionnaires. Often suppliers do not have the knowledge on how to improve sustainability individually and buyers collaborate with those suppliers, which is often argued to be effective to encourage improved supplier performance (Bowen, Cousins, Lamming, & Farukt, 2009; Distelhorst et al., 2016; R. Locke & Romis, 2007; Vachon & Klassen, 2006). *"One strategy is to invest in improving the sustainability of suppliers, so that it can 'weather the storm' and maintain margins despite cost hikes"* (Bouchery, Corbett, Fransoo, & Tan (Eds.), 2017).

4.1 Supplier sustainability assessment approaches

When focal firms evaluate and improve their (first tier) suppliers' sustainability performance, multiple methods exist to do so, e.g. through (self-) assessments or auditing. Focal firms are then able to set up a baseline of performance, and track the supplier's development and compliance over time (Andersen & Skjoett-Larsen, 2009; Kashmanian, 2015). The next step is to come up with initiatives for continuous (sustainability) improvement to meet the goals and objectives (A. J. Van Weele & Vivanco, 2014). The full disclosure of sustainability performance is promoted among suppliers, since evidence shows that good environmental and social performance often relates to and results in increasing economic performance (Coyne, 2006). Within this part the various approaches are reviewed and compared to each other (Lee & Klassen, 2008).

Monitoring-based approaches are often chosen when mitigating supplier risks to the focal firm (Hajmohammad & Vachon, 2014). In short, suppliers' information is collected, criteria are set to score against, and suppliers are asked to report on these dimensions (Seuring & Müller, 2008). To gain the so-called visibility of violations or improved sustainability performance requires establishing mechanisms to detect them. Possible mechanisms can include sharing data, and thus trust and collaboration, direct monitoring, or reporting from interested parties. Different monitoring-based approaches exist and these can be subdivided into: first-, second-, and third-party monitoring approaches, dependable on who conducts the audit.

First-party monitoring relates to monitoring practices executed by the first party (thus the supplier itself), i.e. self-assessments. In the field of sustainability assessment of suppliers it is researched whether the use of self-assessment is feasible and what differences exist between self-assessment and monitoring assessment (Subic, Shabani, Hedayati, & Crossin, 2013). In conclusion, a large gap between these assessments exists which is argued to be the suppliers' misconceptions on its capabilities. However, first-party monitoring is in terms of financial resources and time the most efficient option (Lippmann, 1999). On the other hand, the realism of self-assessments can only be determined when comparing it to other judgements (peer assessments), and differences often exist. Several studies claim that peer assessment is more accurate, due to individuals being (un)able to accurately assess themselves. These errors can be intentional and unintentional, but in the field of sustainability, the intentional errors are the most important. Many of these errors are due to social desirability or window-dressing (to comply with the focal firm's code of conduct and maintain the relationship) (Allen & Van Der Velden, 2005; Brown & Harris, 2014).

Second-party monitoring relates to monitoring practices executed by the second party (thus the focal firm) i.e. auditing. The main idea behind the process of auditing is to 'physically' check the supplier's state of processes and compliance with regulations and the focal firm's code of conduct (Coyne, 2006; Pimenta & Ball, 2015). Since auditing requires resources of the focal firm, this is often limited to those suppliers that are strategically important and/or in a long-term relationship. Another, obvious, alternative is to focus on the suppliers with the highest risk of causing significant damage (Harland, Brenchley, & Walker, 2003). This should start with effective identification of those suppliers, and what incentives should be in place for them (Baden, Harwood, & Woodward, 2009; Green, Morton, & New, 1996; Jiang, 2009). Audits are seen as systematic and reliable tool for supplier compliancy and supplier risk reduction, although audits cost financial resources and time. Others argue that auditing works when states lack systematic factory inspections due to capacity or resource restrictions (Bartley, 2005; O'Rourke, 2003; Rodríguez-Garavito, 2005).

Third-party monitoring relates to monitoring practices executed by the third party (thus an independent auditor), i.e. auditing. Although conflicts of interests are one of the main critics of auditing, one possibility to overcome this problem is by conducting the audit through a third party,

as independent assurance (Coyne, 2006). Besides that, audits by a third party are also useful when no sector-wide approach exists, or when the focal firm has limited financial resources and time (Sodhi & Tang, 2009). Best practices are found in (third party) firms with distinct and separate units per sustainability topics, so that independence and objectivity can be guaranteed. Independent, nonbiased assessments then result in transparency and accurate reports. On the other hand, conflicts of interest are still found in this way of (third party) auditing. The degree of interdependence, i.e. absence of potential conflicts of interest, naturally varies per firm and should be investigated to close the credibility gap.

Regardless of the auditor or audit type, many critics exist on the matter of auditing as sustainability assessment method and whether this is the right way of working. These critics/problems can be summarized and divided into a number of factors: conflicts of interest, auditing fatigue, supplier corruption, and pass/fail mentality. Each factor has its contribution to the critics that exist, although some of them are not solely the result of auditing, but also the result of monitoring on its own. Next to that, different types of audits exist: individual, shared, and joint audits. In other words, the audit results can be used for individual (focal firm) intentions only, but can also be shared across other companies when mutual recognition is found in the conducted audit (Kashmanian, 2015). Another option is joint audits, in which multiple companies jointly conduct the audit.

Within the monitoring-based approaches, there are thus many critics about what to audit, who audits, and how to audit, so the question arises whether auditing is effective for improving sustainability performance. Ian Spaulding, from the company ELEVATE, even states that *"the traditional auditing model is broken, since it does not result in greater visibility, it does not reward good behavior, and it leads to genuine risk mitigation"* (Donaldson, 2014). Previous research shows that third-party monitoring outperforms second-party monitoring, which outperforms first-party monitoring, in terms of accuracy (Darnall & Carmin, 2005). This is obviously influenced by the conflict of interest associated with each approach, and it is stated that sanctions or even removal positively influence the accuracy even more. Even more accurate results would be gained from unannounced audits by third parties with high credibility, since this could prevent suppliers to hide the non-compliances (Egels-Zandén, 2007; O'Rourke, 2003; Teuscher, Grüninger, & Ferdinand, 2006).

First-party monitoring does not have enough accuracy on its own to solely rely on as approach, but might be a feasible method in terms of financial resources, time, and accuracy when corrected with peer assessments (Foerstl et al., 2010; Pimenta & Ball, 2015). The use of this combination is found at both supplier selection and development, so only compliant suppliers enter the focal firm's supply base. Low performance or high risk suppliers automatically trigger actions to follow.

Support-based approaches are often chosen to increase the suppliers' potential and capacity, rather than on an immediate outcome (Vachon & Klassen, 2006). This more direct involvement of the focal firm often leads to increasing the transfer of knowledge, integration, partnership, and collaboration (Cousins, 1999; Lamming & Hampson, 1996). Since the difference in power between the focal firm and supplier is often large, this relationship rather results in support, i.e. compensation, from focal firm to supplier than collaboration "as equals" (Lee & Klassen, 2008). Typical activities performed in this support-based approach are trainings, education programs, sponsoring, sharing information and knowledge, and joint research. Multiple researches propose that the focal firms' support-based approaches positively relate to sustainability improvements (Geffen & Rothenberg, 2000; Grant & Baden-Fuller, 1995; Krause, Scannell, & Calantone, 2000; D. F. Simpson & Power, 2005). By the use of support-based approaches, the limited know-how of suppliers changes to tacit and explicit knowledge and skills in sustainability performance. It should be noted that this collaboration is often funded by the focal firm, government, or both.

The evidence backing up support-based approaches is greater and gaining growing attention in the field (Chen et al., 2017). Collaboration within the supply chain is used for information and knowledge sharing, improving performance, reducing costs and inventories, and ultimately increasing competitive advantage (Soylu, Oruç, Turkay, Fujita, & Asakura, 2006). It is stated that these collaborations are key for a sustainable supply chain management (Lu, Wu, & Kuo, 2007), and that long-term relationships might exceed the issues found in monitoring-based approaches (D. Simpson, Power, & Samson, 2007).

Another, slightly different, field of research is to score suppliers not (only) on their sustainability performance, but their sustainability maturity (or a combination of both). The current motivators for monitoring are often regulatory compliance, risk mitigation, and brand positioning (Bouchery et al., 2017), but when this can be linked to increased (corporate) value, the suppliers are more probable to adopt sustainability reporting on their own. This adoption is often limited because suppliers are still in the lower (CSR) stages of adopting sustainability and do not want to waste their time and money on limited benefits. When this adoption is promoted or focused on, e.g. by determining the suppliers' maturity level and encouraging evolution through these stages of maturity, the benefit for the suppliers increases motivation for sustainability performance.

4.2 Supplier sustainability scoring methods

This part relates to the sustainability scores given to the suppliers and what methods can be used to calculate these scores, which can possibly be the output of the support- or monitor-based approaches previously explained. Different models and frameworks exist on how to assess sustainability, mostly with the aim of making comparisons possible between different companies. Firstly an overview is given about the differences between single and multiple scores, and secondly the methods on how to assess sustainability in a single score are explained.

Krajnc and Glavič (2005) designed a sustainability index, as composite of the "triple bottom line" concept (thus economic, environmental, and social aspects), as a single simplified score. This single score would enable comparisons to different companies or benchmarks within a sector, as well as more efficient assessment and less criteria for the decision-makers. Besides that, they stated the importance of having several indicators as composition to cover the fullest spectrum of sustainability. These result in sustainability sub-indices that reveal the performance of the company compared to other years, other companies or benchmarks. The single score can then be used as information regarding trends, reflection, highlight of opportunities, and early warnings. It is even stated that higher composite scores can be interpreted as a higher likelihood of achieving and remaining sustainable in the future.

Chatterji and Levine (2006) argue there is a difference between relative and absolute performance scores, which are dependent on the overall goal of the index. When using absolute scores (based on fixed pass/fail criteria), mining companies are expected to almost never outperform software companies and might discourage them to improve. On other hand, when using relative scores (based on competition amongst each other or within the sector), the "best" tobacco company could still outperform average software companies, whilst this should be no competition at all in absolute scores (for the average software company, naturally). They conclude that there is no single correct principle, and this is more dependent on context and goal of the index.

To compare single and multiple score assessments on sustainability, Griffiths, Boyle, and Henning (2018) summarized the strengths and weaknesses of sustainability assessment tools. They agreed upon the suggestion of Bartke & Schwarze (2015) that there is no perfect tool and selection of a tool is always a trade-off between adequacy and understandability of the tool. This is indeed in line with the trade-off between simplification and potential loss of visibility, and making sustainability measurable and manageable through single scores. Other common strengths are the multi-dimensionality and criteria-based character which provide a common metric and language, the possibility to encourage striving for higher levels of sustainability performance, and clear communication. On the other hand, common weaknesses are seeking to minimize 'unsustainability' rather than creating something sustainable, the tendency to 'points chase' through requirements and thresholds, and not capturing the entire scope in one single endpoint (Griffiths et al., 2018).

The researcher mostly agrees with López, Garcia, and Rodriguez (2007), McWilliams, Siegel, & Wright (2006), and (Seager & Prado, 2017) that sustainability has no single concept, nor a commonly accepted method of measuring it. As pointed out by Chatterji and Levine (2006), nuance is always needed to approach the best scoring method in its context. Although much effort in put into generating a generalizable approach of sustainability assessment, the best scoring method is always dependent on the goal of the scoring index, its context, the diversity of views,

and trade-off between adequacy and understandability. Since there is no general scoring method without challenges or disadvantages, these scoring methods/tools continue to grow.

Dependent on who uses the sustainability score(s), solely relying on one single score is feasible (not per se infeasible) when correctly approached and used. But since one single score does not capture the whole picture and makes comparisons only slightly possible, using multiple scores seems most efficient. This obviously results in more support for decision-makers and assist better interpretation and transparency, although single scores can contribute to sound and effective decision-making (Kägi et al., 2016). To create this transparency, both midpoint and endpoint indicators (i.e. indicators and composite indices) should be communicated. This could also prevent the compensation (or offset) issue when aggregating, since one single endpoint makes it possible for low midpoints to hide.

The most common procedure of measuring (sustainability) in one single point is the following: selecting indicators, grouping indicators, weighting indicators, judging, normalizing indicators, calculating sub-indices and combining in one single endpoint (Krajnc & Glavič, 2005). In this research the TBL is used to group the indicators, but this could be any other grouping as well. Next to that, when one wants multiple scores (instead of one), this should be possible with the same procedure and steps. Although this literature study does not focus on the specifics of these steps, some insights are given in the weighting, and combining (aggregating) steps.

Based on the literature analysis of when to use what weighting and aggregating methods for sustainability indicators, the most common methods used for weighting can be subdivided in: equal, statistic-based, and expert opinion-based weights (Gan et al., 2017). The use of expert opinion-based weights is more powerful at finer scales than at coarser scales, due to lower costs and because local scales cannot directly be used for larger scales (Van de Kerk & Manuel, 2008). When the purpose of the measurement is to assess strong sustainability, i.e. no compensation is allowed, expert opinion-based weights are recommended. In all other cases (thus for fine scales, weak sustainability, assessing or comparing), all these methods are recommended or applicable.

Although the equal weighting method is simple and straightforward, no insights in indicator relationships can be gained. This transparency is better gained with the use of expert opinion-based methods, which can be used for both quantitative and qualitative data. The main drawbacks of this method are that urgency or concerns are measured rather than importance, and that these opinions are mostly region-specific. At last, the use of statistic-based methods integrates the whole procedure of selecting, weighting, and aggregating and enables the expression of statistical significance when comparing. On the other hand, multiple solutions and multi-collinearity may exist, which influences the sensitivity to outliers.

The most common methods used for aggregating can be subdivided in: additive aggregation, geometric aggregation, and non-compensatory aggregation (Gan et al., 2017). When the purpose of the measurement is to represent strong sustainability, it is argued that one method cannot represent this strong sustainability, so combined aggregation methods should be used. In this way, both threshold values and non-compensability are taken into account.

The use of additive aggregation methods, which sum up the normalized indicators, are by far the most common method in the field, due to its simplicity, transparency, and usability for sensitivity analysis. On the other hand, no synergy or interdependency between the sub-indices may exist among the indicators, which seems unrealistic. Next to that, substitution and compensation are possible with the use of this method. The use of geometric aggregation methods, which multiply the normalized indicators, shares the same abilities, although there exists a limitation on the compensability between indicators and sensitivity analysis cannot be used. This means that trade-offs between indicators still exist in these two methods, but the use of non-compensatory aggregation makes this unacceptable. The Law of the Minimum is often used, which suggests that sustainability is limited by the dimension with the lowest performance, so that no compensation is possible. Computational limitations and loss of information on intensity are seen as potential drawbacks of this method. On the other hand, the maturity level of suppliers often limits their performance, and this should be taken into account when relying on the Law of the Minimum.

4.3 Best practice for Philips

In conclusion, the current best practice is to implement support-based approaches with suppliers, to create shared value and so that information and knowledge can be shared, performance improves, and long-term collaborations are established (A. Van Weele & Van Tubergen, 2017). This is all under the assumption of the focal firm having enough time and financial resources to implement this. When the focal firm is lacking time and financial resources, another approach is needed. It is argued that the use of both monitoring- and support-based approaches provide a synergetic effect, resulting in even higher performance development (Lee & Klassen, 2008). When the focal firm's goal is to improve, this is the best option. When the focal firm's goal is to create as much transparency, i.e. most truthful information, and time and financial resources are lacking, monitoring-based approaches are the alternative best practice. In terms of truthfulness in monitoring-based approaches, the (respectively) best options are third-party, second-party, and first-party monitoring. This choice is not only dependent on truthfulness, but time and financial resources too, but joint and/or shared audits already reduce this burden.

Next to that, it is important to state that the maturity level of both focal firm and suppliers should be taken into account when evaluating and improving the supply chain's sustainability performance. The suppliers' maturity level (and supplier characteristics) influences the expected maximum sustainability performance per supplier, and thus its path to improvement. Besides that,

the focal firm's maturity level is assumed to be high until now, but when this would be low the focal firm might not consider all options given to strive for sustainability, as explained in previous chapters. Low maturity could mean only complying to the law, whilst high maturity could mean joint collaborations, joint responsibility, and/or supplier selection on ability to engage in sustainability issues (thus supplier maturity level) rather than initial sustainability performance (Sanders, Cope, & Pulsipher, 2018; Van Lakerveld & Van Tulder, 2016). The focal firm's maturity level is expected to determine its goal in sustainable supply chain management practices too.

Furthermore, it can be concluded that there is no single concept and/or method of sustainability, and the choice of number of endpoints and procedure steps, especially weighting and aggregating, all depends on the purpose, scales and concepts of the sustainability tool. It is thus most important to know when to use what, because no current best practice exists for all situations. The summary of "when to use what" is best described in the research of Gan et al. (2017), which denotes what methods are recommended, applicable, and not to be used given certain situations. It is suggested there exists no combination of one single weighting and aggregating method that is the best for all. Especially within the weighting methods there exist some differences, thereby taking into account that statistic-based and expert opinion-based weighting methods include multiple possible methods. On the other hand, the combined method can be seen as best practice within the aggregation methods, which can be used for all situations.

Regardless of the methods used, another important decision to make is whether to use one or multiple endpoints in the sustainability measurement, and how to represent this. It can be concluded that it is fairly possible to judge one's sustainability level in one endpoint to assess and compare the suppliers, although midpoints are recommended when making decisions. This is especially the case when the aggregating method allows compensation. The best practice would then be to use a combined method of using multiple midpoints for extra information and circumstances per supplier, as well as one single endpoint to assess and compare. This is all under the assumption that decisions are only made using the endpoint(s). Next to that, promising suggestions have been found in the use of maturity levels to evaluate and improve the supply chain's sustainability performance.

4.4 Conclusion

In terms of Philips, and to answer RQ1, it should be stated how these conclusions relate to the SSP program. Literature suggests to collaborate as much as possible with suppliers, and thus visit them often on-site, for a more tailor-made and personal approach. This is often not possible due to the time and financial resources of the focal firm. One of the most efficient approaches is the use of self-assessments, although these need to be peer assessed to be used as accurate and unbiased data. Philips found an efficient approach for this peer assessment, in the form of requesting evidence documents that support the claims made in the self-assessment. In this way, it seems that Philips

already uses efficient approaches to distribute its limited time and financial resources for supplier collaborations, i.e. site visits. Next to that, the maximum number of resources to collaborate with suppliers could be used for efficiency.

In terms of supplier sustainability scoring, Philips currently uses only risk criteria to select suppliers for collaborations, and mostly uses aggregated single scores to assess and compare suppliers. It is suggested that this should change towards the use of multiple sustainability indicators (topics in the case of Philips), together with its accessory maturity level (in total or per topic). In this way, suppliers can be better assessed and compared, whilst also improving decision making with more information than before. Next to that, if Philips wants to extend and scale the SSP program as standard program for its supply chain, the use of expert opinions might not be the best option. The current expert opinion weights (and influence on the questionnaires) only state what sustainability is for Philips, whilst this might be different per company and its goals.

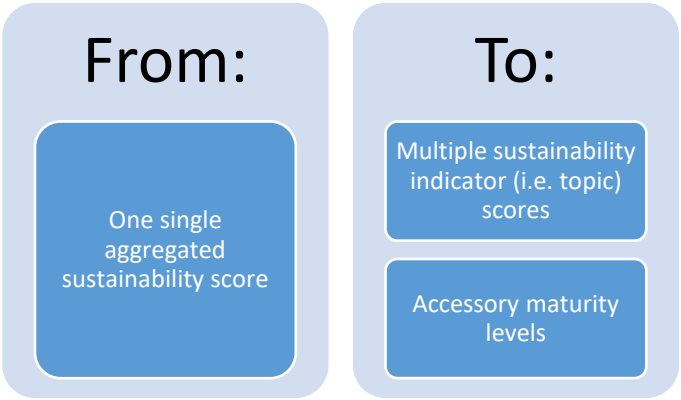

Figure 3: Change needed in terms of supplier sustainability scoring

The gap in the current literature on supplier sustainability assessment and scoring can best be stated in twofold. Firstly, the truthfulness per monitoring-based approach is researched by Darnall & Carmin (2005), but no insights in truthfulness of collaboration or combined methods are stated. Combined methods of monitoring are part of the current best practice, since focal firm often lack time and financial resources to treat all suppliers equally and combined methods release this burden. Secondly, there exists a gap in the field of how to use all the data gained from sustainability assessment and scoring for supplier evaluation models. Current best practices are stated for both assessment and scoring, but the next step on how focal firms should deal with this information and select their suppliers for collaborations seems lacking.

5. Data

The first step in working towards questionnaire validation and supplier evaluation models is to acquire all data needed, regarding questions to be validated and criteria needed to evaluate on (per supplier). This process exists of both data understanding and preparation, i.e. collecting initial data, describing and exploring the data, verifying data quality, and selecting, cleaning and integrating the data for further analysis. The objective is to collect as much useful data possible at once, so that all research questions and future research can depend on this single data set. In this way mindless rework can be avoided. The (second and third stage of the) CRISP-DM process is used to structure the process from acquiring to integrating data.

5.1 Collect initial data

Since the start of the SSP Program, in 2016, each supplier within scope is yearly assessed based on a self-assessment questionnaire (SAQ) and evidence documents (ED). This thus results in a single workbook per assessment per sequence, in which all information regarding questions and scores, dashboards, improvement plan, and supplier feedback are stated. After that, the supplier is classified as either DIY or SSIP for that sequence. In the case that the supplier is classified as SSIP, and their ED are validated on site, this results in a duplicate assessment workbook in which the ED availability and quality are corrected. Each supplier within scope thus has at least one assessment workbook available, depending on their number of sequences (i.e. number of years within scope of the SSP program) and classifications (DIY and/or SSIP).

Besides that, spend per supplier is needed for the supplier evaluation model, which is not part of the assessment workbook, but can be found in the Supplier Management Reporting System (SMRS) database. For the suppliers that are in scope for the supplier evaluation analysis, the spend data of the previous year is needed. This SMRS dataset does not state spend per supplier ID, but per part number, so aggregation is needed to result in useful data. The method how this spend data is aggregated per supplier is further explained in chapter 5.4.

5.2 Describe and explore data

The input for this and future research exists of all available assessment workbooks that are stored on the online platform used by Philips. All useful information from these assessment workbooks is collected by the use of Excel Visual Basic for Application (VBA), which automatically pastes this into one worksheet. In this way all information per supplier can be analysed, instead of only analysing the final scores that are registered on the online (SharePoint) database. At the time of the latest data import, on January 1st 2019, 662 assessment workbooks from 272 unique suppliers are collected. These 662 assessment workbooks consist of 182 workbooks with on-site validation, and 480 workbooks without on-site validation. This means that in 480 cases the suppliers are classified as either SSIP (182 times) or DIY (298 times).

The information collected per assessment workbooks consists of all relevant information for this and future research: workbook characteristics, dashboard scores and weights, focus areas of improvement, SAQ answers and scores, and ED availability and quality.

The total acquired data per assessment workbook is not only relevant for this research, but future research should also be able to use the same dataset. In this way the Excel VBA is only needed to collect the most recent data that can be used by all, so that unnecessary rework is limited. The total acquired data per assessment workbook is thus larger in quantity than needed for this research only. When collecting all relevant information, as stated before, this results in acquiring 2269 fields per assessment workbook. The total acquired dataset thus consists of 662 rows (assessment workbooks) and 2669 columns (relevant information fields). After cleaning and transforming this dataset, the data is explored to deliver an overview of data used within this research.

5.3 Verify data quality

This process of collecting is easy to manage since all assessment workbooks should comply to file naming conventions. Unfortunately, not all file names complied with these file naming conventions, which resulted in multiple corrections in file name after different data imports. Most errors were made in (absence of) spacing, order, and regular typos. By identifying these errors, the file names could be corrected to the file name convention in place. These corrections already resulted in 35 corrected file names. Besides that, 16 workbook assessments were missing on the online platform, of which 14 are found again after file name corrections or uploading. These mistakes on the online platform were found when SSIP suppliers did not have assessment workbooks from both before and after SA or when assessment workbook sequences were missing. By correcting these file names and uploading, the total amount of 662 assessment workbooks is reached. In 73 cases, the assessment workbooks from before and after the SA are the exact same workbook, thus duplicates. This is not a mistake or error, but due to the supplier not delivering the ED (on time) or the Philips team not having enough time before going on site.

In terms of quality of the data and whether it represents the actual sustainability level remains a difficult topic, which is agreed upon in the literature as well (López et al., 2007; McWilliams et al., 2006). There is no single concept of sustainability, and this even differs per industry and per company. To get as close as possible to an agreed upon sustainability questionnaire design that stretches further than the company itself, the SSP program questionnaire is designed by Philips together with the RBA (questionnaire) and various NGOs. In this way the questionnaire aims attention at opinions and importance within the electronic industry. Although that these topics, questions, and weights are thus based on expert-opinion, some error might exist between this sustainability construct and the actual sustainability value. Next to that, the aggregation method used to combine these questionnaire scores might expand this error even further. An initial

comparison between different approaches of weighting and aggregating is further analysed in previous chapter 4.

Besides that, next to the possible errors of the questionnaire design, the sustainability construct depends on two human factors that might affect the data quality. Firstly, the supplier's employee(s) filling in the SAQ might not answer completely truthfully, or not have full knowledge to the questions. Next to that, suppliers might experience "questionnaire fatigue", also "respondent fatigue", when filling in the SAQ, since the approximately 700 questions result in a lengthy process of answering (Kogg & Mont, 2012). Tiredness, boredom, and lack of attention or motivation are common consequences of this fatigue, and might result in more "straight-line" or inaccurate answers, and even lack of answers. All these factors might affect the data quality, but Philips also uses the ED to validate the SAQ. Although this is only done in terms of scores, rather than on question-level, this second opinion might decrease the effects of the previous mentioned human factor. Secondly, Philips employees that validate the ED might be subjective or biased when assigning scores. Although the ED come closer to the actual situation than the SAQ, the SAQ answers are not corrected when finding inconsistencies. Further analysis to these inconsistencies between SAQ and ED is done and explained in chapter 6.1.

5.4 Select and clean data

The previous steps of gathering data resulted in a raw dataset of 662 assessment workbooks with 2669 relevant fields. Since not all information is within scope and usable yet, the dataset should be prepared for further analysis. This is done by selecting the appropriate and relevant assessment workbooks and information, eventually cleaning and transforming the data, and ultimately integrating this all to one single useful and relevant dataset.

To work towards the validation of questionnaires and supplier classifications only a subset of the total dataset is useful. Firstly, 5 workbooks do not have any ED availability and quality stated, so these are left out of scope and omitted. Secondly, potential suppliers (i.e. suppliers that are on the edge of entering the SSP program) are assessed differently than the other suppliers. It is claimed that only a subset of SAQ and ED questions is requested from potential suppliers, so the validation of the questionnaire does not make fully sense. Besides that, in terms of supplier evaluations, potential suppliers always classify as SSIP suppliers in their first sequence, so these 14 workbooks are left out of scope and omitted as well. Thirdly, 2 questions are duplicated in fourfold in the SAQ answers and scores. These 12 duplicates (2 questions x 3 duplicates in both answers and scores) are naturally omitted since they all represent the same question. After data selection the dataset consists of 639 assessment workbooks and 2657 relevant workbook fields.

Within scope of the questionnaire validation (RQ2) are the workbook characteristics, SAQ answers and ED availability. Since several SAQ answers are cross-checked and validated, the general information part and Quality topic remain out of scope. This is done because the ED general information only relies on pictures (e.g. from the main entrance), and because the Quality topic is out of the SSP program's scope per 2019. Within scope of the supplier evaluation model (RQ3) are the workbook characteristics, SAQ and ED characteristics (general activities of supplier, number of PZTs found) and external data like spend and predicted improvements per supplier.

It is essential to transform the selected data into useful data that can be used for further analysis. On one hand, this is not necessary for the workbook characteristics, since these already state appropriate information, like supplier ID, number of sequence, and date. On the other hand, this is necessary for the SAQ answers and ED availability, since these are stored as text when collecting the data from the separate workbooks. Most of these data fields should be transformed in binary values that can be used to analyse. Since there are seven different types of SAQ questions, each type needs a different approach to transform. Next to that, spend data is stated per part number, not per supplier. Because the spend data does not fully cover all suppliers, e.g. due to suppliers with multiple sites, some workbook assessment are omitted. A more detailed overview of these data transformations made can be found in appendix A.

To compare different supplier evaluation models, as further explained in chapter 6.2, missing values are not accepted. Furthermore, since the focus of supplier evaluations is rather on heuristic models than prediction models, missing value imputation is not necessary to increase accuracy of the model. Since the heuristic models are tested against real-life choices made it makes sense to only use information that is complete and real, since estimates do not cover the whole picture. By omitting these assessment workbooks (for the supplier evaluation model only) it is certain that a smaller subset is used for the heuristic models, since incomplete and inaccurate workbooks might lead to shifted results.

Furthermore, as earlier mentioned, predicted improvements per supplier (ID) in 2018 are within scope for the supplier evaluation model. These predicted improvements are provided by a fellow researcher who uses Markov Chain Simulation (MCS) to predict upcoming (ED) scores with an average RMSE of 0.07. These predicted improvements serve as an appropriate insight in how suppliers evolve over time (according to historic data), although more logical predictions should actually be necessary for the supplier evaluation model. Since prediction models are not generated within this research, these predicted improvements provided give an initial insight in usability and opportunities of such prediction models.

5.5 Explore data

To be able to assess and compare suppliers based on multiple scores instead of one single aggregated sustainability score (RQ1), an analysis tool is created that uses the dataset input previously explained. The goal of this analysis tool is to compare the suppliers based on multiple topic scores, instead of the SAQ and ED score only. Next to that, it is possible to change the ratio and to include/exclude certain topics. In this way, it is thus possible to assess and compare suppliers on, for example, Environment topic only with 100% ED score – 0% SAQ score. From 2019 onwards, Philips chose to use the following parameters: 0% SAQ score, 100% ED score, including all topics except for Quality, as earlier explained.

Using these parameters, an average SAQ score of 69% and ED score of 40% are found, with standard deviations of, respectively, 19% and 20%. Finding such high standard deviations is in line with the research done by Locke et al. (2007), in which standard deviations of 16% were found for the whole group of suppliers scoring on average 65%. Besides that, an average ED score of 46% is found after site assessments, which supports the claim that collaborating with suppliers might achieve higher improvements than not collaborating.

Besides the insights on average scores per sequence or per year, the analysis tool provides further insights in the improvements made from one year to the other, and what differences are found in improvements per supplier classification (DIY/SSIP). These improvements can be made from sequence 1 to 2 (from 2016 to 2017, and from 2017 to 2018), and from sequence 2 to 3 (from 2017 to 2018). The most interesting finding of these improvements, as stated in Table 1, is the difference between DIY and SSIP suppliers, especially from sequence 1 to 2. This might indicate the effectiveness of visiting suppliers on-site and collaboratively working together on the improvement plan of actions. The explanation and assumptions made for these comparisons, and summaries of ED score improvements per topic can be found in appendix B.

		Sequence 1 to 2		Sequence 2 to 3	
		%	n	%	n
From 2016	(total)	(+ 22%)	47	(+ 17%)	37
To 2017	DIY	+ 5%	21	+ 16%	24
	SSIP	**+ 36%**	**26**	**+ 19%**	**13**
From 2017	(total)	(+ 28%)	117		
To 2018	DIY	+ 19%	79		
	SSIP	**+ 45%**	**38**		

Table 1: Summary of ED score improvements per sequence

When investigating the difference between ED score before and after the site visit, average improvements between 1 and 21% are found, as stated in Table 2. On average, suppliers improve their ED score because of the site visit, although the difference between 1 and 21% improvement is remarkable. By omitting two outlying cases (e.g. improving from 1 to 71%), the average

improvement found after site visits in sequence 1 (2017) already decreases from 21 to 13%. Next to that, these average improvements are naturally dependent on what suppliers are visited on-site, but it is clearly seen that, on average, suppliers improve their ED score after Philips plans the site-visit and validates the ED. A further look into the contribution of these site visits is also investigated in chapter 7.1.

Year / Sequence	Sequence 1	Sequence 2	Sequence 3
2016	+ 7%	+ 5%	+ 1%
2017	+ 21%	+ 4%	
2018	+ 5%		

Table 2: Summary of ED score improvements before and after SA

In terms of maturity levels, Philips currently uses the nine elements in the dashboard as preliminary maturity levels per topic and in total. Next to that, to make the supplier dashboards more efficient and user-friendly, Philips agrees upon the suggestion that nine elements might be too much since this results in dashboards of 36 fields, excluding the weighted averages. One possible solution lies in another setup of maturity, e.g. combining factors, to decrease this number of fields per dashboard and increase the efficiency while decision making. As found in RQ1, the decision makers should not have too little or too many decision variables, so making the dashboard more efficient would be in line with this finding. The actual implementation of combining the factors is left out of scope for this research.

In conclusion, it can be stated that data quality and sustainability construct can be affected by several human factors like wrong file naming, the suppliers' employees filling in the SAQ and the Philips' employees evaluating the ED. The differences between SAQ and ED scores are already known in terms of score, but deeper examination on question-level seems necessary to evaluate the usability of the questionnaire answers, which underlines the importance of RQ2. Next to that, certain choices have been made with regards to the questionnaire design, and initial insights in the choice of weighting and aggregating are elaborated in the answer to RQ1. The bias that thus exists at both supplier and Philips is thus known and accepted in terms of the impact it has on the SAQ and ED scores. The biggest advantage of the new data import compared to the previous is the accessibility of ED (availability and quality) data. Instead of assuming that the SAQ data is the actual situation, with the use of the (new) ED data it should be possible to have insights whether this is the case and how to make the best use of both. How the whole dataset is used for respectively questionnaire validation and supplier evaluation models is explained in chapter 6.1 and 6.2. All previous steps are taken by the use of Excel (VBA), of which most is done automatically (besides the use of external data). Further steps with the datasets described are taken by the use of programming language R.

6. Models

In the previous chapter it is explained what data is in scope, and how this data is cleaned and eventually transformed towards useful data for this research. This chapter presents two models, namely the inconsistency model in order to validate the suppliers' answers to different questionnaires, and the supplier evaluation model that revises the decision making in line with the supplier sustainability program it is used in.

The main aim of the inconsistency model is to find a method on how to validate two different datasets of questionnaire answers, rather than just looking at the difference in score. It is suggested that the realism of self-assessments can only be determined when comparing it to other judgements and differences often exist. This inconsistency model has a theoretical contribution to the field of questionnaire validations and can potentially be used on both supplier and question level. Besides that, since many focal firms use self-assessments for supplier sustainability selection and assessment, insights in the inconsistency of answering by suppliers contribute to the knowledge and accuracy of supplier sustainability assessment approaches to use. Besides that, the supplier evaluation model mainly aims at revising the decision making process used and aligning it with the (supplier sustainability) program's objectives. For this research it is chosen to make use of the currently chosen supplier evaluation model in terms of possible classifications in the model, but to revise the decision making process towards these classifications. One of the most important extensions of this supplier evaluation model is the use of the maximum number of resources available, as well as two different decision making models that can be implemented. One of them is the TOPSIS method, which is popular in the supplier sustainability field due to its simplicity and adjustability of criteria and objectives.

More specific, both inconsistency and supplier evaluation model relate to the SSP program of Philips, in which it is aimed to increase suppliers' truthfulness and transparency as well as performance over time. On an annual basis, suppliers (in scope) are assessed by filling in a self-assessment questionnaire (SAQ) and providing supporting evidence documents (ED). The SAQ mainly assesses the availability of documents and processes (e.g. policies, procedures, corrective actions), whilst the ED mainly assess the quality of these documents. This naturally results in a large gap between both scores, although a subset of questions have an exact overlap of content. This fact makes it possible to validate both questionnaires as first step towards investigating the truthfulness of suppliers. These connected questions are expected to result in correlations of (nearly) 1, but this does not seem the case. By setting up contingency tables, the most harmful inconsistencies of answering can be highlighted. Next to that, although Philips aims at maximizing the improvements made by suppliers, this objective is currently not used in their supplier evaluation model. Two different methods are proposed to take both risk and improvement into account when selecting suppliers for the SSIP classification (on site assessment), instead of DIY.

6.1 Inconsistency model

The aim of the inconsistency model is thus to research how to validate two datasets of answers, in this case the SAQ and ED answers, besides the current difference in weighted average score. Since there exists an overlap in questions, naturally, the comparison between SAQ and ED answers can easily be made. However, in the current setup of the SSP program there does not exist such a validation between the answers. The main contribution of this model is to be able to give feedback on question-level and gain initial insights in the truthfulness of suppliers.

In short, currently the suppliers answer the SAQ and provide the ED afterwards as validation for their answers, which results in separate SAQ and Validation (ED) scores, respectively, for that sequence. Only in the case that the supplier is classified as SSIP, Philips goes on site to (among others) validate these ED, which results in a 'corrected' Validation (ED) score. The SAQ (score) remains unchanged since no feedback loop exists between SAQ and ED. As previously mentioned, suppliers score on average 69% on the SAQ score, 40% on the Validation (ED) score, and 46% on the 'corrected' Validation (ED) score. When taking the same supplier base into account, so that each average is based on the same amount of connected suppliers (in this case 182 suppliers), these (SSIP) suppliers score on average 68% on the SAQ score, 43% on the Validation (ED) score, and 46% on the 'corrected' Validation (ED) score. The differences between SAQ and (corrected) ED scores seems substantial, and are currently explained as result of the supplier's (intended) overestimation in the SAQ and Philips' scoring methodology for the ED. The SAQ mainly asks for availability of policies, procedures, documents, implementation, and so on, and suppliers are thus scored based on the availability. However, the ED accounts for both availability and quality, and suppliers are mainly scored based on the quality (of their available documents). With the use of ED availability and quality data, a more precise comparison can possibly be made on the overlapping SAQ and ED questions.

To make better comparisons than just on weighted average scores only, and to attempt to measure the suppliers' representability or inconsistency of answering, SAQ answers and ED availability answers is compared. Within the field of supplier sustainability assessment the difference between self-assessment and peer assessment is already researched and is concluded to be large, which seems in line with the difference in SAQ and ED scores within the Philips SSP program (Lamming & Hampson, 1996; Pilbeam, Alvarez, & Wilson, 2012; Subic et al., 2013). It is argued that this is due to suppliers' misconceptions of its capabilities, suppliers being unable to accurately assess themselves, and social desirability or window-dressing (to comply with the focal firm's code of conduct and to maintain the relationship) (Allen & Van Der Velden, 2005; Brown & Harris, 2014). The latter seems the most important cause in terms of potential (image) damage for the focal firm.

As previously stated in chapter 4, the realism of self-assessments can only be determined when comparing it to other judgements (peer assessments). Since several studies claim that peer

assessment is more accurate than self-assessment, and Philips agrees upon this claim[1], the peer assessment is seen as actual situation. Within the Philips SSP program, this thus relates to the ED as peer assessment and actual state, and the SAQ naturally as self-assessment. Several other studies focused on the correlations between self-assessment and peer assessment, and found correlations between -0.05 and 0.82, or with a mean correlation of 0.39 (Allen & Van Der Velden, 2005; Brown & Harris, 2014). This is described as "moderate at best", and thus underlines the need of peer assessments when making use of self-assessment.

The correlation between self-assessment and peer assessment score within the SSP program is found to be higher than that, namely 0.54. This correlation between SAQ score and ED score is based on making use of the most reliable information possible, namely correcting ED scores with validated ED scores when possible. When the subset is narrowed down to suppliers with a SA only, since this is argued to be the most reliable information, a correlation of 0.63 is found. The fact that these correlation are slightly higher than the 0.39 can best be explained by the fact that suppliers nearly always (in 96% of the cases) score themselves higher than the focal firm, in this case Philips, does. Figure 4 below presents this fact clearly in graph: suppliers with low SAQ scores hardly ever have high(er) Validation (ED) scores, whilst suppliers with high SAQ scores have Validation (ED) scores in the range from low to high.

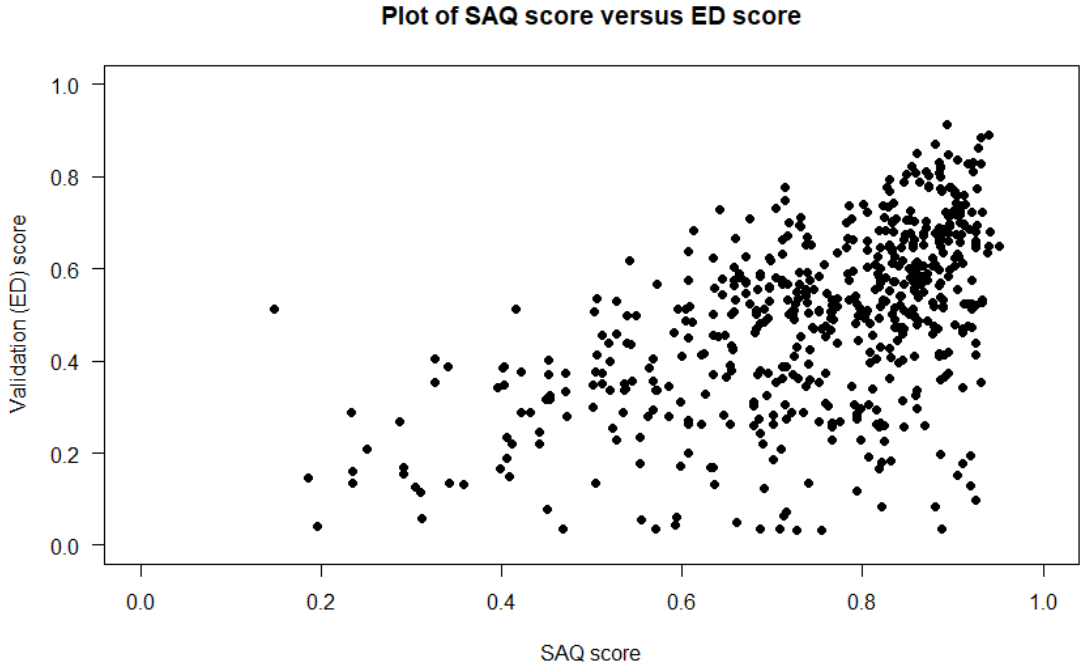

Figure 4: Plot of SAQ score versus Validation (ED) score

[1] In terms of rather using peer assessment (evidence documents, which quality is scored by Philips) than self-assessment (which rather assesses the availability of documents, as earlier explained).

Taking both SAQ and ED scores into account (following the 70-30 ratio) to assess and compare suppliers thus means taking inaccurate scores into account, which might lead to shifted results that affect the whole SSP program (Distelhorst et al., 2016; Gualandris et al., 2015; Yamin et al., 2017). Decisions taken would then depend on inaccurate scores, which lead to the current shift to the 100-0 ratio, taking more reliable information into account. Nevertheless, the one of the main differences between SAQ and ED score remains that the former (mainly) assesses availability rather than quality and the latter (mainly) assess quality. It is expected that the most mature (i.e. nearly perfectly truthful and consistent) and qualitatively best suppliers do not have any difference between SAQ and ED score. In those cases the ratio chosen would not matter. This expectation is recognizably seen in Figure 5 below, in which the difference between SAQ and ED score undoubtedly converges to nearly zero for the more mature suppliers. This seems in line with the suggestions in literature stating that high scoring individuals tend to be realistic (thus SAQ score minus ED score near zero) (Boud & Falchikov, 1989). However, low scoring individuals are expected to be less realistic, which is not clearly the case within the SSP program.

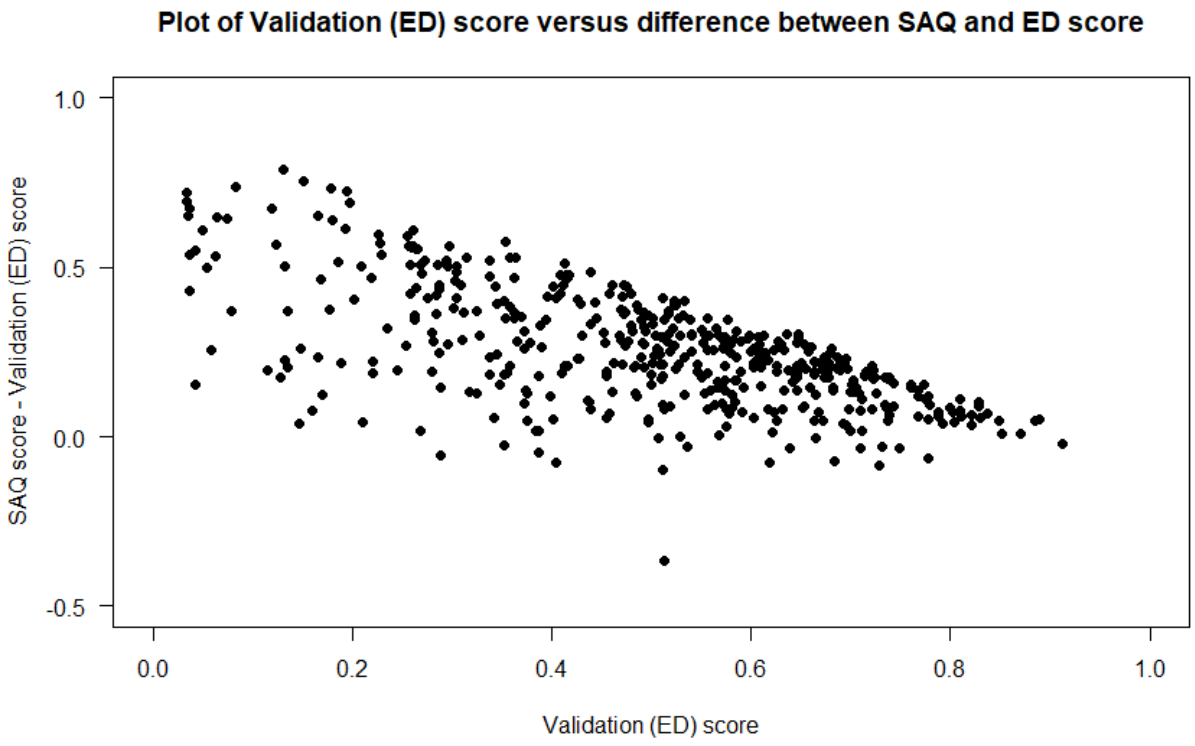

Figure 5: Plot of Validation (ED) score versus difference between SAQ and ED score

Although errors can be intentional and unintentional, in the field of supplier sustainability intentional errors, e.g. window-dressing, are the most important. These intentional errors can best be described as claiming to be sustainable although this is not the case. In terms of document availability this is best translated as claiming to have a document, although this is actually not the case. In other words, such an intentional error within the Philips SSP program would be represented by a SAQ answer of "Yes" (supplier has this document) and an ED availability answer of "No" or "N/A" (the document is not there). The inconsistency model pursues to discover such errors. To find such related questions, the validity between SAQ and ED is researched.

Validations can in fact relate to two types of validity, namely content-related and criterion-related. (Mueller-Hanson, Heggestad, & Thornton, 2003; Terwee et al., 2007) Content-related validity relates to appropriate content and whether it test what it aims to test and if it relates to underlying theoretical concepts. Criterion-related validity relates to the relationship to other measures, and whether the test relates to existing measures and if the test can predict later performance. This inconsistency model should relate to criterion-related validity to validate both questionnaires and whether these relate to each other. This reliability between SAQ and ED thus relates to external consistency, i.e. self-assessment versus peer assessment. The previously researched correlation between SAQ and ED score already indicated a correlation of 0.54 (and 0.63 for suppliers with SA only) on supplier scoring level. The next step, to validate the questionnaire(s), is to research the consistency on question level. Since the SAQ can be seen as self-assessment and ED as peer assessment, and both partly overlap in questions, it seems possible to test external validity.

Because of the length of both questionnaires (together nearly 1000 questions), the researcher chose to focus on one-on-one interconnected questions only. The alternative would have been to investigate whether multiple SAQ questions would correlate with one (or more) ED question(s). In that case the Cronbach's alpha could have been used to state the correlations, but since the focus is on one-on-one relations only another approach is recommended[2]. Although there exist several measurements to state similarity or correlation between two binary variables, the widely-used Pearson correlation coefficient, also Pearson's r, seems sufficient to gain enough insights in (external) consistency (Rajagopalan & Robb, 2005; Zhang & Srihari, 2003). Since both SAQ and ED questions are true dichotomous variables, in terms of availability at least, the Pearson's r can be used to compute these correlations. Since it is mainly aimed to research whether the correlations come close to the expected correlation on 1, the Pearson's r is sufficient for this. The Pearson's r has a value between 1 and -1, of which 1 refers to total positive linear correlation, -1 to total negative linear correlation, and 0 to no linear correlation. The formula for the Pearson's correlation coefficient ρ is:

[2] It is recommended in research and statistics to use the Cronbach's alpha only for three (or more) items.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

In which $cov(X,Y)$ represents the covariance, σ_X the standard deviation of X, and σ_Y the standard deviation of Y. Another method would be to set up a contingency table and compute the correlation based on the cell frequencies, which is further explained. The sequential step is to interconnect the SAQ and ED questions. Since the ED questions come closest to the actual situation it is chosen to start with these questions and find the connected SAQ question. To structure this manual process some strict rules have been set up:

1. Questions regarding the low maturity elements are within scope,
2. Questions regarding the topic Quality are left out of scope,
3. Only connect the SAQ and ED questions that should be consistent for sure,
4. Only connect binary questions.

It is chosen to focus on low maturity topic questions (so questions regarding Policies, Procedures, and Implementation) since these are expected to all be answered by both low and high maturity suppliers. Because of the waterfall structure of the SAQ, questions regarding high maturity topics would most probably not be answered by the low maturity suppliers. To be able to assign significant correlations per dashboard field too, it is chosen to only take those dashboard fields into account that have at least three connected questions. Since the topic Quality is left out of scope, the following dashboard fields remain within scope, as can be seen in Table 3 below.

Dashboard									
Topics	Policy	Procedures	Implementation	Management Responsibility	Communication	Risk control	Target Setting &Tracking	Corrective action approach	Supplier management
Quality									
Environment									
Health and Safety									
Business Ethics									
Human Capital									

Table 3: Dashboard fields within scope for the inconsistency model

Within these dashboard fields the ED questions are compared to the related SAQ questions and connections are made between those questions that are expected to (and should) be consistent. Such an example is ED question *"Valid certification X"* and SAQ question *"Do you have a valid certification X?"*. Or ED question *"Annual management review on Y"* and SAQ question *"Do you have the annual management reviews on Y?"*. All connections that are not as clear as these examples were not taken into account, so that the expected correlations would be 1 for every case. In total, 60 connections were made within scope. All these 60 questions related to binary (with values 0 and 1 after data transformations) and non-constant (so with $\sigma \neq 0$) questions.

After data transformations the SAQ answers "Yes" and "No" are transformed to values 1 and 0, respectively, and ED availability "Yes", "No", and "N/A" is transformed to values 1, 0, and 0, respectively. The only remaining question was how to transform the ED availability "PZT" into such a value. Since the number of PZTs only yields for a small number of cases (only in 1 % of the cases[3]), the transformation approach is not expected to be significant. To identify the intentional errors in the connected SAQ-ED questions, and thus especially the cases in which the SAQ answer has the value 1 and the ED availability has value 0, it is chosen to transform the "PZT" into the value 0. Although this might not be true in all cases, because some suppliers might truthfully commit to violate the code of conduct, transforming "PZT" into the value 1 would make less sense in pursuing to identify damaging errors.

	ED availability = 1	ED availability = 0
SAQ answer = 1	True positive	False positive
SAQ answer = 0	False negative	True negative

Table 4: Setup of contingency table and types of frequencies

	ED availability = 1	ED availability = 0
SAQ answer = 1	Consistent	**Inconsistent**
SAQ answer = 0	**Inconsistent**	Consistent

Table 5: Difference between consistency and inconsistency

	ED availability = 1	ED availability = 0
SAQ answer = 1		**Harmful inconsistent**
SAQ answer = 0	Harmless inconsistent	

Table 6: Difference between harmless and harmful inconsistency

In case that the correlations fail to give enough insights into the expected consistencies and/or errors, it seems wise to put some effort in setting up a contingency table and compute the correlations based on cell frequencies, such as is seen in Table 4. It is argued that there are two different errors/inconsistencies, namely false positives (stating something, e.g. procedure, exists, while it actually does not) and false negatives (stating something does not exist, while it actually does). In other words, window-dressing can thus be seen as such a false positive error. On the other hand, false negatives are argued to be less important in terms of 'document' availability, although this could also indicate intentionally hiding. When setting up such a contingency table the cell frequencies of true positive and true negative together are expected to be equal to the Pearson correlation coefficient, and thus the consistency. The cell frequencies of false positive and false negative together thus reflect the inconsistency. The main aim of this inconsistency model is to focus on harmful inconsistency only, so only taking the false positives into account. The difference between inconsistency and harmful inconsistency for this research is explained in Table 5 and Table 6.

[3] In 108 of the total 106074 cases within scope the ED availability "PZT" was assigned.

Algorithm 1: Contingency table
Input: SAQ answers, ED availability, empty matrix i x j

```
for i = 1 to number of suppliers {
        for j = 1 to number of connection questions {
                if SAQ answer [i,j] = 0 and ED availability [i,j] = 0 {
                        matrix [i,j] = 0 (true negative)
                } else if SAQ answer = 1 and ED availability = 1 {
                        matrix [i,j] = 0 (true positive)
                } else if SAQ answer = 0 and ED availability = 1 {
                        matrix [i,j] = 0 (false negative)
                } else if SAQ answer = 1 and ED availability = 0 {
                        matrix [i,j] = 1 (false positive)
                }
        }
}
```

Output: contingency matrix i x j

Firstly, these inconsistencies are calculated per SAQ-ED connection, as described in algorithm 1, after which the weighted average per connection, dashboard field and/or supplier can be computed. This is done using equal weights (thus stating the proportion of harmfully inconsistent answering) and using expert-opinion weights as used in the SAQ (thus claiming that inconsistent answering of one question could be more harmful than for another). Secondly, as seen below, by aggregating these inconsistencies a harmful inconsistency (HIC) score is calculated per supplier i and compared to different variables such as sequence, year, assessing on site or not, and the suppliers' validation (ED) score. Alternatively, by adding the SAQ weights per connection j as expert-opinion weights (instead of equal weights), this results in the eHIC-score.

$$(e)HIC - score_i = \sum_{j=1}^{60} \frac{w_j * matrix[i,j]}{w_j}$$

In conclusion, both SAQ and ED questionnaires represent different purposes, although some questions are exactly overlapping in content asked for. By connecting these questions, the self-assessment (SAQ) can be validated by peer assessment (ED). Although correlations of 1 are expected within these connected questions, this is not the case. Since inconsistencies can be described as both false positives and negatives, contingency tables are set up to differ between both. By aggregation these (harmful) inconsistencies, a harmful inconsistency (HIC) score can be set up on both question and supplier level. The results of this HIC-score are found in chapter 7.1, which can used by Philips as initial insights in truthfulness and possibility as feedback to suppliers.

6.2 Supplier evaluation model

The main aim of this evaluation model is to revise the objective(s) and rules of the current supplier evaluation model, since this is assumed to be poorly used according to its set rules. By the investigation of the current situation, the aim is to improve and structure the supplier evaluation process. Two different types of models are tested, i.e. risk-based and distance-based heuristic, and compared to each other. After that, by the use of RQ1 (which supplier characteristics to use for decision making) and RQ2 (harmful inconsistency score per supplier as potential risk variable), a recommended setup of the supplier evaluation model is stated for future research. The main contribution of this revised supplier evaluation model is to structure the current process and improve decision making (in terms of selecting suppliers of collaborations) in an optimal manner.

The first step in working towards these goals is to investigate the current supplier evaluation process and research whether the model can be improved. This requires an explanation of the current proposed process and quantitative checks to test its performance. The second step would then be to use the right supplier characteristics and state how the revised supplier evaluation model should and could be implemented within the SSP program. This requires more investigation in different models instead of using heuristics, so the focus lays on the first step.

In the current SSP program suppliers receive a classification after both SAQ and ED are provided and scored by Philips. The possible classifications are BIC, DIY, SSIP, and PZT, as explained in chapter 2. Further investigation towards these classification results in the fact that only DIY and SSIP are currently used. The main difference between both is that SSIP supplier get an on-site assessment (SA), so that the Philips team can recheck the ED, have interviews with staff and employees, and discuss the proposed Improvement Actions, resulting in an Improvement List of Actions. In other words, the main decision to make is: should the supplier be visited on-site or not? Or in terms of the findings of RQ1: which suppliers should be selected for collaboration?

Based on the conversations within Philips the criteria to currently decide upon this question are all only risk-based. The following (risk-) criteria are identified and used for the current heuristic:

- All potential suppliers[4], or
- High spend (e.g. >1 million EUR annually) and/or strategic importance to Philips, or
- Activities focus on painting/coating, mechanic metal processing, cleaning, washing, die-casting, electroplating (i.e. high EHS risk level), or
- (Potential) Zero Tolerance(s) identified, or
- Supplier have been exposed to media and/or NGO attention for sustainability violation (e.g. by IPE in China).

[4] Thus the suppliers that are potentially in scope for the SSP program, and that did not enter Philips supply base yet.

More importantly, it is claimed that suppliers are classified as SSIP if a high risk is identified in any of these five risk criteria. The current heuristic can thus be seen as framework in which all five criteria are binary rated (0 = low risk, 1 = high risk) and the suppliers is classified as SSIP when the 'total risk score' exceeds 0. In other words, suppliers are classified as SSIP regardless of the type of risk and regardless of the number of risks. Besides that, no structure seems present to distribute the suppliers when the maximum number of SSIP suppliers is exceeded. In this way an unlimited number of suppliers could be classified as SSIP, regardless of the time and financial resources of the focal firm (Appendix C).

As literature states and is concluded in chapter 4 and RQ1, focal firms should collaborate as much as possible with its suppliers. In this way, trust, transparency, performance and improvement are gained. Furthermore, it is seen in chapter 5 that SSIP suppliers relatively improve more than DIY suppliers. Since the goal of the SSP program is to improve as much as possible, collaborations should be planned as much as possible. Although focal firms are advised to collaborate as much as possible, this is often limited by the time and financial resources of the focal firm, so only a subset of suppliers can be collaborated with. An efficient manner to assess all suppliers and efficiently distribute the number of collaborations is stated to be the use of self-assessments. Suppliers and focal firms only spend a short time and low financial resources in these self-assessments, and focal firms benefit from this initial insight in sustainability performance of all its suppliers. With the use of these insights the focal firm can efficiently use its limited number of collaborations for its suppliers. This same setup is followed in the SSP program, in which suppliers thus all provide answers to the SAQ and ED, after which Philips evaluates the suppliers accordingly. The problem within this current (heuristic) process is the absence of structure and rules on how to evaluate suppliers and assign classifications.

Firstly, the current heuristic is quantitatively analysed to establish the path towards an improved supplier evaluation model. As stated before, if any of the five risk criteria is classified as high risk, the supplier is classified as SSIP. Therefore, the number of high risks per risk criteria are calculated. It should be noted that not all information regarding the risk criteria is available, such as the strategic importance (defined and computed by another department, with only a small overlap in suppliers), media exposure (are seen as PZTs), and annual spend (not all annual spend of the previous year is collected or known within the same dataset). Besides that, PZTs and media exposure are combined to the risk criteria PZTs. This results in the following risk criteria to be within scope for this research:

- PZTs
- Supplier activities
- Annual spend

Before analysing the current heuristic, it should be stated that the used dataset consists of workbook assessments in 2018 only. Since the planning for site assessments is made per calendar year, a subset of either 2016, 2017, or 2018 only should be used. Since 2016 results in a small subset (37 suppliers), and Philips claims that the supplier evaluation model is better used in 2018 than in 2017 (also Philips is still learning and improving), only assessment workbooks of 2018 are within scope for RQ3. Besides that, potential suppliers are left out of scope since these always get a site assessment, and their assessments are aimed to be slightly different than the others, as previously stated in chapter 5. This results in a dataset of 157 suppliers, of which 60 were classified as SSIP (thus 97 as DIY). This thus means that the fixed number of resources available is assumed to be 60 for this analysis. The potential of relaxing this assumption is further explained in appendix C.

For each of the three risk criteria threshold values are set how to differ between low, medium, and high risk, as currently assessed. The threshold value for PZTs is naturally 0, since having one or more PZTs is seen as high risk. For annual spend this threshold value is set at 1 million Euros, because this addresses 80 % of the total spend and most of these are important to Philips. For supplier activities the risks are classified as high, medium, and low, and another approach is used to classify suppliers. If suppliers have one or more high-risk activities, the supplier activities are classified as high. If not, and the suppliers have one or more medium-risk activities, it is classified as medium. If not, in all other cases, it is classified as low. This thus results in three risks levels per supplier. The total number of high risks found per risk criteria are stated in Table 7. Further analysis, as can be seen in Table 8 below, shows that in one case the supplier was classified with (at least) high risk on risk criteria PZTs, but not classified as SSIP. Since the risk criteria PZTs is argued to be the most important variable in finding ZTs, these inconsistencies should not happen.

Risk criteria	Absolute number of high risks	Relative number of high risks
PZTs	8	5 % (8 out of 157)
Annual spend	116	74 %
Supplier activities	52	33 %

Table 7: Absolute and relative number of high risks as currently assessed

Risk criteria	Number of high risks	Number of high risk suppliers NOT classified as SSIP
PZTs	8	1
Annual spend	116	62
Supplier activities	52	29

Table 8: Number of high risk (criteria) suppliers not classified as SSIP

When the current heuristic would be in place and structurally used, the following distribution of classifications is found: 137 SSIP suppliers, 20 DIY suppliers. This is certainly inconsistent to the actual distribution of classifications, for which the maximum of only 60 suppliers can be in SSIP. In the actual distribution of classifications, 3 of these 20 DIY suppliers and 57 of these 137 SSIP suppliers are classified as SSIP. Besides that, no set of rules exists to take the maximum number of resources into account. In conclusion, the structural implementation of the current heuristic would not make sense in terms of threshold values and/or maximum number of resources.

Further analysis into the current evaluation model shows that the average ED score for suppliers in scope is 54 %. When this is split to SSIP and DIY suppliers, averages of respectively 59 % and 52 % are found, which thus means that (on average) SSIP suppliers already score higher than DIY suppliers before the SA. After the SA, this difference is expected to be even larger. Although this could be explained by the fact that the supplier evaluation model criteria are all risk-based, it would be more logical to take the suppliers' score, maturity, and predicted improvement into account (RQ1). This would thus imply that two objective should be in place to assess supplier as SSIP:

1. Minimize the risk
2. Maximize the predicted improvement

Besides that, the maximum number of resources should restrict the number of SSIP classifications assigned to suppliers, based on time and financial resources of Philips. Since the two objectives result in a trade-off for decision making, there should be a prioritization in place to 'rank' the suppliers in terms of risk and predicted improvement. The aim of this research is to find a revised supplier evaluation model that structures the current process and improves decision making. Two important issues to address when setting up the revised supplier evaluation model, to maintain the strategic and competitive advantage of the focal firm and which are in line with the trade-offs, are (Trapp & Sarkis, 2016):

- Which suppliers should be considered for partnering?
- How can firms effectively allocate resources?

6.2.1 Risk-based heuristic model

The first step towards using both risk- and improvement-based objectives is the use of a heuristic that assesses suppliers on both topics and classifies according to minimizing risk and maximizing predicted improvement. One approach to minimize risk is to classify the suppliers with the highest risk as SSIP and visit them on-site to check the (potential) risk. Minimizing risk thus essentially means maximizing the risk that can be checked on-site. What is needed to use risk as objective is to state risk as latent variable (that cannot be observed) that is inferred from (observable) variables. By the use of risk criteria PZTs, Annual spend, and Supplier activities as observable variables, and stating weights per risk criteria, the latent variable risk can be determined per supplier. This latent

variable can then again be used to state the expected supplier's risk in terms of score (from 0 to 1) or level (low, medium, high).

Although compensation is one of the pitfalls of using weighted averages, the use of appropriate weights and aggregation together with the right threshold values results in an initial risk score that can be used structurally. In terms of weighting the expert-opinion method is chosen because PZTs weigh much heavier than, e.g., activity-based risk. Since the scale is fine (the supplier evaluation model is appropriate for one company), it makes sense to assign weights per criteria. In terms of aggregating the additive (weighted average) method is chosen to both assess and compare suppliers, whilst the use of appropriate threshold values between the high, medium, and low risk level slightly decreases the compensation effect. Based on expert-opinion within Philips, the following weights are chosen per risk criteria:

Risk criteria	Expert-opinion weight
PZTs	60 %
Annual spend	15 %
Supplier activities	25 %

Table 9: Expert-opinion weights per risk criteria

Since PZTs and Annual spend are binary values (in terms of high and low risk, respectively 1 and 0 in score), and Supplier activities can be multiple values (high, medium, and low risk, respectively 0, 0.5, and 1 in score), the minimum score for supplier with at least one high risk criteria is 0.15. On the other hand, when a supplier has a PZT and it receives a high risk for this criteria, its risk score is at least 0.60. Therefore, these two values serve as threshold values between high, medium, and low risk in total. This results in the following risk levels per risk score X:

Risk score = X	Risk level	Number of suppliers in risk level	Number of suppliers classified as SSIP
$X \geq 0.60$	High	8	7
$0.15 \leq X < 0.60$	Medium	119	48
$X < 0.15$	Low	30	5

Table 10: Risk level based on risk score X

In terms of improvement, the predicted relative improvements from a fellow researcher are used, as stated in chapter 4, to have an initial insight in the usability of such variables for this supplier evaluation model. This thus results in a risk score and level, and predicted improvement per supplier to decide upon. The next step is to find a simple heuristic that classifies supplier according to both risk (score or level) and improvement, whilst taking the maximum number of resources into account. Since three possible risk levels exist, it seems logical to give more importance (when classifying) to high risk suppliers than to medium risk suppliers than to low risk suppliers. In this manner, suppliers can be added to the SSIP classification until the maximum number is reached. When the whole risk level group cannot fit into the available spots for SSIP, another variable

should determine priority within this group. Since the risk level already takes care of the risk objective, the predicted improvements can now be used to take care of the improvement objective. In other words, by ranking the suppliers on the predicted improvement percent (from large to small), a possible prioritization is stated within each (risk level) group of suppliers. In conclusion, Algorithm 2 explains this risk-based heuristic in other words, and Figure 6 illustratively:

Algorithm 2: Risk-based heuristic model

Input: riskLevel and predictedImprovement for all suppliers

availableSSIP = 60

numHighRisk = sum(riskLevel = high)

numMediumRisk = sum(riskLevel = medium)

numLowRisk = sum(riskLevel = low)

if numHighRisk > availableSSIP {

    distribute SSIP over availableSSIP suppliers with riskLevel = high in order of predicted improvements, others DIY

} else if numHighRisk + numMediumRisk > availableSSIP {

    classify all suppliers with riskLevel = high as SSIP

    distribute SSIP over (availableSSIP – numHigh) suppliers with riskLevel = medium in order of

    predicted improvements, others DIY

} else if numHighRisk + numMediumRisk + numLowRisk > availableSSIP {

    classify all suppliers with riskLevel = high or medium as SSIP

    distribute SSIP over (availableSSIP – numHigh – numMedium) suppliers with riskLevel = low in order of predicted improvements, others DIY

} else {

    classify all suppliers as SSIP

}

Output: Suppliers selected for collaboration/SSIP

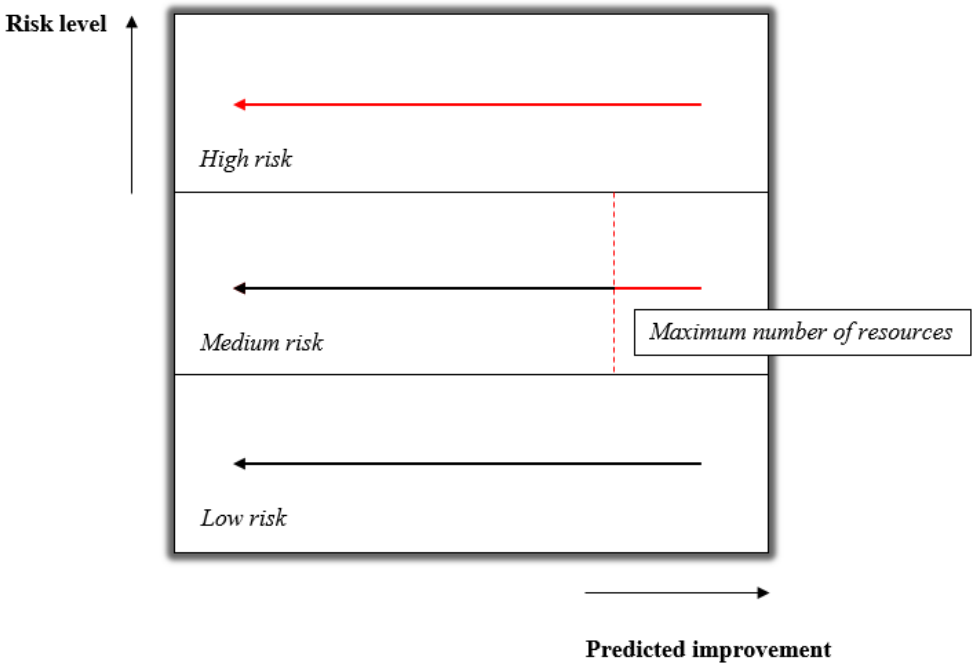

Figure 6: Heuristic evaluation model illustration

6.2.2 Distance-based heuristic model

The second step towards using both risk- and improvement-based objectives is the use of another heuristic that computes the trade-off between minimizing risk and maximizing improvement. By the use of such a model, suppliers should be ranked according to this trade-off, so that the most appropriate suppliers are classified as SSIP. Firstly, this is done for the same two variables as used in the heuristic (thus one risk score and one predicted improvement). Secondly, the difference is shown when using four variables (the three separate risk criteria scores and one predicted improvement). This one model should be able to use varying input, so that any number of objectives, criteria, or weights can be implemented and compared.

Within the field of supplier sustainability multiple models exist (Mavi, Goh, & Mavi, 2016; Rao, Goh, & Zheng, 2017), all with their own benefits and pitfalls. Since the main goal of this research is to show the added value of such models within supplier sustainability classification (compared to the current heuristic, and most probably to the risk-based heuristic model), it is mainly aimed to set a first step towards (mathematical) decision support models in the field of supplier selection and improvement. Therefore, it is chosen to make use of a popular and supported model within the field of supplier sustainability (Mokhtarian & Hadi-Vencheh, 2012; Olson, 2004), which makes the trade-off between any number of objectives and ranks the suppliers accordingly: the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) (Hwang & Yoon, 1981).

The input for the TOPSIS model are the decision (variables) matrix, objectives (per variable), and weights per objective. After that, the decision matrix is normalized and weighted, and the best and worst solution are stated, so that the Euclidian distance of every supplier to this best and worst solution can be calculated. Next, the R-index is calculated, which represents the relative distance of the supplier between this best and worst solution, and which is used to rank the suppliers by 'similarity to the ideal solution'. Depending on what is seen as 'ideal solution', the R-indexes close to 0 or 1 represent the suppliers that should be classified as SSIP. Since it is aimed to use the model for supplier classifications, the ideal solution represents the case with the least risk and least predicted improvement (thus the best case) (Tzeng & Huang, 2011). In this way, the suppliers closest to this best case are classified as DIY and the suppliers farthest away as SSIP. This method is further mathematically explained in Algorithm 3, and illustratively in Figure 7:

Algorithm 3: Distance-based heuristic model

Input: decision matrix D [n x m], m objectives [cb] and m weights [w]

$availableSSIP = 60$

$$r_{ij}(d) = \frac{d_{ij}}{\sqrt{\sum_{i=1}^{n} d_{ij}^2}}, i = 1, \dots, n; j = 1, \dots, m$$

$$v_{ij}(d) = w_j r_{ij}(d), i = 1, \dots, n; j = 1, \dots, m$$

$$PIS = \{v_1^+(d), v_2^+(d), \dots, v_m^+(d)\}$$

$$PIS = \{(\max_i v_{ij}(d) | \, j \in J_1), (\min_i v_{ij}(d) | \, j \in J_2) | \, i = 1, \dots, n\}$$

$$NIS = \{v_1^-(d), v_2^-(d), \dots, v_m^-(d)\}$$

$$NIS = \{(\min_i v_{ij}(d) | \, j \in J_1), (\max_i v_{ij}(d) | \, j \in J_2) | \, i = 1, \dots, n\}$$

$$D_i^+ = \sqrt{\sum_{j=1}^{m} \left[v_{ij}(d) - v_j^+(d)\right]^2}, i = 1, \dots, n$$

$$D_i^- = \sqrt{\sum_{j=1}^{m} \left[v_{ij}(d) - v_j^-(d)\right]^2}, i = 1, \dots, n$$

$$C_i^+ = \frac{D_i^-}{D_i^+ + D_i^-}, i = 1, \dots, n$$

Choose best $availableSSIP$ alternatives, in ascending order of C_i^+

Output: Suppliers selected for collaboration/SSIP

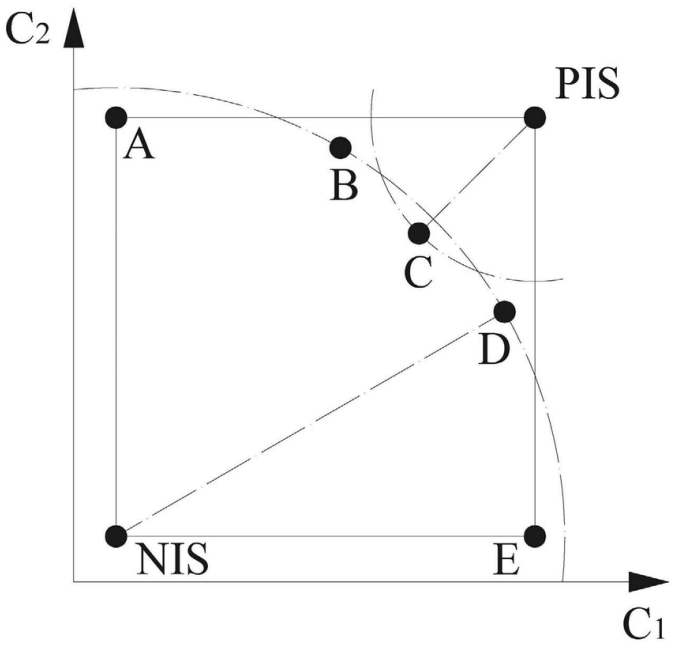

Figure 7: TOPSIS concept of PIS and NIS

The main benefit of TOPSIS is stated to be its admissibility (in terms of the possibility to use different criteria with different ranges), and simplicity (Rai & Kumar, 2016). Since the TOPSIS method is closely related with the theory that consumers base their evaluations on differences relative to references, i.e. the reference-dependence theory, the distances to PIS and NIS seem straightforward and logical as measure (Kahneman & Tversky, 1984). On the other hand, examples exist in which this relative distance to both PIS and NIS, ranks alternatives with larger distances to PIS higher than some alternatives with smaller distances to PIS (Tzeng & Huang, 2011). Another disadvantage of the TOPSIS method is that it is, naturally, dependent on several weights given per criteria, which might influence the output. However, this means that the model can be adjusted by slightly differing the weights. By the use of sensitivity analysis this suggestion is tested in chapter 7.2.

One distance-based model (A) is tested for comparisons with the current heuristic and risk-based heuristic model, and another distance-based model (B) is tested to show the potential and possibilities regarding the TOPSIS method. For distance-based model A, with two variables, the weights are chosen with the use of expert-opinion, as seen in Table 11. For distance-based model B, with four variables, the weights are calculated by multiplying the weights per objective (Table 11) with the weights per risk variable (Table 9), as seen in Table 12. Since each variable has its own best and worst case in this model, the PIS and NIS might be different than model A. In this way, it is expected that this influences the output of model B. Next to that, these variables are stated as their actual values, so that their whole range of values is taken into account, instead of their risk level in terms of low, (medium,) or high.

Objective	Expert-opinion weight
Minimizing risk	67 %
Maximizing predicted improvement	33 %

Table 11: Expert-opinion weights per distance-based model objective (A)

Objective	Expert-opinion weight
Minimizing PZTs	67 % * 60 % = 40 %
Minimizing Activity-based risk	67 % * 25 % = 17 %
Minimizing Annual spend	67 % * 15 % = 10 %
Maximizing predicted improvement	33 %

Table 12: Expert-opinion weights per distance-based model objective (B)

In conclusion, three different (heuristic) models are set up for supplier evaluation and differ in the decision making process for selecting suppliers (for collaborations). Firstly, the current model represents the current situation, which only uses risk criteria. Secondly, the risk-based model represents the currently used risk objective, together with the predicted improvements per supplier. Thirdly, the distance-based model represents the trade-off between both risk (as aggregated value, or separate values) and improvement by using the TOPSIS method. Showing the added value of these (heuristic) models serves as the first step towards the potential use of mathematical models for decision support in the SSP program, by pointing out that structure and the maximum number of resources align the evaluation model to the SSP program's objectives. The results of these models is further explained in chapter 7.2.

7. Results

The previous chapter explained in twofold what models are used for research, and this chapter explains the results of those models. Within this chapter the most important findings and conclusions are stated in twofold again, thus per model. Firstly, the results of the inconsistency model are researched and applied to the whole dataset. Secondly, the results of the supplier evaluation model are investigated and tested to a subset (one calendar year). This chapter represents the findings of implementing both models to the dataset of Philips' SSP program.

At first, by implementation of the inconsistency model the harmful inconsistency score is calculated by the use of equal weights (HIC-score) and expert-opinion weights (eHIC-score). In this way an initial insights can be given in the truthfulness of suppliers when conducting the self-assessments. Literature states that self-assessments are not accurate enough on its own, and are in need of peer assessments to bridge this gap. Perfectly truthful suppliers are expected to score a (e)HIC-score of zero, whilst any inconsistency between both questionnaires results in a (e)HIC-score above zero. The more inconsistent while answering the questionnaires, the higher the (e)HIC-score. Next to that, since multiple HIC-scores are presented per supplier, further analysis is done to how these scores relate to the suppliers' sustainability scores, over time, and before and after visiting the supplier on site. More specifically, when implementing this inconsistency model at Philips, the SAQ and ED are validated to each other.

At last, the implementation of the evaluation models is stated by comparing the current heuristic, risk-based heuristic, and distance-based heuristic. All models besides the current heuristic make use of a structured set of rules and maximum number of resources available for the focal firm. These models thus relate to the decision making process of the supplier evaluation model, i.e. which suppliers should be selected for (on site) collaborations. More specifically, the implementation of these models is done within the Philips' SSP program, by selecting suppliers for either SSIP or DIY.

7.1 Inconsistency model

Since correlations fail to give enough insights (appendix D), the HIC score is used for implementation. The results and evaluation of the inconsistency model can essentially be stated with two different weighting methods, namely using equal weights or expert-opinion weights. This results in the so-called HIC-score and eHIC-score. Using the former method would result in a HIC-score that represents the proportion of questions being answered harmfully inconsistent, i.e. false positive answers. The latter method would result in a weighted HIC-score (eHIC-score) in which answering some questions inconsistent is more harmful than for others. This can be done by using the same weights as given in the SAQ, thus based on expert-opinion.

	ED availability = 1	ED availability = 0
SAQ answer = 1	True positive (‘tp’) (20671x, 54%)	**False positive (‘fp’) (8765x, 23%)**
SAQ answer = 0	False negative (‘fn’) (1633x, 4%)	True negative (‘tn’) (7271x, 19%)

Table 13: Cell frequencies of contingency table

The HIC-score makes use of equal weights, so that it represent the proportion of questions harmfully inconsistent answered. Although this proportion is based on the 60 selected connections between SAQ and ED, as stated in chapter 6.1, it gives an initial insight the inconsistency of answering. For every connection it is then stated whether it turned out to be a false positive (value of 1) or not (value of 0). As Table 13 states, the percentage of inconsistent answers is 27%, of which most (23%) are harmful. By the aggregation of these values, the HIC-score per supplier, per dashboard field, and per selected question can be calculated.

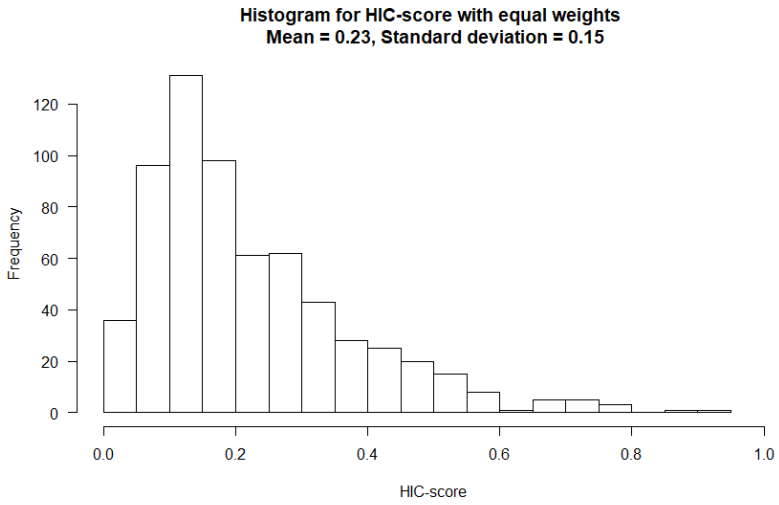

Figure 8: Histogram for HIC-score with equal weights

Firstly, this is done per supplier to gain more insights in the difference between SAQ and ED. The average HIC-score for suppliers is naturally equal to the proportion of false positives, since equal weights are used when aggregating. This thus means that suppliers are on average more consistent, but in more than a quarter (27%) of the questions the suppliers remain inconsistent. One expected reason for this is the document impracticality. Other reasons might be found in window-dressing, social desirability, unawareness, misinterpretation of the questions, etc. Although in some of these cases the suppliers are harmless inconsistent, this results in an average HIC-score of 23%. When

making use of the expert-opinion weights per question and aggregating these results per supplier, an average eHIC-score of 22% is found. Since the eHIC-score is less than the original HIC-score, this implies that the high weighing questions are (on average) filled in less harmfully inconsistent. Secondly, when all connections are aggregated per question and the five questions with the highest weights are investigated, four out of these five HIC-scores are found to be below average (3, 6, 8, and 18%) and only one above average (31%). This thus seems in line with the suggestion that the most important questions (based on expert-opinion) are filled in less harmfully inconsistent. In other words, suppliers tend to be more harmfully inconsistent for less important questions than for more important questions.

This average proportion of suppliers being harmfully inconsistent in 23% of the connections is naturally one of the reasons that explains the difference between SAQ and ED score. How these HIC-scores relate to the SAQ and ED scores can be seen in Figure 9, which illustrates a clear finding: the higher the HIC-score, the larger the difference between SAQ and ED score. This is obviously logical, since being harmfully inconsistent means that points are earned in the SAQ and not in the ED. This remains the same when using the eHIC-score. Since the HIC-score only takes a subset of overlapping questions into account, and not all inconsistent connections are used (only the false positives, not the false negatives), the whole situation cannot be captured. Next to that, multiple reasons for these inconsistencies remain unknown.

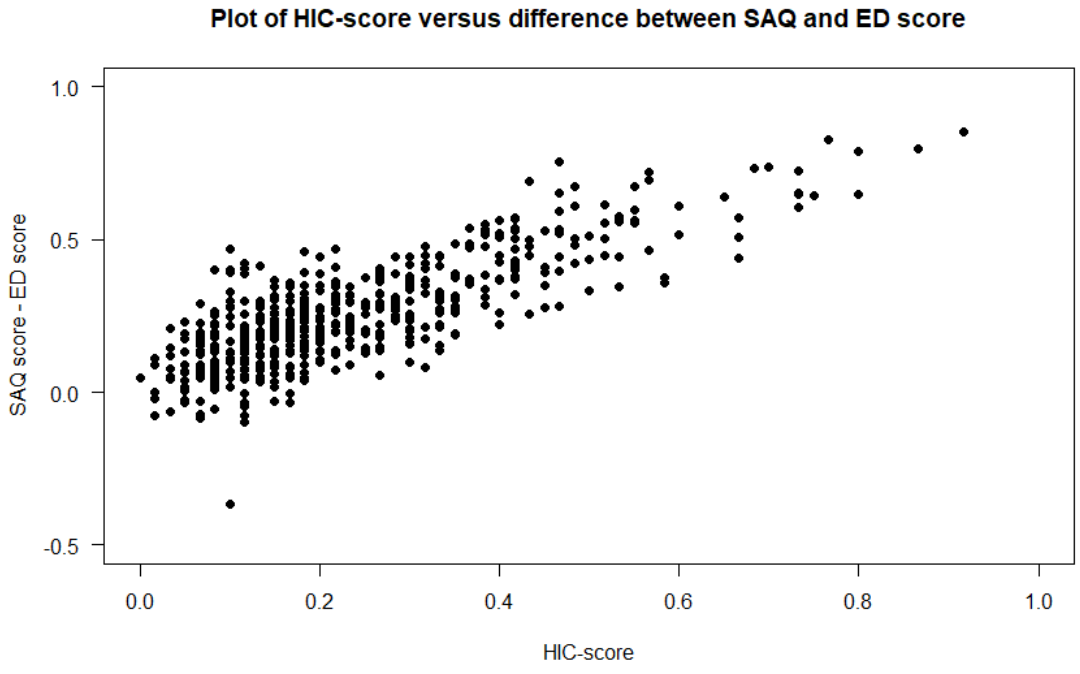

Figure 9: Plot of HIC-score versus difference between SAQ and ED score

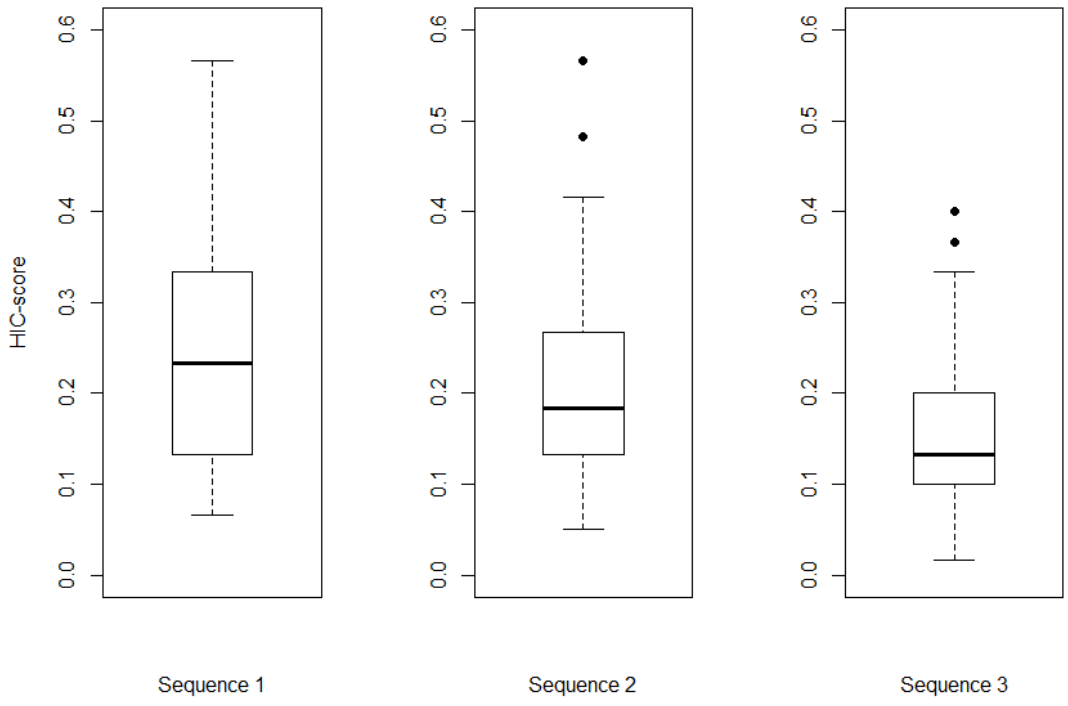

Figure 10: Boxplots of HIC-score per sequence (1, 2, and 3)

Further analysis into the usability of the HIC-score is done to assess the SSP program and whether its goals are reached. The first goal of the program is to make suppliers learn, not only in score but also in transparency and openness towards Philips, thus less (harmful) inconsistency. This analysis can be done for suppliers with two and three sequences. In the case of suppliers with three sequences, in 37 cases, the HIC-score decreases from 24 % to 21 % to 16 %, as seen in Figure 10. In the case of suppliers with two sequences only, in 123 cases, the HIC-score decreases from 27 % to 21 %, as seen in Figure 11. The claim that suppliers learn over time and try to be more transparent and open towards Philips, seems correct, although these cannot be statistically supported in terms of significant difference. The suggestion that assessors learn over time is assumed to be most clear when assessing the quality of the ED, rather than the availability, but this might moderate the previous claim. Next to that, it is seen that the widespread distribution of HIC-score decreases from sequence 1 to 3. The HIC-scores above the upper quartile, i.e. in the upper whisker, decrease and the HIC-scores in the inter-quartile range are shifting more towards the (decreasing) median.

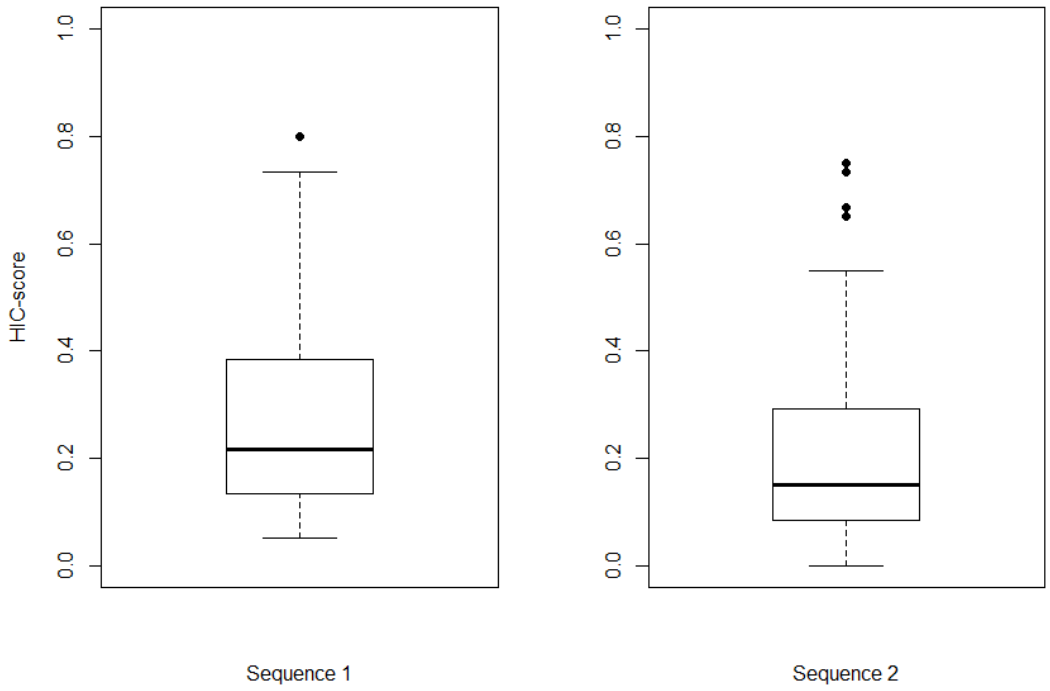

Figure 11: Boxplots of HIC-score per sequence (1 and 2)

The second claim of the SSP program is that the most truthful information is still found on-site, thus during site assessments (SAs). Although some ED are too large (or impractical) for suppliers to send to Philips, the actual availability of ED can be found at the supplier on-site and then eventually corrected. By comparing the HIC-score before and after SAs, this claim can be supported by actual numbers. In these 174 cases, the average HIC-score decreases from 22 % to 17 % after going on-site. Although 71 duplicates exist within these cases, the averages do not change after omitted these cases. This average decrease in score can thus be explained by the change in ED availability, although some (9) cases exist in which the HIC-score increases after SA. In these cases, some ED are found to be not available on-site, whilst the provided ED before SA stated they would be available. This could be the case in terms of falsified records or wrong assessment of the ED by Philips. Although these 9 cases are equally distributed over the years 2016, 2017, and 2018 (thus 3, 3, and 3), the proportion of such cases found over the years is 16 %, 4 % and 3 %, respectively. The power of a SA can be clearly seen in Figure 12, in which the difference in HIC-score is stated before and after a SA. Next to the suggestion that the HIC-score decreases on average, it is illustrative that the boxplot before SA shows high outliers (such as the value 0.92) whilst the boxplot after SA only shows one outlier (with value 0.47).

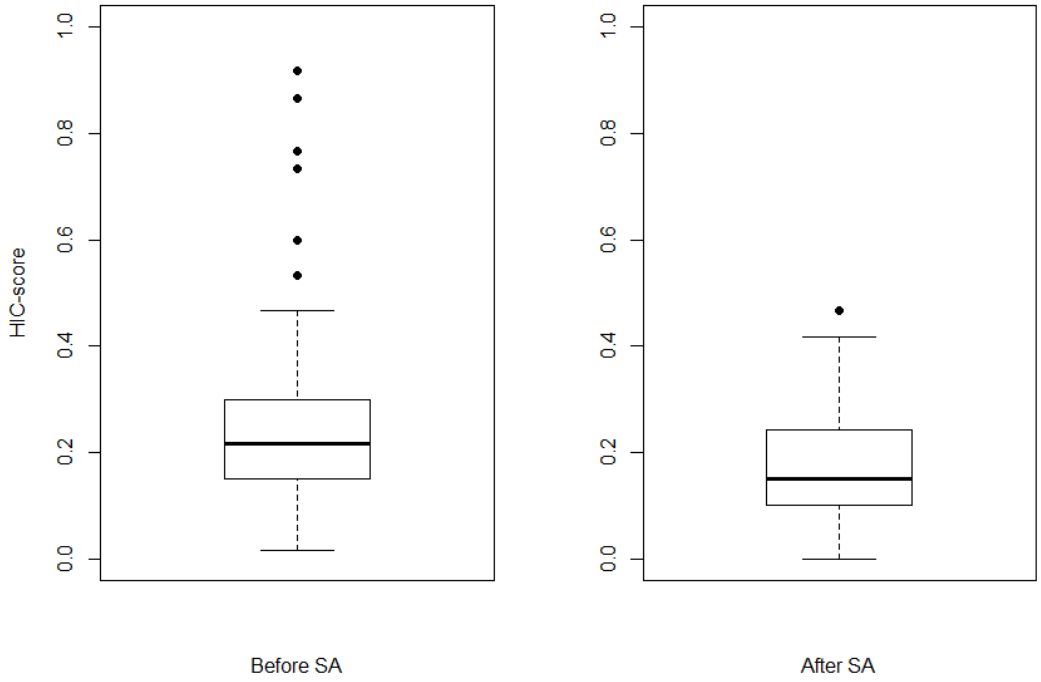

Figure 12: Boxplots of HIC-score before and after SA

It can thus be stated that suppliers (and possibly assessors) do learn over time, and that site assessments (SAs) still create a higher level of transparency and consistency in terms of document availability. On the other hand, this cannot be statistically supported. As explained, the decrease might also be due to document impracticality (e.g. thousands of pages). So despite the fact that the difference between SAQ and ED score is explained by the change in the level of quality, it also results from the documents being (un)available in the first place. This might also (partly) contribute to the suggestion that SSIP suppliers reveal a higher ED score after the SA than before.

When aggregating on question-level, the questions with the highest HIC-scores might indicate the problematic questions (for many reasons, such as impractical documents, falsified records, misjudgement, or finding a PZT) which could be focussed on-site and in further sequences. To indicate the most inconsistent connected questions without sharing confidential information: 10 connected questions have a higher HIC-score than $\mu + \sigma$, and 1 connected question has a higher HIC-score than $\mu + 2\sigma$. Further analysis on question-level to the differences in document availability would be valuable to indicate which documents are most often 'found' (change in availability from "Yes" to "No"), or result in a PZT (change from "Yes" to "PZT"). This information could also serve as input for predicting ZTs on question- and/or supplier-level.

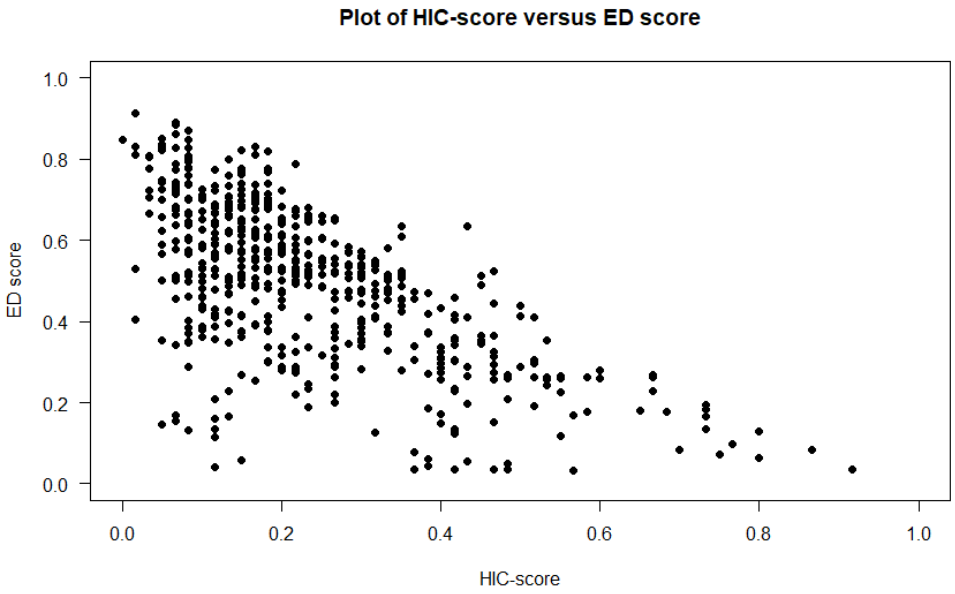

Figure 13: Plot of HIC-score versus ED score

In conclusion, the HIC-score represents the difference between document availability as claimed in the SAQ and later provided ED, whilst focusing on the false positives only since these are seen as the most harmful inconsistencies. By creating a subset of connection questions, which exactly ask for the same document, an initial insight is created in the transparency and openness of the suppliers and the difference between SAQ and ED score. Although this difference is mainly caused by the diverging focus of the SAQ (mainly assessing availability) and ED (mainly assessing quality), Figure 9 and Figure 13**Error! Reference source not found.** still illustrate the relation between HIC-score and difference in scores. It is thereby found that high HIC-scores relate to large differences, which is partly logical due to points scored in the SAQ whilst scoring no points in the ED. It is also found that the HIC-score correlates with this difference ($r = 0.80$, p-value < 0.05). Potentially more importantly, it is found that low HIC-scores result in both low and high ED values (Figure 13), which implies that not only high-scoring suppliers are consistent and transparent in answering, but (some) low-scoring suppliers too. Next to that, further analysis is done to aggregating the connection questions on question-level, and using a weighted HIC-score that gives more importance to certain (connected) questions based on expert-opinion. Furthermore, the HIC-score supports the claims that suppliers (and assessors) are learning over time and that going on-site remains the best method to gain transparency and openness at suppliers. The HIC-score can potentially be used as truthfulness or risk indicator, as well as predictor for future ED- or HIC-scores, which is further explained in Appendix E. Next to that, Philips should be able to focus on

these inconsistencies and potentially inform suppliers about their HIC-score or give feedback on question-level.

7.2 Supplier evaluation model

The output of each model (current, risk-based, and distance-based model) is the selection of SSIP suppliers. So for each supplier, their characteristics, risk criteria, predicted improvement, and model SSIP/DIY classification are now known. Each model is constrained by the choices in weights and threshold values, whilst on the other hand this results in great adjustability and usability in general. For this research the weights and threshold values are chosen by expert-opinion from Philips, but since sustainability has no single concept this can easily be changed when applying the models within another company or industry. It is chosen to compare the different models on various characteristics and variables, such as:

- Average risk score of selected SSIP supplier (the higher, the more focused on risk)
- Average predicted improvement in of selected SSIP suppliers (the higher, the more focused on improvements)

7.2.1 Supplier evaluation models with risk as one aggregated score

When comparing these variables to the total group of suppliers and to each other, a first insight in differences and performances of the models is given. Besides that, as previously stated in chapter 6.2, the current heuristic already showed some absence of structure, which resulted in one supplier with PZTs not being classified as SSIP (Table 8) and several suppliers with a high risk score not being classified as SSIP (Table 10) as well. The setup of these tables is again used to check whether structured models do not result in any high risk suppliers to be classified as DIY instead of SSIP, as is expected. When applying the models, each one results in a vector with 60 SSIP classifications and 97 DIY classifications. The comparisons are summarized in Table 14 below.

Comparison variables *(average or total for all 157 suppliers)*	Current	Risk-based	Distance-based A	Distance-based A+	Distance-based B	Distance-based B+
Average risk score in SSIP $\mu = 0.28$ *(total)*	0.35	0.37	0.44	0.42	0.43	0.37
Average predicted improvement in SSIP $\mu = 19\%$ *(total)*	15%	32%	17%	30%	19%	35%
Number of high risk suppliers in SSIP *(8 in total)*	7	8	8	8	8	8
Number of medium risk suppliers in SSIP *(119 in total)*	48	52	52	49	52	42
Number of low risk suppliers in SSIP *(30 in total)*	5	0	0	3	0	10

Table 14: Overview of supplier evaluation models and accessory comparison variables

As can be seen in the summarizing table above, the current model is compared with the risk-based and distance-based heuristics as explained in chapter 6.2. The distance-based model is divided in models A (risk as one aggregate value) and B (risk as three separate criteria), in which models A+ and B+ indicate the inclusion of the improvement objective. In other words, models A and B only assess risk (so that it can be compared to the current heuristic), whilst model A+ and B+ work towards aligning the model with the program's objective (thus the trade-off between risk and improvement).

That fact that the current model is only risk-based can easily be seen by the average risk score of SSIP suppliers of 0.35 (compared to 0.28 for all suppliers), whilst the average predicted improvement is 15% (compared to 19% for all suppliers). This seems in line with the findings in chapter 5 because suppliers with SSIP classifications have, on average, higher ED scores than suppliers with DIY classifications. Due to the Markov Chain Simulation used to predict the improvements, lower ED scores result in higher predicted improvements.

The use of model A results in a higher risk score (0.44) and higher predicted improvement (17%), compared to the current model. Since both take only risk into account the use of the model A outperforms the current model in terms of risk, improvement, and structure. The use of the risk-based model instead of the current model would mean an increase of both average risk score (0.37) and predicted improvement (32%) in SSIP scope, because of selecting those suppliers that are less similar to the (positive) ideal situation. This seems logical since the risk-based heuristic works from high to low risk level and from large to small predicted improvement. The use of model A+ results in a higher risk score (0.42) but lower predicted improvement (30%), compared to this heuristic model, but outperforms the current model as well.

It can thus be said that all models outperform the current model, in terms of risk, improvement, and structure. Differences within the risk-based model and distance-based model A(+) still exist, and show the importance to the trade-off to be made. As can be seen in the number of high, medium, and low risk suppliers within SSIP scope (per model), the main difference between the risk-based and distance-based models is made in how to divide the suppliers with medium and low risk (52 and 0, versus 49 and 3, respectively). In the case of the model A+, there are thus at least three low risk suppliers that have such a high predicted improvement (potential) that these are preferred over other medium risk suppliers with poor predicted improvement. This is naturally due to the weights assigned to the objective of maximizing predicted improvement, and this might shift when changing the weights between the two objectives. In terms of SSIP classifications given, the risk-based model and distance-based model A+ agree with each other in 69% of the suppliers, and thus differ in 31%.

7.2.2 Supplier evaluation models with risk as separate criteria

Until now it is assumed that risk is defined as one single aggregated score of the risk criteria, which should be minimized. When this assumption is modified by the assumption that the three risk criteria should be used separately, each with their own objective, the distance-based model naturally changes. The PIS and NIS most probably change by this modification, and another classification distribution is expected. This results in distance-based models B and B+.

The use of distance-based model B+ results in an average risk score of (0.37) and predicted improvement of (35%). Compared to the distance-based model A+, this model B+ takes the predicted improvements somewhat more into account, which can possibly be explained by the method used to calculate its weights. Although the three risk variables combined represent 67% of the total weight, these are now all split and do not represent risk aggregately but their own separate variable, whilst on the other hand the predicted improvement weight remains 33%. This is just slightly lower than the separate weight of the risk variable PZTs (40%), but clearly higher than the other two risk variables (17% and 10%). Using these weights it is again seen that more low risk suppliers are classified as SSIP, which is expected to be the result of the relative high weight on predicted improvements. The difference between model B and B+ shows the difference of taking the improvement objective into account or not.

Another comparison can be made by investigating the DIY suppliers, as stated in Table 15 below. These variables show the lowest ED score, highest risk score, and highest predicted improvement not classified as SSIP, but as DIY. For example, the highest risk score of 0.40 found is worse than the 0.28, since the objective is to minimize risk. In this comparison it can again be stated that distance-based model B+ focuses more on predicted improvement (and ED score), whilst distance-based model A+ focuses more on risk, which is in line with the findings in Table 14.

Comparison variable	Distance-based model A+	Distance-based model B+
Lowest ED score in DIY	19%	**27%**
Highest risk score in DIY	**0.28**	0.40
Highest predicted improvement in DIY	57%	**25%**

Table 15: Overview of comparison between distance-based models

7.2.3 Sensitivity analysis

Since it is aimed to build a supplier evaluation model that fits all objectives necessary, and in which weights can still be adjusted according to expert-opinion or statistics, it is advised to choose the distance-based (TOPSIS) model. For the sensitivity analysis distance-based model B+ is used. By splitting the risk variables the compensation effect decreases and less loss of (important decision

making) information takes place, as concluded in chapter 4 (RQ1). However, the high weight of risk criteria PZTs (60%) cannot be compensated by the other two risk criteria. Given these weights assigned to the risk criteria by expert-opinion (Table 9), it is chosen to perform a sensitivity analysis on the weights assigned to the objectives (Table 11). The weights chosen for the sensitivity analysis are all percentages from 0 to 100% with steps of 10%, so that insights are given in distance-based models with full focus on risk (100% risk, 0% improvement) and on improvements (0% risk, 100% improvement), and everything in between. By differing these weights, the output of the model changes as well in terms of classification distribution. The comparison variables remain the same, as previously stated in Table 14.

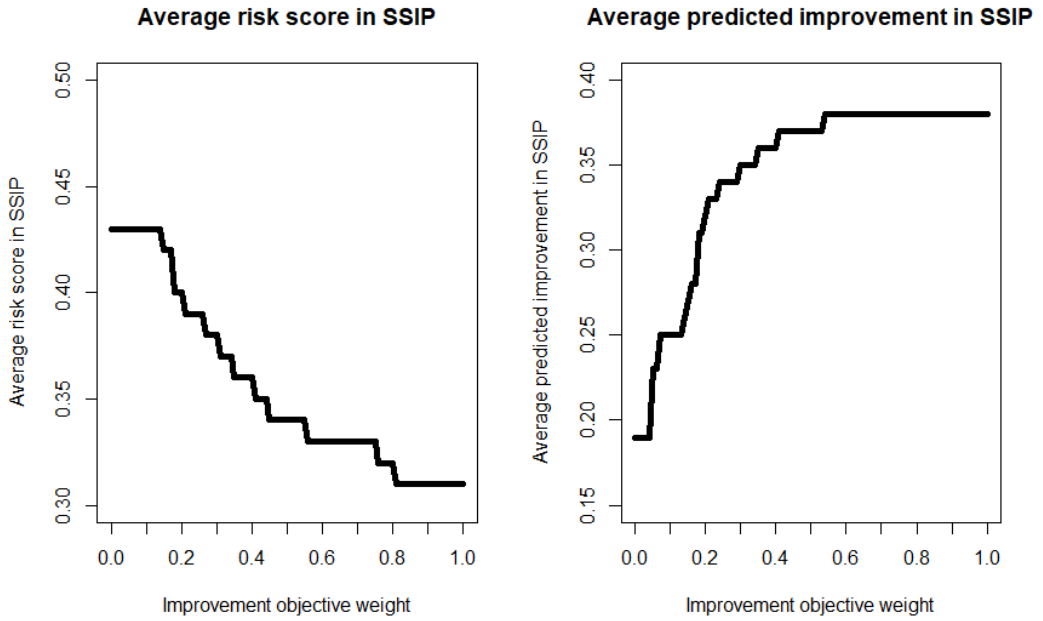

Figure 14: Sensitivity analysis on average risk score and average predicted improvement

As can be clearly seen, the more weighted the improvement objective becomes (and thus the less weighted the risk objective becomes) the average risk score decreases and the average predicted improvement increases. This is according to expectations, and stresses out the sensitivity to the choice in trade-off that needs to be made. In Figure 14 it is seen that the steepest lines (either as increase or decrease) are found between 0 and 20 % as improvement objective weights. Next to that, it is seen that these lines only slightly change from 60 % onwards. It is therefore expected that the optimal value is found between the 20% and 60%. Furthermore, the average risk score of SSIP suppliers always remains above the average risk score of all suppliers (0.28). The average predicted improvement of SSIP suppliers only remained above the average predicted improvement of all suppliers (19%) with weights higher than 0%. Besides that, to create a better model than the

current situation in terms of risk, the distance-based model should have a higher average risk score of SSIP suppliers than 0.35, which is only the case with improvement objective weights from 0 to 40%. It can thus be concluded that improvement objective weights between approximately 20 and 40% (and thus risk objective weights between approximately 60 and 80%) create the best model in terms of both risk and improvement. Depending on what focus the distance-based model should have, expert-opinion weights can be chosen in this range.

Another approach to find the optimal weight(s) is to use Multi-Objective Optimization to find the solution that makes the best trade-off between average risk score and predicted improvement. Although it is possible to define four separate weights for the four criteria, it is hereby focussed to find the best weights for risk and improvement, which sequentially result in the four criteria weights. At first, the expert-opinion weights (67% risk, 33% improvement) were used as to calculate these four criteria weights. Another possibility is to use these two weights for the Weighted Sum Method, which adds all objectives into one single objective and is used to find the optimal solution. This would results in the following objective:

$$\max_{w} F(w) = 0.33 * \mu_{predicted\ improvement}(w) + 0.67 * \mu_{risk\ score}(w)$$

When using the expert-opinion weights in this formula would results in a score of $F(33\%, 67\%) = 0.3634$. A simple check of all possibilities of integer percentages shows that the optimum value for this objective is found in the following case: $F(14\%, 86\%) = 0.3739$. The average risk score and predicted improvement of this optimal salutation are, respectively, 0.43 and 26%. Compared to the results as stated in Table 14 the model is more risk-focussed (0.43 versus 0.35) and less improvement-focused (26% versus 35%). Based on the expert-opinion about risk and improvement, there are thus two possibilities: 1) use these weights as input for the model, 2) use these weights to define the trade-off between average risk score and predicted improvement of the SSIP suppliers. Since expert-opinion is used for all weights and threshold values, the choice in model can best be made according to expert-opinion too.

In the conclusion, by making use of the currently used risk criteria (thus annual spend 2017, activity-based risk via the SAQ and PZTS via the ED) the current evaluation model is tested and shows the lack of structure and absence of set of rules to decide upon supplier classifications. By making use of a heuristic models this already improves the structure. Next to that, the use of the maximum number of resources and improvement objective (besides the risk objective) improves the supplier evaluation model as a whole. In this way suppliers with high risk (potential) and high predicted improvement (potential) are visited on-site so that Philips can collaborate and start the conversation with the supplier to make the best plan in how to improve, in terms of both risk and improvement. The best model is argued to be the distance-based model, since it allows for multiple objectives and adjustment of the weights per objective, in case of the TOPSIS method.

It is mainly aimed to firstly find the best model for supplier evaluations, which is thus the distance-based model, and secondly to state what information should be used as input. Based on the conclusions from chapter 4 and 6.1 (RQ1 and RQ2), it is recommended to add variables that reveal more about the objectives for the decision makers than currently done. This includes the ED score (in total and per topic), the maturity level (in total and per topic), number of employees, HIC-score (RQ2), root causes for ZTs (research in progress by a fellow researcher) and learning curve performance (research in progress by a fellow researcher). What variables are used depends on the model implemented and the number of variables to be taken into account, since this can increase the compensation effect. An overview of the variables that decision makers are recommended to take into account within Philips are stated in Table 16 below, which is thus based on knowledge from RQ1 and RQ2, own observations, and future (broader TKI Dinalog) research directions.

	Risk-based	Improvement-based
PZTs	X	X
Supplier activities	X	
Annual spend	X	
Strategic importance	X	X
Predicted improvement		X
ED scores (in total and per topic)	X	X
Maturity level (in total and per topic)	X	X
Number of employees		X
HIC-score	X	
Root causes for ZTs	X	
Learning curve performance	X	X

Table 16: Overview of recommended variables to research in evaluation models

In terms of implementation of the supplier evaluation model it is thus proposed to use the distance-based model B+, so that multiple variables can easily be added. Since Philips has more than 200 active suppliers, it is not efficient to run the model after each supplier provided its assessment. One promising direction of this model is that predictive variables can be used that are known in the beginning of the year when the planning is made. By the use of these variables, the planning for the upcoming year can already be predicted and used as initial SSIP selection. When new information is gathered via assessments, and new risk or improvement potential arises, the decision can still be made to change the SSIP selection. Since the TOPSIS method ranks suppliers, it can be easily seen which supplier switches places with the other. For example, when one supplier is predicted to be DIY, but then one or multiple PZTs are found, this supplier should be classified as SSIP. The SSIP supplier with the lowest rank therefore becomes DIY, and the DIY supplier with the PZT(s) becomes SSIP. In this way, the model can still be overruled when the actual situation differs from the prediction, or when exceptions need to be made.

8. Conclusion

In the field of supplier sustainability assessment one of the concerns remains the untruthfulness of suppliers that leads to inaccurate and biased results, which are further used by decision makers of the focal firm, regardless of the approach used (Esbenshade, 2005; Pruett, Merk, Zeldenrust, & de Haan, 2005; Rodríguez-Garavito, 2005). Although many focal firms shift away from second-party monitoring approaches, such as auditing, this inaccuracy (slightly) remains when making use of first-party monitoring, such as self-assessments (Foerstl et al., 2010; Pimenta & Ball, 2015). The main aim of this research is to create a method on how to validate these self-assessment questionnaires with peer assessment, and to revise the decision making process behind supplier evaluation models that selects which supplies should be collaborated with.

Firstly, what approach should be used to assess suppliers and establish actual sustainability scores comes forward from a literature research, which concludes on the suggestion that focal firms should collaborate with their suppliers as much as possible, in order to increase the transparency, openness, sustainability performance, and shared value (A. Van Weele & Van Tubergen, 2017). One of the most frequently used approaches remains auditing, although many critics exist. Since focal firms often lack time and financial resources to collaborate with its suppliers, efficient use of monitoring should be used to collaborate with the most risky and/or low performing suppliers (Baden et al., 2009; Green et al., 1996; Harland et al., 2003; Jiang, 2009; Lippmann, 1999). The most efficient monitoring approach is suggested to be self-assessments, when corrected with peer assessments (Foerstl et al., 2010; Pimenta & Ball, 2015). This knowledge contributes to the field of supplier sustainability programs and assessment methods, in which auditing is thus still popular. Since sustainability is not one single concept that fits all, this assessment might differ per industry or company (Chatterji & Levine, 2006). Next to that, the focal firm's goal, scale and type of sustainability are important parameters to choose the best approach to assess and compare (Gan et al., 2017). It can be concluded that it is fairly possible to judge one's sustainability level in one endpoint to assess and compare the suppliers, although more midpoints are recommended when making decisions on the supplier (Kägi et al., 2016; Krajnc & Glavič, 2005). This is especially the case when the aggregating method allows compensation. It is suggested there exists no combination of one weighting and aggregating method that is the best for all. These suggestions contribute to how to establish sustainability scores that best reflect the actual situation and are in line with the supplier sustainability program and decisions made.

Secondly, a method is created to bridge the gap between self- and peer assessments and thus to validate the suppliers' data (Lamming & Hampson, 1996; Pilbeam et al., 2012; Subic et al., 2013). Instead of only comparing these results in terms of scores, the self-assessment answers can actually be validated by peer assessment, even though the questions only partly overlap. By connecting the self- and peer assessment questions an expected correlation of 1 should be found, although

inconsistencies always occur. To differ between false positives and negatives, the inconsistency of self-assessing is investigated by using contingency tables and focus on the false positives. These harmful inconsistencies might indicate window-dressing or social desirability, but could also indicate unawareness or misinterpretation (Allen & Van Der Velden, 2005; Brown & Harris, 2014). By aggregation a harmful inconsistency (HIC) score can be set up on both supplier and question level, which can serve as initial insight in truthfulness and feedback for suppliers in order to increase transparency and truthfulness. Since supplier information should be as accurate and unbiased, this finding contributes to reducing inaccurate scores that might lead to shifted results (Distelhorst et al., 2016; Gualandris et al., 2015; Yamin et al., 2017). Within the SSP program, self-assessment questions and supporting evidence documents are connected and validated, and result in an average HIC-score of 23%. In other words, for less of a quarter of the connected questions the evidence documents were not provided. On the other hand, it is found that suppliers that answer consistently (and are expected to be more truthful and transparent) can score both low and high (ED) scores, and it is thus found that not only the high scoring suppliers are consistent in answering. Furthermore, it is seen that suppliers are learning over time in terms of consistency and these can even be increased by collaborating and visiting the supplier on-site. However, these findings cannot be statistically supported, although the SSP program is just a few years in place.

Thirdly, the decision making process of supplier evaluation models is revised and structured to align with focal firms' goals and objectives. Next to that, the maximum number of resources available to the focal firm should be in place when selecting with which suppliers to collaborate with (Trapp & Sarkis, 2016). Two heuristic models are set up to minimize risk and maximize improvement. It is suggested to use the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) method, because to its simplicity and adjustability (Hwang & Yoon, 1981). This distance-based model selects those suppliers farthest away from the best case supplier (no risk, no improvement potential) to collaborate with and visit on site, whilst taking the maximum number of resources into account. Within the SSP program, the use of either heuristic model already improves the current situation in terms of both risk and improvement potential, of which the latter was not used beforehand.

It is thus concluded that heuristic models already structure the evaluation process according to the trade-off made (between risk and improvement at Philips), and can be used by subsequently ranking the suppliers to choose between visiting on-site or not. Since only the evidence documents are currently used for scoring, the HIC-scores can also serve as risk criterion. Besides that, it gains feedback on supplier and question level and possibly serves as predictor. The average HIC-score of 23% however shows that the validation of questionnaires remains an important issue for accurate and unbiased data. Since the TOPSIS method allows for adjustability of the criteria, findings of these inconsistencies can be added as initial factor of accuracy, truthfulness, or risk.

9. Implementation at Philips

In terms of recommendations for Philips to implement the conclusions and findings of this research, suggestions are stated per research question and with regards to the data independently. In the following chapters the limitations and future research directions are listed, so these are out of scope for this chapter. Within scope are mainly the steps and issues with regards to this research and how to implement it within the Philips environment and SSP program.

The recommendations in terms of data can essentially be stated in twofold: create standard file naming conventions and take away the subjectivity of assessing. Instead of checking the inconsistencies manually, one effortless preventive solution for this file naming convention is to create a macro in the assessment (Excel) workbooks that automatically generates the file name and saves the file accordingly. In this way, most mistakes cannot occur anymore, it releases the burden of manual checking, and incorrect file naming are not expected to happen. Furthermore, the quality of the evidence documents is suggested to be subjective from time to time and might relate too much to the score of last year, and assessor of the documents. By setting a general quality scoring standard, and/or cross-checking the documents among the assessors, the quality and subsequently ED score is expected to be more accurate and unbiased.

The recommendations in terms of the research questions can be stated in threefold: the use multiple endpoints instead of one, the HIC-score per supplier as feedback between SAQ and ED, and implement the distance-based model with overruling possibilities for exceptions. Firstly, it is suggested to be important that decisions are not made using one single sustainability score per supplier, but rather multiple endpoints (sustainability topic scores) and accessory maturity level(s). Since there is an actual difference between one supplier having equal scores per topic (e.g. 80%) and another supplier having three perfect scores and one low score (e.g. three times 100% and one 20%), it is critical to differ between these two exemplary suppliers (since both would approximately result in the same single score). The same goes for two suppliers with almost exactly the same topic scores, but different improvement potentials. Secondly, the HIC-scores should be used as feedback for suppliers between the SAQ and ED, as suggested by the research from Smouter (2018). By doing so, suppliers learn more effectively and this is expected to result in more accurate and unbiased data. This is in line with the goals of the SSP program, namely to increase openness, transparency, and truthfulness among suppliers. Thirdly, the implementation of the distance-based model should go along with the possibility of overruling some classifications in case of exceptions, as earlier mentioned in chapter 7.2.3. In this way the model still serves as guiding supplier evaluation model, but it allows for exceptions and is able to learn from inefficiencies. When overruling one classification, the ranking of the distance-based model should automatically be able to select the next supplier on the list.

10. Limitations

The models presented in chapter 6 give promising results, as seen in chapter 7, in terms of validating the suppliers' data in order to work towards accurate and unbiased data, and revising the supplier evaluation model in order to structurally take both SSP program objectives into account. On the other hand, both models can be used in different manners and still have sufficient room for improvement. This chapter elaborates on the assumptions made and limitations of this research as a whole, although it is more focused on the models. These limitations naturally result in future research directions, in which these limitations are not in place or overcome.

One of the first limitations faced within this research is that it is dependent on the SSP program of Philips, and the decisions made by experts regarding its setup and design. At first, this relates to the design of using self-assessment questionnaires and peer assessment (supporting evidence documents) to assess and compare suppliers, using on-site visits as extra validation, and the program's objective to improve suppliers the utmost. At second, all decisions made regarding the self-assessment questions and supporting evidence documents asked, weights assigned, scoping and threshold values are dependent on expert-opinion. Although the SSP program is designed with regards to the RBA and international standards as well, it might not be representable for all companies outside Philips. This is logical since there is no single concept of sustainability, and each industry or company has its own view and objectives behind sustainability (programs). The inconsistency model of this research is thus dependent on the questions asked by Philips and its quality. For both self-assessment questions and supporting evidence documents requested the interpretation of the suppliers might play a role in the answers and/or provided documents. Unfortunately, this (mis)interpretation is not researched and quantifiable. On the other hand, misinterpretations due to wrongly translated questions from English to Chinese might appear to all questionnaires and could result in consistent misinterpretations. Next to that, yearly trainings are organized so that suppliers align with the SSP program and increase their understandability.

Besides that, another limitation in terms of data is the waterfall structure used by Philips, which might result subsets of questions not visible to the supplier. The answers to these (invisible) questions are thus missing values, although missing values might also appear due to resistance, unawareness, or questionnaire fatigue. The difference between one and the other missing value is not visible in terms of data, although differences between resistance and waterfall structure missing values might be important and potentially indicate inconsistency and/or untruthfulness. However, missing values being transformed to zeroes are not expected to impact the harmful inconsistency scores largely, since it is assumed that the combination of positive SAQ answers (thus non-zeroes) and absence of evidence documents might indicate harmful inconsistency.

Another limitation of the inconsistency model might lie in the suggestion that the quality of the evidence documents is occasionally rated subjectively. The assessors can score each evidence document as 0, 0.5 and 1, but there is no clear structure for all documents, so each document is also compared to the quality of last year. In this way, a score of 0.5 for one supplier might not be of the same value as 0.5 for another supplier. Next to that, this might differ per assessor. If this same 'subjectivity' is also used when rating the availability of the evidence documents, this might affect the inconsistency model. Because availability should be objective, obviously, this research makes use of these availabilities as correct and actual information.

Next to that, the literature review is naturally subject to the subset of literature in scope of this research, and the subset of literature chosen and investigated by the researcher. Since the literature regarding sustainability is gaining growing attention, not all literature could be investigated and new insights might appear daily. In terms of sustainability assessment approaches, systematic literature reviews might result in more complete summaries than the current literature review, which is more evaluative from nature.

In terms of limitations to the supplier evaluation model(s), it should initially be stated that it is mainly aimed to investigate the added value of (heuristic) models instead of the currently used model. As earlier stated, no future research is done in the many mathematical models, but the frequently used and simplistic TOPSIS method is chosen as exemplary model. Since the TOPSIS method allows adjustability as well, different models can easily be set up and compared.

Another limitation within the supplier evaluation models is that limited information is used as risk criteria. Since not all information is known for each risk criteria (e.g. annual spend, strategic importance), this limits the number of suppliers in scope. Because of this decision 157 of the 208 suppliers are used, although accuracy of the models might increase when all information is available. In terms of predicted relative improvements, this research is dependent on the data provided by a fellow researcher that achieved a RMSE of 0.07. Since improvement potential can be stated in multiple ways, these predictions already give a sufficient insights for this research.

At last, the use of expert-opinion for threshold values and weights of the improvement and risk criteria are based on experts within Philips, although this might differ per industry or company. It is even possible that other criteria might be more relevant for others, but the chosen criteria (and its weights) for this research thus relate to Philips. Further sensitivity analysis towards these weights and threshold values might give further insights, although the current sensitivity analysis done already demonstrates the trade-off between risk and improvement.

11. Future research

Although this research already delivers two models that support supplier sustainability assessment programs, there are some potential research directions that should be noted and could deliver important results. These future research directions are noted per step in this research, but are again mostly focused on the models provided.

Firstly, as frequently mentioned before, it is valuable to know the reasons behind the inconsistencies between self-assessment and supporting evidence, so that the next step towards truthfulness can be made. Currently inconsistencies can potentially be the result of impracticality, unawareness, window-dressing or social desirability(Jia, Zuluaga-Cardona, Bailey, & Rueda, 2018), questionnaire fatigue (Jenkins, 2001; Kogg & Mont, 2012; O'Rourke, 2003), missing values in the data, assessors (Esbenshade, 2005; Pruett et al., 2005) and so on. When further investigating these reasons, and researching why the HIC-score (i.e. document availability) changes after site visits, more knowledge is gained in truthfulness.

Secondly, when using the distance-based model of this research, sensitivity analysis in the weights and threshold values might give valuable insights in the trade-off made between risk and improvement (Olson, 2004). Although this might differ per industry or company, the optimal solution for this trade-off might lie in the weights and/or threshold values chosen. Another important topic when making this trade-off is what predicted improvement potential exactly means. Within this research the predicted improvement for the next sequence is used as best available improvement potential per supplier, although it might be better to predict what the improvement would be when selecting that supplier for collaboration and visiting on-site.

Thirdly, in terms of recommended criteria to use for the distance-based model, it should first be researched how to assess the supplier's maturity level and if it is necessary to investigate the maximum potential per supplier (type) (Correia, Carvalho, Azevedo, & Govindan, 2017; Foerstl et al., 2010; Klimko, 2001). For example, a transportation company would most probably not need waste water documents and never score points for these documents, thus limiting its score. When maturity levels result from the scores, maturity levels are limited as well. When maturity levels result from subjectivity or another method, this limitation might are be taken care off.

Fourthly, the use of mathematical models should be further investigated in terms of what decisions to make with these models, and what models are the most appropriate. This is in line with the following phases of the TKI Dinalog consortium, i.e. the use of mathematical models for decision support is expected to be researched in the future already. For the supplier evaluation decisions made, other models (instead of TOPSIS) might be of better support. Different extensions of the TOPSIS method are known too, and might be the next step towards these mathematical models (Mavi et al., 2016; Olson, 2004).

Bibliography

Ahi, P., & Searcy, C. (2013). A comparative literature analysis of definitions for green and sustainable supply chain management. *Journal of Cleaner Production*, *52*, 329–341. https://doi.org/10.1016/j.jclepro.2013.02.018

Allen, J., & Van Der Velden, R. (2005). The Role of Self-Assessment in Measuring Skills. *REFLEX Working Paper*, (March), 1–24. https://doi.org/10.1016/j.ajog.2007.02.051

Amaeshi, K. M., Osuji, O. K., & Nnodim, P. (2008, August 19). Corporate social responsibility in supply chains of global brands: A boundaryless responsibility? Clarifications, exceptions and implications. *Journal of Business Ethics*. Springer Netherlands. https://doi.org/10.1007/s10551-007-9490-5

Andersen, M., & Skjoett-Larsen, T. (2009). Corporate social responsibility in global supply chains. *Supply Chain Management: An International Journal*, *14*(2), 75–86. https://doi.org/10.1108/13598540910941948

Baden, D. A., Harwood, I. A., & Woodward, D. G. (2009). The effect of buyer pressure on suppliers in SMEs to demonstrate CSR practices: An added incentive or counter productive? *European Management Journal*, *27*(6), 429–441. https://doi.org/10.1016/J.EMJ.2008.10.004

Bartke, S., & Schwarze, R. (2015). No perfect tools: Trade-offs of sustainability principles and user requirements in designing support tools for land-use decisions between greenfields and brownfields. *Journal of Environmental Management*, *153*, 11–24. https://doi.org/10.1016/j.jenvman.2015.01.040

Bartley, T. (2005). Corporate Accountability and the Privatization of Labor Standards: Struggles Over Codes of Conduct In The Apparel Industry. *Polities and the Corporation Rewarch in Politicnl Sociology*, *1*(1), 211–244. https://doi.org/10.1016/S0895.9935(05)1JU07-8

Bouchery, Y., Corbett, C. J., Fransoo, J. C., & Tan (Eds.), T. (2017). Sustainable Supply Chains A Research-Based Textbook on Operations and Strategy, *4*. Retrieved from http://www.springer.com/series/13081

Boud, D., & Falchikov, N. (1989). Quantitative studies of student self-assessment in higher education: a critical analysis of findings. *Higher Education*, *18*(5), 529–549. https://doi.org/10.1007/BF00138746

Bowen, F. E., Cousins, P. D., Lamming, R. C., & Farukt, A. C. (2009). The Role of Supply Management Capabilities in Green Supply. *Production and Operations Management*, *10*(2), 174–189. https://doi.org/10.1111/j.1937-5956.2001.tb00077.x

Brown, G., & Harris, L. (2014). The future of self-assessment in classroom practice: Reframing self- assessment as a core competency. *Frontline Learning Research*, *3*, 22–30. https://doi.org/10.14786/flr.v2i1.24

Carter, C. R., & Jennings, M. M. (2002). LOGISTICS SOCIAL RESPONSIBILITY: AN INTEGRATIVE FRAMEWORK. *Journal of Business Logistics*, *23*(1), 145–180.

https://doi.org/10.1002/j.2158-1592.2002.tb00020.x

Carter, C. R., & Jennings, M. M. (2004). The role of purchasing in corporate social responsibility: a structural equation analysis. *Journal of Business Logistics*, *25*(1), 145–186. https://doi.org/10.1002/j.2158-1592.2004.tb00173.x

Carter, C. R., & Rogers, D. S. (2008). A framework of sustainable supply chain management: moving toward new theory. *International Journal of Physical Distribution & Logistics Management*, *38*(5), 360–387. https://doi.org/10.1108/09600030810882816

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). Crisp-Dm 1.0: Step-by-step data mining guide. *CRISP-DM Consortium*. https://doi.org/10.1109/ICETET.2008.239

Chatterji, A., & Levine, D. (2006). Breaking Down the Wall of Codes: Evaluating Non-Financial Performance Measurement. *California Management Review*, *48*(2), 29–52. Retrieved from http://faculty.haas.berkeley.edu/levine/papers/Chatterji& Levine Wall of Codes CMR.pdf

Chen, L., Zhao, X., Tang, O., Price, L., Zhang, S., & Zhu, W. (2017). Supply chain collaboration for sustainability: A literature review and future research agenda. *International Journal of Production Economics*, *194*, 73–87. https://doi.org/10.1016/j.ijpe.2017.04.005

Correia, E., Carvalho, H., Azevedo, S. G., & Govindan, K. (2017). Maturity models in supply chain sustainability: A systematic literature review. *Sustainability (Switzerland)*, *9*(1), 1–26. https://doi.org/10.3390/su9010064

Cousins, P. D. (1999). Supply base rationalisation: myth or reality? *European Journal of Purchasing & Supply Management*, *5*(3–4), 143–155. https://doi.org/10.1016/S0969-7012(99)00019-2

Coyne, K. L. (2006). Sustainability auditing. *Environmental Quality Management*. Wiley-Blackwell. https://doi.org/10.1002/tqem.20119

Darnall, N., & Carmin, J. (2005). Greener and cleaner? The signaling accuracy of U.S. voluntary environmental programs. *Policy Sciences*, *38*(2–3), 71–90. https://doi.org/10.1007/s11077-005-6591-9

Distelhorst, G., Hainmueller, J., & Locke, R. (2016). *Does Lean Improve Labor Standards? Management and Social Performance in the Nike Supply Chain*. *SSRN*. https://doi.org/10.2139/ssrn.2337601

Dobrovolskienė, N., Tamošiūnienė, R., Banaitis, A., Ferreira, F. A. F., Banaitienė, N., Taujanskaitė, K., & Meidutė-Kavaliauskienė, I. (2017). Developing a composite sustainability index for real estate projects using multiple criteria decision making. *Operational Research*. https://doi.org/10.1007/s12351-017-0365-y

Donaldson, T. (2014). Bangladesh Resetting the Bar on Compliance Standards – Sourcing Journal. Retrieved December 12, 2018, from https://sourcingjournal.com/topics/compliance/bangladesh-resetting-bar-compliance-standards-18607/

Dyllick, T., & Hockerts, K. (2002). Beyond the business case for corporate sustainability. *Business*

Strategy & the Environment, *11*(2), 130–141. https://doi.org/10.1002/bse.323

Egels-Zandén, N. (2007). Suppliers' Compliance with MNCs' Codes of Conduct: Behind the Scenes at Chinese Toy Suppliers. *Journal of Business Ethics*, *75*(1), 45–62. https://doi.org/10.1007/s10551-006-9237-8

Elkington, J. (1998). Partnerships from cannibals with forks: The triple bottom line of 21st-century business. *Environmental Quality Management*, *8*(1), 37–51. https://doi.org/10.1002/tqem.3310080106

Esbenshade, J. (2005). Monitoring Sweatshops: Workers, Consumers, and the Global Apparel Industry. *ILR Review*, *59*(1). Retrieved from https://digitalcommons.ilr.cornell.edu/ilrreview/vol59/iss1/84

Foerstl, K., Reuter, C., Hartmann, E., & Blome, C. (2010). Managing supplier sustainability risks in a dynamically changing environment-Sustainable supplier management in the chemical industry. *Journal of Purchasing and Supply Management*, *16*(2), 118–130. https://doi.org/10.1016/j.pursup.2010.03.011

Gan, X., Fernandez, I. C., Guo, J., Wilson, M., Zhao, Y., Zhou, B., & Wu, J. (2017). When to use what: Methods for weighting and aggregating sustainability indicators. *Ecological Indicators*. https://doi.org/10.1016/j.ecolind.2017.05.068

Geffen, C. A., & Rothenberg, S. (2000). Suppliers and environmental innovation. *International Journal of Operations & Production Management*, *20*(2), 166–186. https://doi.org/10.1108/01443570010304242

Grant, R. M., & Baden-Fuller, C. (1995). A Knowledge-Based Theory of Inter-Firm Collaboration. *Academy of Management Proceedings*, *1995*(1), 17–21. https://doi.org/10.5465/ambpp.1995.17536229

Green, K., Morton, B., & New, S. (1996). Purchasing and Environmental Management: Interactions, Policies and Opportunities. *Business Strategy and the Environment*, *5*(3), 188–197. https://doi.org/10.1002/(SICI)1099-0836(199609)5:3<188::AID-BSE60>3.0.CO;2-P

GRI. (2014). About GRI. Retrieved September 29, 2018, from https://www.globalreporting.org/information/about-gri/Pages/default.aspx

Griffiths, K., Boyle, C., & Henning, T. F. P. (2018). Beyond the certification badge-How infrastructure sustainability rating tools impact on individual, organizational, and industry practice. *Sustainability (Switzerland)*, *10*(4). https://doi.org/10.3390/su10041038

Gualandris, J., Klassen, R. D., Vachon, S., & Kalchschmidt, M. (2015). Sustainable evaluation and verification in supply chains: Aligning and leveraging accountability to stakeholders. *Journal of Operations Management*, *38*, 1–13. https://doi.org/10.1016/j.jom.2015.06.002

Hahn, T., & Scheermesser, M. (2006). Approaches to corporate sustainability among German companies. *Corporate Social Responsibility and Environmental Management*. https://doi.org/10.1002/csr.100

Hajmohammad, S., & Vachon, S. (2014). Managing Supplier Sustainability Risk: Strategies and Predictors. *Academy of Management Proceedings*, *52*(2), 48–65.

https://doi.org/10.5465/AMBPP.2014.14266abstract

Harland, C., Brenchley, R., & Walker, H. (2003). Risk in supply networks. *Journal of Purchasing and Supply Management*, *9*(2), 51–62. https://doi.org/10.1016/S1478-4092(03)00004-9

Hoejmose, S. U., Roehrich, J. K., & Grosvold, J. (2014). Is doing more doing better? The relationship between responsible supply chain management and corporate reputation. *Industrial Marketing Management*, *43*(1), 77–90. https://doi.org/10.1016/j.indmarman.2013.10.002

Hwang, C.-L., & Yoon, K. (1981). *Multiple Attribute Decision Making* (Vol. 186). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-48318-9

Jenkins, R. (2001). Code of conduct: Self regulation in a global economy. *United Nations Research Institute for Social Develoment*, (2), 1–35. Retrieved from http://digitalcommons.ilr.cornell.edu/codes

Jia, F., Zuluaga-Cardona, L., Bailey, A., & Rueda, X. (2018). Sustainable supply chain management in developing countries: An analysis of the literature. *Journal of Cleaner Production*, *189*, 263–278. https://doi.org/10.1016/J.JCLEPRO.2018.03.248

Jiang, B. (2009). The effects of interorganizational governance on supplier's compliance with SCC: An empirical examination of compliant and non-compliant suppliers. *Journal of Operations Management*, *27*(4), 267–280. https://doi.org/10.1016/J.JOM.2008.09.005

Kägi, T., Dinkel, F., Frischknecht, R., Humbert, S., Lindberg, J., De Mester, S., … Schenker, U. W. (2016). Session "Midpoint, endpoint or single score for decision-making?"—SETAC Europe 25th Annual Meeting, May 5th, 2015. *International Journal of Life Cycle Assessment*, *21*(1), 129–132. https://doi.org/10.1007/s11367-015-0998-0

Kahneman, D., & Tversky, A. (1984). Choices, Values, and Frames. *American Psychologist*, *39*(4), 341–350. Retrieved from https://pdfs.semanticscholar.org/44ea/b3013cb6c63a534570994c9cffe3935ec7ed.pdf

Kashmanian, R. M. (2015). Building a Sustainable Supply Chain: Key Elements. *Environmental Quality Management*, *24*(3), 17–41. https://doi.org/10.1002/tqem.21393

Klimko, G. (2001). Knowledge Management and Maturity Models: Building Common Understanding. In *Proceeding of the 2nd European Conference on Knowledge Management* (pp. 269–278). https://doi.org/10.1109/PVSC.2011.6186270

Kogg, B., & Mont, O. (2012). Environmental and social responsibility in supply chains: The practise of choice and inter-organisational management. *Ecological Economics*, *83*, 154–163. https://doi.org/10.1016/J.ECOLECON.2011.08.023

Krajnc, D., & Glavič, P. (2005). How to compare companies on relevant dimensions of sustainability. *Ecological Economics*, *55*(4), 551–563. https://doi.org/10.1016/j.ecolecon.2004.12.011

Krause, D. R., Scannell, T. V, & Calantone, R. J. (2000). A Structural Analysis of the Effectiveness of Buying Firm's Strategies to Improve Supplier Improvement. *Decision Sciences*, *31*(1). Retrieved from https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-5915.2000.tb00923.x

Kwan, S. C., & Hashim, J. H. (2016). A review on co-benefits of mass public transportation in climate change mitigation. *Sustainable Cities and Society*, *22*, 11–18. https://doi.org/10.1016/J.SCS.2016.01.004

Ladd, S., & Badurdeen, F. (2010). Supplier sustainability evaluation and selection. In *IIE Annual Conference and Expo 2010 Proceedings*. Retrieved from https://search-proquest-com.dianus.libr.tue.nl/docview/734714835/fulltextPDF/5CADE084EAD64D6CPQ/1?accountid=27128

Lamming, R., & Hampson, J. (1996). The Environment as a Supply Chain Management Issue. *British Journal of Management*, *7*(s1), S45–S62. https://doi.org/10.1111/j.1467-8551.1996.tb00147.x

Lee, S. Y., & Klassen, R. D. (2008). Drivers and enablers that foster environmental management capabilities in small- and medium-sized suppliers in supply chains. *Production and Operations Management*, *17*(6), 573–586. https://doi.org/10.3401/poms.1080.0063

Lintukangas, K., Hallikas, J., & Kähkönen, A. K. (2015). The Role of Green Supply Management in the Development of Sustainable Supply Chain. *Corporate Social Responsibility and Environmental Management*, *22*(6), 321–333. https://doi.org/10.1002/csr.1348

Lippmann, S. (1999). Supply chain environmental management: Elements for success. *Corporate Environmental Strategy*, *6*(2), 175–182. https://doi.org/10.1016/S1066-7938(00)80027-5

Locke, R. M., Qin, F., & Brause, A. (2007). Does Monitoring Improve Labor Standards? Lessons From Nike. *ILR Review*, *61*(1), 3–31. https://doi.org/10.1177/001979390706100101

Locke, R., & Romis, M. (2007). *Improving working conditions in garment supply chains*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.575.2938&rep=rep1&type=pdf

López, M. V., Garcia, A., & Rodriguez, L. (2007). Sustainable development and corporate performance: A study based on the Dow Jones sustainability index. *Journal of Business Ethics*, *75*(3), 285–300. https://doi.org/10.1007/s10551-006-9253-8

Lu, L. Y. Y., Wu, C. H., & Kuo, T.-C. (2007). Environmental principles applicable to green supplier evaluation by using multi-objective decision analysis. *International Journal of Production Research*, *45*(18–19), 4317–4331. https://doi.org/10.1080/00207540701472694

Matthyssens, P., & Faes, W. (2013). Green offerings and buyer-supplier collaboration in value chains. In A. Lindgreen, F. Maon, J. Vanhamme, & S. Sen (Eds.), *Sustainable value chain management: a research anthology*. Aldershot: Gower Publishing Ltd.

Mavi, R. K., Goh, M., & Mavi, N. K. (2016). Supplier Selection with Shannon Entropy and Fuzzy TOPSIS in the Context of Supply Chain Risk Management. *Procedia - Social and Behavioral Sciences*, *235*, 216–225. https://doi.org/10.1016/j.sbspro.2016.11.017

McWilliams, A., Siegel, D. S., & Wright, P. M. (2006, January 1). Corporate social responsibility: Strategic implications. *Journal of Management Studies*. Wiley/Blackwell (10.1111). https://doi.org/10.1111/j.1467-6486.2006.00580.x

Mokhtarian, M. N., & Hadi-Vencheh, A. (2012). A new fuzzy TOPSIS method based on left and

right scores: An application for determining an industrial zone for dairy products factory. *Applied Soft Computing Journal*, *12*(8), 2496–2505. https://doi.org/10.1016/j.asoc.2012.03.042

Mueller-Hanson, R., Heggestad, E. D., & Thornton, G. C. (2003). Faking and selection: Considering the use of personality from select-in and select-out perspectives. *Journal of Applied Psychology*, *88*(2), 348–355. https://doi.org/10.1037/0021-9010.88.2.348

Nidumolu, R., Prahalad, C. K., & Rangaswami, M. R. (2009). Why sustainability is now the key driver of innovation. *Harvard Business Review*, *87*(9), 57–64. https://doi.org/10.1109/EMR.2013.6601104

O'Rourke, D. (2003). Outsourcing regulation: Non- governmental systems of labor standards and monitoring. *Policy Studies Journal*, *31*(1), 1–29. https://doi.org/10.1111/1541-0072.00001

Olson, D. L. (2004). Comparison of weights in TOPSIS models. *Mathematical and Computer Modelling*, *40*(7–8), 721–727. https://doi.org/10.1016/j.mcm.2004.10.003

Philips. (2017). *Supplier Sustainability Performance: Beyond auditing*.

Pilbeam, C., Alvarez, G., & Wilson, H. (2012). The governance of supply networks: a systematic literature review. *Supply Chain Management: An International Journal*, *17*(4), 358–376. https://doi.org/10.1108/13598541211246512

Pimenta, H. C. D., & Ball, P. D. (2015). Analysis of environmental sustainability practices across upstream supply chain management. In *Procedia CIRP* (Vol. 26, pp. 677–682). https://doi.org/10.1016/j.procir.2014.07.036

Pruett, D., Merk, J., Zeldenrust, I., & de Haan, E. (2005). Looking for a quick fix: how weak social auditing is keeping workers in sweatshops. Retrieved from http://eprints.lse.ac.uk/56743/

Rai, D., & Kumar, P. (2016). Instance based Multi Criteria Decision Model for Cloud Service Selection using TOPSIS and VIKOR. *International Journal of Computer Engineering and Technology*.

Rajagopalan, S., & Robb, R. (2005). Assessment of similarity indices to quantify segmentation accuracy of scaffold images for tissue engineering. In J. M. Fitzpatrick & J. M. Reinhardt (Eds.) (Vol. 5747, p. 1636). International Society for Optics and Photonics. https://doi.org/10.1117/12.594654

Rao, C., Goh, M., & Zheng, J. (2017). Decision Mechanism for Supplier Selection Under Sustainability. *International Journal of Information Technology & Decision Making*, *16*(01), 87–115. https://doi.org/10.1142/S0219622016500450

Rodríguez-Garavito, C. A. (2005). Global Governance and Labor Rights: Codes of Conduct and Anti-Sweatshop Struggles in Global Apparel Factories in Mexico and Guatemala. *Politics & Society*, *33*(2), 203–333. https://doi.org/10.1177/0032329205275191

Sanders, S., Cope, M., & Pulsipher, E. (2018). Do Factory Audits Improve International Labor Standards? An Examination of Voluntary Corporate Labor Regulations in Global Production Networks. *Social Sciences*, *7*(6), 84. https://doi.org/10.3390/socsci7060084

Seager, T. P., & Prado, V. (2017). Letter to the Editor on "Weighting and Aggregation in Life Cycle Assessment: Do Present Aggregated Single Scores Provide Correct Decision Support?" *Journal of Industrial Ecology*, *21*(6), 1601–1602. https://doi.org/10.1111/jiec.12559

Seuring, S., & Müller, M. (2008). [41] From a literature review to a conceptual framework for sustainable supply chain management. *Journal of Cleaner Production*, *16*(15), 1699–1710. https://doi.org/10.2308

Simpson, D. F., & Power, D. J. (2005). Use the supply relationship to develop lean and green suppliers. *Supply Chain Management: An International Journal*, *10*(1), 60–68. https://doi.org/10.1108/13598540510578388

Simpson, D., Power, D., & Samson, D. (2007). Greening the automotive supply chain: a relationship perspective. *International Journal of Operations & Production Management*, *27*(1), 28–48. https://doi.org/10.1108/01443570710714529

Singh, R. K., Murty, H. R., Gupta, S. K., & Dikshit, A. K. (2008). An overview of sustainability assessment methodologies. *Ecological Indicators*, *9*, 189–212. https://doi.org/10.1016/j.ecolind.2008.05.011

Smouter, R. (2018). *Supplier Sustainability Prediction Moving from Reactive to Proactive Assessments*. Eindhoven: Eindhoven University of Technology.

Sodhi, M. S., & Tang, C. S. (2009). *Sustainable supply chains*. *Computer Aided Chemical Engineering* (Vol. 27).

Soylu, A., Oruç, C., Turkay, M., Fujita, K., & Asakura, T. (2006). Synergy analysis of collaborative supply chain management in energy systems using multi-period MILP. *European Journal of Operational Research*, *174*(1), 387–403. https://doi.org/10.1016/J.EJOR.2005.02.042

Subic, A., Shabani, B., Hedayati, M., & Crossin, E. (2013). Performance analysis of the capability assessment tool for sustainable manufacturing. *Sustainability (Switzerland)*, *5*(8), 3543–3561. https://doi.org/10.3390/su5083543

Tachizawa, E. M., & Wong, C. Y. (2015). The Performance of Green Supply Chain Management Governance Mechanisms: A Supply Network and Complexity Perspective. *Journal of Supply Chain Management*, *51*(3), 18–32. https://doi.org/10.1111/jscm.12072

Terwee, C. B., Bot, S. D. M., de Boer, M. R., van der Windt, D. A. W. M., Knol, D. L., Dekker, J., … de Vet, H. C. W. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*, *60*(1), 34–42. https://doi.org/10.1016/j.jclinepi.2006.03.012

Teuscher, P., Grüninger, B., & Ferdinand, N. (2006). Risk management in sustainable supply chain management (SSCM): lessons learnt from the case of GMO-free soybeans. *Corporate Social Responsibility and Environmental Management*, *13*(1), 1–10. https://doi.org/10.1002/csr.81

Trapp, A. C., & Sarkis, J. (2016). Identifying Robust portfolios of suppliers: A sustainability selection and development perspective. *Journal of Cleaner Production*, *112*, 1–45. https://doi.org/10.1016/j.jclepro.2014.09.062

Tzeng, G.-H., & Huang, J.-J. (2011). *Multi Attribute Decision Making*. Taylor & Francis Group.

United Nations Global Compact (UNGC). (2018). The Ten Principles | UN Global Compact. Retrieved August 1, 2018, from https://www.unglobalcompact.org/what-is-gc/mission/principles

Vachon, S., & Klassen, R. D. (2006). Extending green practices across the supply chain. *International Journal of Operations & Production Management*, *26*(7), 795–821. https://doi.org/10.1108/01443570610672248

Van de Kerk, G., & Manuel, A. R. (2008). A comprehensive index for a sustainable society: The SSI - the Sustainable Society Index. *Ecological Economics*, *66*(2–3), 228–242. https://doi.org/10.1016/j.ecolecon.2008.01.029

Van Lakerveld, A., & Van Tulder, R. (2016). *Managing the Transition to Sustainable Supply Chain Management Practices Evidence from Dutch leader firms in the Philippines*. Retrieved from http://emit-project.net/wp-content/uploads/2016/06/1Tulder_Lackerveld_2016rsm10.pdf

Van Tulder, R., Van Tilburg, R., Francken, M., & Da Rosa, A. (2013). *Managing the transition to a sustainable enterprise: Lessons from frontrunner companies*.

Van Weele, A. J., & Vivanco, L. (2014). Corporate social responsibility: moving from compliance to value creation in value chain relationships. In *The value chain shift: seven future challenges facing top executives* (pp. 123–137). IMD Value Chain 2020 Project. Retrieved from https://research.tue.nl/en/publications/corporate-social-responsibility-moving-from-compliance-to-value-c

Van Weele, A., & Van Tubergen, K. (2017). Responsible Purchasing: Moving from Compliance to Value Creation in Supplier Relationships. In *Sustainable Supply Chains A Research-Based Textbook on Operations and Strategy* (pp. 257–278).

Varma, S., Wadhwa, S., & Deshmukh, S. G. (2006). Implementing supply chain management in a firm: issues and remedies. *Asia Pacific Journal of Marketing and Logistics*, *18*(3), 223–243. https://doi.org/10.1108/13555850610675670

Wilhelm, M. M., Blome, C., Bhakoo, V., & Paulraj, A. (2016). Sustainability in multi-tier supply chains: Understanding the double agency role of the first-tier supplier. *Journal of Operations Management*, *41*, 42–60. https://doi.org/10.1016/j.jom.2015.11.001

World Commission on Environment and Development (WCED). (1987). Our common future. *Oxford and New York: Oxford University Press*. Retrieved from https://scholar.google.com/scholar_lookup?hl=en&publication_year=1987&author=World+Commission+on+Environment+and+Development+%28WCED%29&title=+Our+Common+Future+

Wu, K.-J., Liao, C.-J., Tseng, M., & Chiu, K. K.-S. (2016). Multi-attribute approach to sustainable supply chain management under uncertainty. *Industrial Management & Data Systems*, *116*(4), 777–800. https://doi.org/10.1108/IMDS-08-2015-0327

Wu, K.-J., Zhu, Y., Tseng, M.-L., Lim, M. K., & Xue, B. (2018). Developing a hierarchical structure of the co-benefits of the triple bottom line under uncertainty. *Journal of Cleaner*

Production, *195*, 908–918. https://doi.org/10.1016/j.jclepro.2018.05.264

Yamin, S. C., Parker, D. L., Xi, M., & Stanley, R. (2017). Self-audit of lockout/tagout in manufacturing workplaces: A pilot study. *American Journal of Industrial Medicine*, *60*(5), 504–509. https://doi.org/10.1002/ajim.22715

Zadek, S. (2004). The Path to Corporate Responsibility: Best Practice. *Harvard Business Review*, *Vol. 82 Is*, p.125-132. https://doi.org/10.1007/978-3-540-70818-6_13

Zhang, B., & Srihari, S. N. (2003). Properties of binary vector dissimilarity measures. In *Proc. JCIS Int'l Conf. Computer Vision, Pattern Recognition, and Image Processing* (pp. 1–4). https://doi.org/10.1117/12.473347

Zhou, X., & Xu, Z. (2018). An Integrated Sustainable Supplier Selection Approach Based on Hybrid Information Aggregation. *Sustainability*, *10*(7), 2543. https://doi.org/10.3390/su10072543

Zimmerli, W. C., Holzinger, M., & Richter, K. (2007). *Corporate Ethics and Corporate Governance*. *Springer*. https://doi.org/10.1007/978-3-540-70818-6

A. Data transformations

Since the objective of the questionnaire validation is to verify whether all documents that are stated to be available are actually available when checking the ED, the only SAQ questions in scope are those that ask for documents, policies, or procedures (that are stated in the ED as well). The question types date, free text, number, RB (also text), and SC (also text) do not state any confirmation of document availability, so these SAQ questions are omitted. The question types multiple choice and yes/no do potentially possess this information. In the case of the yes/no type of questions, the transformation is easy: the answers "Yes" are transformed to 1 and the answers "No" are transformed to 0. In the case of the multiple choice questions, which are stored as dichotomous questions (each possible multiple choice answer is stored as separate question, so that there is a clear distinction between answers given), are transformed to 1 if the dichotomous answer contains text, and transformed to 0 if the dichotomous answer contains "-".

SAQ question type	In or out of scope	Transformation approach
Date	Out of scope	
Free text	Out of scope	
Multiple choice	In scope, transform to binary	1, if answer is "Yes"
		0, if answer is "No"
Number	Out of scope	
RB	Out of scope	
SC	Out of scope	
Yes/No	In scope, transform to binary	1, if answer contains text
		0, if answer is "-"

Table 17: Transformation approaches per SAQ question type

Since the SAQ is constructed as a waterfall structure and some suppliers do not fill in all questions, some missing values exist within the remaining questions in scope. This can thus be the case when depending on previous questions (e.g. when the multiple choice answer "other" is not chosen and the follow-up question, an explanation in text, is thus not answered), or when suppliers choose or forget to not answer questions. Since these missing values are not missing at random (NMAR), and the logic behind those missing values is thus known, these missing values can easily be imputed. In those cases, it is chosen to observe these (missing) answers as "No" and transform it to 0. This is also in line with the concept of the questionnaire validation, which is elaborated on in the following chapter 6.1.

On the other hand, valuable information (the difference between an actual "No" and missing value) might lie in these answers when investigating the way of filling in the SAQ. Since this research rather focuses on those questions in which the suppliers answered "Yes" in the SAQ, the imputation of missing values as "No" is not expected to play a significant role.

Besides that, to validate the binary SAQ answers to the ED, the availability per ED should be transformed too. The availability is answered with: yes, no, N/A, or PZT. Before transforming these answers to binary, the number of PZTs per assessment workbook is calculated and saved for further analysis. After that, the answer "Yes" is transformed to 1, the answers "No" and "N/A" are transformed to 0, and the answer "PZT" remains the same for further analysis. The answer "PZT" is given when the supplier does not comply with the code of conduct and/or violates the Zero Tolerances stated by Philips, which does not imply whether the document is or is not available. Further analysis to this matter is explained in the following chapter 7.1, since this decision should be in line with the concept of the questionnaire validation.

ED availability	Transformation approach
Yes	1
No	0
N/A	0
PZT	PZT

Table 18: Transformation approaches per ED availability answer

Next to that, as earlier explained, the spend data is currently known per part number, not per supplier (ID). To aggregate this spend per supplier, multiple step are taken. At first, all spend per month is aggregated per part number to determine spend per year. Secondly, each part number is connected to its (supplier) GU ID, so that spend per GU ID is known. Since this GU ID is different than the supplier ID, each GU ID is connected to its supplier ID, so that spend per supplier ID is known. The only problem with these connections is that suppliers with multiple sites (and so with multiple supplier IDs) only have an aggregated spend on company-level, not on site-level. The SMRS database does not give any insights in these spends on company-level, so 81 suppliers (of which 24 supplier are in scope since 2018) remain with unknown spend. Since these missing values are not missing at random (NMAR), since the logic behind those missing values is thus known. This requires advanced methods, since spend per supplier is known, but the logic behind it remains unclear. Due to the importance of the accuracy of this data (for decision making in SSIP/DIY), it is chosen to continue without these cases and only rely on accurate data. In this way, the current situation can actually be compared to the models, instead of relying on possibly shifted results.

B. Explanation analysis tool

The analysis tool is mainly created to give further insights in supplier sustainability improvements rather than just comparing one average per sequence. The tool is currently set up to give further insights in the final and topic scores, but this can potentially also be the maturity elements or even on question level. Besides that, the user can set its own parameters, regarding the ratio between SAQ and ED (i.e. 0 – 100% as chosen by Philips), and inclusion/exclusion of certain topics (or maturity elements, or questions). Based on the parameters set, the weighted averages are calculated accordingly, which results in insights in supplier sustainability improvements. Although both absolute and relative improvement are stated, it should be noted that all cases conforming the assumptions below are taken into account. This means that extremely low scoring suppliers might increase the average improvement due to their low baseline score, i.e. suppliers that score 5% in sequence 1 and 20% in sequence 2.

These improvements are not only stated for the whole group of suppliers, but also split per supplier evaluation model classification (i.e. DIY and SSIP as by Philips). In this manner the effectiveness of the supplier evaluation model and accessory characteristics can easily be seen and analysed. The most interesting finding of this analysis tool is that SSIP supplier (which are thus visited by Philips on-site) improve more than the DIY suppliers (which are expected to be mature enough and to improve sufficient). These SSIP suppliers score higher in both relative and absolute improvements. Next to that, it can be seen that the improvement from sequence 1 to 2 (thus new suppliers in the program) is larger than the improvement from sequence 2 to 3.

This can be supported by the following example:

47 suppliers started (the first sequence) in 2016 and had the second sequence the year after, and 37 of these suppliers had the third sequence the year after. These 47 suppliers had an average absolute improvement of 8% and relative improvement of 22%. The 21 DIY suppliers (without site assessment) had an average absolute improvement of 2% and relative improvement of 5%. The 26 SSIP suppliers (with site assessment) had an average absolute improvement of 13% and relative improvement of 36%.

The following assumptions are made to support this analysis tool:

- Average improvements are calculated based on same supplier bases (to prevent inaccurate comparisons),
- Suppliers with scores that are negative or zero (due to PZTs or no applicable questions) are left out of scope if needed,
- Use the exact scores instead of rounding up to integers in between (per dashboard field),
- Using data until 31-12-2018.

C. Maximum number of resources

Within the research of revising the supplier evaluation model, and its decision making process behind it, one valuable characteristic lays in the use of the maximum number of resources to use for the focal firm. One of the reasons why focal firms cannot heavily collaborate with all its suppliers is due to impracticality in terms of location, time, and financial resources. When suppliers conducted their self-assessment and the focal firm has to select a subset of suppliers to collaborate with and visit on site, there is thus a maximum number of (time and/or financial) resources available. Within this research it is assumed that the maximum number of resources is fixed, and every supplier takes the same amount of time per on site assessment. This is in line with the current structure of Philips' site assessments, which are assumed to all have the same time spent per site assessment. On the other hand, the real time spent on site can differ per supplier, due to site size, process complexity, risk level, sustainability maturity level, travelling time, and so on. Since there are no set of rules currently used and most of the characteristics are not known, the assumption of equal time spent per supplier is set.

It seems logical that there is a huge potential in relaxing this assumption and differing suppliers from each other. Low scoring, low matured, and risky suppliers might need and benefit most from more time spent than the high matured and risk-free suppliers, although it is expected that this latter group is not probable to be chosen for collaborations. When only focussing on the group of selected suppliers for site assessment (thus the SSIP suppliers), there might be a difference between the 'worst' and 'best' supplier in this group. For example, if the 10 best SSIP suppliers are expected to need less collaboration and thus need less days on site (e.g. one day less), there are more resources available again for collaborations. In this case, it can be chosen to use these available resources to spend more time on site at the worst SSIP suppliers, or to add more suppliers to the SSIP group.

Another promising direction lays in the efficiency of planning the site visits in terms of location. Since travelling time is part of the number of days spend per supplier, it might be the case that suppliers with a large amount of travelling time receive less real time spend on site. When suppliers with site visits can be combined in terms of travelling (so that the assessor only has to fly once instead of multiple times), more resources are available again. In practical terms, this might be a difficult option since suppliers have their own dedicated assessor (although this can be fixed by rearranging suppliers over the assessors) and sites close to each other might have a different timing in terms of when their assessment yearly starts (e.g. one supplier is yearly assessed around March, whilst the other around September). This is all dependent on the program's choices in the past, but rearranging suppliers in terms of assessors and timing is potentially a promising solution in making more time and financial resources available for collaborations.

D. Expected correlations

If all suppliers fill in their SAQ and ED correctly and fully accurate, this should be visible when investigating the connected questions (as explained in chapter 6.1). Although some documents might be impractical to deliver online (such as large documents of hundreds of pages), the expected correlations of these connected questions should be near 1. A correlation of 1 implies that there is a perfect positive relationship in place, since one variable moving would mean that the other variable moves in the same direction. For these connected questions, which asks for the exact same document (e.g. policy, procedure, and so on), the correlation is thus expected to be 1. When the supplier states in the SAQ that they possess a certain document, they should have this document, and should thus be able to provide this document for the ED. It is therefore expected that all connected questions have (expected) correlations of 1.

Unfortunately, this is not the case, since not a single connected question achieves this correlation of 1. The highest correlation found corresponds to 0.88, whilst the average correlation found is 0.32. Although it is known that Pearson's r is not perfect for stating correlations between two dichotomous variables, the correlations should still be (close to) 1 when suppliers would fill in their SAQ and ED perfectly. In case of the lowest correlation found (of 0.10), the connected questions were filling in correctly in almost 80% of the cases, whilst 20% of the cases showed a difference between SAQ and ED.

Since there is a clear difference between the inconsistencies (false positives and false negatives), it is chosen to not use the correlations for further use, but investigate the contingency table per connected question. Only in this way there can be differed between the false positives (thus suppliers claiming they have a document, which they do not provide afterwards) and the false negatives (suppliers claiming they do not have a document, whilst they do provide it afterwards). Further assumptions and methods regarding these contingency tables are explained in chapter 6.1.

E. HIC-score as predictor

Although the focus of this research is not in prediction models, an initial investigation is done to the HIC-score as predictor. It is chosen to only investigate the potential use of the HIC-score as predictor in linear models, so that a positive outcome can be further investigated with mathematical models or machine learning in a later stage of the broader research project.

Linear model 1:

$$ED\ score\ (before\ site\ visit) = -0.57 + 0.75 * SAQ\ score + 0.72 * HIC - score\ (before\ site\ visit)$$

Linear model 2:

$$ED\ score\ (after\ site\ visit) = 0.25 + 0.17 * SAQ\ score + 0.79 * ED\ score\ (before\ site\ visit) - 0.30 * HIC - score\ (before\ site\ visit)$$

In two linear models (1 and 2) the HIC-score was seen as significant variable, which are stated above. In both cases the ED score (respectively, (1) before and (2) after the site assessment) is the dependent variable, whilst the HIC-score is the independent variable. The linear regression models result in a R-squared of, respectively, 0.66 and 0.67, whilst their MAE is 0.08 and 0.07, and their RMSE is 0.11 and 0.10. This does not seem so bad with linear regression only, so this might give some interesting insights when using it with other (mathematical) prediction models.