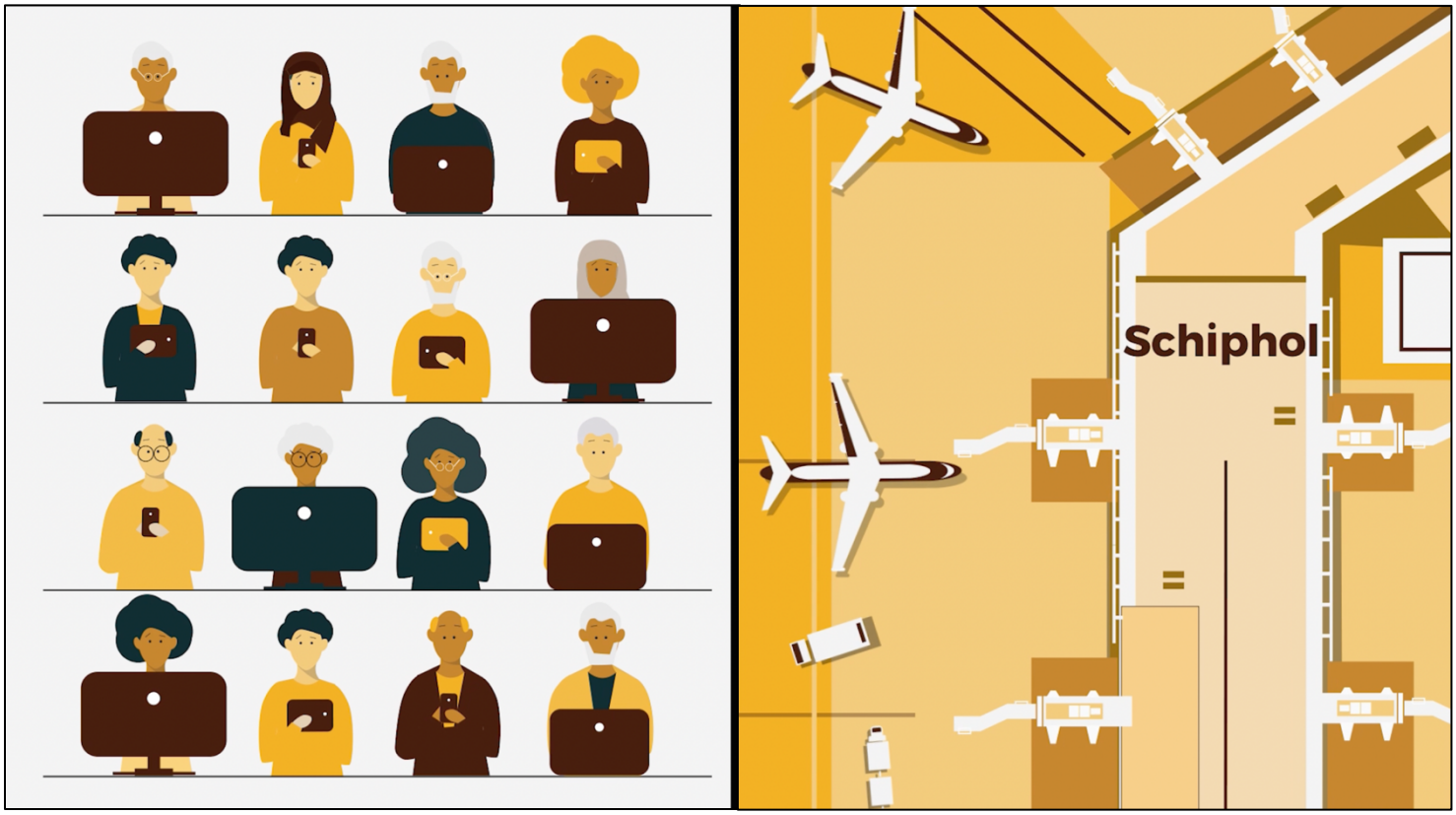
The face validity of the Participatory Value Evaluation method

Assessing the face validity of the Participatory Value Evaluation method by applying an established framework in the case study of the Schiphol Environmental Council



Charlotte Tuit
Master thesis
Complex Systems Engineering and Management

TU Delft

Frontpage image designed by Marcel Janssen (www.ziehoe.nl)

The face validity of the Participatory Value Evaluation method

Master thesis submitted to Delft University of Technology

in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in **Complex Systems Engineering and Management**

Faculty of Technology, Policy and Management

by

Charlotte Tuit

Student number: 4697073

To be defended in public on September 23, 2022

Graduation committee

Chairperson	: Dr. U. Pesch, Section Ethics/Philosophy of Technology
First Supervisor	: Dr. mr. N. Mouter, Section Transport and Logistics
Advisor	: A.M. de Ruijter, Section Transport and Logistics

Preface

This report presents the results of my master thesis. This master thesis is the final step to complete the master's programme Complex Systems Engineering and Management at Delft University of Technology.

The past half year have been a process of which this thesis is the end result. I look back on a very instructive period. Not only because the master thesis is my first research experience, but also because I was allowed to combine this experience with an internship which gave me experience in practice as well. This was a very valuable combination to me.

However, achieving this end product would not have been possible without a number of people whom I would like to thank. First of all, I would like to thank my graduation committee. Udo, thank you for all the feedback on my thesis. Annamarie, thanks for sparring with me about shaping the subject and all the analyses that came with it. I would like to give a special thanks to Niek for all the inspiration about my thesis, for involving me in the Schiphol Environmental Council project and for the opportunities at Populytics that you have offered me.

Second, I would like to thank all the interviewees for the insight and wisdom you have given me. Furthermore, I would like to thank all my colleagues at Populytics. The Populytics team is young, hardworking and ambitious. I have learned a lot from you.

Finally, I would also like to thank Arjan and my family for reviewing my report and for giving me support and motivation.

Charlotte Tuit
Delft, August 22nd 2022

Summary

Public participation, the process that revolves around involving citizens in public decision-making, has become an integral part of Dutch society (Dang, 2020). The aim of public participation is to create policy that is more effective, due to the fact that preferences of residents are taken into account during decision making. There are many ways in which a public participation approach can be designed. A methodology that can be stated as a digital participatory approach is the Participatory Value Evaluation (PVE). The PVE is an online method that enables participation for a large group of participants. During the completion of a PVE consultation, the respondents are given the opportunity to select policy options. Due to a maximum budget, which is a constraint that the respondents are allowed to divide, they are faced with trade-offs. This means that the PVE can also serve the goal as an evaluation method. In this research, the focus is on the measurement of preferences, or the PVE as an evaluation method. This perspective is chosen because the participation process is often broader than the PVE consultation itself. Therefore, measuring preferences is more feasible.

Since the PVE is a relatively new method, researching the validity of the PVE method is a path that not many researchers have yet explored. However, in order for a research method to eventually be wide applied in studies, it is considered necessary to guarantee validity and reliability. Field (2005) defines validity as: “measure what is intendent to be measured.” For the PVE method this means that the respondents are enabled to experience the entire complex decision-making situation of a policymaker and to issue an advice about this complex situation. In this research, the case study of the Schiphol Environmental Council (ORS) is used as the complex decision-making situation. The goal of this case study is to research by means of a PVE consultations how to improve the participation and information facilities for local residents, which are the respondents, around Schiphol Airport. There are many involved parties with different interests in this case study, which are the stakeholders. In this research, stakeholders are representatives of the Noord-Holland Environmental Federation, Schiphol Group, Air Traffic Control the Netherlands, residents’ representatives per runway, employer’s organization VNO-NCW West and policy officers of the municipalities Haarlemmermeer and Ouder-Amstel. These stakeholders are all part or have been part of the ORS.

The concept of validity as defined above is an umbrella term. Different types of validity can be identified, one of which is face validity. A definition of face validity can be found in the article by Taherdoost (2016): “Face validity is the degree to which a measure appears to be related to a specific construct.” Regarding the PVE method, face validity concerns that the choice options and their information appear genuine to respondents. Anastasi and Urbina (2007) state that if the content of an instrument, like a PVE consultation, lacks face validity, there is a likelihood that the results obtained from this instrument provide false information and lead to decisions that are misleading. Certainly with regard to the PVE method, face validity is important. It is possible that policymakers base policy on the results of a PVE consultation. Moreover, the information used to make decisions is increasingly complex and difficult to comprehend in today’s world. Precisely for a PVE method that put citizens in the shoes of policymakers with regard to a complex situation, it is important that the citizen is able to fully experience the policymaker’s problem. A lack of face validity may also cause commotion and resistance from stakeholders, such as with the Amsterdam-Wind consultation (Populytics, 2021). All together this leads to the following main research question: *How to measure face*

validity regarding the PVE method and how do respondents and stakeholders evaluate the face validity of a PVE consultation?

To answer the main research question, it is required to go through a number of steps that are embedded in a mixed methods approach. A total of four steps are included in this research, each of which is linked to a sub research question. An overview of the applied methods per step of the research is presented in table 0.1.

Table 0.1: Overview of applied methods per step of the research

	Literature review	Expert interviews	PVE method	Descriptive statistics	Mann-Whitney U test	One-way MANOVA	Factor analysis	Multiple regression analysis	Multinomial logistic regression	Document analysis	Field research	In-depth interviews
Step 1: Identifying a framework to assess face validity in the PVE method	X	X										
Step 2: Analyzing the results of the face validity framework in a case study regarding two types of a choice task			X	X	X	X	X	X	X			
Step 3: Comparing the evaluation of face validity with previous PVE consultations on the basis of properties and benchmarks					X							
Step 4: Comparing the framework from literature with concerns about face validity in practice										X	X	X

The first step of this research consists of identifying a framework to assess face validity in the PVE method. The sub question related to this step is defined as follows: *how to design a framework that is able to measure face validity regarding the PVE method?* To answer this question a literature review is performed. This review gave insights in the importance of face validity, guidelines to set up a face validity assessment and categories of face validity. In total, twenty-five articles were selected which included ten categories of face validity. After those twenty-five a saturation point was reached. The literature review is complemented with expert interviews. Conducting the expert interviews has two aims. The first aim is to select the five most important categories of face validity, since the consultation of the ORS in which the framework is applied should not become too long. The second aim is to select the five categories that are most important with regard to the PVE method. For this purpose, both

validity experts and PVE experts were interviewed. The expert interviews are conducted in a structured manner, because the objective of the expert interviews is concrete. For each selected category a statement was set up which together forms the framework.

From the literature review it became clear that when setting up a face validity assessment it must first be determined if: the raters of face validity are experts or laypeople, if the assessment is a questionnaire, a 'think aloud' interview or a focus group, in which stage of the development the assessment will take place and whether it is an absolute or a relative assessment. Thereby, if the content of a PVE consultation is not face valid, there is a likelihood that the results obtained from this instrument provide false information and decisions of respondents that are misleading for policymakers. On the other hand, misleading information in the consultation can also cause respondents to make choices that they would not make otherwise. Furthermore, when evaluating the face validity of the PVE as an evaluation method, these are the five most recommended categories regarding the outcomes of the literature review and the expert interviews: clarity, unambiguity, relevance, readability and completeness. Based on these categories, specific statements have been drawn up that form the framework as presented in table 0.2.

Table 0.2: Overview of five face validity categories and statements that form the face validity framework

Category	Statement
Clarity	I have received sufficient information to make a choice about the possible tasks of the Schiphol Social Council
Unambiguity	I found it was clear what was meant by each task with regard to the possible tasks of the Schiphol Social Council
Relevance	I think this research is a good way to give my opinion about the Schiphol Social Council and the Environmental House
Readability	I found the questions asked to me in this study understandable
Completeness	I felt I could give all my opinions on how citizens should be involved in decision-making about Schiphol and how information should be provided

In the second step of this research the face validity framework is applied in the case study of the ORS. With this application, a face validity experiment is carried out in which the two existing types of choice tasks are applied. A choice task is an overview in which respondents are asked to choose between policy options while they may not exceed certain constraints. The sub question related to this step is defined as follows: *what are the similarities and differences of two different types of a choice task regarding a PVE consultation on the evaluation of face validity?* The aim of this experiment is to investigate in a PVE consultation whether two different choice tasks influence the assessment of face validity of the PVE method. Because face validity concerns how valid or genuine a consultation appears and the design of a consultation affects the appearance, this study experiments with the design of a PVE consultation.

The two types are the 'sliders' choice task and the 'points' choice task. The design of both types is shown in the figures 0.1 and 0.2. The difference between the 'sliders' and the 'points' choice task is that the effort of each option is also shown in the 'sliders' choice task. Compared to the 'points' choice task, the 'sliders' choice task can give the respondents more difficulty to choose. It can be confronting for the respondents to mention that not every option can be executed to the maximum. For this experiment, each of the choice tasks is filled in by half of the total sample. Further information is exactly the same in both cases. Ultimately, the 'sliders'

experiment is completed by 648 respondents. The ‘points’ experiment is completed by 582 respondents. It is decided to only include completes in order to minimize the chance of bias. Another choice made to reduce the chance of bias, is to let the respondents answer the face validity statements on an odd five-point Likert scale. Furthermore, the respondents were required to be eighteen years or older and live in one of the 54 municipalities surrounding Schiphol.

Figure 0.1: ‘Sliders’ choice task

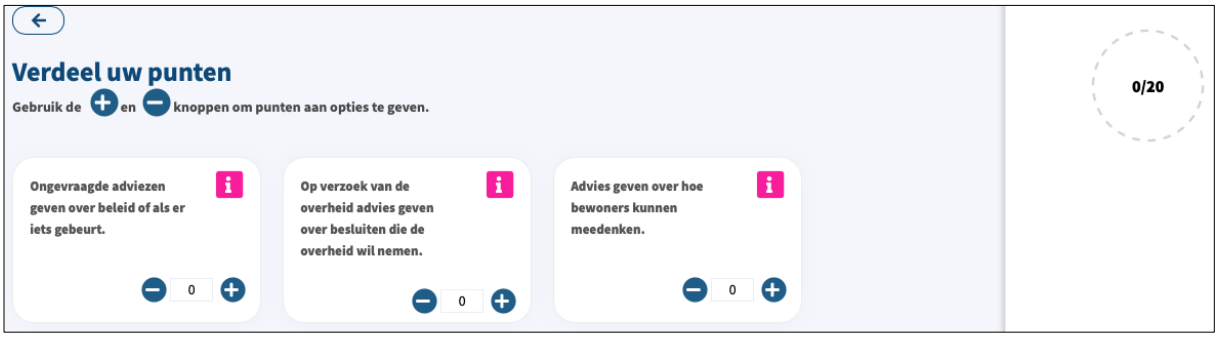


Figure 0.2: ‘Points’ choice task

Several statistical tests are performed to investigate the differences and similarities between the two types of choice tasks. First, descriptive statistics provided a global understanding into how each face validity statement of the established framework scores. These statistics show that clarity and unambiguity are rated the lowest of the five categories in the framework across the total sample. Second, factor analyses are performed to analyze whether the statements of face validity measure a common variable. The factor analyses show that there are three common variables per experiment which consists of the face validity of the specific choice task, the face validity of the PVE consultation in general and the face validity of the PVE method. Moreover, the Mann-Whitney U test examines whether there is a significant difference between the two experiments of the PVE consultation based on a single statement regarding face validity. However, this test does not provide multivariate results using information among multiple dependent variables which are the multiple face validity statements in this study. Therefore, one-way MANOVA tests are also performed. It appears from these tests that there is no significant difference in the evaluation of face validity between the two types of choice tasks. Furthermore, multiple regression analyses are performed with the aim to research whether and which characteristics of respondents have influence on the evaluation of the face validity common variables (as identified with the factor analyses). Thereby, multinomial regression analyses are performed to examine whether and which characteristics of respondents influence the evaluation of the five face validity

categories separately per experiment. From both types of regression analyses, it follows that more demographic characteristics have influence on the evaluation of face validity in the 'points' choice task than in the 'sliders' choice task. Finally, latent class cluster analyses (LCCAs) are performed. While a regression analysis focuses on which characteristics explain the assessment of a face validity category, the LCCA focuses on whether certain groups of people can be identified who share the same characteristics and who collectively score high or low on certain categories of face validity. From the results of the latent class cluster analyses, it is remarkable that in experiment one the majority of the respondents (55,64%) rated the face validity highly. In experiment two, the clusters are more equally distributed.

In addition to the Schiphol Environmental Council case, there are also previous PVE consultations that have included elements of the face validity framework established in the first step of this research. The different case studies have different properties. Furthermore, the various case studies make it possible to set benchmarks for a face validity category that indicates when face validity is evaluated relatively high or low. Therefore, the sub question of the third research step is defined as follows: *which properties influence the differences in the evaluation of face validity between different case studies and what are the benchmarks of those differences?*

In previous PVE consultations, three categories of face validity are identified that also appear in the framework drawn up in this research. The first category is clarity. This category emerges in the climate consultation (Mouter et al., 2021a), in the heat transition vision in Utrecht (Mouter et al., 2020) and in the case of the long-term Corona policy (Geijsen et al., 2022). In the cases of heat transition vision of Utrecht and the long-term Corona policy, the relevance category emerges as well. Relevance is also included in another case study that concerns sustainable energy in the Foodvalley region (Spruit & Mouter, 2021). The third category is completeness which emerges in the case of the Foodvalley region as well. The above case studies each have their own properties. In order to obtain an overview of these characteristics, a typology has been drawn up. The case studies are compared with each other on the basis of the corresponding face validity categories. The method applied to compare the evaluation of a face validity category between two case studies is the Mann-Whitney U test. Here, 'sliders' PVE consultations are compared with 'sliders' consultations and 'points' PVE consultations with 'points' PVE consultations to reduce bias.

It results from the third step that the four properties impact on personal life, the respondents, the platform and who is in charge influence the differences in the evaluation of face validity between different case studies. Furthermore, for the case studies in which clarity is included, it appears that the benchmark regarding the assessment of clarity lies between the average scores 3,53 and 3,84 on a five-point Likert scale. On this scale, 1 stands for totally disagree and 5 stands for totally agree. For the case studies in which relevancy is included, it appears that the benchmark regarding the assessment of relevancy lies between the average scores 3,53 and 3,95. For the case studies in which completeness is included, the benchmark regarding the evaluation of completeness lies between the average scores 3,89 and 4,11.

The first three steps in this research are based on the perspective of the literature. Therefore, the fourth and final step of this research takes a different perspective, namely the perspective of practice. The aim of this step is to identify which concerns stakeholders and citizens have

in practice when it comes to face validity and whether these concerns are sufficiently covered in the established face validity framework. The corresponding sub question of this fourth research step is as follows: *to what extent do the concerns of citizens and stakeholders with regard to face validity correspond in practice with the established framework?*

In order to study the concerns in practice about face validity, four perspectives have been divided to create a complete overview of concerns. The first perspective concerns the local residents and respondents. The statistical analyses performed in this research concern the panel consultation. In addition to this panel, an open consultation is conducted. Every person aged eighteen or older is allowed to participate in the open consultation. In this consultation the respondents are asked if they could explain why they assessed the face validity categories with a certain answer on the Likert scale. These open answers showed that respondents also cited other face validity categories that had not been included in the consultation. That is why the answers on these questions are analysed to study the concerns about face validity in practice based on a document analysis. Following Bowen (2009), a document analysis is "a form of research in which documents are interpreted by the researcher to give voice and meaning around an assessment topic". The second perspective consists of stakeholders involved in the process of designing the PVE consultation about Schiphol. After analysing the data from the consultation, a draft report is drawn up. The stakeholders involved in the design of the consultation were given the opportunity to provide feedback on the draft report before it became public. Their feedback is analysed by means of a document analysis. Subsequently, field research is performed. Field research is a method in the qualitative domain in which people are observed in their natural settings (Burgess, 2002). After the final report has been completed, the results are presented on the 'Regioforum'. The meetings of the 'Regioforum' are open to members, deputies, supporters and electors of the residents' organizations registered with the Schiphol Environmental Council. During this 'Regioforum', it is analysed whether the attendees make comments that fall under face validity. The third perspective consists of a stakeholder who is not involved in the process of designing the PVE consultation. The insights of this perspective are collected by means of an in-depth interview with an employee of the mainport strategy department of the airline KLM. The last perspective is the perspective of the client that is also collected with in-depth interviews. The client is ministry of Infrastructure and Water Management combined with the ORS. Therefore, two policy officers are interviewed who work at the ministry and a project manager of the ORS are interviewed.

It results from the fourth step that the most concerns are about clarity, completeness, unambiguity and the feasibility across the local residents and respondents, stakeholders involved and not involved in the process of designing the PVE consultation and the client. Feasibility is the only one not included in the framework designed in step one of this research.

Concluding, it is recommended to focus in the following PVE consultations in particular on clarity and unambiguity if the goal is to increase face validity. Furthermore, if the aim is to measure the assessment of face validity, more than one category of face validity should be questioned. Next, it can be concluded that there is no significant difference in the evaluation of face validity between the two experiments. However, respondents are more indifferent to a 'sliders' choice task than to a 'points' choice task. The opinions with regard to face validity in the experiment with the 'points' choice task differ more widely. Moreover, if the goal is to

score high on the clarity category, the aim should be to achieve an average evaluation score above 3,67 for clarity, above 3,88 for relevancy and around or higher than 4,11 for completeness on a five-point Likert scale. Be aware that if the impact on personal life is in the short term, this will result in a lower average evaluation score regarding clarity than impact on the long term. Finally, when evaluating the face validity of the PVE as an evaluation method, these are the six most recommended categories from the perspectives of literature and practice: clarity, unambiguity, relevancy, readability, completeness and feasibility.

This research contributes to exploring the face validity of the PVE method. A framework specifically aimed at the PVE method has been drawn up to measure face validity and this has been supplemented with concerns from practice. Furthermore, the current status of face validity has been investigated on the basis of previous case studies. Benchmarks have been drawn up from these previous case studies. However, there are only a few case studies in which categories of face validity have been included. This results in uncertainty of the benchmarks. To reduce this uncertainty, it is recommended to add face validity categories in future consultations. Moreover, it is concluded that respondents are more indifferent to a 'sliders' choice task than a 'points' choice task. A hypothesis for this conclusion is that respondents are more familiar with a 'points' PVE as it resembles a survey and therefore have more extreme opinions. Respondents are less familiar with a 'sliders' PVE. This hypothesis gives rise to further research. Another hypothesis that follows from the results of this research is that people who live close to a problem situation, such as at Schiphol, will rate the face validity lower. A recommendation is to further investigate this proximity effect and to test this hypothesis. Finally, it is remarkable from the results that 18- to 34-year-olds and people with a low level of education generally rate the face validity categories lower. Therefore, another suggestion for further research is to investigate how the face validity of these groups can possibly be increased and what their needs are regarding face validity.

Contents

1. Introduction

Public participation, the process that revolves around involving citizens in public decision-making, has become an integral part of Dutch society (Dang, 2020). Public participation ensures that policymaking is no longer a top-down process, but is a bottom-up process as well. This attention to public participation is mainly based on the ideology that citizens should be able to gain more influence on decisions that affect them (Burton, 2009). The main aim of public participation is to create policy that is more effective, due to the fact that preferences of residents are taken into account during decision making. That public participation has become an integral part of society is partly due to the dissatisfaction with the traditional democratic representation and its mechanisms (Ianniello et al., 2019). However, the trust of citizens in national government organization in the Netherlands has fallen from seven out of ten people in April 2020 to three out of ten people in September 2021 (Engbersen et al., 2021). The vision that arises, states that it cannot be assumed that the government has enough skills or knowledge to come to an optimal policy decision. Therefore, the government is dependent on other parties that do have these skills and knowledge. Examples of these other parties are companies, citizens and interest groups. Furthermore, the bottom-up public participation approach has three main advantages over the top-down approach of policymaking. First, the bottom-up public participation approach closes close the gap between the skills and knowledge that the government does not possess but other parties do. This makes policymaking more efficient (Edelenbos et al., 2006). Second, it gives citizens the opportunity to voice their opinion about specific policy problems instead of solely during elections. Third, the use of participatory approaches in itself leads to more legitimate policy decisions (Buijs & Boonstra, 2020).

There are many ways in which a public participation approach can be designed. In the midst of digitalization, digital participatory approaches appear to contribute to closing the gap. Digital approaches make it easier to inform citizens, to empower citizens, to reduce costs of the policymaking process and to communicate with citizens (Zheng & Schachter, 2016). A methodology that can be stated as a digital participatory approach is the Participatory Value Evaluation (PVE).

This section discusses what this PVE method entails. Thereafter, the PVE method is described from the perspective of an evaluation method and from the perspective of a participation method. After describing the PVE method, a link is made with the validity of this method with a focus on face validity. This link leads to the knowledge gap and the main research question. Subsequently, the scientific and societal relevance of this research are discussed. This is followed by the research approach and the sub questions. Finally, the reading guide is described.

1.1 Participatory Value Evaluation

The process of setting up a PVE can be divided into five phases which together form a PVE-process as identified by Bouwmeester (2021). The first phase consists of defining the policy problem with the facilitator and/or the policy maker. A policy problem is a collective challenge or problem. The policy maker has to deal with this policy problem. The facilitator can be defined as the organization of person that facilitates this process of participation (Nouws,

2020). In phase two, policy options are defined, which are proposals to tackle the policy problem. The policy options contain of different properties. These are the effects of the policy options. Any constraints are also determined, which are limited properties that the respondents may not exceed such as the maximum public budget that the respondents may use in the PVE. Phase three deals with the design of the PVE. Within the design of a PVE, the policy options of a specific issue are put together in an overview. Such an overview in which respondents are asked to choose between policy options while they may not exceed certain constraints is also referred to as a choice task. In phase four the citizens participate by completing the PVE. The citizens who complete the PVE are also referred to as respondents. This phase is also referred to as a PVE consultation, or "a distinct PVE around some topic" (Bouwmeester, 2021). The phases before phase four are in preparation of the PVE consultation and in the phase after phase four the consultation is processed. Finally, there is phase five in which the results of the PVE are analysed and reported to the policy maker. The data provided by the PVE provide insights into citizens' preferences. For example, the PVE provides information about how often a policy option has been chosen. Moreover, the PVE delivers qualitative information that reflects the motivation of respondents to choose a particular policy option. In addition, the collected data can be used to determine how respondents rate different properties of the policy options (Dartee, 2018). This shows that the PVE is both an evaluation method for policy options and a method that facilitates citizen participation (Mouter et al., 2021b).

When the PVE process is organized as above, the respondents are put in the shoes of the policymakers. The respondents are given the opportunity to select policy options. During the completion of the PVE consultation, the respondents receive information about the policy options and the context of the decision-making process. This information also includes the properties of the policy options. Due to the maximum budget which is a constraint that the respondents are allowed to divide, they are faced with trade-offs. The preferences that respondents assign budget reflect the value that respondents assign to the options. After allocating the budget to options, respondents are asked to justify why they selected these policy options.

1.1.1 Participatory Value Evaluation and the evaluation of policy options

Because the preferences of the respondents with regard to policy options are questioned in a PVE consultation, the PVE can be stated as an evaluation method. All individual selections of policy options by respondents together lead to an optimal set of policy options in terms of social value (Dekker et al, 2019).

1.1.2 Participatory Value Evaluation and public participation

As the PVE is a method for public or citizen participation, the public participation can be defined as "active involvement of a broad range of stakeholders in decision-making and action" (Few et al., 2007). Where respondents are only involved in the PVE consultation phase, stakeholders consist of the parties that are involved in the PVE preparation and consultation process.

Arnstein (1969) created a framework to categorize levels of influence in public participation. The PVE would fit into the tokenism category of this framework, which means that public participation within the PVE mainly revolves around informing and consulting the citizens. In

a PVE, the respondents are informed by the facilitator about the possible policy options and their properties. Subsequently, the respondents are asked to provide advice based on this information about which policy option(s) to implement and why.

Another framework is the one of Rowe and Frewer (2005) in which three forms of participation are defined based on the flow of information that takes place between respondents and the facilitator. When there is one-sided information from the facilitator, this is called public communication. When information only goes from the respondents to the facilitator, this can be defined as public consultation. When there is an information flow from both sides, there is public participation, as is the case with the PVE.

In general, in order to state that public participation has been successful or unsuccessful in a particular case, Rosener (1978) indicates that it is important to assess the effectiveness. Since a PVE consultation is in principle initiated before a policy decision is made, it can be stated that a PVE is carried out in the policy preparation phase (Openbaar Bestuur, 2005, p.10). The PVE can be deployed in several ways within this phase. For example, in a case in Súdwest-Fryslân, citizens were involved in selecting policy options that were included in the PVE consultation. In the same case, after the PVE consultation, citizens were also involved in the processing in order to provide more explanation and substantiation of the results. Furthermore, the aim was to arrive at recommendations for policymakers (Spruit and Mouter, 2020). Therefore, it is a requirement for the effectiveness of public participation that the facilitator and the citizens are aware of the goals and objects of the participation. Although the effectiveness is regarded in the literature as an overarching concept, it can be stated that effectiveness is about the extent to which an instrument or process adequately measured the goal concept (Rowe & Frewer, 2004). This definition essentially touches on the concept of validity.

1.1.3 Participatory Value Evaluation and (face) validity

Following Drost (2011): “validity is concerned with the meaningfulness of research components.” In other words, are researchers actually measuring what they intended to measure? This corresponds to Field’s (2005) definition of validity: “measure what is intendent to be measured.” For the PVE method this means that the respondents are enabled to experience the entire complex decision-making situation of a policymaker and to issue an advice about this complex situation. However, researching the validity of the PVE method is a path that not many researchers have yet explored. This is the case while it is considered necessary for research methods to guarantee validity, but also reliability, in order to eventually be wide applied in research (Drost, 2011). Following the master thesis of de Geus (2019), no empirical research has been conducted into the validity of the PVE method. This was the reason for de Geus (2019) to focus on possible cognitive bias effects of framing. He concluded that emphasis framing has an effect on respondents’ decision-making in a PVE consultation, which influences the validity of the method.

The concept of validity as defined above is an umbrella term. Within the concept of validity, different types of validity can be identified. For example, Drost (2011) identifies six types of validity. Face validity is one of these types. A definition of face validity can be found in the article by Taherdoost (2016): “Face validity is the degree to which a measure appears to be

related to a specific construct." Regarding the PVE method, face validity concerns that the choice options and their information the respondents see, appear genuine.

Anastasi & Urbina (2007) state that face validity is a desirable feature of instruments. If the content of an instrument, like a PVE survey, is inappropriate or irrelevant, there is a likelihood that the results obtained from this instrument provide false information and lead to decisions that are misleading. Bannigan & Watson (2009) also indicate that an assessment of face validity is important. When respondents and stakeholders consider an instrument to be face-valid this ensures acceptance of the instrument which increases its usefulness. Certainly with regard to the PVE method, it is important that the information that follows from the PVE consultation is not perceived as false and that the consultation is perceived as useful and meaningful by stakeholders. It is possible that the facilitator, often a policy maker, bases policy decisions on the results of a PVE consultation. Moreover, the information used to make decisions is increasingly complex and difficult to comprehend in today's world. Precisely for a PVE method that put citizens in the shoes of policymakers with regard to a complex situation, it is important that the citizen is able to fully experience the policymaker's problem. This argument shows that face validity is an important type of validity for the PVE method. A final addition is that Dempsey & Dempsey (1992) describe face validity as the fastest type of determining validity. It can be added that when validity tests are not performed due to the complexity of the research or due to time constraints, it is recommended to at least assess face validity. It follows that the focus on the face validity type can be feasible for a master thesis project.

Furthermore, the following real-life example of the Amsterdam-Wind PVE consultation shows the importance of face validity for the PVE method. The aim of the Amsterdam-Wind consultation was to find out what the residents of Amsterdam thought of windmills and what they noticed as the advantages and disadvantages of windmills. Around this topic of windmills in Amsterdam there were action groups of citizens who had organized themselves. These action groups were against the installation of windmills. However, during the process of designing the consultation, these stakeholders are only involved in the final stage of the PVE design. As a result, the action groups did not agree with the content of the consultation. One of the arguments they used was that face validity was a fundamental problem. The action groups did not hesitate to speak out against the consultation in the media. This is apparent from an article of Bakker which is a member of an action group who wrote the following in Het Parool (2022): "This digital questionnaire allowed participants to choose: wind power generation versus noise, pollution, health, loss of natural values, economic benefit, etc. You could not indicate that wind power generation in the city might not be feasible at all, precisely in view of the loss of health, natural values or enjoyment of living."

1.1.4 Knowledge gap and main research question

So, the PVE method is a promising research method that enables public participation. When evaluating public participation, Rosener (1978) points out that it is important to focus on effectiveness. This effectiveness touches on validity. De Geus (2019) indicates that no empirical research has yet been conducted into the validity of the PVE method. However, validity is a umbrella concept. The face validity type is a type that specifically addresses the operationalisation of a construct (Taherdoost, 2016). A lack of face validity can lead to false information and therefore misleading decisions of both respondents and policy makers. A lack

can also cause commotion and resistance from stakeholders, such as with the Amsterdam-Wind consultation. Especially for the PVE method face validity is important since respondents are put in the shoes of policymakers and their complex decision situation. Therefore, it is important to research how to measure face validity within the PVE method and how involved citizens and stakeholders evaluate the face validity within this method. However, because the PVE method serves multiple purposes as a participation method and an evaluation method, the question arises in relation to what face validity is measured. In this research, the focus is on the measurement of preferences, or the PVE as an evaluation method. This perspective is chosen because the participation process is often broader than the PVE consultation itself. Therefore, measuring preferences as a goal in relation to a face validity assessment is more feasible and more easy to define. The lack of research regarding the validity of the PVE method and the importance and feasibility of research into face validity regarding the PVE method as an evaluation method, leads to the following main research question: *How to measure face validity regarding the PVE method and how do respondents evaluate the face validity of a PVE consultation?*

1.2 Relevance of the research

Regarding the scientific relevance, this research contributes to this aspect by researching the validity of the PVE method. The validity of the PVE method is not widely researched yet. Moreover, by examining the face validity of the PVE method, more insight is gained into the demands of citizens about public participation within the PVE. From the perspective of citizens, this allows them to advice the policymakers in a more optimal way. Therefore, an aim of this research is to design a framework that measures face validity that can be applied in multiple PVE consultations. In this way, it is possible to continue the monitoring of face validity in order to better enable citizens to give advice and in that sense to increase the validity of the PVE method. From this follows that the scientific contribution of this study is two-sided. On the one hand, it adds value to the validation of this relatively new research method. On the other hand, it adds value to the field of public participation.

Besides the two-sided scientific relevance, this study has also a two-folded aspect with regard to a societal point of view. First, this research contributes to investigating the face validity within the PVE method from the perspectives of citizens, which will lead to a better understanding and therefore a higher satisfaction of the PVE among citizens. As a result, citizens are better able to give advice to policymakers. This is important since their advice can be incorporated into policymaking by policymakers. Secondly, this research focuses on a case study about the Schiphol Environmental Council. By measuring face validity in this case, it can be stated how residents have experienced this consultation and whether they have points for attention for a possible subsequent consultation. This puts them in a better position to give their opinion and/or advice in the situation around Schiphol Airport in which they are involved.

Finally, there is the relevance of this research in relation to the master's program 'Complex Systems Engineering and Management' (CoSEM). When complex systems are invoked, it is important that a researcher is aware of how this complexity comes across to the respondents. In today's world, decision information is becoming increasingly complex. The PVE method fulfills a need to collect input from the citizens and in particular the values that citizens evoke. However, it is important for a citizen who completes a PVE consultation to understand the complexity of a system in order to be able to advise the policymaker. However, if the situation

is too complex for citizens, this may lead to less trust in the government. Therefore, it is important in the PVE method but also in further research within CoSEM that a balance is found between complexity and comprehensibility towards the citizen. This research, which focuses on face validity, contributes to how the complexity of systems appears to citizens.

1.3 Research approach and sub questions

1.3.1 Research approach

To answer the main research question, it is required to go through a number of steps that are embedded in a mixed methods approach. So, qualitative and quantitative research will be combined (Creswell & Plano Clark, 2017). The steps taken in this research are presented in figure 1.1.

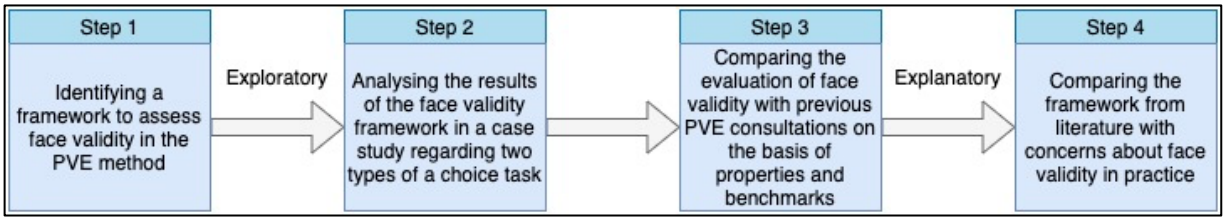


Figure 1.1: Overview of steps in research approach

The first step of this research concerns identifying a framework, which is a set of questions or statements that is able to measure the face validity regarding the PVE method, in a qualitative manner. The output of this step is first used to evaluate face validity of respondents within a PVE consultation for a case study, which is the second step of this research. According to Creswell and Plano Clark (2017), the transition from step one to step two involves exploratory research, as the qualitative data collected leads to the quantitative data outcomes of the PVE-survey.

Furthermore, an experiment is performed in step two on the basis of the evaluation framework of face validity. In this experiment, two different approaches of a choice task are tested using a face validity assessment. The aim of this experiment is to investigate in a PVE consultation whether these different approaches influence the assessment of face validity of the PVE method. Because face validity concerns how valid or genuine a consultation appears and the design of a consultation affects the appearance, this study experiments with the design of a PVE consultation. This PVE consultation presents, among other questions, policy options to the respondents who can express their preferences for these options by assigning points to these options. The PVE consultation applied in this research is a case study on participation and the provision of information around Schiphol Airport on behalf of the Schiphol Environmental Council.

The face validity evaluation in this research specifically focuses on the case study of the Schiphol Environmental Council. However, in previous case study consultations some categories of face validity are also included. Therefore, the results of the assessment of face validity of these studies will be compared with each other. The aim is to research whether the different properties of these case studies lead to a difference in the evaluation of face validity and how these assessments relate to each other in a benchmark.

The first three steps in this research are based on the perspective of the literature. A framework is drawn up from the literature in which face validity can be assessed and this is applied in a case study with an experiment. Therefore, the fourth and final step of this research takes a different perspective, namely the perspective of practice. The aim of this step is to identify which concerns stakeholders and citizens have in practice when it comes to face validity and whether these concerns are sufficiently covered in the established framework. This step has been performed in a qualitative way. According to Creswell and Plano Clark (2017), the transition from quantitative data, which is about evaluating face validity, to qualitative data can be described as explanatory.

The mixed methods approach brings several advantages. First, a mixed methods approach is appropriate for answering a main research question that neither qualitative nor quantitative methods could answer alone, as is the case when first a set of questions or statements is identified to evaluate face validity, after which this set is used to actually measure face validity in a case study. Furthermore, a mixed methods approach is able to provide a strong voice to stakeholders, as they are enabled to motivate their choices and suggestions for improvement regarding face validity of the PVE method (Shorten & Smith, 2017).

1.3.2 Sub questions

One sub research question is linked to each of the four steps in the research approach. The sub research questions together answer the main research question.

1. *How to design a framework that is able to measure face validity regarding the PVE method?*

The aim of this sub question is to identify a framework, in other words a set of questions or statements, to measure face validity in a PVE consultation. Therefore, is it important to investigate what face validity exactly contains and how face validity has been measured or questioned in other studies. To answer this sub question, a literature review is performed. Complementary to the literature review, expert interviews are conducted as well. In more detail, experts gave their advice on the selection of questions to be adopted to evaluate face validity regarding the PVE method.

2. *What are the similarities and differences of two different types of a choice task regarding a PVE consultation on the evaluation of face validity?*

The insights of the first sub question are applied in the PVE consultation of the Schiphol Environmental Council. The aim of this sub question is to analyze the results of the questions in the face validity framework between two different types of choice tasks. To this end, a PVE consultation is set up with two experiments. With the quantitative data collected regarding face validity, it is possible to perform certain statistical analyses. First, descriptive statistics are applied. Descriptive statistics provide a global insight on the quantitative data of the consultation. For example, it provide insight into how each question/statement scores on average in terms of face validity. Subsequently, Mann-Whitney U tests and one-way MANOVA tests are performed to check whether the results of the face validity assessments significantly differ between the two types of the choice tasks in the PVE consultation of the case study. The Mann-Whitney U test examines whether there is a significant difference between the two experiments of the PVE consultation based on a single statement of face validity. From this test can be concluded what the effect is of a face validity category on the difference in

assessment between the two types of choice tasks. However, this test does not provide multivariate results using information among multiple face validity categories. Therefore, the one-way MANOVA test is executed. From this test can be concluded what the effect is of a composition of face validity categories on the difference in assessment between the two types of choice tasks. In addition, a factor analysis is performed. The aim of this analysis is to find out whether the different questions that should measure face validity all measure the same latent variable or whether there are different latent variables that measure aspects of face validity (Moores et al., 2012). The hypothesis is that three latent variables can be identified. A first latent variable is expected to consist of the two face validity categories that are queried immediately after the first choice task of the PVE consultation. A second latent variable is expected to consist of a general view on the face validity of the PVE consultation. This latent variable consists of three face validity categories that are set at the end of the PVE consultation. Finally, a latent variable is expected that consists of all face validity categories that are included in this study, because all these categories are part of the concept face validity. Section 3.4 discusses when the categories of face validity that are part of the framework are stated in the consultation. Furthermore, multiple regression analysis can be performed with the goal to identify which demographic characteristics have influence on the evaluation of face validity. Because also the different questions/statements of face validity instead of the latent variable(s) is researched, a multinomial logistic regression is deployed. Finally, a latent class cluster analysis (LCCA) is performed. The aim of this LCCA is to investigate whether there are homogeneous groups of participants who score high or low on face validity. A follow-up question is what these respondents have in common with regard to demographic characteristics within a cluster.

3. *Which properties influence the differences in the evaluation of face validity between different case studies and what are the benchmarks of those differences?*

Additionally to the Schiphol Environmental Council case, there are also previous PVE consultations that have included elements of the established framework in question one. The different case studies have different properties that can be represented by a typology. Comparing the estimations of elements of face validity between different case studies lead to conclusions about the influence of different properties on the evaluation of different face validity categories. A method for these comparisons is the Mann-Whitney U test. Furthermore, the various case studies make it possible to set benchmarks for a face validity category that indicates when face validity is evaluated relatively high or low.

4. *To what extent do the concerns of citizens and stakeholders with regard to face validity correspond in practice with the established framework?*

Three methods are applied to answer the fourth sub question. First, a document analysis is applied. The stakeholders defined in section 2.3 are given the opportunity to provide feedback on the draft report which is the last step before the end result of the client is finalized. Moreover, respondents of the open consultation are asked to argue why they assessed the face validity categories with a certain answer option. The aim of the document analysis is to analyze the feedback and answers that touches on face validity. After the final report is finalized, the results are presented in a forum. Therefore, the second method is field research to analyze this conversation in order to identify concerns about face validity in practice. Third, in-depth interviews are held. The aim is to clarify the extent to which the framework from the literature and the concerns from practice overlap with each other.

An overview of all the methods that are used, is presented in figure 1.2. A more detailed overview is presented in Appendix A.

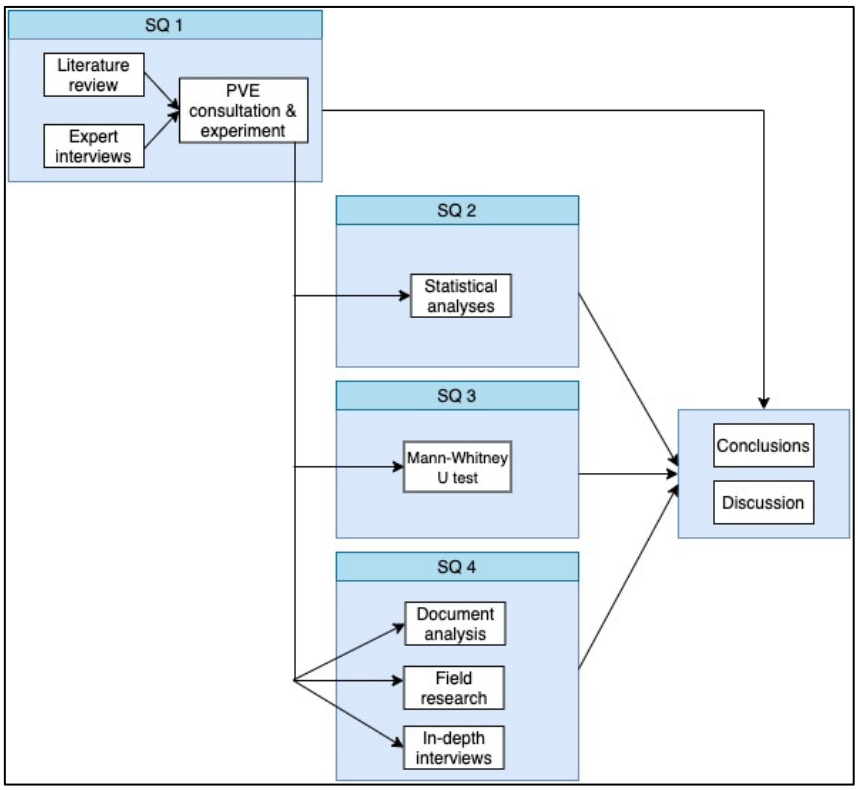


Figure 1.2: Research flow diagram

1.4 Structure of the report

The structure of this report is as follows. The next chapter, Chapter 2, elaborates on the research methods applied in this research. In Chapter 3, a framework is designed for assessing the face validity of the PVE method. Then, in Chapter 4, the results of the face validity assessment and experiment among respondents of a PVE consultation are analysed. Thereafter, Chapter 5 elaborates on the benchmarks of the evaluation of face validity categories and the influence of case study properties. Chapter 6 is not reasoned from the perspective of the literature, but discusses the concerns surrounding face validity in practice. Finally, Chapter 7 presents the conclusions of this research and Chapter 8 consists of the discussion and reflection.

2. Methodology

This chapter describes the methods that are applied in this research. The structure of this chapter follows the order in which the different methods have been applied. Therefore, the literature review is first described, after which the expert interviews are discussed. Subsequently, the case study is introduced. Thereafter follow the description of the PVE method, the face validity experiment and the statistical analyses that have been applied. This is followed by the Mann-Whitney U test and the benchmarks that are used to answer the third sub question. Finally, the methods are described that identify concerns about face validity in practice. This entails field research, document analysis and semi-structured in-depth interviews.

2.1 Literature review

A literature review is performed to answer the first sub question. On the basis of this review, a framework is designed that is able to assess face validity. The framework consists of a set of statement questions that together are able to measure face validity. In order to arrive at a framework, there are a number of steps that are followed. Therefore, the focus is first on the history of face validity and its importance. Then the framework of Nevo (1985) plays a role because this framework provides guidelines for setting up a face validity assessment. Publications selected for this part of the literature review all address part of the Nevo framework (1985) or other elements that need to be defined for an assessment of face validity. Finally, twenty-five case studies and leading articles with regard to face validity are selected. These articles indicate the elements (or categories) that make up face validity. This selecting procedure ended with twenty-five articles, because a saturation point was reached. From then on, no new categories of face validity could be identified. Based on these categories, specific statements have been drawn up that form the framework. The publications selected for this section of the literature review provide specifications of face validity. Most of the time these are case studies in which the face validity of a certain instrument is assessed. Moreover, the selection of articles is based on relevance, i.e., the number of citations.

The databases Google Scholar and Scopus are used as the tools to search for literature. Appendix B lists the keywords used during this literature review. Furthermore, when searching and selecting articles, both backward snowballing and forward snowballing are used. Backward snowballing is used as some articles had interesting references for this review. Forward snowballing has been used because it was found that some articles published in the 1980s can be seen as building blocks of face validity. The main example is the article by Nevo (1985). An advantage of a literature review is that literature provide a broad coverage. In addition, documents are stable, which is also an advantage (Bowen, 2009).

2.2 Expert interviews

Expert interviews are suitable to complement a literature study (Pfadenhauer, 2009). The literature review together with expert interviews answer the first sub research question. Conducting the expert interviews has two aims. The first aim is to select the most important categories of face validity. In this research a PVE consultation is conducted for a specific case study. This case study is described in the next section. However, there is limited space to ask five statements and thus five categories of face validity. This maximum has been set to ensure

that the consultation does not become too long. With regard to the substantive questions, the length of this consultation is estimated at twenty minutes which is relatively long for a consultation. The second aim touches on the PVE method. The categories of face validity identified from the literature are not specific to the PVE method. However, it is a goal of this research to design a framework to assess face validity for the PVE method. Therefore, it is important that experts select the categories specifically with regard to the PVE method.

Qualitative expert interviews are usually performed in the exploratory phase of a research. Bogner et al. (2009) argue that expert interviews can be conducted when the field of research is poorly defined. The goal is then to collect contextual information. It can be stated that this is in line with the aim of the expert interviews in this study. In more detail, experts can give their advice on the selection of questions to be adopted to evaluate face validity regarding the PVE method. Because of the limited space to ask statements about face validity in the case study consultation, it is important to get confirmation of experts in advance that the set of questions will be useful and as complete as possible to measure face validity.

Because of the two aims of the expert interviews, experts in two different research fields are interviewed. Experts in this context are researchers into validity or researchers who specifically focus on the PVE method. In total, five experts were interviewed. Two of the interviewees are experts in the field of the PVE method. Two other interviewees are experts in (face) validity and one interviewee is specialized in both fields of research. The experts interviewed are presented in table 2.1. When selecting experts, the constructivist definition is applied. This definition indicates that it is the researcher who determines what an expert is in the context of his or her research or of processes in society.

Table 2.1: Interviewees of the expert interviews

Expert	Function/expertise
1. PVE expert 1	Co-founder and director of projects at a company that conducts PVE consultations. This expert is also involved as a researcher in the development of the PVE method.
2. PVE expert 2	Project leader at a company that conducts PVE consultations and PhD candidate in the field of the PVE method at Delft University of Technology.
3. Validity expert 1	Postdoctoral researcher in health economics at the Erasmus School of Health Policy and Management. This expert focuses on the validation of an instrument to measure well-being in the adult population.
4. Validity expert 2	Associate Professor of open data at Delft University of Technology. This expert has a background in criminology/psychology and is aware of the existence of the PVE method.
5. PVE and validity expert	PhD candidate in the field of the PVE method and its validation at Delft University of Technology.

Overall, the expert interview is known as an efficient and quick qualitative method (Bogner et al., 2009). However, a limitation of expert interviews is that the perception of an interviewee

has impact on the outcomes of the interview (Pfadenhauer, 2009). Conducting multiple interviews and having them complemented by the literature review overcome this limitation.

The expert interviews are conducted in a structured manner. A structured set-up has been chosen because the objective of the expert interviews is concrete. Moreover, the answers of different experts are easy to compare with a structured set-up. The expert interview protocol is presented in Appendix C. Furthermore, Appendix C also contains a code list. This list is related to the analysis of the expert interviews. First, the audio recordings of the interviews were transcribed. Interesting passages from these interviews are coded with this code list per question from the expert interview protocol (Meuser & Nagel, 2009). The results of the expert interviews are combined with the results of the literature review. Collectively, this leads to a framework in which five categories of face validity are questioned on the basis of statements with regard to the PVE method.

2.3 Case study: the Schiphol Environmental Council

The case study of this research is about the Schiphol Environmental Council (ORS). The ORS is the council where local residents can go with questions, comments and requests about the various public interests of the Schiphol Airport. Members of the ORS are representatives of the Noord-Holland Environmental Federation, Schiphol Group, Air Traffic Control the Netherlands, KLM, the employers' organization VNO-NCW West and residents' representatives per runway. In the report of the van Geel committee, the committee has issued advice on the ORS. Van Geel (2020) notes that the polder model for decision-making in the ORS is no longer effective. Recently, the focus in consultations with residents has been too much on Schiphol's growth opportunities. This is related to the absence of mutual trust between the various parties within the ORS (Berenschot, 2020). This bottleneck manifests itself in four different ways. First of all, the parties in the ORS no longer reach agreements. Secondly, the discussions between parties in the ORS are mainly about the formal aspects of the cooperation and not about substantive issues. Third, the parties operate outside the ORS. As a result, there is no longer a joint contribution to policy-making. Finally, the parties jointly fail to implement concrete measures (Berenschot, 2020).

As a result of this bottleneck, a common goal of Schiphol and residents is missing. Therefore, van Geel (2020) proposed two new entities that mainly contribute to more intensive and broad participation and improved information provision. These entities are a Schiphol Social Council and an Environmental House. Following van Geel (2020), it is important to involve citizens in shaping the precise functions of those entities. Regarding the Schiphol Social Council, the question arises what the citizens' needs for participation are. The aim of the Schiphol Social Council is that residents living in the vicinity of Schiphol can contribute to policy and its implementation, so that governments and implementing bodies can benefit from this. Van Geel (2020) had defined seven functions that the Schiphol Social Council should fulfill: dialogue function, representative function, knowledge function, service function, advisory functions, repetitive participation functions and social signaling function. With regard to the Environmental House, the usefulness and necessity of strengthening the information and service provision is endorsed by all parties of the ORS. Despite recent efforts, bottlenecks remain. In the report by Berenschot (2020), an Environmental House is sketched with the following functions in the most extensive scenario: data collection, measurement and monitoring, information dissemination and knowledge development, meeting place with

dialogue function, handling of requests based on regulations, complaint registration and handling, mediator function and a visitor center. One of the main underlying questions is what are the needs of citizens.

However, the reports by Berenschot (2020) and van Geel (2020) give limited prioritization of the functions of both entities and can also be regarded as vague. Therefore, the aim of a PVE consultation is to prioritize and concretize the various functions by having residents in the vicinity of Schiphol issue advice. Various stakeholders and the client which is the ministry of Infrastructure and Water Management together with the Schiphol Environmental Council are involved during the setting up of the PVE consultation and the preparation of the final report. More specific, the stakeholders involved are five residents' representatives of the Schiphol Social Environmental Council, the public affairs manager of the employers' organization VNO-NCW West, the director of Natuur en Milieufederatie Noord-Holland, the public and community affairs manager of Schiphol Group, the account/-issue manager and strategy expert of Air Traffic Control the Netherlands, the strategic advisor airport affairs of the municipality of Haarlemmermeer and a process manager of the municipality Ouder-Amstel. A number of steps have been taken with these stakeholders to optimize the validity of the consultation. First, exploratory discussions were held with the stakeholders to gain knowledge about the problems they experience. During these discussions, attention was also paid to questions such as: what do you think is the core dilemma, what would you like to know from citizens, what choices must be made and what do think of the idea of a Schiphol Social Council and an Environmental House? The aim of this first phase was to establish objectives and preconditions for the consultation. Second, a draft of the consultation was shared with the stakeholders on which they could provide feedback. Meetings were also scheduled for this with the stakeholders and they were also allowed to provide a document with feedback. Third, stakeholders were given a final opportunity to provide feedback on the 99% version of the consultation. In both feedback phases it was explained to the stakeholders why and how their feedback was or was not included in the consultation. Once there was consensus on the consultation, it was distributed to residents in the Schiphol area. The stakeholders were also given the opportunity to provide feedback on the draft report presenting the results of the PVE consultation.

Moreover, the Schiphol Environmental Council case study has the following four properties. First of all, this case study focuses on the short-term impact on personal life. It is the intention that the Schiphol Social Council and the Environmental House will be established in the near future. Second, this case study consists of both a panel and an open consultation. The aim of the panel consultation is to obtain a representative group of respondents via a data panel. Anyone over the age of eighteen was allowed to fill in the open consultation. Third, this consultation has been carried out in the new version of the online platform. In addition to this new version, there is also an old version. Fourth, this is a consultation where the stakeholders are in charge of the design of the consultation.

The controversial subject of the PVE consultation of the ORS makes this case study interesting for research (Zainal, 2007). Values and interests are in conflict in a controversial topic. In a controversial case, opinions are fundamentally divided which can lead to strong emotions. In this case about Schiphol, the interests with regard to nuisance are diametrically opposed to the interests with (economic) advantages. These extremes lead to strong opinions. When face

validity is evaluated in a controversial case, it is expected that extremes will also arise. This means that a wide spectrum of opinions can be analysed. In addition, if the face validity is in order in a controversial case, it will also be in order in non-controversial cases (Zainal, 2007).

2.4 PVE method and face validity experiment

As described in the introduction (chapter 1), the PVE method can be applied for participatory purposes as well as for evaluation respondents' preferences. In this research, the focus is on the evaluation of preferences in this online webtool. In short, respondents in a PVE consultation are presented with various options from which they can choose. However, there is a constraint. A maximum number of points can be divided among the options or a maximum budget. Respondents divide the points or the budget among the various options and thereby express their preferences.

As an evaluation method, the PVE method shares the same advantages as the Cost-Benefit Analysis (CBA) on which it is based. Both methods result in a policy advice. To this end, different policy options can be compared in both methods and an optimal portfolio of policy options can be calculated. In addition, both methods are based on a theoretical framework (Mouter et al., 2021c). However, the PVE method has succeeded in tackling a number of disadvantages of the CBA. The CBA method makes use of Willingness to Pay (WTP). WTP is a stated preference method where citizens can highlight their preferences by making choices with their private resources in scenarios that are hypothetical. However, WTP is not a good measure to ask individuals what they would spend public money on since WTP is about private resources according to scholars. Therefore, Willingness to Allocate Public Budget (WTAPB) is developed. An extension of this WTAPB is the PVE method. Both in WTAPB experiments and in the PVE method, it is up to respondents to make choices about the distribution of public budget. The main difference between the two is that in the PVE method the respondent is not obliged to divide the (entire) public budget (Mouter et al., 2021c). In addition, the PVE method makes it possible to focus on social and ethical goals, while the CBA focuses on quantifiable goals. Therefore, the PVE method is capable of recording soft goals, such as those relating to the environment or health. It follows that the CBA is able to identify a generalized picture of preferences, while applying the PVE method it is also possible to interpret outcomes in the context in which a PVE consultation has been conducted. Finally, the CBA uses standardized values, while the PVE method is able to retrieve local knowledge (Mouter et al., 2021c).

In this research, the PVE method is applied in the case study of the Schiphol Environmental Council. In this PVE consultation, attention is paid to both new entities drawn up by van Geel (2020). A choice task has been included about the Schiphol Social Council and another choice task about the Environmental House. Before these two choice tasks, the focus is on the question of how respondents would like to participate around Schiphol. After each choice task, in-depth questions are asked about or participation principles or information provision. Finally, the focus is on demographic characteristics. A more detailed structure of this PVE consultation is described in section 3.4. The PVE consultation is set up on the basis of the framework by Peeters (2020). His framework consists of three phases in which the first phase focusses on the research design, the policy options and further questions. In the second phase, the PVE consultation is tested and filled in by the respondents. The third and final phase consists of analyzing the results and writing a report about the outcomes.

The PVE method has several types of a choice task that can be included in a consultation. To answer the second sub question, an experiment is set up to research the similarities and the differences regarding the evaluation of face validity between the two different types of choice tasks. Testing different versions of a consultation is a well-known principle in discrete choice experiments (DCEs). These DCEs capture people's trade-offs in choice situations. The PVE method and the DCE method are similar in that both are methods of stated preference (Rotteveel et al., 2022). Of the DCEs published between 2009 and 2012, face validity is included in approximately sixty percent (Clark et al., 2014). The main reasons for including multiple versions for face validity testing in a DCE is to research "whether the choice task accounts for important preference attributes and whether results are consistent with a priori preference expectations" (Janssen et al., 2017).

The face validity experiment consists of two types of choice tasks. These two types together form the face validity experiment in which the different types are referred to as experiments. The 'sliders' experiment consists of the 'sliders' choice task of the Schiphol Social Council and the 'points' experiment of the 'points' choice task. Both experiments are discussed in detail in section 3.4. Furthermore, the content of the two experiments is equal. In these two experiments the face validity framework, which was drawn up on the basis of the literature review and the expert interviews, is included. The PVE consultation has been drawn up jointly with stakeholders. These two types of choice tasks lead to consensus about the consultation among the stakeholders.

The response scale on which the respondents answer the statements is a Likert scale. Randall and Fernandes (1991) argue that an odd point Likert scale, and thus offering a neutral option, reduces the chance of bias in the respondents' answers. Therefore, it is recommended to apply an odd point Likert scale. On the other hand, when an odd scale is applied instead of an even scale there is some loss of information since a neutral option does not provide much information about the opinion of a respondent. So, the choice between an even and an odd Likert scale is a trade-off. In general about odd scales, empirical studies agree that 5- to 7-point scales increase reliability and validity over coarser point scales. More detailed point scales do not improve the reliability and validity (Dawes, 2008). Because it takes longer to complete statements with a 7-point scale than a 5-point scale (Matell & Jacoby, 1972) and a 5-point scale is more often used in face validity assessments (e.g., Nevo, 1985; Moores et al., 2012) a 5-point scale is applied for the face validity statements.

2.5 Statistical analyses

With the quantitative data collected regarding face validity in the PVE consultation of the case study, it is possible to perform certain statistical analyses. Statistical analyses are performed to answer the second sub question. One statistical analysis, the Mann-Whitney U test is also performed to answer the third sub question. Therefore, it is necessary that a sufficient number of respondents complete the questionnaire (in science the minimum sample size of thirty respondents is mostly used (Knofczynski & Mundfrom, 2007)). In the studies by Dartee (2018) and Nouws (2020), this turned out to be a challenge. Therefore, a data panel is used to achieve a representative sample on the basis of age, gender and education. Furthermore, the respondents were required to be eighteen years or older and live in one of the 54 municipalities surrounding Schiphol. These are also the municipalities that are part of the

Schiphol Administrative Board (BRS). The 54 municipalities are specified in Appendix F. These requirements have been agreed in advance with the data panel.

In total, 2572 respondents started filling in the PVE consultation between April 22, 2022, and May 15, 2022. However, not every respondent completed the consultation. When performing the data analyses, it is decided to include the answers of all the respondents who fulfill the consultation completely. This entails that the respondents answered the five face validity statements and answered all questions about generic and case-specific demographic characteristics (as listed in section 3.4). The answers 'I rather not say' or 'no opinion' are regarded as missing values. It is decided to only include completes in order to minimize the chance of bias. Ultimately, the 'sliders' experiment is completed by 648 respondents. The 'points' experiment is completed by 582 respondents. SPSS can be used as a data analysis tool to perform the statistical analyses. Furthermore, a level of significance of 0,05 is applied for all statistical analyses. The statistical analyses performed are discussed below.

2.5.1 Descriptive statistics

First of all, descriptive statistics are performed. Descriptive statistics provide a global understanding on the quantitative data of the PVE consultation. Moreover, this method provide insight into how each face validity statement of the established framework scores. Descriptive statistics also provide insight in the distribution of age, gender and education level of the whole sample and of the both experiments of the face validity experiment. To research whether the results of the sample are generalizable to all residents in the 54 municipalities around Schiphol, chi-square tests are performed with the variables age, gender and educational level.

2.5.2 Mann-Whitney U test and one-way MANOVA

In both experiments of the Schiphol Social Council choice task the established framework of face validity is included. Based on these five statements, it is possible to compare the face validity of both experiments using this framework. The aim is to research whether the assessment of face validity of the two experiments differs significantly from each other. To this end, five Mann-Whitney U tests are first performed. This means that a test is performed for each of the five face validity categories. The Mann-Whitney U test is an appropriate test as it does not assume a normal distribution of the dependent variable. The five face validity categories for both experiments are included in a test of normality of which the results are presented in Appendix G. These tests show that none of the categories is normally distributed. The Mann-Whitney U test is also considered the non-parametric equivalent of the independent samples t-test. The dependent variable should be measured on a continuous or an ordinal scale. In this case, the statements of the face validity categories are answered on a Likert scale (ordinal). The independent variables consist of two groups, i.e., the two different experiments.

The Mann-Whitney U test examines whether there is a significant difference between the two experiments of the PVE consultation based on a single statement regarding face validity. Thus, this test does not provide multivariate results using information among multiple dependent variables which are the multiple face validity statements in this study. Because these tests do not take into account the correlations between the dependent variables, these tests can be considered less powerful (UCLA, 2021). A test in SPSS that takes into account the correlations

between multiple dependent variables is the one-way MANOVA test. This one-way MANOVA test is applied to determine differences between independent groups on two or more dependent variables that are measured on a continuous scale. In this case, the two experiments of the PVE consultation form two independent groups in the independent variable. In this test there are multiple dependent variables which consist of a composition of the face validity statements on an interval ratio. With a one-way MANOVA test a number of assumptions apply. First, there must be an independence of observation. Those who made the first experiment did not also complete the second experiment of the PVE consultation. Second, there must be an adequate sample size. There is a minimum that the number of completes in each experiment must be greater than the number of dependent variables being analyzed. This assumption is achieved by using a data panel. In addition, there should be no univariate or multivariate outliers. This is also not the case since the dependent variables are measured on a Likert scale. Furthermore, there should be no question of multicollinearity. Finally, a normal distribution of the data is assumed. Despite it being found that all categories in both experiments are not normally distributed (Appendix G) and the one-way MANOVA test assumes a normal distribution of data, one-way MANOVA tests have been performed. Research by Ito (2980) has shown that a one-way MANOVA test is robust when it comes to the assumption of a normal distribution of data. Therefore, he recommends the one-way MANOVA test over other statistical analysis tests for data that is not normally distributed.

A total of three one-way MANOVA tests are performed. This has to do with the different face validity statements included in the PVE consultation. There are two statements that are placed immediately after the Schiphol Social Council choice task. Therefore, a test is performed with these two statements as dependent variables since these two statements together can state something about the face validity of this choice task. The other three statements assess the face validity about the PVE consultation in general. This is the second MANOVA test. The third test contains all five statements as dependent variables, since all face can state something about the face validity of the PVE method. It is arguable that the three statements made at the end of the consultation and the five statements together both say something about the face validity of the PVE consultation. However, a drop-off rate of approximately forty percent has been established among the participants of the data panel. This means that more respondents completed the first two face validity statements that the last three statements, which is probably related to the face validity. For this reason, it is valuable to include the three statements at the end separately in a one-way MANOVA test.

2.5.3 Factor analysis

A factor analysis is performed to analyze whether the statements of face validity measure a common variable. This common variable is called a latent variable. Factor analysis states as data requirements that there is no perfect multicollinearity between the variables, that the variables are measured on a continuous or categorical scale, that there is a linear relationship between variables and that the residuals are normally distributed.

As with the one-way MANOVA test, various factor analyzes are performed. Three factor analyses are executed for each experiment. First, a factor analysis is performed with the two statements that are placed directly after the Schiphol Social Council choice task. When a latent variable is identified, it addresses the face validity of the specific choice task. Second, a factor analysis is performed with the three statements that are questioned at the end of the PVE

consultation. When a latent variable is identified with those three statements, it addresses the face validity of the PVE consultation in general. Third, a factor analysis is performed with all the five face validity statements. If a latent variable is found with all those five statements, it addresses the face validity of the PVE method. Otherwise, a latent variable measures a specific aspect of face validity. The reason of the drop-off rate also plays a role in performing the three factor analyses, just like with the one-way MANOVA tests. The latent variables that are identified by the factor analyses are included in further statistical analyses that may reveal differences between the 'sliders' choice task and the 'point' choice task of the experiments, such as differences in demographic characteristics that influence the assessment of face validity. For all six performed factor analyses the simple structure is approximated with a skewed rotation using direct oblimin. Moreover, the latent variables are represented by mean of the sum score.

2.5.4 Multiple regression analysis and multinomial logistic regression

The latent variables that are identified in the factor analyses, are the dependent variables in the following analysis. The latent variables are analyzed by means of a multiple regression analysis. The aim is to research whether and which characteristics of respondents have influence on the evaluation of the face validity latent variables.

Regression analysis requires a linear relationship between the dependent and independent variables, that the residuals are normally distributed and that there is no multicollinearity. Furthermore, this analysis requires that the dependent variable has a continuous scale. This requirement is met since the sum scores of the latent variables are continuous. The independent variable requires a continuous or categorical scale. This requirement is met as well since all the characteristics have a categorical scale. Dummy coding is applied for the independent categorical variable(s) as is presented in Appendix H. General demographic characteristics can be distinguished as well as case-specific characteristics. Moreover, this analysis is part of exploratory research as it has not previously been researched whether and which demographic characteristics affect face validity in the PVE method. That is why many demographic characteristics are included in this analysis in order to be able to research whether there are characteristics that have an effect on the face validity assessment.

With regard to the multiple regressions, a hierarchical structure is applied to identify the demographic characteristics that are statistically significant. The structure of this hierarchical regression model is as follows. In the first step, only the general demographics are added. The general demographics are the 'simplest' variables of all the characteristics. These are the closest to the respondent. In the next step, the case-specific characteristics are added to this. The structure of the hierarchical model is presented in table 2.2.

Table 2.2: Structure of the hierarchical model

	Step 1	Step 2
General demographics	X	X
Case-specific characteristics		X

The multiple regressions are applied to research whether and which characteristics of the respondents influence the evaluation of the latent variables. The following method specifically examines whether and which characteristics of respondents influence the evaluation of the

five face validity categories separately per experiment. The method applied is the multinomial logistic regression.

In total, this multinomial logistic regression is performed ten times, i.e. a regression per face validity category per experiment. This multinomial logistic regression requires that the dependent variable is measured at a nominal or ordinal scale. This requirement is achieved since the categories of face validity are measured on a Likert-scale (ordinal). The independent variables should be on a continuous, ordinal or nominal scale. This requirement is met as well since all the characteristics have a categorical scale. Furthermore, there should be no multicollinearity and there should be no outliers. Similar to the multiple regressions, the hierarchical model of table 2.2 is applied for the multinomial logistic regressions.

When the dependent variable is ordinal, it is preferable to apply an ordinal logistic regression instead of the multinomial logistic regression. However, this does not apply if the parallel regression assumption is not met (Liang et al., 2020). The parallel regression assumption implies that there should be a linear relationship between any independent variable and the logit of the dependent variable. This assumption can be tested with the test of parallel lines in SPSS. The null hypothesis states that the slope coefficients of all the response categories are the same. In Appendix E the results are presented of the tests of parallel lines. It follows that the assumption parallel regression assumption is not met. Therefore, the multinomial logistic regression is performed since this regression does not require this assumption to be met as contrary to the ordinal logistic regression.

Within a multinomial regression analysis, the different answer options of the face validity categories are compared to each other. It is most interesting to research if a low, neutral or high assessment of a face validity category differs in terms of the characteristics that influence this evaluation. Therefore, the answer options are recoded into three categories: (totally) disagree, neutral and (totally) agree. Additionally, the categorical dependent characteristics are recoded as well. The recoding is presented in Appendix J. This appendix also shows which categories form the reference categories.

2.5.5 Latent class cluster analysis

In addition to the regression analyses, a latent class cluster analysis (LCCA) is performed. While a regression analysis focuses on which characteristics explain the assessment of a face validity category, the LCCA focuses on whether certain groups of people can be identified who share the same characteristics and who collectively score high or low on certain categories of face validity. The LCCA is also adapted in a PVE consultation on climate policy among Dutch people by analyzing the combination of chosen policy options and characteristics of respondents like gender or income (Mouter et al., 2021a). A data requirement for the LCCA is that de data level of the indicators should be categorical, ordinal, continuous or count. Furthermore, it is assumed that the indicators in a LCCA are independent of each other. This assumption is also called the local independence assumption. Regarding the LCCA, SPSS is not used as a data analysis tool. The tool Latent Gold is used instead.

A total of two LCCA's are performed, one for each experiment. The general demographics and the case-specific characteristics are not researched in a separate LCCA, as the results are not significantly different than a LCCA where all characteristics are included simultaneously. The

characteristics form the covariates in the LCCA's. Since it is particularly interesting to research whether there are homogeneous groups that rate the face validity high or low, the answer options of the face validity categories are divided into three groups. These are (totally) disagree, neutral and (totally) agree. The face validity categories from the indicators in the LCCA's. Within the LCCA's, the Wald test is interpreted. If the p-value of the Wald test is significant, the coefficients of the indicators or covariates are not equal to zero.

2.6 Mann-Whitney U test (to answer the third sub question)

In this research, the focus so far has been on analysing face validity of the PVE consultation of the Schiphol Environmental Council. However, there are previous PVE consultation case studies in which face validity statements have been included. These previous case studies have not included face validity in its entirety but have included specific categories of face validity. Based on the framework established in the literature review and expert interviews, it is determined which other case studies included questions from this framework as categories of face validity. In previous PVE consultations, three categories of face validity are identified that also appear in the framework drawn up in this research. The first category is clarity. This category emerges in the climate consultation (Mouter et al., 2021a), in the heat transition vision in Utrecht (Mouter et al., 2020) and in the case of the long-term Corona policy (Geijsen et al., 2022). In the cases of heat transition vision of Utrecht and the long-term Corona policy, the relevance category emerges as well. Relevance is also included in another case study that concerns sustainable energy in the Foodvalley region (Spruit & Mouter, 2021). The third category is completeness which emerges in the case of the Foodvalley region as well.

The above case studies each have their own properties. In order to obtain an overview of these characteristics, a typology has been drawn up. The case studies are compared with each other on the basis of the corresponding face validity categories. Here, 'sliders' PVE consultations are compared with 'sliders' consultations and 'points' PVE consultations with 'points' PVE consultations. The aim is to research whether the assessment of a face validity category differs significantly between case studies and which characteristics influence this. A method to compare the evaluation of a face validity category between two case studies is the Mann-Whitney U test. This method is suitable since the dependent variable may have a continuous or an ordinal scale. In this situation, the statements are answered on a Likert scale (ordinal). The independent variables consist of two groups, i.e., the two different case studies. Moreover, the Mann-Whitney U test does not demand a normal distribution of the categories. The tests of normality show that none of the included variables are normally distributed (Appendix M). SPSS can be used as a data analysis tool to perform the Mann-Whitney U test. This test is performed to answer the third sub question.

In addition to performing the Mann-Whitney U tests, a benchmark has been drawn up for the categories clarity, relevance and completeness. This is also part of answering the third sub question. Bandwidths show the range of the minimum and maximum average scores of a face validity category of multiple case studies. Within these bandwidths a benchmark, or reference goal, is provided when face validity is in order compared to other case studies. The average score gives a general picture when most respondents are convinced that the face validity is in order.

2.7 Document analysis and field research

The methods document analysis and field research are applied together to research the perspective of practice, which is related to the fourth sub question. With these two methods, it is possible to address the concerns about face validity of the PVE method that exist in practice. Therefore, this perspective is separate from the previous sub questions, in which a framework has been built up from the literature with which face validity can be assessed.

First, the document analysis is performed. A document analysis is a qualitative method. Following Bowen (2009), a document analysis is "a form of research in which documents are interpreted by the researcher to give voice and meaning around an assessment topic". In the Schiphol Environmental Council case study, the end product is a report. Before the final report is published, the stakeholders defined in section 2.3 are first given the opportunity to comment on the draft report. The feedback given on the draft report is analysed. More specifically, comments related to face validity are analysed. Moreover, an open consultation is also conducted in the Schiphol Environmental Council next to the panel consultation. Every person aged eighteen or older is allowed to participate in the open consultation. In this consultation the respondents are asked if they could explain why they assessed the face validity categories with a certain answer on the Likert scale. These open answers showed that respondents also cited other face validity categories that had not been included in the consultation. Therefore, these quotes of respondents have been analysed. According to Bowen (2009), this document analysis falls under a content analysis. This is because a search is made for passages that concern a specific subject which is in this case face validity. These passages are described in the results of the document analysis.

An advantage of document analysis is that no reflexivity is required. There is no interaction with, for example, respondents. So, the process with the stakeholders is not influenced by this method. In addition, documents are stable, which is also an advantage (Bowen, 2009). A limitation of document analysis is that this method is often difficult to repeat because of a lack of transparency about the execution process (Bowen, 2009). This limitation is addressed by the clear indication that certain passages are included because they are linked to face validity. It is also made clear that only the feedback from the stakeholders is included as documents.

Second, the field research is performed. Field research is a method in the qualitative domain in which people are observed in their natural settings (Burgess, 2002). After the final report has been completed, the results are presented on the 'Regioforum'. The meetings of the 'Regioforum' are open to members, deputies, supporters and electors of the residents' organizations registered with the Schiphol Environmental Council. This forum took place on June 24, 2022, in Hoofddorp. During this 'Regioforum', it is analysed whether the attendees make comments that fall under face validity.

As with document analysis, no reflexivity is required in field research. There is no interaction between the researchers and the respondents or stakeholders, as the researcher is observing from the side lines. As a result, it is possible that respondents or stakeholders make comments about face validity while they are not aware of this. From this practical perspective, the respondents or stakeholders are therefore not guided by the researcher. However, a limitation is that detailed information is obtained during field research. This makes it impossible to collect information among large groups of people (Meredith, 1998). This is also

because of the time limitation of this research. Therefore, field research is carried out together with a document analysis to overcome this limitation.

2.8 In-depth interviews

Together with the document analysis and the field research, the in-depth interviews form the third method that serves to answer the fourth sub question. The document analysis and the field research uncover the concerns about face validity in practice of the stakeholders that are involved in the process of designing the PVE consultation at the 'Regioforum' and of the local residents and respondents. The aim of the in-depth interviews is to uncover the concerns about face validity in practice of a stakeholder who is not involved in the process of designing the PVE consultation and of the client which is the ministry of Infrastructure and Water Management together with the Schiphol Environmental Council. Furthermore, the aim of the in-depth interviews is to establish the link between the practice and the literature. This clarifies whether the framework from literature is satisfactory compared to the concerns in practice. Table 2.3 shows who has been interviewed.

Table 2.3: Interviewees of the in-depth interviews

Interviewee	Function
1. Stakeholder that is not involved	Employee of the Mainport Strategy department at KLM.
2. Client 1	Policy officer at the Ministry of Infrastructure and Water Management in the Directorate General Aviation and Maritime Affairs.
3. Client 2	Project manager at the Schiphol Environmental Council.
4. Client 3	Policy officer at the Ministry of Infrastructure and Water Management in the Directorate General Aviation and Maritime Affairs.

The in-depth interviews are conducted in a semi-structured manner. In this way the concerns can be traced, but the interviewees are also free to share their experiences. The interview protocols are presented in Appendix N and Appendix O. Furthermore, these appendices also contain a code list. This list is related to the analysis of the in-depth interviews. First, the audio recordings of the interviews were transcribed. Interesting passages from these interviews are coded with this code list per question from the expert interview protocol (Meuser & Nagel, 2009). The results of the in-depth interviews are combined with the results of the document analysis and the field research. Collectively, this leads to possible adjustments from practice on the face validity framework for the PVE method that has been drawn up in this research.

3. A framework to assess face validity in the PVE method

This chapter elaborates on the designed framework which is capable of measuring face validity regarding the PVE method. The framework consists of statements. The results of the literature review are discussed in the first section of this chapter. The second part of this section emphasises on the expert interviews. Finally, the framework of statements is applied in PVE consultation of the case study of the Schiphol Environmental Council.

3.1 Results literature review

This section elaborates on the concept of face validity and all aspects involved in a measurement or assessment of face validity. The aim of this literature review is first of all to clarify the concept of face validity with regard to the history and importance of a face validity measurement. Second, the aim is to discuss and clarify the preconditions of a face validity assessment for this study. The framework of Nevo (1985) forms the basis of these preconditions. This framework indicates that it is important for an assessment to clearly map who will assess face validity, what that person will assess, which measurement approach will be used and how. Third, the aim is to identify categories of face validity that serve as guidelines for setting up an assessment in the form of statements. To identify these categories, twenty-five articles are included in the literature review. These are articles in which the concept of face validity is defined, but also in particular case studies. In these case studies the face validity is assessed of, among other things, student exams, health care instruments or questionnaires on various themes. Overall, the literature review contributes to answering the first sub question.

The structure of the literature review is as follows. First, face validity is defined and the history of this concept is discussed. Subsequently, the importance of face validity is discussed. This is followed by the definition of the raters of face validity, the method to measure face validity and at what moment the measurement is taken. Thereafter is discussed which measurement approach is used. Finally, it is assessed which categories of face validity can be identified and with which statements these categories can be measured.

3.1.1 Face validity and its history

As stated in the introduction (chapter 1), validity is a umbrella concept. With regard to validity tests, two broad domains can be identified. The first domain is internal validity, in which questions are asked about the correctness of conclusions drawn by the researcher. The second domain is external validity which is meant to research whether conclusions from one case are generalizable to another case. Face validity is part of the domain of internal validity (Gaber & Gaber, 2010). Nunnally and Bernstein (1994) define face validity as: “reflecting the extent to which a measure reflects what it is intended to measure.” Sartori and Pasini (2007) add to this definition that face validity is concerned with the appearance of a measure or procedure test. Moreover, a test has face validity when persons agree that the test appears to be valid regarding the kind of measurement that is to be done. Roberts (2000) describes face validity as: “making a decision about the appropriateness of use of some particular measuring instrument in a given assessment situation through the process of simple inspection of that instrument.” From this quote it can be conducted that face validity does not depend on established theories. Therefore, the “simple inspection” of Roberts (2000) gives reason to regard face validity as subjective. From the simple inspection of Roberts, Drost (2011) defined

the following definition with regard to face validity: “a subjective judgment on the operationalization of a construct.” This definition shows that face validity refers to the looks and the feel of the measuring instrument. It is an assessment that asks the question: “on the face of things, does this research make sense?” (Alderson et al., 1995).

The considered subjectivity of face validity has led to an interesting history. During the 1940s and the early 1950s, face validity was developed and used by various researchers. In particular educators, content analysis researchers and psychologists have contributed to the development of face validity (Nevo, 1985). The development of face validity has been captured by many researchers in various research areas. In a short period of time, face validity became a widely used type of validity. However, due to its use in numerous different research areas, different conclusions emerged. A discussion raised about the value of testing face validity and about the ambiguity of testing face validity (Moiser, 1947). The ongoing discussion caused confusion about the relevance of face validity, which ended up in a discontinuation of the use of face validity. Therefore, researchers in the 1960s disapproved face validity. Instead, researchers have moved to more advanced validity procedures. Those procedures are based on established theories and can also be statistically substantiated. Researchers were searching for facts (Freeman, 1963, p.90).

The silence regarding face validity was broken by an article of Nevo (1985). The purpose of this article is to list and clarify the confusion surrounding face validity. Moreover, Nevo (1985) encourages other researchers to take face validity seriously. A point of confusion regarding face validity that is researched by Gaber and Gaber (2010) concerns the significance of the face validity test. There are researchers who argue that the face validity test is insignificant. Gaber and Gaber (2010) indicate that this is due to “its observations are not based on any empirically verifiable testing procedure”. However, more recent studies show that face validity can indeed have significant value when a “common sense” of research results is applied. Regarding this “common sense”, the question of Alderson et al. (1995) is a useful example: “on the face of things, does this research make sense?” Gaber and Gaber (2010) indicate that it is a matter of balance. Basing test evaluations on statistical analysis has the advantage of appearing scientific. The disadvantage is that it affects the intuitive judgments of the ordinary community.

Secolsky (1987) came up with a counterreaction to Nevo’s article (1985). Nevertheless, a discussion about face validity gradually started to emerge. Since then, articles have been published by both proponents (e.g., Roberts (2000)) and opponents (e.g., Newfields (2002)) of including the face validity test in a validity procedure. However, it is notable that the application of face validity tests has taken off from the 2000s. According Maginn (2006, p.2), this growth of applying face validity tests can be explained by a growing interest in investigating public participation among researchers using a qualitative approach. Subsequently, Gaber and Gaber (2010) propose the face validity test as a tool to recognize when comments of the public can be interpreted as grounded qualitative observations. Since the PVE method is both an evaluation method for policy options and a method that facilitates public participation (Mouter et al., 2021b), a face validity test is important to test whether this form of public participation appears to make sense (Gaber & Gaber, 2010). A follow-up question is why testing face validity is important.

3.1.2 The importance of face validity

In order to discuss the importance of face validity, it is necessary to first take a step back. As mentioned in the introduction, the effectiveness of policies is an important element of its evaluation (Rosener, 1978). It can be stated that the effectiveness of a policy is the extent to which an instrument or process adequately measured the goal concept (Rowe & Frewer, 2004). It is difficult to implement and achieve the intended effects of policies without support. From a policy perspective, support can also be regarded as the instrumental variant of policy legitimacy (van Damme et al., 2017). Therefore, the pursuit of policy effectiveness is a crucial reason for the government to increase the legitimacy of its policy (Buijs & Boonstra, 2020).

In this research, legitimacy will be defined as acceptability, which corresponds with the definition of Bouwhuis (2011). Following Bokhorst (2014), the concept of legitimacy can be based on laws, political processes and social support. Therefore, legitimacy can be divided into legal legitimacy, political legitimacy and social legitimacy. In this literature review, the political and social legitimacy are discussed, because a connection with (face) validity can be made from these two. Within these two types of legitimacy, the PVE is presented as an isolated method. This means that the PVE is not placed within the broader repertoire of methods that help policymakers to gain insight into what citizens think. The choice of this approach coincides with the choice to focus in this study on measuring preferences of respondents, which is part of the isolated PVE method.

3.1.2.1 Political legitimacy

The Netherlands is a country in which several groups of people who think differently about the organization of their life, live with and next to each other. Because of this difference in thinking, there are different views on what problems are and what solutions to these problems can be in terms of policies (Bouwhuis, 2011). One concept that covers these different views is pluralism. According to Mouffe (2005), the Netherlands can be regarded as a pluralistic country. She argues that it is a danger to deny pluralism. That is why politics should accept and focus on pluralism and thereby aim to guarantee the stability of the society. Democracy is stated by Mouffe (2005) as necessary to be able to recognize pluralism, since in a democracy compromises can be made between different points of view that may lead to a decision. Bouwhuis (2011) links democracy to participation. According to his research, participation is a condition for taking legitimate decisions in a pluralistic country. Quick and Bryson (2016) argue the same as Bouwhuis (2011): “one of the compelling reasons for public participation is to ensure that government policy and program choices are legitimate in terms of being acceptable to and addressing the needs of the public”. Also, Barnes et al. (2003) share this argument.

A problem recognized in literature is that it is difficult to decide who or what determines that a policy decision is legitimate. A possible manner to ensure legitimate political decisions is to set up an independent measure. However, everyone has the right to develop their own independent measure. A second manner to ensure legitimate political decision making is to involve citizens in decision-making. Citizens’ consent to the procedure and the content of the policy decision is then the source of legitimacy. It follows that participation in itself can be seen as a way of establishing legitimacy, as Bouwhuis (2011) argues. A case study by Mazepus (2017) confirms this finding by Bouwhuis. In this study, the question is raised when political authority can be considered legitimate. One of the answers of scholars is that citizen

participation should be given a role in the political decision-making of the authority to increase legitimacy (Mazepus, 2017). Buijs and Boonstra (2020) add that the legitimacy of the government as the ruler through elections does not necessarily mean that the content of a specific decision and the decision-making process for a policy are automatically regarded as legitimate.

Another problem described in literature is that it is difficult to determine what makes a decision acceptable. Christiano (2004) is one of many authors who address this problem. He splits legitimacy into two parts. The first component is also called procedural legitimacy, which addresses the acceptability of how decisions are made. The second component is substantive legitimacy, which refers to the acceptability of the content of a decision. According to Christiano (2004), a combination of both substantive and procedural legitimacy can be regarded as desirable.

Political legitimacy mainly focuses on the process of political decision-making. According to Buijs and Boonstra (2020), this type of legitimacy concerns all decision-making within the democracy. It is about the principles that are at the heart of democracy such as transparency and participation. Participation is about: "inclusivity through political equality in the presence and influence of diverse groups, the influence of citizens in various phases of the decision-making process, transparency of the process and room to arrive at a well-considered judgment based on the right information, deliberation and room for reflection." It can be concluded from this that participation can be regarded as a means of increasing the legitimacy of a political choice, but participation in itself must also be legitimate. Ozawa (2012) argues that if participation is not seen as legitimate, it can estrange the public from the government and it can disrupt the realization of policy decisions.

Since political legitimacy mainly focuses on the decision-making process, this process can be divided into three dimensions: input, throughput and output (e.g., Buijs & Boonstra, 2020; Mazepus, 2017). Input legitimacy mainly concerns the openness of decision-making to the public. According to Scharpf (1997, p.19), it is about the possibility that society is able to create its own interests known. This means that it might be possible to link the political decisions and the preferences of the society. The quality of the exchange is information is important in this case. The participants are expected to be able to explain themselves well and to use logical arguments. To this end, it is important that participants are given the opportunity to add useful arguments (Jacobs et al., 2009). Throughput-legitimacy includes the way the decision-making is shaped. Buijs and Boonstra (2020) indicate that: "policy is seen as legitimate from this dimension when the process of policy making and implementation is so well organized that it leads to good quality decision-making." For good quality decision-making there has to be transparency, attention to the concerns of stakeholders and openness to input from citizens, but also all relevant information must be available. Moreover, is important to realize that stakeholders are more likely to accept a decision of which they are convinced that this policy decision was taken in a procedurally just manner, even if it is not their preferred decision (Quick & Bryson, 2016). Output-legitimacy concerns the extent to which the policy decision increases the well-being of the society. The efficiency and effectiveness of the process and of the policy decision itself are important (Buijs & Boonstra, 2020).

The elements of political legitimacy as mentioned above can be scaled under the normative basis of legitimacy since these elements, such as transparency, are social norms. These norms justify a political decision. From the perspective of deliberative democracy, a process or decision is legitimate if it leads to the attainment of these social norms. From the perspective of participatory democracy, a process or decision is legitimate if people believe it to be legitimate. This is about acceptance. Social legitimacy is in line with the vision of participatory democracy.

3.1.2.2 Social legitimacy

Societal legitimacy is about how citizens and other actors experience the legitimacy of a political choice. On the basis of this experience, the political decision is accepted or not. Societal legitimacy therefore follows from societal debates and processes. It is because of the social debates that the details of the normative basis used in political legitimacy to determine whether policy is legitimate may change over time, person and place. Social legitimacy can also ensure that even though a political decision meets all social norms, the problems the policy decisions give in practice cannot be prevented. In some political decisions, many actors and citizens have an opinion about the legitimacy and desirability of the policy that has been proposed (Buijs & Boonstra, 2020). In addition, de Bruijn et al. (2002) argue that in the case of very complex political issues, a policy solution or process can never be objectified by all actors involved. If the policy decision is nevertheless accepted by various actors, the decision can be regarded as authoritative. This argumentation shows that it is impossible to strive for complete legitimacy in complex policy decisions, when actors cannot objectively substantiate the legitimacy of the political decision and process. Therefore, the question of whether a political decision or process is legitimate cannot be answered with a yes or no. However, legitimacy can be rated from a gradual scale, whereby the aim may be to amass more legitimacy.

3.1.2.3 Linking political and social legitimacy to face validity

The PVE method is a form of public participation. Therefore, this method is a way to ensure the legitimacy of a political decision (Bouwhuis, 2011). However, it has become apparent that the PVE method itself must legitimate as well in order to arrive at a legitimate decision. It can be stated that face validity is a part of testing the legitimacy of the PVE method.

From the perspective of political legitimacy, it appears that the quality of the information exchange is important. With regard to the PVE, it is necessary that citizens receive sufficient information in order to be able to give good advice to the politician. So, a good flow of information to the citizens provides a useful flow of information back in terms of advice from the citizens to the politicians. If the aim is to obtain relevant information in the interest of good politicians, it is important that the information provided in the PVE is face-valid. Taherdoost (2016) means by this that the information provided must be clear, legible, relevant for the citizen and consistent. Another factor for political legitimacy is the ability of citizens to express themselves well, as mentioned above. Therefore, the PVE has to offer enough space to allow citizens to express their opinion. Face validity fits well in this, because several authors focus on the completeness of the instrument and the relevance for the users of the instrument (e.g., Moores et al. (2012)) which is in this case the PVE consultation. Furthermore, if the content of the PVE consultation is inappropriate or irrelevant, there is a likelihood that the results obtained from this instrument provide false information and decisions of respondents

that are misleading for policymakers (Anastasi & Urbina, 2007). These are the consequences of an instrument that is not face valid. For example, respondents may become frustrated while completing a consultation and as a result deliberately give false answers. It follows that if the PVE consultation is not face-valid, there is a chance that citizens will not agree with the decisions that the politicians make based on the results of a PVE consultation. On the other hand, misleading information in the consultation can also cause respondents to make choices that they would not make otherwise. So, a non-face-valid PVE may lead to infectiveness of policy decisions. Effectiveness is precisely a normative value of political legitimacy.

From the perspective of social legitimacy, legitimacy deals with how citizens and other actors experience the legitimacy of a political decision. It is about the appearance of a policy decision. Face validity is about the appearance of an instrument, or the subjective judgment as Drost (2011) defines it. It can be argued that social legitimacy and face validity have the aspect of subjectivity in common. Measuring face validity can offer an opportunity to give the social perspective a place within a participation instrument. When respondents and stakeholders consider an instrument to be face-valid, this ensures acceptance of the instrument, which increases its usefulness. However, in some political decisions, many actors and citizens have an opinion about the legitimacy of a policy decision (Buijs & Boonstra, 2020). The case of the ORS can be stated as controversial and complex. This leads to the expectation that not all actors perceive the PVE method as a legitimate step in the process. According to de Bruijn et al. (2002), this also makes it impossible to expect an objective assessment from all stakeholders and citizens. Therefore, it is an impossible goal to aim for complete legitimacy. A lack of legitimacy is no longer a reason to not assign value to the results. The aim to achieve the highest possible level of legitimacy on a gradual scale is a goal to establish the authoritativeness of the PVE method.

3.1.2.4 Further arguments about the importance of face validity

Besides the argument of legitimacy, more arguments are given in the literature about why face validity is important or at least desirable. In his article, Nevo (1985) lists five reasons why a test with high face validity is preferred over a test with low face validity. A test with high face validity has first the advantage of evoking positive motivation among respondents. Secondly, a high face validity is important for attracting potential respondents. This is important one since it is necessary to have enough respondents to come to representative conclusions. A high assessment of face validity can also lead to less dissatisfaction. A fourth argument is that a high face validity will convince policy makers to carry out the test. The final argument of Nevo (1985) is that a high face validity improves public relations.

In addition to Nevo, Sartori and Pasini (2006) give three more arguments for the importance of face validity. First, there are often many opportunities within an instrument in terms of questions to ask. However, not all questions can be asked to avoid a long-lasting instrument. Therefore, face validity is useful to prioritize questions in an argument. The aim is to develop an instrument that is experienced as useful. Bannigan & Watson (2009) also indicate that an assessment of face validity is important. When respondents and stakeholders consider an instrument to be face-valid, this ensures acceptance of the instrument, which increases its usefulness. Certainly, it is important that the information that follows from the PVE consultation is not perceived as false and that the consultation is perceived as useful and meaningful by stakeholders with regard to the PVE method. It is possible that the facilitator,

often a policy maker, bases policy decisions on the results of a PVE consultation. Second, face validity can ensure that respondents' assessment needs can be better met. For example, a face validity test can provide information about certain caveats. The aim is to make the instrument more meaningful. Third, a face validity test is also a quick method to find out whether the instrument fulfills its intended purpose or not. Dempsey and Dempsey (1992) also state that a face validity test is the fastest type of determining validity of an instrument.

Shotland et al. (1998) give five arguments for the importance of face validity from the perspective of job-relevant selection tools. Because of the more positive motivation as argued by Nevo (1985), Shotland et al. (1998) add that this more positive motivation can be linked to better test results. As for the job selection tools, face validity can also increase the attractiveness of the company. The transparency and job relevance can ensure that the respondent does not feel that the purpose of the test is being withheld. Third, face validity can lead to realistic examples. The case of Shotland et al. (1998) concerns concrete job examples. A fourth argument is that within the case of job-relevant selection tools the managers within a company describe more comfort and support for more relevant selection tools. Finally, face valid selection tools are often more easily defendable when brought to court.

A concrete example of the importance of face validity is researched in the article by Sato and Ikeda (2015) about the face validity of a student exam. They argue that the test-takers' performance is lower when the face validity is also lower. The reason is that if the students state that the test is irrelevant in comparison with the study material, they will put less effort into the exam. A consequence is that the exam grades will not test the ability about the study material. This was the result of a case study in Japan and Korea. This case study has shown indeed that the results of an English exam were lower when the perception of face validity is also lower.

Opponents argue that these earlier mentioned arguments about the importance of face validity are not based on support theories, but are only subjective. It is stated by an opponent that a face validity test may be interesting, but may not be scaled under the term validity (Newfields, 2002). Despite these caveats, Sartori (2010) argues that face validity retains its own utility. Respondents should simply not think about an experiment: "what on earth is this item for?" Such items may lead to the irritation of respondents. Thereby, if the content of an instrument like a PVE survey is inappropriate or irrelevant, there is a likelihood that the results obtained from this instrument provide false information and decisions that are misleading (Anastasi & Urbina, 2007).

3.1.3 Who are the raters of face validity?

In the literature there is a point of discussion about who determines if an item or instrument is face valid. The question is whether experts or laypeople, which are people from the community, should perform the face validity assessments (Gaber & Gaber, 2010).

Proponents of using experts to perform face validity assessments argue that experts have substantive knowledge about the subject of research. In addition, experts have technical knowledge about validity testing. This knowledge can provide insights into the design and development of face valid research projects (Stallard & Rayner, 2005).

On the opposite of the proponents of using experts are the opponents. These opponents are in favour of assessing face validity by laypeople. Laypeople are also called respondents or citizens by whom the instrument is intended to be completed. These opponents are convinced that citizens who feel involved or affected are able to share observations that are of a non-technical nature. From the perspective of practical research, the observations can be regarded as relevant. Gaber and Gaber (2010) argue that the assertation of the relevance of observations from the laypeople is based on the Dewey ontology. The essence is that people's perceptions and their intelligence are secondary to people's experience. It follows from this perspective that experts cannot add more value than laypeople on the basis of their knowledge or intelligence in the area of non-technical observations. In addition, the example can be given that if someone is going to describe an item or instrument, the description will be based on someone's experience. Therefore, an objective description about the non-technical observations of an item or instrument does not exist.

Another argument that, according to Gaber and Gaber (2010), provides a foundation for using laypeople for assessing face validity is the emic-etic approach. Emic denotes a group member's point of view. This is the point of view from the 'inside'. Etic refers to the objective point of view of someone who is 'outside' the group and who researches the group (Plano Clark & Creswell, 2008, p. 290). Translated into an assessment of face validity, a respondent is a group member. A respondent, or a citizen, is part of the sample that includes all respondents. In addition, the citizens form the group that feels involved and/or affected by the subject of the instrument. In this case of experts and laypeople, the people from 'outside' the group can be referred to as experts. If there is a clear 'inside' group, it is justified for research to use an emic approach. This means that the citizens who feel involved and/or affected can be examined by means of an instrument in order to better understand the behaviour of the group from the view of the 'outsiders'. Face validity is a type of validity that can help to better understand the interaction and reciprocity between the respondents and the instrument. From this approach it makes no sense for experts to determine a face validity assessment by experts when those experts are still investigating the group of citizens itself.

A concrete example of an assessment of face validity by laypeople is reflected in the article by Connell et al. (2018). The aim of this article is to identify items within the Recovering Quality of Life instrument for people that struggle with mental health. By identifying the items, the service users of this instruments were asked which items they think are important to include. Connell et al. (2018) state that including the service users in the development of an instrument is recognised as important. However, including service users is often not done or not reported. Therefore, their article focusses on the perspective of the service user. An important argument is that what may be regarded by a researcher as a good outcome, may differ from the perspective of service uses of what is a good outcome or what is important. Only service users are able to determine whether the instrument captures the outcomes in a favourable manner. The input of service users may improve the acceptability, the relevance and also the quality of the research and the instrument itself. It can be deduced from this example that when it concerns Connell et al. (2018), the experience of the service users cannot be estimated or simulated by experts. This article presented that experts were sometimes surprised by the opinions or statements of the service users. In this case, the service users are equal to the

users of the instruments. In other words, the service users are equal to the respondents and therefore equal to the citizens.

The commitment of laypeople for the assessment of face validity took a turn in terms of civilization in the mid-2000s. Gaber and Gaber (2010) give three reasons for the noticeable turn. They present these three reasons from a plan making process perspective. First, there are more and more studies presenting how citizens' insights into policies can add value. Secondly, there is an increasing attention for mechanisms that ensure that the input for citizens lead to added value in a process. These studies focus no longer at the general value of laypeople's observations. Thirdly, researchers are focussing on applied methodological questions in evaluation the comments from citizens. With face validity it is possible to address such a methodological question. In general, this noticeable turn has resulted in an increasing number of laypeople being used to assess the face validity of an instrument.

Nevo (1985) argues that a rater of face validity must be a layperson. However, Nevo distinguishes three types of laypeople. The attitude of these three groups towards the face validity of an item or instrument would be of interest. Nevo defines that the raters are able to rate an item, an instrument or multiple instruments in terms of face validity. An instrument consists of several items. The three groups of raters are the testees, the nonprofessional users and the interested individuals. The testees include the persons who fill in the instrument(s) or item. Examples of testees are participants in experiments, job applicants or students taking a test. The nonprofessional users are the persons who work with the results of the item or instrument(s). Examples of the nonprofessional users are admissions officers, psychiatrists, but also employers. The interested individuals are also referred to as the general public. Examples of the general public are politicians, but also journalists and judges.

Moreover, Nevo (1985) explicitly argued that the term face validity must be avoided when items or instruments are rated by experts. Nevo (1985) states that: "when psychologists rate personality questionnaire items as to their subtlety or when language testing specialists consider the relevance of oral examinations to language proficiency testing, content validity seems the appropriate term." Content validity, as well as face validity, have been defined by multiple researchers before. An example of a definition of content validity by Nunnally and Bernstein (1994) is: "the degree to which a measure's items represent a proper sample of the theoretical content domain of a construct." However, in the literature is a noticeable tension between content validity and face validity. These two types of validity have been used vice-versa by some researchers. These researchers, like Nunnally and Bernstein (1994), state that the items need to be face valid in order to meet the criterion of content validity by the pool of items. This confusion between the two types of validity also creates confusion about whether a rater should be an expert or a layperson.

Despite the confusion, there is a conceptual difference between content and face validity (Hardesty & Bearden, 2004). To clarify this conceptual difference, a comparison can be made with a dartboard. The dartboard is then equated with the domain of a construct that is measured on the basis of items in an instrument. To establish the criterion of content validity, the darts must land randomly all over the dartboard. Only then there will be a valid representation of the construct being measured. Suppose that all darts land on the right half of the dartboard, only half of the construct is measured by items within an instrument. In this

case, the instrument cannot be considered as content valid. Another example is that all darts only land in the innermost rings of the dartboard. Even then, the entire construct is not fully covered and all the items are too much alike. The instrument cannot be regarded as content valid. There is agreement in the literature that it is up to the experts to assess content validity (Hardesty & Bearden, 2004).

This comparison with the dartboard can also be interpreted from the view of face validity. By doing so, an item is face valid when a dart hits the board and when the board is missed the studied item is not face valid. This is consistent with Nevo's (1985) definition of face validity. He defines face validity as the degree to which raters (the three layperson groups) think the items and instruments are appropriate to reach the construction goals. Despite the conceptual difference between content and face validity, some researchers still regard face validity as part of content validity. As a result, both experts and laypeople are used in the literature to assess the face validity of an item or instrument.

In this research, the assessment of face validity of a PVE consultation is left to laypeople. More specifically, face validity will be assessed by respondents of the PVE consultation. Laypeople were chosen because the article by Connell et al. (2018) has concluded that there can be a difference between a good outcome for experts and for service users. Moreover, since the face validity of the PVE has not yet been researched, it is important that laypeople assess because they form the 'inside' group according to Gaber and Gaber (2010). Therefore, the experts should first investigate the inside group. Another reason why respondents were specifically chosen is because it is important for the trustworthiness of the results that the items of the instrument are face valid in their view, as shown in section 3.1.2. Additionally, the stakeholders who work or has to deal with the results of the item or instrument(s) will also assess face validity in this case about the ORS. Although Nevo (1985) does not specifically name the stakeholders as raters, it can be argued that stakeholders are part of the general public. The stakeholders are, just like the general public, interested in the results of the PVE consultation. These stakeholders are involved in the case because of their interests. There are also stakeholders who are involved because of their obligations, which is the case with authorities.

3.1.4 Which methods can be used to assess face validity?

Various methods are applied in the literature to measure face validity. The three main methods are a questionnaire, a 'think aloud' interview and a focus group. These methods are further discussed below.

3.1.4.1 Questionnaire

The first method concerns a questionnaire for assessing face validity. In his article, Nevo (1985) recommends the use of a questionnaire. Nevo used respondents to rate an item or an instrument based on a 5-point Likert scale in a questionnaire. The points on the 5-point Likert scale have the following meaning: "5-the test is extremely suitable for a given purpose; 4-the test is very suitable for that purpose; 3-the test is adequate; 2-the test is inadequate; and 1-the test is irrelevant and therefore unsuitable" (Nevo, 1985). Moores et al. (2012) also apply a 5-point Likert scale to assess face validity in a questionnaire. Their article is about the QQ-10 questionnaire. The QQ-10 contains ten key themes that have been translated into ten statements about the opinion of patients about the use of questionnaires. The QQ-10 is an

instrument to assess face validity, feasibility and utility (Moores et al., 2012). Another example where a questionnaire is applied, is the article by Desai and Patel (2020). Unlike the examples above, this article does not use a 5-point Likert scale but instead include statements that needs to be answered with yes or no. Ten criteria have been drawn up in this article that jointly measure face validity.

An advantage of a questionnaire is that the researcher is able to ask specific questions. Therefore, the researcher is able to guarantee that the entire concept of face validity is assessed (Marshall, 2005). Another advantage is that the questionnaire is a less intensive method than the two methods discussed in the following paragraphs. A disadvantage of a questionnaire is that the respondents do not explain their answers in real depth (Morgan, 1996). As a result, a questionnaire generates more global results. However, it depends on the purpose of a research whether this is a real disadvantage.

3.1.4.2 'Think aloud' interview

The second method concerns a 'think aloud' interview. In a 'think aloud' interview, the participant is asked to literally think aloud while assessing the items of an instrument. This means that the respondent verbalizes his or her thoughts. The thoughts normally remain unspoken during the process of completing items. Therefore, a 'think aloud' interview is led by the participant and not by the researcher (Horwood et al., 2014). In this way, the dynamics of filling in the items are not disturbed and that is why filling in the instrument with a 'think aloud' interview can be compared to a situation in which the instrument is filled in without the interview taking place. Ericsson and Simon (1993) agree that a 'think aloud' interview should not change the performance of the task. Due to a 'think aloud' interview, performing a task may at most take a bit longer because of expressing thoughts by respondents. That is why the researcher and the respondent are not in the same room when conducting the interview. So, there is no interaction between the respondent and the researcher. Additionally, the participants are told not to plan what they say, but to speak freely. Furthermore, the respondents are not asked to justify what they are doing and to explain their strategies (Horwood et al, 2014).

A disadvantage of the lack of interaction between respondent and researcher is that there is no possibility to discuss the issues that respondents experience in depth (Austin & Delaney, 1998). Subsequently, a 'think aloud' interview is of a qualitative nature. Therefore, it is possible to identify the nature and existence of issues. However, it is not possible to provide information about the impact and the extent of the issues mentioned in the instrument. Another disadvantage is that a 'think aloud' interview relies on the participants verbalizing their thoughts. It is not possible to detect issues that are encountered by respondents but there are not verbalized by them (Horwood et al., 2014). Contrary to these drawbacks, the 'think aloud' interview is described as useful in identifying the face validity of an instruments and possible improvements. Moreover, the 'think aloud' study provide insights into the thinking process of the respondents (Kaklamanou et al., 2013).

3.1.4.3 Focus group

The third method concerns a focus group. Compared to the questionnaire and the 'think aloud' interview, the focus group is relatively the least applied. A focus group can be defined as "a research technique which collects data through group interaction on a topic determined

by the researcher" (Morgan, 1996). A focus group is usually used in combination with another method, such as a questionnaire or interviews. A concrete example of a case in which focus groups are used in combination with interviews is presented in an article by Connell et al. (2018). In this article, focus groups are used to clarify an justify the results of the individual interviews. In the article of Belone et al. (2016) focus groups were used not only to determine where respondents experience issues in the instrument but also to discuss solutions to these issues. In this method, the researcher takes an active role, which differ from the 'think aloud' interview. Another comparison with the 'think aloud' interview is that in the interview the existence of issues is interpreted. In a focus group, the main point of interest is the sources of certain motivations and therefore also on the impact of possible issues. This is a strength of a focus group. Another strength is that the researcher is able to make an observation about the diversity and consensus among participants. Furthermore, it is a strength that the researcher is able to ask the respondents themselves for comparisons among issues instead of aggregating data of individuals or speculation whether the individual data differ from each other (Morgan, 1996).

In addition to the advantages, the focus group also has disadvantages. The main weakness of a focus group is that a polarization effect may arise. This means that parties within the focus group become even further apart due to a difference in opinion. It is up to the moderator to determine how to deal with those differences. Another disadvantage is that the moderator has an important role in what is and is not discussed within the focus group. This may lead to individuals within this group not feeling fully heard (Morgan, 1996).

3.1.4.4 Method to assess face validity in this research

In order to determine which method is used to assess face validity, it is important to consider the purpose of a study. The results of a questionnaire can provide a global picture of how face validity is assessed. A 'think aloud' interview, on the other hand, is useful for taking a step-by-step look at a respondent in order to find out how someone experiences the PVE consultation. With focus groups, the focus is on the underlying sources of certain motivations. These groups can provide a deeper understanding of the respondents' experience. In this research, the face validity of the PVE method is assessed within a case study. Since the face validity of the PVE method has not been researched before, it is first important to gain a global picture of the current state of face validity of the PVE method. Therefore, in this study a questionnaire is used to create a global picture of the assessment of face validity of the PVE method.

3.1.5 When to assess face validity?

According to psychology, it is one of the biggest challenges to validate an instrument or questionnaire. The instrument has to be assembled in a way that it is psychometrically sound (Tsang et al., 2017). Validated questionnaires or instruments are also often used in healthcare. An example of an invalid questionnaire is about a questionnaire that measures the food intake of people. This list can be invalid because it measures what people say they have eaten and not actually what people have eaten (Boynton & Greenhalgh, 2004).

From psychology, several steps have been identified to validate an instrument or questionnaire. The following three steps are the most important when it comes to validating. When an instrument or questionnaire has been drawn up by researcher, it first is assessed by an expert committee. It is up to the expert reviewers to rate the items within the instrument

or questionnaire. An important part is that the experts assess whether items can be experienced as biased by respondents. After the expert committee, there is preliminary pilot testing. This preliminary testing focuses on a small sample of the intended participants. These preliminary tests can be used to determine whether there is confusion about certain items and whether respondents have ideas for improvements. An image can also be obtained of the response distribution per item. After the preliminary pilot testing, the items are revised. Several consecutive rounds may be required before proceeding to the next step, which is a pilot test for initial validation. The pilot test for initial validation contains more intended participants than the preliminary pilot testing. Furthermore, the pilot test for initial validation is the phase in which types of validity and reliability are initially measured (Tsang et al., 2017).

Although the third step is specifically about validity and reliability, it can be argued that face validity is related to the goals that emerge in the preliminary pilot testing following the definition of face validity by Taherdoost (2016). He states that face validity "evaluates the appearance of the questionnaire in terms of feasibility, readability, consistency of style and formatting, and the clarity of the language used." In other words, face validity is about assessing items, which is also the main purpose in the preliminary pilot testing. A concrete example of the assessment of face validity in the preliminary pilot testing can be found in the development of the Child Oral Health Impact Profile (COHIP) questionnaire (Broder et al., 2007). After the initial pool of items has been drawn up by the researchers, experts are approached, as well as a sample of intended participants, to assess face validity. Ultimately, two rounds are needed to achieve face validity. After these two rounds, the next phase for the questionnaire follows, namely the initial validation by a pilot test.

However, Moores et al. (2012) state that their QQ-10 instrument to measure face validity can be applied during the development of an instrument, but also during or after the implementation. This means that face validity cannot only be assessed during the preliminary pilot testing, but also after the introduction of the instrument. Also, Del Greco et al. (1987) state that it is important to considerate to measure face validity for both the pilot test and the final instrument. A concrete example of retrospective assessment of face validity can be found in the article by Tweed and Cookson (2001). In this article, medical students and their examiners are enabled to assess the face validity of a professional final exam after they took or graded this exam.

In this research, the face validity will be assessed afterwards. This is due to the limited time available for this research. There is no time to complete several consecutive rounds of preliminary pilot testing to improve the face validity, should the need arise. Therefore, it is decided to evaluate the face validity afterwards in order to be able to provide a global picture of the assessment of face validity by respondents in the PVE method.

3.1.6 The measurement of face validity

If the goal is to measure face validity, there are two general approaches that can be applied. In the 'absolute' approach it is the task of the rater to assess an instrument or an item on face validity. In the 'relative' approach the rater is asked to judge the face validity of multiple instruments simultaneously. In this way, the different instruments can be compared to each other. A condition is that the rater is familiar with all the instruments that are being compared (Nevo, 1985). An 'absolute' approach will be applied in the research. Therefore, the raters are

presented with one instrument that they fill in and of which they assess the face validity. This ‘absolute’ approach is applied because in this case the respondents will complete one PVE consultation.

However, specifying to an ‘absolute’ approach does not indicate how face validity can be specifically questioned. In the literature, categories can be distinguished within face validity. In this way, the concept of face validity can be concretised. Table 3.1 presents the different categories that can be identified from the literature and shows which categories can be identified in which article. When there is a cross in a cell of the table, it means that this category occurs in that article.

Table 3.1: Categories of face validity identified from the literature

Article ↓; category →	Clarity	Relevance	Readability	Appropriateness of difficulty level	Unambiguity	Aesthetics	Completeness	Feasibility	Efficacy	Insensitivity	Familiarity
Banna et al., 2010	X		X		X	X					X
Bannigan & Watson, 2009	X	X		X							
Broder et al., 2007	X	X	X				X				
Chabrol et al., 2005	X				X		X				
Connell et al., 2018	X	X			X				X	X	
Del Greco et al., 1987						X					
Desai & Patel, 2020	X	X	X	X	X	X					
Drost, 2011			X	X							
Engström et al., 2018	X	X	X				X				
Frantz & Holmgren, 2019		X	X	X							
Hardesty & Bearden, 2004	X	X	X		X						X
Hojat & Gonnella, 2011	X	X					X				
Holloway et al., 2014	X	X	X	X	X			X			
Kennedy et al., 2019	X	X	X					X			
Maithel et al., 2006		X				X	X		X		
Moores et al., 2012	X	X		X			X	X		X	
Mousazadeh et al., 2017			X	X	X						
Nevo, 1985		X									
Oluwatayo, 2012	X	X	X	X	X	X					
Pelet et al., 2012	X		X	X							
Sarantopoulos et al., 2017	X										
Sartori, 2010						X			X		
Sato & Ikeda, 2015	X				X						
Taherdoost, 2016	X	X	X			X		X			
Tweed & Cookson, 2001	X	X		X				X			
Number of times identified:	*18*	*16*	*13*	*10*	*9*	*7*	*6*	*5*	*3*	*2*	*2*

3.1.6.1 Categories and statements of face validity

The categories, as identified in table 3.1, serve to concretize the concept of face validity. A total of twenty-five articles have been included in this table. In order to arrive at a framework, i.e. a set of questions or statements, the categories must be further specified in specific statements. Despite the fact that categories of face validity can be identified from the literature, the specific inquiry in the form of questions or statements is often missing (Desai & Patel, 2020). In addition, the aim is to formulate a set of questions or statements to specifically measure face validity regarding the PVE method. Therefore, the identifiable categories are first explained below, after which a statement is drawn up for each category that focuses specifically on the PVE method. If it appears necessary to provide case-specific information about the ORS case for clarification, this will be added to the statements.

Clarity

The category clarity emerges eighteen times in the twenty-five articles included in this literature review to identify categories. This makes clarity the most common category of face validity in these articles. state that clarity means that an instrument or item must be understandable (Pelet et al., 2012). This includes that a proper instruction must be available (Desai and Patel, 2020). Also, Hardesty and Bearden (2004) state that an item must be deleted if this question is not clear. A case study on a new assessment tool to examine future doctors indicates that when assessing the face validity of this new tool, the clarity of the examiners' instructions should be questioned to the students (Tweed & Cookson, 2001). Furthermore, Oluwatayo (2012) has formulated ten criteria that may be included in an assessment of face validity in his research. It is striking that Oluwatayo has included clarity and unambiguity in the same criterium. Unambiguity will be explained later in this research. Moreover, another criterion of Oluwatayo addresses the adequacy of the instruction, which overlaps with the clarity Desai and Patel (2020). So, clarity can be about whether an item or instrument is understandable, but according to Oluwatayo (2012) this has a lot of overlap with unambiguity. In addition, clarity can concern the adequacy of instruction. Instruction can be interpreted as information that is given to help completing an item or instrument, such as in the case of the future doctors. For the PVE method specifically it is important to provide respondents with sufficient information about the options they can choose in a choice task in order to be able to provide a meaningful advice (Nouws, 2019). Therefore, the following statement can be set up to measure the category clarity in case of the PVE method: I have received sufficient information to make a choice.

Relevance

The category relevance emerges in sixteen of the twenty-five articles included in the process of identifying categories of face validity. It turns out from the literature that relevance can be interpreted in three ways. The first way is apparent from the article by Kennedy et al. (2019) in which a food skills questionnaire is developed, validated and the reliability is tested. In the validity test, 85% of the respondents indicated that the questions in the questionnaire were about food skills. In other words, the questions were found to be relevant by the respondents with regard to the purpose of the questionnaire, which is to measure food skills. Some articles, e.g. Hojat and Gonnella (2011), give respondents the possibility in a pilot test to delete items if they think those items are not necessary. The second way emerges in the article by Chabrol et al. (2005), examining the face validity of the Defense Style Questionnaire (DSQ). In this study, respondents were asked to link items from the questionnaire to defense mechanisms.

When the respondents, which were clinicians, had responded with different defense mechanisms, this item was removed. According to the researchers, the item was not realistic or relevant enough to include. When the items are realistic, they can be defined as relevant items. The third way emerges in the article by Moores et al. (2012) in which a ten-item instrument was set up to assess the patient's view on the use of questionnaires in health care. Two of the items are to what extent a questionnaire has helped to improve communication with the treating physician and to what extent the questionnaire is valuable to their treatment. If the patient has the idea that the questionnaire is helpful for the treatment, it can be regarded as relevant. A concrete example of a statement that addresses the third perspective of relevance can be found in the article by Tweed & Cookson (2012) about the exam of future doctors: "The exam was a good assessment of my competence as a future doctor." In case of the PVE method and this case study, the questions contribute to the goal of the consultation as these have been drawn up together with the client. Furthermore, there is by definition no question of a realistic experiment since the choice tasks that the respondents are given will take place in the future. But the third perspective of relevance is important. In previous PVE consultations questions were asked that are similar to this relevance. This leads to the following statement: I think this is a good method for expressing my opinion on how citizens should be involved in decision-making about Schiphol.

Readability

Readability is included as a category in thirteen articles out of twenty-five. Following Mousazadeh et al. (2017) readability can be regarded as using proper terms and proper grammar. Engström et al. (2018) indicate that readability is about the wording. It can be added that the wording must be free of technical jargon to improve the readability (Holloway et al., 2014). However, Oluwatayo (2012) distinguishes between two criteria that both relate to legibility. One criterium is about the correct spelling of difficult words and the other criterium is about the readability in general. Because there is only limited space for measuring face validity in the PVE consultation that is used as a case study in this research, it is decided to focus on the readability in general. The aim is to obtain a global view of the assessment of face validity of the PVE method. The statement is as follows: I found the choice task understandable.

Appropriateness of difficulty level

The category appropriateness of difficulty level emerges ten times in the twenty-five articles included in this literature review to identify categories. The appropriateness of difficulty level is also described in the literature as easy to response, easy to answer or easy to complete (Connell et al., 2018; Frantz & Holmgren, 2019; Moores et al., 2012). When items are not easy to answer, there is hindrance of understanding or hindrance of completions. This is often the result of items that are too abstract. Items in an instrument are then vaguely or insufficiently defined (Holloway et al., 2014). As a result, the items cost too much thought, which means that respondents are sometimes unable to finish the items. The PVE method focuses on making choices in a choice task. That is why the statement in this category deals with the ease of making choices: I found it easy to make a choice.

Unambiguity

Unambiguity is included as a category in ten articles out of twenty-five. Several articles included in this review describe ambiguity as a factor that decreases face validity (e.g., Banna

et al., 2010; Hardesty & Bearden, 2004; Mousazadeh et al., 2017). Therefore, the aim is to achieve unambiguity in an item or instrument. However, this category is not further explained in these articles. But Connell et al. (2018) do specify ambiguity. According to them, there is ambiguity if item can be interpreted in several ways. A concrete example of ambiguity is described in the article by Sato and Ikeda (2015). In a language test, test takers regarded some questions as items intended to measure the indirect writing ability as items to measure the reading ability. With regard to the PVE method, it is expected that it is a difficult statement to assess ambiguity when it is about the entire consultation. This is inadvisable with regard to the appropriateness of the difficulty level because of too much thought. That is why the statement regarding ambiguity is made specifically for a choice task: I found it clear with the choice task what was meant by each task.

Aesthetics

The next category identified is aesthetics. This category is mentioned in seven out of the twenty-five articles in total. First, this category can contain the interplay of the text and the images within the instrument. When images do not provide sufficient support for the text, the aesthetic value of the instrument decreases (Banna et al., 2010). Second, the aesthetic value of an instrument or item in the literature is linked to its professional appearance according to Del Greco et al. (1987). In addition, when an instrument looks professional, it provokes serious answers from the respondents. Third, the attractiveness of an instrument or item appears to play a role with regard to face validity. Attractiveness is one of the ten criteria that Oluwatayo (2012) has drawn up to measure face validity. Sartori (2012) also states that attractiveness plays a role. He compared projective techniques with psychometric techniques with regard to face validity. It turned out that projective techniques are perceived as more interesting, mysterious, beautiful and attractive by respondents. Since attractiveness was most frequently mentioned in the twenty-five articles of this literature review in terms of aesthetic value, the statement for this category will focus on that. The statement is as follows: I thought the platform in which I made the choice task looked attractive.

Completeness

Completeness has been identified six times as a category of face validity in the twenty-five articles included in this review. According to Maithel et al. (2006), completeness is about the ability to give feedback on the items. Is the instrument complete or are there items missing? For example, it appears more often in the literature that respondents are offered the option of adding items to the instrument in the (preliminary) pilot testing (e.g., Chabrol et al., 2005; Broder et al., 2007). Since this study measures the face validity of the end product, i.e. the final PVE consultation, the perspective of adding items does not seem appropriate. However, Holloway et al. (2014) indicate that completeness is about whether an instrument included all respondents' opinions and concerns. The statement regarding completeness focuses on the opinions, because the intention is that respondents can give advice to the decision-makers on the basis of the PVE method. Therefore, the statement is as follows: I felt that I could give all my opinions on how citizens are involved in decision-making about Schiphol and how information should be provided. The information provision refers to the Environmental House.

Feasibility

The category feasibility emerges five times in the twenty-five articles included in this literature review to identify categories. It is striking that in the included literature the term 'feasibility' is mentioned, but in most of the articles not further explained (e.g., Taherdoost, 2016). However, Tweed and Cookson (2001) touched upon the feasibility category by adding a statement in their face validity assessment questionnaire that asked whether there was enough time for examinees to demonstrate what they wanted. In the article by Kennedy et al. (2019) on the food skills questionnaire, respondents were asked whether the length of the questionnaire was appropriate. So, both articles discuss the feasibility of the length of a questionnaire or exam. Since this is the only perspective of feasibility that has been identified, the statement in this study is about the same perspective. Therefore, the statement of feasibility is as follows: It was doable to complete the survey within twenty minutes. At the start of the consultation, it is indicated that it will take approximately twenty minutes to complete the PVE consultation.

Efficacy

Efficacy is a category that occurs significantly less often in the literature, only three times in the twenty-five articles. Sartori (2010) is the only researcher that provides more explanation for this category. In his article, Sartori (2010) looks for the differences in face validity between projective techniques and psychometric techniques. It turns out that psychometric techniques are perceived by respondents as more scientific, more transparent, more repeatable, more credible and more prone to forgery. Since the PVE method is a scientific research method, the statement of this category focuses on the scientific level. The statement is as follows: I thought that the choice task is of a scientific level. The statement concerns a choice task as it touches on the scientific core of the PVE method and because the assessment then remains manageable for the respondent with regard to avoiding too much thought.

Insensitivity

Insensitivity has been identified two times as a category of face validity in the twenty-five articles included in this review. Connell et al. (2018) indicate that one of the most common reasons to object an item is sensitivity. Some items can be perceived as too positive or too negative by respondents, resulting in upset feelings. The aim should be to make an instrument appear as 'neutral' or as insensitive as possible. Moores et al. (2012) echo this by confirming that when respondents have a negative experience while completing an instrument, the results are also more negative. A negative experience could be that the instrument was embarrassing or upsetting. The negative feelings that an item or instrument can evoke are in the center of the statement for this category: I felt that the choice task evoked too many negative feelings in me.

Familiarity

Like the insensitivity category, the familiarity category has been identified twice as a category of face validity in the twenty-five articles included in this review. Moreover, familiarity is the last category of face validity that is identified and included in this review. Mousazadeh et al. (2017) and Hardesty and Bearden (2004) both indicate that familiarity with the concepts that appear in an instrument influence the face validity of that instrument. More familiarity would lead to higher face validity. The statement for this category is as follows: Before I filled in this

choice tasks, I was already familiar with the idea of setting up a Schiphol Social Council and an Environmental House.

Relativity and acceptability

Relativity and acceptability are two categories of face validity that have been identified in the twenty-five articles, but are not included further on in this study. Relativity was only mentioned in the article by Mousazadeh et al. (2017). However, this category was not given any further explanation. As a result, the category remained to vague to include it in this research. Acceptability is a category that was more common mentioned in literature, but is always strongly linked to one of the other categories. For example, Holloway et al. (2014) deals with the acceptability of wording, which is closely related to readability, completeness etc. Bannigan and Watson (2009) mention the acceptance of the text. This can be related to the relevance, appropriateness of the difficulty level, readability, etc. Because of this overlap, it is decided to leave the acceptability out.

3.1.6.2 Overview of categories linked to statements of face validity

The categories identified above with their specific statements, taking into account the PVE method and the ORS case study, are presented below in table 3.2.

Table 3.2: Categories of face validity with their statements

Category of face validity	Statement
Clarity	I have received sufficient information to make a choice.
Relevance	I think this is a good method for expressing my opinion on how citizens should be involved in decision-making about Schiphol.
Readability	I found the choice task understandable.
Appropriateness of difficulty level	I found it easy to make a choice.
Unambiguity	I found it clear with the choice task what was meant by each task.
Aesthetics	I thought the platform in which I made the choice task looked attractive.
Completeness	I felt that I could give all my opinions on how citizens are involved in decision-making about Schiphol and how information should be provided.
Feasibility	It was doable to complete the survey within twenty minutes.
Efficacy	I thought that the choice task is of a scientific level.
Insensitivity	I felt that the choice task evoked too many negative feelings in me.
Familiarity	Before I filled in this choice tasks, I was already familiar with the idea of setting up a Schiphol Social Council and an Environmental House.

3.2 Results expert interviews

The purpose of the expert interviews is to prioritize the categories identified in the literature review. Since the ORS case study, which is used in this research, is commissioned by a client, there is no place to measure all categories. Five categories can be measured on the basis of statements in the PVE consultation of this case study. Therefore, the experts are asked to

select the five most important categories for measuring face validity for the PVE method specifically. The literature from which all the categories have been identified do not specifically focus on the PVE method, while one of the goals of this study is to measure face validity of the PVE method specifically. Like the literature review, the expert interviews also contribute to answering the first research sub-question.

The expert interviews protocol consisted of three parts (see Appendix C). These three parts are also maintained in the structure of this section in which the results of the expert interviews are discussed. First, the completeness of the identified categories from the literature review is discussed. This is followed by the argumentation of the prioritizing of the categories by the experts. Thereafter, the remarks and comments on the statements of the prioritized categories are discussed.

3.2.1 Completeness of the categories

Three of the five experts interviewed indicate that the list of categories identified from the literature review (as presented in table 3.2) is complete.

However, one of the five experts suggested a category that does not appear in the already identified categories from the literature review. It concerns the acceptability category, or "the extent to which people are okay with it". According to the expert, acceptability can be interpreted in two ways. On the one hand, acceptability is about whether people agree with using this method, which is the PVE method. On the other hand, acceptability is about "whether people agree with the advice that they have given, by means of this method". As far as the expert is concerned, acceptability is more of a "background characteristic category". So, acceptability is not one of the most important categories, according to the expert, but it is important not to overlook it. Furthermore, the expert indicates that acceptability overlaps with other identified categories, such as aesthetics. The expert states that the respondent accepts the aesthetics if "you see the PVE in a way that you would like to express your opinion and do not feel pushed in a certain direction". This is about the acceptance of a design. Moreover, acceptability can also be related to the insensitivity category: "suppose the PVE evokes a lot of negative feelings, you probably find the PVE unacceptable." Here it is about the acceptance of the feelings that the PVE evokes.

Based on the interview with this expert, it is decided not to include acceptability as a category of face validity in this research. The main argument is that acceptability overlaps with the other categories. When it comes to acceptability, the following question has to be asked: acceptability in relation to what? The interpretation of 'what' consists of the categories that are included in the list of identified categories (as presented in table 3.2).

The same expert also indicated that the category legitimacy lacks. Legitimacy, according to this expert, is about "people feeling that the PVE gives them the opportunity to contribute in some meaningful way while they are in the shoes of the policymakers." If respondents feel that their advice does not lead to an impact on the decision-making, the respondents will not feel the urgence to participate in a consultation. According to this, it can be stated that the category of relevance which is included in the identified categories from the literature review, is closely related to legitimacy. Another expert confirmed that legitimacy matters and that relevance and legitimacy are closely related. However, a difference between the two is that

relevance can be linked to appropriateness and legitimacy to meaningfulness. Due to the limitation of the number of statements that may be included in the case study, one expert advised to include either relevance or legitimacy. The other expert that mentioned legitimacy as a missing category, advised to include the relevance. The literature refers more often to relevance than to legitimacy.

Based on the interviews with these two experts, it is decided not to include legitimacy in this research because of the overlap with relevance. Following the experts, it would be better to ask for other categories than to include both relevance and legitimacy. Since relevance is mentioned in the literature and legitimacy often appears in the background, relevance is the one included as a category.

3.2.2 Argumentation of the prioritizing of the categories

In this section, the focus is on the argumentation of the prioritizing of the categories. The five categories that are selected by each expert are presented in chapter 3.3. Strikingly, all five experts started their argumentation by stating that every category identified in the literature review concerns something that is important.

With regard to clarity, it is in any case important to ask: “have you received enough material or information to be able to express your preferences and to make a choice?” If participants have the feeling that there is information missing to make a certain reasoning or choice, the conclusion can be drawn that the PVE method does not provide the information well. It can be considered a tolerated reason if someone states that if information is lacking, when the person has a look at the PVE, and therefore concludes that the PVE method is not face valid. “In that respect, face validity and clarity are easily linked.” There is one expert who states that the statement with regard to clarity is complicated because the need for information provision differs per respondent. “For example, a distinction can be made between very involved citizens who always feel that there is too limited information and the less involved citizens who are quickly overwhelmed by the amount of information.” According to this expert, this category would be important if a PVE is designed for a specific target group and not for the general public. Otherwise, the question has to be asked whether it is at all feasible to achieve clarity in the view of every respondent.

The two experts who mentioned legitimacy as an extra category above, are of the opinion that relevance should be included because of the relationship between legitimacy and relevance. “People should also really see for themselves the relevance of what they participate in or what they give their opinion about. Precisely because citizens are asked for their advice, they should be convinced that this method is a good way to give their advice.” Another expert indicates that relevance is important from a practical point of view, since policymakers want to know whether the PVE is a good method for allowing citizens to express their opinion.

When the readability is questioned, this provides information “whether the entire PVE or a choice task was easy or difficult to read and interpret”. According to an expert, readability can be about the difficulty of the words, the length of the sentences, the grammar of the sentences, etc. If one or more items of this legibility do not seem to make sense at first sight, the respondent may be disturbed. This is a reason for a lower rating of face validity. When the aim is to get a global picture of the face validity of the PVE, it is recommended by an expert to

question the readability of the entire consultation. Moreover, it turns out from the experience of another expert that people often comment on readability in relation to another person. The expert describes this as follow: “At PVE, many people have an opinion about what is readable by other people. People indicate that they have the confidence in themselves that they are able to complete the PVE in a good manner, but their neighbour is not able to do that.” A third expert who is engaged in the validation of an instrument indicates that readability has been a “hassle” within her instrument. Because her instrument and also the PVE method are intended for a broad audience, the complexity must be reduced to keep the instrument legible for everyone.

The same argument about the incapability of a neighbour that came to the fore in readability also plays a role in the appropriateness of the difficulty level. This is a category that has already been questioned in previous PVE consultations, as “policymakers are concerned about the difficulty level”. This category is important from a practical nature. It is stated by an expert that the appropriateness of the difficulty level is related to readability and clarity. When this category is added with the associated statement, information is retrieved that contains elements of multiple other categories like clarity. Therefore, the appropriateness of the difficulty level is more inclusive than other categories, but this causes obstacles for further analyses. The clarity category, on the other hand, is more demarcated. Further analyses with the results will have fewer obstacles. That is why an expert opts for clarity instead of appropriateness of the difficulty level.

Unambiguity is considered important by the majority of the experts. Certainly with regard to the PVE in which participants make an integral choice. This means that many options have to be considered at the same time, which creates the danger that the participant loses the overview. Moreover, when a question is ambiguous, this leads to irritation among the respondents. As a result, respondents may not complete the consultation. Another reason to include ambiguity is that “respondents often already have an idea or an association with something. People quickly classify things under something that you as a researcher did not intend.”

The argument aesthetics was selected by an expert is because the previous experience of this expert shows that respondents sometimes feel framed. For example, in previous PVE consultations, respondents stated that certain examples were incorrect. Another example is about an image: “one tree was greener than the other as a result of which respondents were linking the greener tree to the idea that an option was more environmentally friendly. This was actually not the case.” So, if the aesthetics are not right according to the participants, they will make choices that they would not make otherwise. This is closely related to the appearance of the PVE, “the design of the PVE should look neat but also neutral.” Another expert picks up on this by stating that the aesthetics do not necessarily have to be “very fancy”. “It is simplicity that positively influences the credibility of an instrument.” For example, “if you make the appearance too complex, it can feel like a threat to some respondents.” Moreover, if the aesthetics become too playful, respondents take the PVE consultation less seriously. However, aesthetics feels like an afterthought to the other experts.

Completeness is strongly related to clarity in the eyes of an expert. Therefore, this expert indicates that it is important for clarity to focus on sufficient information and for completeness

to be comprehensive with regard to concerns or opinions of respondents. In this way, there is sufficient distinction between the categories completeness and clarity. This expert points out that from a lot of answers that participants have given in other PVE consultations, it appears that people become frustrated when they are not enabled to express all their concerns or opinions. “Sometimes respondents feel backed into a corner with respect to the options they can choose from.” From the perspective of the PVE method, this has two reasons. First, a PVE consultation is often not a “point zero”. Possible policies are already on the table of the policymakers. These policies are then included in the PVE consultation. Second, a respondent is unable to make well-founded choices when too many choice options are presented. Moreover, another expert confirms that there are many discussions about the completeness of a PVE consultation. These discussions especially take place because of the many different perspectives on the level of detail that should be applied to ensure completeness. A third expert states that “a danger with completeness is that researchers frame their research too much. Studies have shown that expert knowledge sometimes influenced the results too much.” However, it has not yet been tested in a previous PVE consultation whether respondents could express all their opinions or concerns. Suppose that according to citizens the consultation is incomplete, this could be a reason for them to object that the results are skewed. “If other or more options had been included, the entire outcome would have been different.”

As far as feasibility is concerned, “it does not matter much if a respondent takes longer than twenty minutes if this respondent is fine with that”. On the other hand, “if a PVE takes too long, citizens do not complete the PVE.” In addition, an expert indicated that twenty minutes is quite long and someone would be more inclined to participate in a consultation that takes fifteen minutes.

Efficacy is initially linked by an expert to the question “what the person asking the questions in the PVE wants to hear.” With regard to the scientific perspective of efficacy, the researcher needs to be careful when a narrow research question is addressed in a PVE consultation. Respondents can experience that the consultation is missing relevant questions. So, it can be argued that social efficacy is separate from scientific efficacy. A concrete example is set in the province of Gelderland, where the province is currently working out a climate plan. There are many options within this climate plan. That is why has been decided to only present the options in a PVE that are still under discussion. This approach is effective, but respondents will “undoubtedly respond that the consultation is incomplete”. Because of the tension between the social and scientific perspective of efficacy, it is important that this category is included in this research, according to an expert. Another argument to include efficacy is that this statement can be answered well by a broad audience. The argument about the separation of social and scientific efficacy is a reason for another expert to not include efficacy. This expert states: “it is less important for respondents whether a consultation is of a scientific level than being able to influence the results.”

Since the PVE method is mainly used in the context of participation processes, it is important that every respondent feels heard in such an instrument as the PVE method. An example of this happened in a case of an expert where “someone was insensitive to the sentiment of people who were against windmills.” As a result, people broke off this research on windmills. “When a consultation does not evoke negative feelings, everyone is able to participate.” When

certain groups of respondents do not participate because of the negative feelings, the results will be less reliable. This is a difficult task since the PVE is mainly applied in controversial cases. Therefore, another expert states that there are by definition negative feelings around a PVE consultation. Nevertheless, according to a third expert, it remains important to give space to all the feelings that a consultation can evoke and to strive for insensitivity.

An expert selected familiarity as one of five categories. According to this expert, familiarity is related to clarity. “If you are more familiar with the problem, you will already have more information about the problem. So, the clarity category plays a less important role.”
An argument given by another expert for not selecting the familiarity category is that it may be of interest to researcher, but it is not relevant or interested for the respondents. In addition, according to an expert, extra information is already given in a PVE consultation to inform people who were not yet familiar with a concept.

In general, most of the categories are chosen by the experts because in the case that these categories are not fulfilled, this can lead to an outcry among the respondents. This could lead to the consultation not being completed or that the results obtained with the consultation are skewed or could contain incorrect information.

3.2.3 Remarks and comments on the statements of the categories

The experts were initially asked to critique the statements belonging to their five chosen categories. However, it turned out that some experts were interested in criticizing the other statements as well. These comments and remarks are also included below.

In general, it was recommended by the experts to omit technical jargon. An example of jargon is the choice task. Despite the fact that the consultation explains what a choice task is, there is a chance that some respondents will not understand this term. A similar comment was made about the statement of appropriateness of the difficulty level. The statement states that it is about a choice. It was advised to further specify this choice.

With regard to the statement about clarity, all experts confirm that they would also ask for sufficient information. However, this statement can be applied at several levels: “about an individual question, about a component such as a choice task or about the PVE in general.”

The statement of relevance as it was initially drafted, leads to confusion among the experts. According the experts, this statement questions several goals of that a PVE may have. It is their advice to focus on one clear goal. Examples of goals that were listed by the experts to focus on were involving citizens in decision-making or asking citizens about their preferences. It is important to clarify this goal, because respondents also fill in the PVE consultation on the basis of this goal. Otherwise, it remains unknown to which part of the statement a respondent answers. It seems that the statement as formulated in the first instance focuses on the goal of participation, i.e. involving citizens in decision-making. When asking a statement about citizens’ preferences, the aim is to evaluate the chosen options in a choice task and to ask about preferences. The PVE method can be good at one goal, but not good at another goal from the participant’s perspective. An expert advises to focus in particular on the quality of the choices that respondents make, since the participation process can be broader than the

PVE itself. A proposition of a statement is: “I was able to communicate my choices about this theme in a good way to the decisionmakers who will look at the results of this PVE.”

For readability, the comprehensibility can be questioned. According to an expert, it is also an option to ask whether the PVE consultation is easy to interpret. Another expert suggests as a statement: “the consultation was easy to read and was understandable.” According to this expert, there are two successive steps. First it must be legible and only then a person is able to assess whether it is understandable.

With regard to the statement about the appropriateness of the difficulty level, an expert is convinced that it is more important to question whether someone is convinced of his or her choices instead of making choices easily. This expert states: “policymakers also find it difficult to make a choice sometimes.” In addition, a PVE puts citizens in the shoes of a policymaker in order to experience what it is like to make policy choices. Therefore, finding it difficult to make choices is linked to the complexity of the theme and that should not be the intention of this statement.

In the statement of unambiguity, it must be clear that it is about each task in itself, as commented by an expert.

The statement regarding aesthetics is in line with the expert’s idea. Questioning neutrality of the PVE, as emerged in the argumentation of selecting aesthetics, is considered too complex by an expert. An addition from another expert is that this statement can focus on the professionalism or the support of images regarding the text.

Regarding the statement about completeness, some experts question the opinions. Instead of opinions, it can also be about concerns or ideas. According to the experts, there is no right or wrong, but it is important to realize that this is a choice to make.

The experts who have dealt with those categories have no comments or criticisms regarding the statements of the categories efficacy, insensitivity or familiarity.

3.3 Selection of categories based on literature review and expert interviews

This section focuses on the selection of the five categories that are included in the PVE consultation of the ORS. First, five categories are selected on the basis of the literature review and the expert interviews. Thereafter, the five revised statements that are linked to the selected categories included in the PVE consultation of the ORS are presented.

3.3.1 Selection of five categories

With regard to the selection of five categories of face validity, the literature review and the expert interviews are consulted. Eleven categories of face validity have been identified from the literature review, as described in section 3.1.6. However, one category is identified in more articles than another category. Table 3.3 provides an overview of how often each category is identified in the twenty-five articles included in the literature review. Based on these numbers, it can be concluded that clarity, relevance, readability, appropriateness of difficulty level and unambiguity should be included in the PVE consultation of the ORS.

Table 3.3: How many times a category is identified in the twenty-five articles from the literature review

Category	Number of times identified in 25 articles
Clarity	18
Relevance	16
Readability	13
Appropriateness of difficulty level	10
Unambiguity	9
Aesthetics	7
Completeness	6
Feasibility	5
Efficacy	3
Insensitivity	2
Familiarity	2
Relativity	1

In the expert interviews, each expert was first asked to select five categories that they consider most important in terms of face validity for the PVE method. After these five categories were selected, the experts were also asked to rank the five categories by importance. In this way, each expert has designed his or her own top five of categories.

Since the literature does not specifically focus on the PVE method, but the experts do, both are included in the selection of categories. A point scale has been applied for this selection. The most important of the five categories is awarded five points. The least important of the five categories is awarded one point. The categories that fall in between receive four, three or two points respectively. The categories outside the selection of five receive no points. The five categories that added up the most points are selected. Table 3.4 shows the distribution of points and the total number of points per category. It is remarkable from this table that there are contradictions between the literature and the experts regarding the value attached to face validity categories. This difference is most apparent in the unambiguity category where four experts attach relatively much value to unambiguity and the literature relatively small value. Overall, this table shows that the following five categories are included in the PVE consultation of the ORS: unambiguity, readability, relevance, clarity and completeness.

Table 3.4: Distribution of points and total points per category

Category	Literature	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	*Total points per category*
Unambiguity	1	-	5	5	4	4	19
Readability	3	4	-	-	5	5	17
Relevance	4	3	1	1	3	3	15
Clarity	5	5	3	-	-	-	13
Completeness	-	2	2	3	-	2	9
Appropriateness of difficulty level	2	-	-	4	-	-	6
Efficacy	-	-	4	-	-	-	4
Insensitivity	-	-	-	2	-	-	2
Familiarity	-	-	-	-	2	-	2
Feasibility	-	-	-	-	1	1	2
Aesthetics	-	1	-	-	-	-	1

3.3.2 Revision of the statements of the five selected categories

The statements of the five categories selected above are revised based on the remarks and comments of the experts during the expert interviews. These five revised statements will be included in the PVE consultation of the ORS.

Compared to the first experiment of the unambiguity statement, the jargon has been removed from the statement. The statement relates to the possible tasks of the Schiphol Social Council. It was specifically chosen to place this statement directly after the MRS choice task, because in the discussions with stakeholders it appeared that this choice task caused the most fuss. The revised statement is as follows: I found it was clear what was meant by each task with regard to the possible tasks of the Schiphol Social Council.

The jargon has also been removed from the statement for readability. Furthermore, the readability is questioned for the entire consultation instead of a choice task on the advice of an expert. In this way, a global picture of readability as part of face validity can be obtained. The revised statement of readability is as follows: I found the questions asked to me in this study understandable.

In the revised statement of relevance, one goal is included instead of two goals. This goal focuses on the decision making by respondents. The jargon has also been removed. The revised statement is as follows: I think this research is a good way to give my opinion about the Schiphol Social Council and the Environmental House.

In the statement of clarity, ‘a choice’ as is stated in the first experiment of the statement is more specified on advice of the experts. The revised statement of clarity is as follows: I have received sufficient information to make a choice about the possible tasks of the Schiphol Social Council. Because of the specification, it is decided to question this category directly after the Schiphol Social Council choice task. In this way, the statement remains manageable for the respondents. In addition, it appears that this choice task caused the most fuss.

The last statement is about completeness. Nothing has changes about this statement. Because the intention is that respondents can give advice to policymakers on the basis of the PVE method, the statement focuses on the opinions of the respondents. The statement of completeness is as follows: I felt I could give all my opinions on how citizens should be involved in decision-making about Schiphol and how information should be provided. Table 3.5 provides an overview of the final statements that are included in the PVE consultation.

Table 3.5: Overview of five face validity categories and statements that are included in the consultation

Category	Place	Statement
Clarity	After choice task	I have received sufficient information to make a choice about the possible tasks of the Schiphol Social Council
Unambiguity	After choice task	I found it was clear what was meant by each task with regard to the possible tasks of the Schiphol Social Council
Relevance	At the end of consultation	I think this research is a good way to give my opinion about the Schiphol Social Council and the Environmental House
Readability	At the end of consultation	I found the questions asked to me in this study understandable
Completeness	At the end of consultation	I felt I could give all my opinions on how citizens should be involved in decision-making about Schiphol and how information should be provided

3.4 Design of the PVE consultation

As presented in the description of the case study of the ORS (as described in section 2.3), the functioning of the ORS is ineffective. The polder model that is applied, has failed because the shared ownership has disappeared. The main reason is that the focus has been too much on the growth of the Schiphol Airport in recent years. That is why van Geel (2019) has proposed two new entities that contribute to more intensive and broader participation and improved information provision. These two entities are called the Schiphol Social Council and the Environmental House. According to van Geel (2019), it is important to represent these two entities together with local residents.

The report by van Geel (2019) mentions several functions of the Schiphol Social Council and the Environmental House, but there is no prioritization given to these functions in this report. It follows that policymakers do not have a clear picture of what the priorities of these two entities should be, nor where to start. In addition, there is the question of how citizens themselves would like to participate and what local residents consider important participation principles. The aim of this PVE consultation is to find the answers of the above questions.

To achieve this aim, the PVE consultation has the following structure. After an introduction and an instruction of the consultation, questions are asked about how respondents want to be involved in decisions about Schiphol. This is followed by a first choice task with possible functions of the Schiphol Social Council, which are also known as options. After this choice task, respondents are asked to argue why they chose certain options and why not other options. The online webtool indicates which options have been chosen by the respondent and with how many points. The qualitative answers that follow from this facilitates the interpretation of the quantitative answers from this choice task. This motivation of answers is followed by in-depth questions about the Schiphol Social Council. These questions mainly concern which parties should be part of the Schiphol Social Council and which participation

principles respondents consider important with regard to citizen involvement. Thereafter, a second choice task follows. This choice task concerns the possible functions of the Environmental House. Immediately after this choice task, the respondents are again asked for a motivation for the chosen options and also why they did not choose the other options. This choice task is followed by questions about the provision of information that deal with, for example, the reliability of information. Finally, the last general questions follow. In these general questions, respondents are asked about demographic characteristics. A detailed elaboration of the entire PVE consultation can be found in Appendix D. Moreover, this structure of the PVE consultation is designed with input from stakeholders and the client (as described in chapter 2.3).

With regard to the feasibility of this research and the limited space to measure face validity, this research focuses on the choice task of the Schiphol Social Council and the last general questions. These two parts will be further explained below. Furthermore, the testing of the PVE consultation will be discussed.

3.4.1 Schiphol Social Council choice task and the face validity experiment

In this section the design of the Schiphol Social Council choice task is first discussed, after which the face validity experiment is further discussed.

3.4.1.1 Schiphol Social Council choice task

The Schiphol Social Council is the successor to the ORS. It can be stated that the Schiphol Social Council stands for broader and more intensive participation (van Geel, 2019). Some decisions surround the Schiphol Airport have consequences for the environment and the health of local residents in the area. This is the focus of the Schiphol Social Council. In the report by van Geel (2019), but also in the supplementary report by Berenschot (2020), various functions are mentioned that the Schiphol Social Council could fulfil.

Following van Geel (2019), a first function of the Schiphol Social Council is the dialogue function. The Schiphol Social Council is the party that makes the voices from society as a whole heard to the government and the operational organizations. A second function is the representative function. Van Geel's (2019) idea is that the Schiphol Social Council does not only consist of resident representatives, but also of other interest groups such as "environmental organisations, employer groups, employee groups, village and neighbourhood councils and young people". Then there is the knowledge function. This means that the Schiphol Social Council forms the basis for knowledge building. There must also be room for scientific disciplines to share their knowledge and to add scientific content to the dialogue. The service function means that the Schiphol Social Council helps other parties to find their way in the complex situation of interests surrounding the Schiphol Airport. This ensures that stakeholders are not overlooked. The fifth function is the advisory function. The Schiphol Social Council has the option of issuing advice to the competent parties. It is up to these competent parties to weigh up these advices. Another function is the repeating participation function. In many subjects related to Schiphol, a sequential degree of participation is advisable, following van Geel (2019). The last function is the social signalling function. The Schiphol Social Council is able to inform the competent parties early on about issues that arise in society.

However, it appears from the stakeholder discussions that the functions as drawn up by van Geel (2019) do not all correspond to the current picture of the Schiphol Social Council. In addition, it appears that the functions are regarded as vague. That is why the functions of van Geel (2019) have been tightened up and adjusted. In the webtool, the functions of the Schiphol Social Council are presented in such a way that at first only the function itself is visible. When the respondent clicks on the information button, more information about this function appears. The functions, or options, and the additional information which are included in the Schiphol Social Council choice task are presented in table 3.6.

Table 3.6 also includes the effect of each function. In consultation with the client and the stakeholders, it is decided to link each function to an effort effect. An effect in terms of money was rejected, because it then seems as if the customer does not want to spend money on affected local residents. Since one option requires more effort than another option, effort is divided into three attribute levels. The level '+' costs the least effort and the level '+++' costs the most effort. The level '++' is exactly in between. The abstractness of these levels of effort is a limitation as relatively little information is available about these options. It is not known yet what these functions will cost and how long it will take before these functions are set up. The three levels of efforts are therefore rough estimates. A constraint is imposed on the maximum effort that the respondents are allowed to divide over the various functions. This constraint touches on the goal of prioritizing these options, or functions, of the Schiphol Social Council.

Table 3.6: Functions of the Schiphol Social Council with information and effort

What are the possible tasks of the Schiphol Social Council?	What does this function mean? (Information button)	How much effort does it take? (+, ++ or +++)
Thinking along about the effects of air traffic on people's daily lives.	The Schiphol Social Council makes proposals to improve the lives of residents. This concerns, for example, flight routes, measures such as insulation and improvements of the living environment.	++
Devise and carry out own research.	The Schiphol Social Council can itself commission research into the effects of the policy related to Schiphol on the environment. The Council has money to conduct investigations.	+++
Thinking along about research.	Does the minister or Schiphol want to research the effects of Schiphol on the environment? They can ask the Schiphol Social Council to contribute ideas about the design of this research. This means, for example, that the Schiphol Social Council can be involved in the selection of the research agency and also advise researchers.	+
Conduct a second opinion.	Has there already been research into the effects of Schiphol on the environment? Then the Schiphol Social Council can have this research checked by other scientists who are not involved in the research.	++
Organize that residents are allowed to think along.	Are residents allowed to contribute ideas about a decision about the environmental effects of Schiphol? Then the Schiphol Social Council will help. This council ensure that the interests of social groups count. The council can request the opinions of residents via the internet or with living room conversations in which members of the council visit the residents at home.	+++
Giving advice on how residents can think along.	Does the government involve residents in a certain decision about Schiphol? The Schiphol Social Council then gives advice on how the approach could possibly be improved. The council ensures that various social groups involved in a decision are allowed to contribute ideas.	+
Provide advice at the request of the government on decisions that the government wants to take.	If the government plans to make a decision, they can ask the Schiphol Social Council to advise on it.	++
Giving unsolicited advice about policy or when something happens.	Do residents think about something a lot? Or are they concerned? Then the Schiphol Social Council can inform the government about this.	++

3.4.1.2 Face validity experiment

Within the PVE method there are different types of choice tasks that can be applied. Regarding face validity, it is a research question whether respondents' evaluation of face validity differs between the types. Therefore, the closed consultation of participants that are living in the region around Schiphol is split up. Therefore, half of respondents are shown one experiment and the other half of respondents are presented the second experiment. These two approaches give cause to research the face validity between the two different question methods. Moreover, these two experiments are randomly distributed among the respondents of the data panel.

In the meetings with the stakeholders, the situation arose in which the stakeholders could not agree on the type of choice task for the Schiphol Social Council. This is in contrast to the choice task of the Environmental House where there is consensus about the specific request. Therefore, the experiment of the two types of choice tasks is applied to the choice task of the Schiphol Social Council. This is why this research focuses on the Schiphol Social Council choice task.

When conducting an experiment, it is important that all other factors are constant to isolate the effect of the experiment. In this case, the specific query of the choice task of the Schiphol Social Council in the form of the two approaches is the only element that differs in the consultation. Within the Schiphol Social Council choice task, the respondents can also choose between the same options and receive the same information about the content of these options.

Hence, the face validity experiment consists of two types of the Schiphol Social Council choice task. The first approach to address the Schiphol Social Council is called the 'sliders' choice task. As the name implies, sliders are used as the way for respondents to show how much attention should be given to each option by the Schiphol Social Council. The difference between the 'sliders' and the 'points' choice task is that the effort of each option is also shown in the 'sliders' choice task. Taking this effort into account, the respondent is able to consider how much attention should be given to each option by moving the slider more to the right. Compared to the 'points' choice task, the 'sliders' choice task can give the respondents more choice pain. It can be confronting for the respondents to mention that not every option can be executed to the maximum. A constraint is imposed on the maximum effort to be deployed. This can be regarded as a disadvantage of the 'sliders' choice task. On the other hand, it can be stated as an advantage that the 'sliders' choice task forces the respondents to prioritize the options in which they include the effort of each option in their advice. More information can be provided about prioritizing the options.

Figure 3.1 presents the 'sliders' choice task for the Schiphol Social Council in the webtool. The respondents who get this approach, are presented the choice task in this way. A counter is displayed on the right side of this figure. A constraint of sixty points has been set that the respondents can divide among the different options with their different efforts. The respondent is not obliged to allocate all points, but is allowed to allocate fewer than sixty points as well. The possibility is left for a respondent to give advice from which it appears that a respondent does not consider the options as important. Furthermore, in the middle of the figure are the eight options of functions of the Schiphol Social Council in a row. When the pink

i-button is pressed, more information about each selection task is displayed. This additional information corresponds to the middle column in table 3.6.

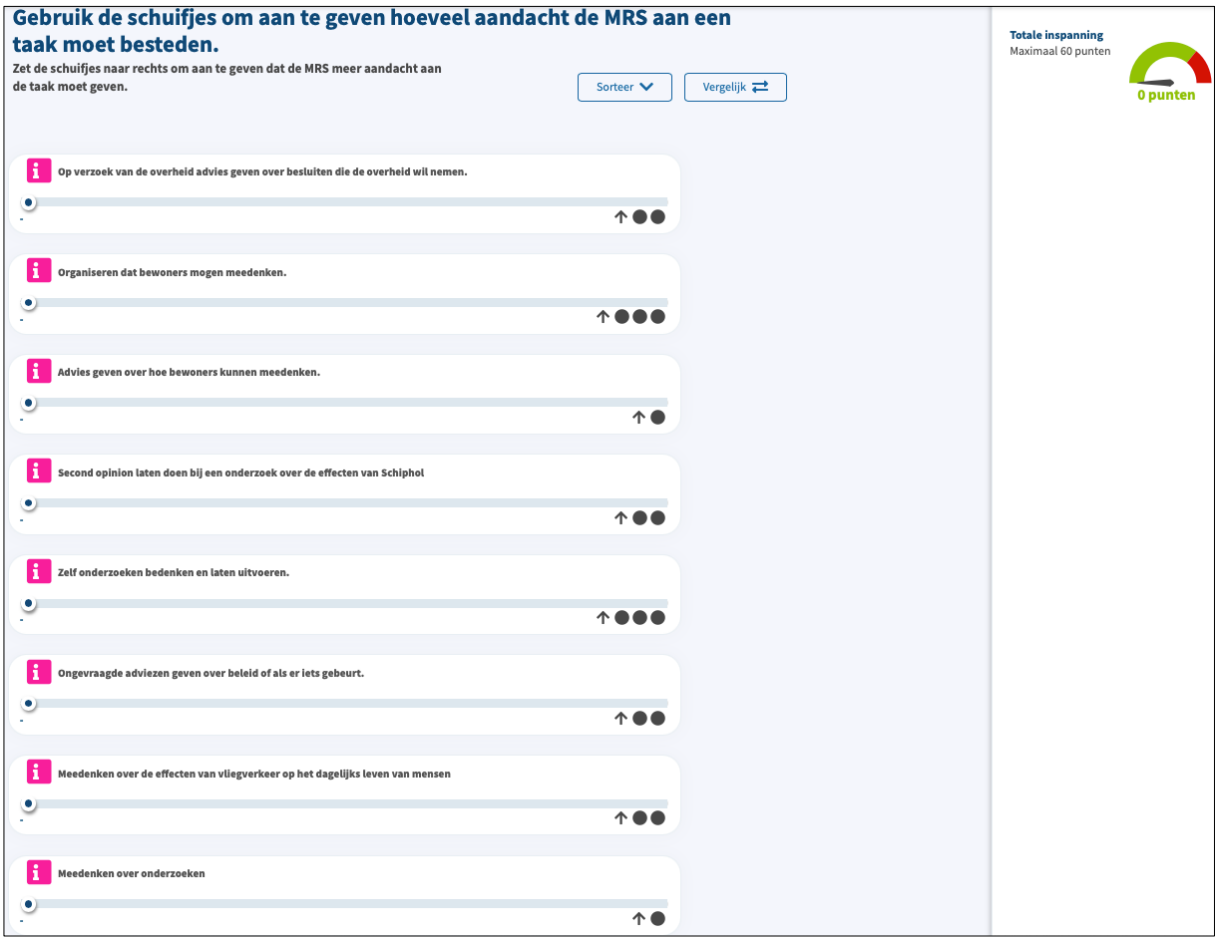


Figure 3.1: The 'sliders' choice task about the Schiphol Social Council

The effort of an option is represented by dots instead of a '+' in figure 3.1. When it concerns a slider of an option with a single dot as effort, this slider consists of ten steps. One point is added to the counter for each step to the right. It follows that an option with the most minimal effort can add up to ten points to the counter. For an option with two dots as effort, the slider is divided into twenty steps. Each step adds one point to the counter. So, an option with two dots as an effort can add a maximum of twenty points to the counter. In the same way, a slider with three dots as an effort consists of thirty steps that can be taken to the right. This option can add up to thirty points.

The second approach to address the Schiphol Social Council is called the 'points' choice task. As the name implies, the respondents are given the opportunity to divide points among the options of the choice task to indicate how important they think the different options are. An advantage of this approach is that the respondent can avoid the pain of choosing by distributing the points fairly among all options as much as possible. However, this can also be regarded as a disadvantage since a 'points' choice task provides less insight into the prioritization of options compared to a 'sliders' choice task. The 'points' choice task does not include the efforts of the various options.

Figure 3.2 presents the 'points' choice task for the Schiphol Social Council in the webtool. The respondents who get this approach, are presented the choice task in this way. As in the first approach, the pink information button provides additional information about each option. This is the same additional information as in the first approach. Furthermore, in this second approach a maximum of sixty points cannot be distributed, but twenty points is the maximum. To the right of the figure is the counter that keeps track of the total number of distributed points. By means of the plus and the minus for each option, the respondents can add or remove points for a certain option.

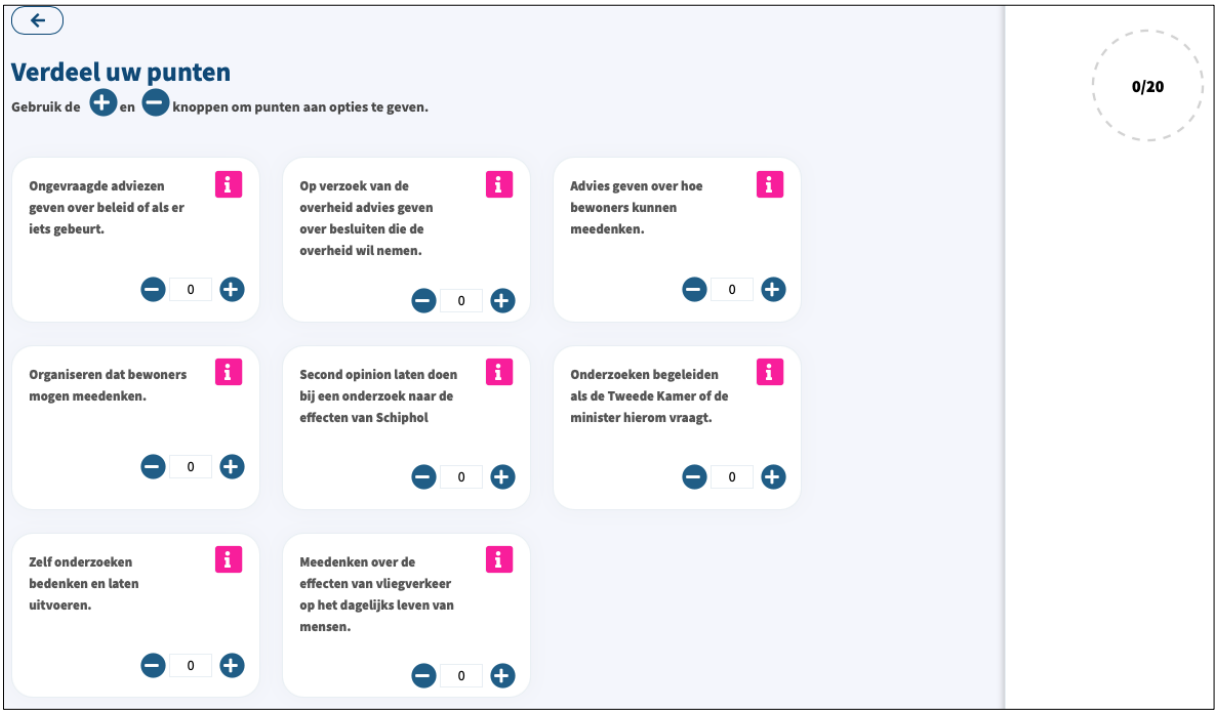


Figure 3.2: The 'points' choice task about the Schiphol Social Council

3.4.1.3 Statements of face validity directly after the experiment

After the participants have completed one of the two experiments of the Schiphol Social Council, they are first asked to motivate why they awarded points to certain options and, if applicable, why not to other options. Immediately after this motivation, the face validity statements of the categories clarity and unambiguity are questioned.

3.4.2 Last general questions

First of all, these last general questions deal with the face validity statements that do not focus on the choice task of the Schiphol Social Council specifically, but on the consultation as a whole. It concerns the statements of the categories relevance, readability and completeness.

Furthermore, demographic characteristics are questioned in the last general questions. These characteristics can be divided into general demographic characteristics and case-specific demographic characteristics. General demographic characteristics include characteristics that are also more common in other studies and that are expected to affect the assessment of face validity. These general characteristics are gender, age, educational level and relational status (Broder et al., 2007; Connell et al., 2018). Furthermore, there are a couple of general demographics that are included in the nuisance perception survey conducted by the ORS in

2017. These demographics have in included in this research as this influence the valuation of noise nuisance at Schiphol. As a result, these demographics may also influence the evaluation of the face validity. These demographics are working life, living environment, if there are children living at home, whether the house a participant is living in is an owner-occupied house or a rental house and if the municipality where the participant lives is part of the 'inner' area with 58 dB(A) nuisance or in the 'outside' area with 48 dB(A) nuisance (ORS, 2017). The expectation is that respondents living an owner-occupied house in the 'inner' area rate the face validity lower than respondents living in a rental house in the 'outside' area. These respondents will experience more nuisance from Schiphol and therefore fill in the consultation with a more negative view.

With regard to the case-specific demographic characteristics, there are a number of characteristics of people that may influence this person's attitude towards Schiphol and the functions of the Schiphol Social Council and the Environmental House that they prefer. These case-specific demographics are included in the consultation based on input from stakeholders. These characteristics concern: whether a respondent lives near a flight route of Schiphol, if a respondent works for Schiphol or works for an organization that works closely with Schiphol, if a respondent is a Schiphol customer or traveller, if a participant is member of a citizen organization, if a respondent is inconvenienced by Schiphol, if a respondent is spending free time near Schiphol, if a respondent is satisfied with the way in which decisions about Schiphol can be influenced, what a respondent thinks about the reliability of the information that can be found about the effects of Schiphol on the environment and what a respondent thinks about the independence of the information that can be found about the effects of Schiphol on the environment. An overview of all the demographic characteristics is presented in Appendix E.

3.4.3 Testing the PVE consultation

This PVE consultation of the ORS has been tested in a number of rounds. First of all, the consultation was set up by a number of employees of Populytics. Because several employees were involved, there was the opportunity to provide feedback on each other's input. This feedback was provided both on a textual and a substantive level. Based on this feedback, it was decided to place the choice task of the Schiphol Social Council before the choice task of the Environmental House. While setting up the PVE consultation, some employees noticed that this caused more controversy among the stakeholders involved. The employees also complemented each other. For example, the tasks of the Schiphol Social Council are based on the knowledge of a number of different employees who gave input in a number of different rounds. Other questions in the consultation were also added or removed based on feedback. In addition, the language use was adapted to easier language use in a feedback round by employees.

Second, a language agency has been engaged. After the design of the consultation has been drawn up by the researchers, a language agency was asked to rewrite the consultation to so called B1-level language. The B1-level has been chosen because approximately eighty percent of the Dutch population understands this language level. It is important that the PVE consultation is legible, since a wide sample of respondents participate in the consultation. As a result of this check by the language agency, words such as institution and living environment

have been replaced. Instead, organisation and everyday life have been added. Moreover, the language agency has also divided long sentences in shorter sentences to increase legibility.

Third, the consultation was tested by a test panel. This test panel consists of a part of the respondents designated by the data panel. Since there are two different experiments of the PVE consultation, the experiment with the 'sliders' choice task and the experiment with the 'points' choice task of the Schiphol Social Council, two test panels have been set up. Both test panels consist of 125 respondents, which represents ten percent of the total desired number of respondents per experiment. These test panels were used to test whether the data obtained from the consultation could be used for further data analysis. Moreover, the data of the test respondents is tested on bugs in the software. In addition, the answers to the five face validity statements were also analysed to find out whether the consultation was, among other things, legible and clear. No problems arose here. Finally, these test panels tested how many respondents who started the consultation, also completed it. The data panel has taken this information into account when recruiting the rest of the respondents. During these tests, it turned out that there was a dropout of test respondents of about forty percent.

3.4.4 The attraction of respondents

Section 2.5 describes which requirements the respondent sample of the PVE consultation must meet. A data panel is involved to recruit the respondents. It is the responsibility of this panel to have a sufficient number of respondents participate in this PVE consultation. Therefore, the attraction of the respondents for the PVE consultation has been outsourced. The data panel informs their panel members about this PVE consultation by means of an e-mail.

4. Evaluation of face validity by respondents

In this chapter the results are presented of the statistical analyses that are performed to answer the second sub question. The second sub question is formulated in the following way: *how do respondents rate and evaluate the face validity of a PVE consultation regarding the Schiphol Environmental Council?* First of all, this chapter elaborates on the characteristics of the sample. This is followed by the descriptive results of the statements regarding face validity. Thereafter, the differences between the two different experiments of the PVE consultation are explained. Furthermore, the results of the factor analyses are presented. The results of the multiple regression analyses and the multinomial logistics regressions, which examine the demographic characteristics that influence the assessment of the face validity statements, are presented as well. Finally, the results of the Latent Class Cluster Analyses (LCCA's) are evaluated.

4.1 Sample characteristics

Table 4.1 and 4.2 present an overview of the main characteristics of the sample. Table 4.1 shows the number of people and percentages for the entire sample and table 4.2 shows the number of people and the percentages for the 'sliders' experiment and the 'points' experiment separately. These main characteristics gender, age and educational level are provided to the staff member of the data panel as requirements for a representative sample. It is notable in table 4.2 that the percentage of men in the 'sliders' experiment (53,3%) is higher than in the 'points' experiment (50,2%). Furthermore, it is noticeable that in the experiment with the 'sliders' choice task there are relatively more respondents with a higher education (46,7%) and in the second experiment with the 'points' choice task relatively more respondents with a medium education (39,9%).

Next to the main sample characteristics, table 4.1 and table 4.2 present the percentages of the Dutch population and the results of the chi-square tests. Data of the CBS (2021) is used for this purpose. Since there is no overview of the educational level per municipality, the data for the Netherlands as a whole is used for this characteristic. Moreover, the categories in the data from CBS (2021) differ from those surveyed in this consultation. Therefore, the CBS categories have been rebalanced. These chi-square tests show that the difference between the total sample and the respondents of both experiments separately with the Dutch population is significant regarding age and educational level. With regard to gender, there is no significant difference between the total sample, the respondents of the 'sliders' experiment and the 'points' experiment with the Dutch population. In order to draw conclusions for the population despite the significant differences in age and educational level, the data has been weighted on the characteristics of gender, age and educational level. As a result, more weight has been assigned to the answers of participants who are underrepresented in the sample and less weight to the participant groups that are overrepresented on the basis of a weighting factor.

In addition to these main demographic characteristics, there are other demographic characteristics that are part of this research. An overview of all generic and case-specific characteristics and their categorization is presented in Appendix F.

Table 4.1: Main characteristics of the total sample

	Total respondents ('sliders' experiment and 'points' experiment)	Percentage of Dutch population aged 15 years and older (CBS,2021)	Chi-square test (2-tailed)
Gender			
Man	637 (51,8%)	49,5%	0,108
Woman	593 (48,2%)	50,5%	
Age			
18-34 years	395 (32,1%)	20,2%	0,000
35-64 years	609 (49,5%)	46,8%	
65 years and older	226 (18,4%)	23,1%	
Educational level			
High	555 (45,1%)	34,4%	0,000
Medium	464 (37,7%)	36,6%	
Low	211 (17,2%)	29,0%	

Table 4.2: Main characteristics of the respondents of 'sliders' experiment and of 'points' experiment

	'Sliders' experiment	'Points' experiment	Percentage of Dutch population aged 15 years and older (CBS,2021)	Chi-square test (2-tailed)*
Gender				
Man	345 (53,3%)	292 (50,2%)	49,5%	1. 0,056
Woman	303 (46,7%)	290 (49,8%)	50,5%	2. 0,746
Age				
18-34 years	206 (31,8%)	188 (32,3%)	20,2%	1. 0,000
35-64 years	323 (49,8%)	286 (49,1%)	46,8%	2. 0,000
65 years and older	119 (18,3%)	108 (18,6%)	23,1%	
Educational level				
High	303 (46,7%)	252 (43,3%)	34,4%	1. 0,000
Medium	231 (35,7%)	232 (39,9%)	36,6%	2. 0,000
Low	114 (17,5%)	98 (16,8%)	29,0%	

**1: respondents of 'sliders' experiment. 2: respondents of 'points' experiment.*

4.2 Descriptive results

To provide insight into how each face validity statement scores, descriptive results have been applied. Table 4.3 shows the descriptive results of the total sample. It is remarkable that for each category statement most respondents chose the answer option agree. Furthermore, 73,8% of the total sample agrees with the statement about readability. On the other hand, 9,0% of the total sample disagrees with the statement of unambiguity about how clear the options are. Moreover, the statement of clarity has the most neutral assessments (36,6%) compared to the other statements. Overall, clarity has the lowest average score (3,60) and readability the highest average score (3,89) over the whole sample.

Table 4.3: Descriptive results of the total sample

	Frequencies total sample					Total Sample	
	Totally disagree	Disagree	Neutral	Agree	Totally agree	Mean	St.Dev
Clarity	11 (0,9%)	71 (5,8%)	450 (36,6%)	571 (46,4%)	127 (10,3%)	3,60	0,786
Unambiguity	14 (1,1%)	97 (7,9%)	384 (31,2%)	600 (48,8%)	135 (11,0%)	3,61	0,828
Relevance	14 (1,1%)	39 (3,2%)	302 (24,6%)	632 (51,4%)	243 (19,8%)	3,85	0,807
Readability	14 (1,1%)	44 (3,6%)	264 (21,5%)	653 (53,1%)	255 (20,7%)	3,89	0,809
Completeness	16 (1,3%)	32 (2,6%)	314 (25,5%)	622 (50,6%)	246 (20,0%)	3,85	0,810

In the tables 4.4 and 4.5 the descriptive results of the total sample are split into the descriptive results of the 'sliders' experiment and the 'points' experiment. In both experiments, most of the respondents agreed with the statement of readability compared with other statements. In the 'sliders' experiment 73,0% agreed and in the 'points' experiment 74,8% agreed. On the other hand, most of the respondents disagreed with the statement of unambiguity. In the 'sliders' experiment 10,8% disagreed and in the 'points' experiment 7,0% disagreed. In both experiments, most respondents opted for neutral for the clarity category compared to the scores of neutral in the other categories with 37,2% in the 'sliders' experiment and 35,9% in the 'points' experiment. Overall, all five categories are assessed lower in the 'sliders' experiment than in the 'points' experiment.

Table 4.4: Descriptive results of 'sliders' experiment

	Frequencies experiment 1: 'sliders' choice task					Experiment 1	
	Totally disagree	Disagree	Neutral	Agree	Totally agree	Mean	St.Dev
Clarity	7 (1,1%)	35 (5,4%)	241 (37,2%)	296 (45,7%)	69 (10,6%)	3,59	0,792
Unambiguity	7 (1,1%)	63 (9,7%)	189 (29,2%)	316 (48,8%)	73 (11,3%)	3,59	0,853
Relevance	10 (1,5%)	23 (3,5%)	163 (25,2%)	324 (50,0%)	128 (19,8%)	3,83	0,838
Readability	6 (0,9%)	24 (3,7%)	145 (22,4%)	339 (52,3%)	134 (20,7%)	3,88	0,806
Completeness	10 (1,5%)	19 (2,9%)	175 (27,0%)	319 (49,2%)	125 (19,3%)	3,82	0,829

Table 4.5: Descriptive results of 'points' experiment

	Frequencies experiment 2: 'points' choice task					Experiment 2	
	Totally disagree	Disagree	Neutral	Agree	Totally agree	Mean	St.Dev
Clarity	4 (0,7%)	36 (6,2%)	209 (35,9%)	275 (47,3%)	58 (10,0%)	3,60	0,778
Unambiguity	7 (1,2%)	34 (5,8%)	195 (33,5%)	284 (48,8%)	62 (10,7%)	3,62	0,799
Relevance	4 (0,7%)	16 (2,7%)	139 (23,9%)	308 (52,9%)	115 (19,8%)	3,88	0,772
Readability	8 (1,4%)	20 (3,4%)	119 (20,4%)	314 (54,0%)	121 (20,8%)	3,89	0,814
Completeness	6 (1,0%)	13 (2,2%)	139 (23,9%)	303 (52,1%)	121 (20,8%)	3,89	0,786

In figure 4.1 to 4.10 the descriptive results per category and per experiment are presented in histograms. The x-axes of the histograms show the five answer options for the statements of face validity categories. The meaning of the values at the x-axes are the same as in the previous tables. The y-axes show the frequencies of the five response options. The histograms clarify that the vast majority of respondents in all categories in both experiments chose the

fourth answer option agree. Moreover, the answer option totally disagree is chosen the least by the respondents in all categories in both experiments.

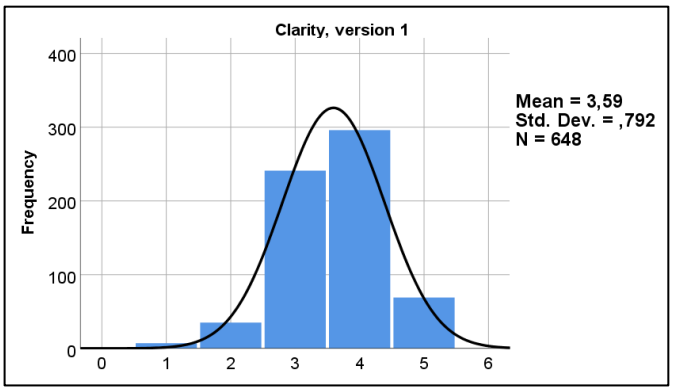


Figure 4.1: Histogram clarity 'sliders' experiment

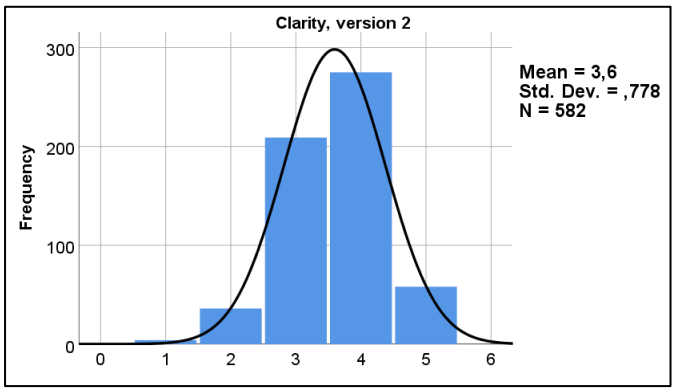


Figure 4.2: Histogram clarity 'points' experiment

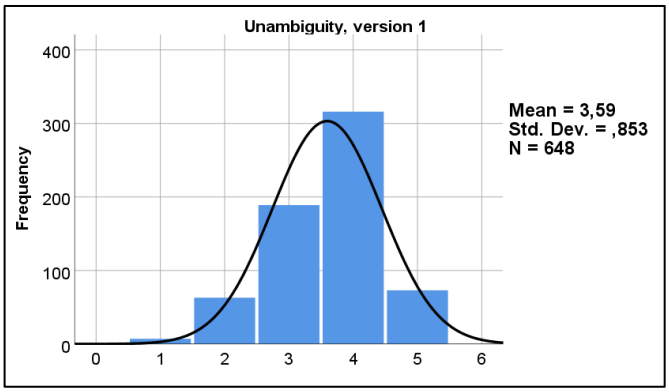


Figure 4.3: Histogram unambiguity 'sliders' experiment

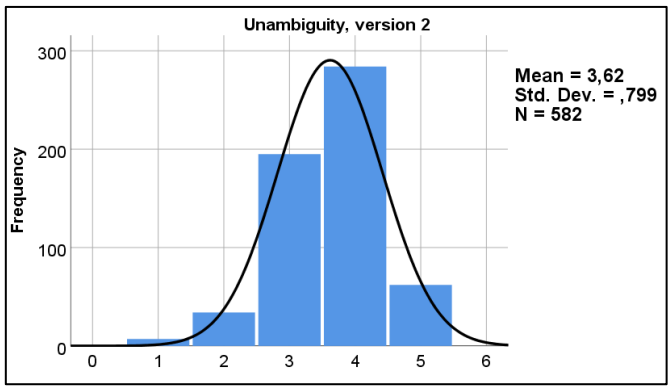


Figure 4.4: Histogram unambiguity 'points' experiment

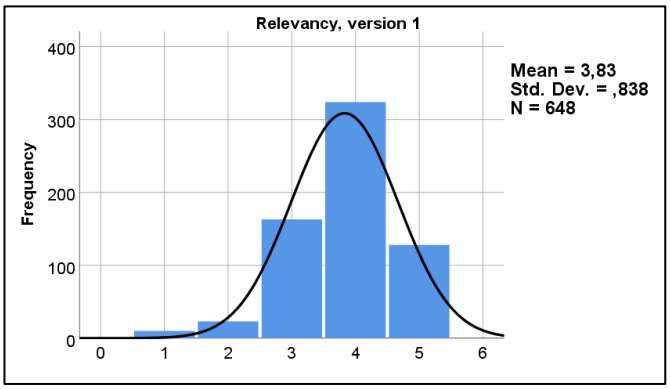


Figure 4.5: Histogram relevance 'sliders' experiment

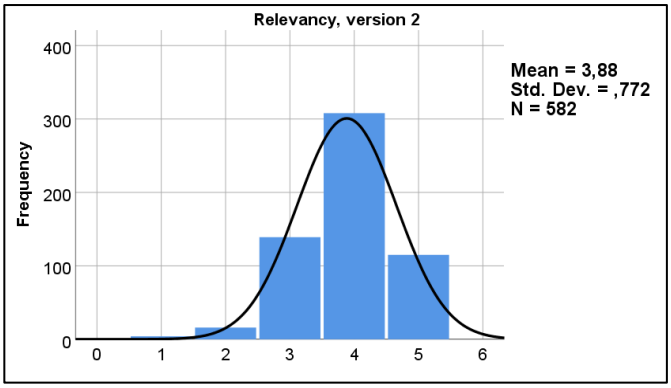


Figure 4.6: Histogram relevance 'points' experiment

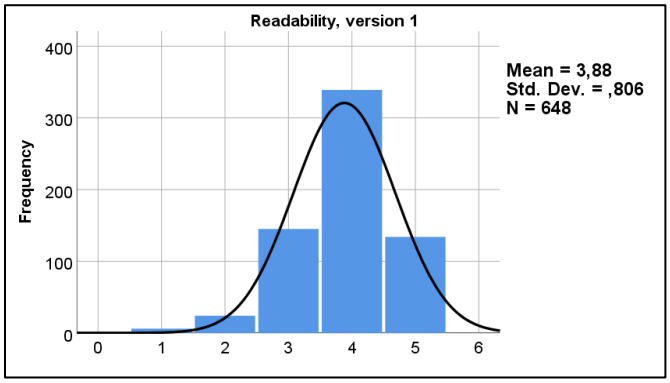


Figure 4.7: Histogram readability 'sliders' experiment

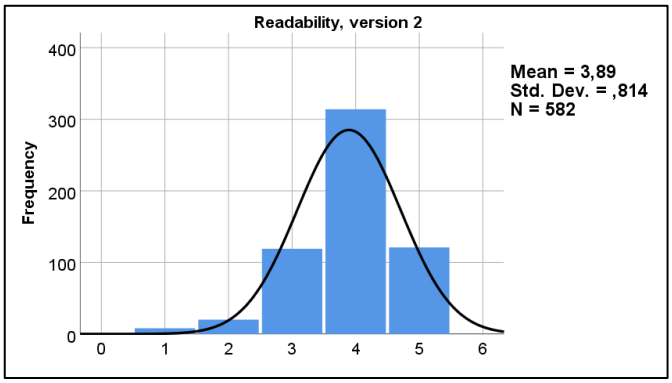


Figure 4.8: Histogram readability 'points' experiment

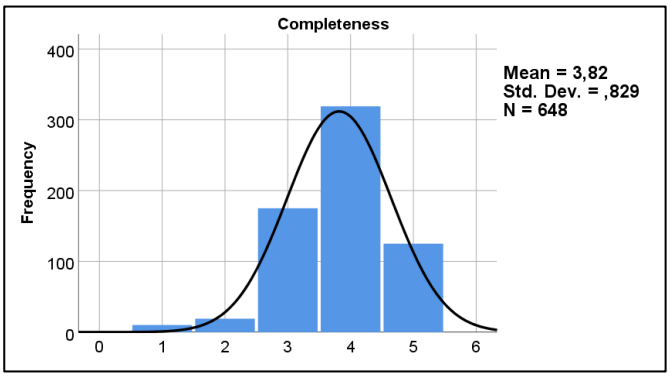


Figure 4.9: Histogram completeness 'sliders' experiment

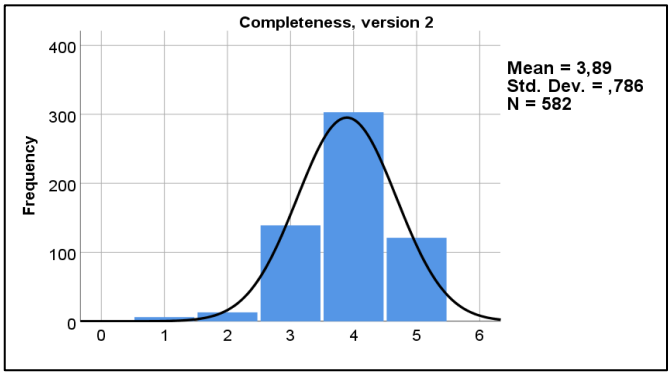


Figure 4.10: Histogram completeness 'points' experiment

4.3 Differences between the two experiments regarding face validity

To research whether the two experiments of the PVE consultation differ from each other regarding the evaluation of face validity, Mann-Whitney U tests and one-way MANOVA tests are performed.

First of all, Mann-Whitney U tests are carried out in order to make a comparison between the two experiments per category. This entails that it is tested, for example, whether there is a significant difference in the scores that the respondents gave to the clarity statement between both experiments. Therefore, ten variables have undergone a test of normality. These variables consist of the scores of each category for both experiments. The results of the tests of normality are presented in Appendix G. These tests of normality show that none of the variables is significant and therefore is not normally distributed. This is also reflected in the histograms in section 4.2 which show that the fourth answer option 'agree' is filled in the most in all categories in both experiments. Table 4.6 presents the results of the Mann-Whitney U tests.

Table 4.6: Results of Mann-Whitney U test to research the differences of face validity evaluation per category

	Experiment	Number	Mann-Whitney U	P-value
Clarity	'Sliders'	648	188966,50	0,945
	'Points'	582		
Unambiguity	'Sliders'	648	189752,00	0,837
	'Points'	582		
Relevance	'Sliders'	648	363356,50	0,368
	'Points'	582		
Readability	'Sliders'	648	191055,50	0,661
	'Points'	582		
Completeness	'Sliders'	648	197364,00	0,120
	'Points'	582		

Table 4.6 shows that there are no significant differences for each category separately between the assessment of face validity in both experiments. The p-values of every Mann-Whitney U test are higher than a value of 0,05.

Second, three different one-way MANOVA test are performed. In the one-way MANOVA tests multiple categories are tested simultaneously (in a multivariate test) while they are also controlled for each other (in the between-subjects effects). In the first one-way MANOVA test, the variables clarity and unambiguity are included. These two categories are asked immediately after the choice task of the Schiphol Social Council. Therefore, these two categories may take a position on specifically the face validity of the Schiphol Social Council. The results of this one-way MANOVA test of clarity and unambiguity are presented in table 4.7.

Table 4.7: Results of one-way MANOVA test of clarity and unambiguity

	Between-subjects effects			Multivariate test	
	R-squared	F-value	P-vale	Wilk's lambda	P-value
Clarity	0,000	0,002	0,963	1,000	0,841
Unambiguity	0,001	0,267	0,606		

Table 4.7 shows under the multivariate test that the p-value is higher than 0,05. From table 4.7 it becomes clear that the categories clarity and unambiguity together do not constitute a significant difference in the evaluation of face validity of the Schiphol Social Council choice task between the two experiments. Moreover, the between-subjects effects show that neither category produces a significant difference in the assessment of face validity of the Schiphol Social Council choice task between the two experiments when it is controlled by the other category.

In the second one-way MANOVA test, the variables relevance, readability and completeness are included. These three categories are questioned at the end of the PVE consultation. These three categories thus provide sense of the evaluation of face validity about the consultation in general. Table 4.8 presents the results of the one-way MANOVA of relevance, readability and completeness.

Table 4.8: Results of one-way MANOVA test of relevance, readability and completeness

	Between-subjects effects			Multivariate test	
	R-squared	F-value	P-vale	Wilk's lambda	P-value
Relevance	0,001	1,396	0,238	0,997	0,271
Readability	0,000	0,071	0,790		
Completeness	0,002	2,675	0,102		

As with the previous Mann-Whitney U test, the p-value of the multivariate test is greater than 0,05. The categories relevance, readability and completeness together do not provide a significant difference in the general evaluation of face validity of the consultation between the two experiments based on these three categories. The between-subjects effects show that none of the categories produces a significant difference in the assessment of face validity of the consultation in general between the two experiments given that a category is controlled by the other two categories.

The third one-way MANOVA test consists of all five the face validity categories included in this study. These five categories form all part of face validity and therefore examines the face validity of the PVE method. Since this third one-way MANOVA test contains more categories of face validity than the second test, this test provides a more comprehensive picture of the possible differences in the evaluation of face validity between the two experiments. Table 4.9 presents the results of the one-way MANOVA test of the five categories clarity, unambiguity, relevance, readability and completeness.

Table 4.9: Results of one-way MANOVA test of clarity, unambiguity, relevance, readability and completeness

	Between-subjects effects			Multivariate test	
	R-squared	F-value	P-vale	Wilk's lambda	P-value
Clarity	0,000	0,002	0,963	0,997	0,508
Unambiguity	0,000	0,267	0,606		
Relevance	0,001	1,396	0,238		
Readability	0,000	0,071	0,790		
Completeness	0,002	2,675	0,102		

The multivariate test with a p-value higher than 0,05 shows that the five categories together provide no significant difference in the evaluation of face validity between the two experiments. Furthermore, the between-subjects effects show that none of the five categories separately produces a significant difference in the assessment of face validity between the two experiments, when a category is controlled by the other four categories.

4.4 Results of the factor analyses

The aim of the factor analyses is to research whether the categories of face validity jointly load on a latent variable. These latent variables are included in further statistical analyses that may reveal differences between the 'sliders' choice task and the 'points' choice task, such as demographic characteristics that influence the assessment of face validity. Despite the fact that no significant differences are detected between the two experiments of the consultation with regard to the evaluation of face validity in section 4.3, separate factor analyses are performed for both experiments. In this manner, further possible differences between the

two experiments may be uncovered. For both experiments of the consultation, three factor analyses are performed.

The first analysis executed is a factor analysis in which the categories clarity and unambiguity are included. In the second factor analysis the variables relevance, readability and completeness are included. A third factor analysis is performed with all five face validity categories included in this research. If this latent variable exists, it deals with the face validity of the whole consultation, i.e. the PVE method. Since it contains more categories than the second factor analysis, this factor analysis provides a more comprehensive view on the face validity of the PVE method instead of the consultation itself as is the case with the second factor analysis.

Table 4.10 presents the results of the three factor analyses performed for the 'sliders' experiment of the PVE consultation of the ORS. For all three analyses, the results per analysis contain one factor that has an eigenvalue above 1,00. This means that one latent variable can be distinguished per factor analysis. The expectations regarding the latent variables of the three factor analyses as presented in section 1.3.2 are fulfilled. So, there is first a latent variable identified which deals with the face validity of the Schiphol Social Council 'sliders' choice task. Another latent variable deals with the face validity of the PVE consultation in general. A third latent variable deals with a more comprehensive view on the face validity of the PVE method. Moreover, all three factor analyses have a Cronbach's alpha greater than 0,700. This means that the scale of the latent variables can be considered reliable.

Table 4.10: Results of factor analyses of 'sliders' experiment

Factor analyses 'sliders' experiment	Factor loading
Factor analysis with two categories (Cronbach's alpha = 0,718)	
Clarity	0,748
Unambiguity	0,748
Factor analysis with three categories (Cronbach's alpha = 0,771)	
Relevance	0,737
Readability	0,769
Completeness	0,677
Factor analysis with five categories (Cronbach's alpha = 0,779)	
Clarity	0,578
Unambiguity	0,571
Relevance	0,717
Readability	0,739
Completeness	0,612

Table 4.11 shows the results of the three factor analyses performed for the 'points' experiment of the ORS consultation. These results are consistent with the results of the factor analyses of the 'sliders' experiment. For all three analyses, the results per analysis contain one factor that has an eigenvalue above 1,00. This means that one latent variable can be distinguished per factor analysis. It is notable that the scale for the latent variables of the analyses with three and with five categories is reliable, since the Cronbach's alpha is higher than 0,700. The Cronbach's alpha of the factor analysis with two categories is 0,694, which implies that this latent variable does not have a reliable scale. However, because this latent

variable has reliable scale in the 'sliders' experiment and because both experiments do not differ significantly in the evaluation of face validity (section 4.3), the latent variable is still included in this research. So, the factor analyses of the 'points' experiment results in a latent variable which deals with the face validity of the Schiphol Social Council 'points' choice task, another which deals with the face validity of the PVE consultation in general and a third which deals with a more comprehensive view on the face validity of the PVE method. It follows that the expectations regarding the latent variables of the three factor analyses as described in section 1.3.2 are fulfilled.

Table 4.11: Results of factor analyses of 'points' experiment

Factor analyses 'points' experiment	Factor loading
Factor analysis with two categories (Cronbach's alpha = 0,694)	
Clarity	0,729
Unambiguity	0,729
Factor analysis with three categories (Cronbach's alpha = 0,784)	
Relevance	0,644
Readability	0,806
Completeness	0,775
Factor analysis with five categories (Cronbach's alpha = 0,774)	
Clarity	0,507
Unambiguity	0,556
Relevance	0,669
Readability	0,731
Completeness	0,725

4.5 Results explanatory personal characteristics of the latent variables

To research whether and which characteristics of respondents influence the assessment of the latent variables, multiple regression analyses are performed. Section 4.4 resulted in six latent variables (three per experiment of the consultation) that are all included separately as dependent variables in a multiple regression. This results in six multiple regressions of which outcomes are presented below. Dummy coding of the independent variables, i.e. the demographic and case-specific characteristics of the respondents, is applied to set up the multiple regressions. An overview of how this dummy coding is applied is presented in Appendix H. The tables presented below are abbreviated. Only the variables with a p-value less than 0,100 are presented. The variables with a p-value less than 0,050 are shown in blue. The complete results of the multiple regressions are shown in Appendix I. Because hierarchical multiple regression analyses are applied, the result tables consist of a step one and a step two. These steps have been drawn up in the methodology (chapter 2). To answer the sub research question, the focus is on the significant regression coefficient in step two. Furthermore, the coefficient presented are the standardized regression coefficients. In section 4.5.1 the multiple regression analyses of the 'sliders' experiment are discussed and in section 4.5.2 the multiple regression analyses of the 'points' experiment are discussed.

4.5.1 Results multiple regression analyses of the 'sliders' experiment

First, a multiple regression analysis is performed with the latent variable that corresponds to the face validity of the Schiphol Social Council choice task specifically. The results of this

multiple regression analysis are shown in table 4.12. With regard to this latent variable, there is one coefficient that loads significantly in step two while controlling for the other independent variables. The constant also significant. The variable that loads significantly is being a member of a citizen organization. When a person is a member of a citizen organization, the face validity of this person regarding the Schiphol Social Council 'sliders' choice task will be higher than a respondent who is not a member of a citizen organization (coefficient = 0,088).

The adjusted R square has a negative value in both steps. This means that the model has not explained any variance of the face validity of the Schiphol Social Council 'sliders' task. Furthermore, the partial F test is also not significant in both steps. Therefore, adding a subgroup of variables in each step of this regression model does not significantly improve the model.

Table 4.12: Results of multiple regression of face validity of Schiphol Social Council 'sliders' choice task ('sliders' experiment)

	Step 1		Step 2	
Independent variables	Coefficient	P-value	Coefficient	P-value
General demographics				
Gender – man	-0,067	0,114	-0,076	0,080
Case-specific characteristics				
Member of a citizen organization – yes			0,088	0,039
Model information				
Constant	3,707	0,000	3,813	0,000
Partial F test	0,844	0,620	0,769	0,794
Adjusted R square	-0,003		-0,010	

Second, a multiple regression analysis is performed with the latent variable that corresponds to the face validity of the PVE consultation in general. The outcomes are presented in table 4.13. Regarding this latent variable, there is no variable in step two that loads significantly when controlled for the other independent variables. In step one, the highly educated dummy variable loads significantly. This regression coefficient can be interpreted as follows. When a respondent is highly educated, this respondent will generally rate the face validity of the PVE consultation lower than when a respondent is medium or poorly educated (coefficient = -0,115). Additionally, the constants also load significantly.

The adjusted R square shows that when the general demographics are added, the model explains 0,3 percent variance of the face validity of the PVE consultation in general. When the case-specific characteristics are also added, the model does not explain any variance at all. Moreover, both partial F tests are not significant and thus adding the subgroups of general demographics and case-specific characteristics does not significantly explain the model.

Table 4.13: Results of multiple regression of face validity of PVE consultation in general ('sliders' experiment)

Independent variables	Step 1		Step 2	
	Coefficient	P-value	Coefficient	P-value
General demographics				
Dummy – high educated	-0,115	0,048	-0,099	0,094
Type of house – owner-occupied house	0,073	0,091	0,057	0,201
Case-specific characteristics				
Dummy – neutral satisfied with influence			-0,089	0,075
Model information				
Constant	3,947	0,000	4,016	0,000
Partial F test	0,025	0,313	0,963	0,520
Adjusted R square	0,003		-0,002	

As follow up, a multiple regression analysis is performed with the latent variable that consists of all the five categories of face validity included in this research. Table 4.14 presents the results of this multiple regression analysis. With regard to this latent variable, there is no coefficient of the general demographics or case-specific characteristics that loads significantly in both steps while controlling for the other independent variables. Only the constants load significantly. It follows that this regression model cannot explain any variance of the face validity of the PVE method. Moreover, the partial F tests are not significant as well. So, adding groups does not lead to a significant improvement of the model.

Table 4.14: Results of multiple regression of face validity of PVE method ('sliders' experiment)

Independent variables	Step 1		Step 2	
	Coefficient	P-value	Coefficient	P-value
General demographics				
Gender – man	-0,069	0,105	-0,072	0,097
Dummy – high educated	-0,101	0,081	-0,087	0,144
Model information				
Constant	3,851	0,000	3,395	0,000
Partial F test	1,010	0,441	0,832	0,711
Adjusted R square	0,000		-0,007	

4.5.2 Results multiple regression analyses of the 'points' experiment

This section follows the same structure as section 4.5.1, but focuses on the multiple regression analyses of the 'points' experiment. A multiple regression analysis is first performed with the latent variable that corresponds to the face validity of the Schiphol Social Council 'points' choice task specifically. The results of this multiple regression are shown in table 4.15. With regard to this latent variable, there are seven coefficients that load significant in step two, while controlling for the other independent variables. The constant is also significant.

First, there is the dummy variable living together. When someone lives together, this person will rate the face validity of the Schiphol Social Council 'points' choice task lower than a someone who is single, does not live together, is a widow(er) or is divorced. When a person lives in an owner-occupied house, this person rates the latent variable with the coefficient of 0,087 higher based on a five-point Likert scale compared to a respondent who lives in a rental house. The regression coefficients of being (totally) unsatisfied or neutral with the influence

on decisions about Schiphol can be interpreted as follows. When someone is (totally) unsatisfied with his or her influence or neutral, this person will rate the face validity of the Schiphol Social Council 'points' choice task lower than someone who is (totally) satisfied (coefficient (totally) unsatisfied = -0,129; coefficient neutral satisfied = -0,197). Next to that, there is the regression coefficient of the dummy variable neutral reliability of information about Schiphol. If someone considers the reliability of information about Schiphol to be neutral, this person will score lower on the latent variable than someone who considers the information about Schiphol as (totally) reliable (coefficient = -0,150). Finally, it follows from table 4.15 that if a person thinks the information about Schiphol is (totally) dependent or neutral dependent, this person will rate the face validity of the 'points' choice task about the Schiphol Social Council lower than someone who thinks the information is (totally) independent (coefficient (totally) dependent = -0,133; coefficient neutral independence = -0,205).

The adjusted R square shows that when all variables are added, the model explains 11,8 percent variation of the latent variable. Of the explained variance of the entire model, 1,2 percent is explained by the general demographics alone. Furthermore, the results of the partial F tests are shown at the bottom of table 4.14. In step two this coefficient is significant. This means that adding case-specific characteristics leads to a significant improvement of the model (coefficient = -0,094).

Table 4.15: Results of multiple regression of face validity of Schiphol Social Council 'points' choice task ('points' experiment)

Independent variables	Step 1		Step 2	
	Coefficient	P-value	Coefficient	P-value
General demographics				
Dummy – high educated	0,100	0,108	0,103	0,087
Dummy – living together	-0,094	0,057	-0,094	0,048
Type of house – owner-occupied house	0,068	0,125	0,087	0,039
Municipality – 'inner' area 58 dB(A)	-0,095	0,026	-0,075	0,068
Case-specific characteristics				
Spending free time near Schiphol – yes			0,073	0,079
Dummy – (totally) unsatisfied with influence			-0,129	0,019
Dummy – neutral satisfied with influence			-0,197	0,000
Dummy – neutral reliability of information			-0,150	0,002
Dummy – (totally) dependent information			-0,133	0,032
Dummy – neutral independence of information			-0,205	0,000
Model information				
Constant	3,433	0,000	3,943	0,000
Partial F test	1,523	0,098	3,886	0,000
Adjusted R square	0,012		0,118	

Second, a multiple regression analysis is performed with the latent variable that corresponds to the face validity of the PVE consultation in general. The outcomes are presented in table 4.16. Regarding this latent variable, there are six variables in step two that loads significantly when they are controlled for the other independent variables. Also, the constant of step two is significant as shown in the model information of table 4.16.

The first significant regression coefficient is of the dummy variable 18-34 years. This means that respondent with an age between 18- and 34-years old rate the face validity of the PVE consultation in general lower than people who are 65 year or older (coefficient = -0,216). Furthermore, if someone is highly educated, this person will rate the face validity of the PVE consultation in general higher than a person who has a low education level (coefficient = 0,146). The regression coefficient of working at Schiphol can be interpreted as follows. When a person is employed at Schiphol, this person will rate the face validity of the PVE consultation in general lower than someone who is not employed at Schiphol (coefficient = -0,096). Then there is the dummy variable neutral satisfied with influence on decisions about Schiphol. It follows that if a person considers his or her influence on decisions to be neutral, this person will rate the latent variable lower than a person who is (totally) satisfied with his or her influence (coefficient = -0,139). The last two significant regression coefficients relate to the reliability of information about Schiphol. If someone considers the information around Schiphol to be (totally) unreliable or neutral with regard to this reliability, this person will rate the face validity of the PVE consultation in general lower than a person who considers the information around Schiphol to be (totally) reliable (coefficient (totally) unreliable = -0,149; coefficient neutral reliable = -0,115).

The adjusted R square shows that when all variables are fed in, the model explains 7,5 percent variance of the latent variable. Of the explained variance of the entire model, 2,5 percent is explained by the general demographics. Moreover, the coefficients of the partial F tests are significant in both steps. This means that adding a subgroup of variables in each step of this regression model leads to a significant improvement of the model.

Table 4.16: Results of multiple regression of face validity of PVE consultation in general ('points' experiment)

	Step 1		Step 2	
Independent variables	Coefficient	P-value	Coefficient	P-value
General demographics				
Dummy – 18-34 years	-0,173	0,013	-0,216	0,002
Dummy – high educated	0,174	0,005	0,146	0,017
Dummy – living together	-0,097	0,047	-0,084	0,084
Case-specific characteristics				
Working for Schiphol – yes			-0,096	0,028
Inconvenienced by Schiphol – yes			0,077	0,090
Dummy – neutral satisfied with influence			-0,139	0,009
Dummy – (totally) unreliable information			-0,149	0,003
Dummy – neutral reliability of information			-0,115	0,017
Model information				
Constant	3,868	0,000	4,116	0,000
Partial F test	2,077	0,012	2,748	0,000
Adjusted R square	0,025		0,075	

Finally, a multiple regression analysis is performed with the latent variable that consists of all the five categories of face validity included in this research. Table 4.17 presents the results. With regard to this latent variable, there are seven significant regression coefficients in step

two when they are controlled for the other independent variables. The constant of step two is also significant.

As in the previous multiple regression analysis, the coefficients of the dummy variables 18-34 years and high educated are significant. A person between the ages of 18 and 34 years rate the face validity of the PVE consultation in general lower than people who are 65 year or older (coefficient = -0,202). If someone is highly educated, this person will rate the face validity of the PVE consultation in general higher than someone who has a low education level (coefficient = 0,151). Furthermore, when a person lives together, this person will rate the face validity of the PVE method lower than a person who is single, does not live together, is a widow(er) or is divorced (coefficient = -0,104). Then there is the dummy variable neutral satisfied with influence on decisions about Schiphol. It follows that if a person considers his or her influence on decisions to be neutral, this person will rate the latent variable lower than a person who is (totally) satisfied with his or her influence (coefficient = -0,192). The next two significant regression coefficients relate to the reliability of information about Schiphol. If someone considers the information around Schiphol to be (totally) unreliable or neutral with regard to this reliability, this person will rate the face validity of the PVE consultation in general lower than a person who considers the information around Schiphol to be (totally) reliable (coefficient (totally) unreliable = -0,138; coefficient neutral reliable = -0,152). Finally, it follows from table 4.17 that if people think neutral about the information about Schiphol, these people will rate the face validity of the PVE method lower than someone who thinks the information is (totally) independent (coefficient neutral independence = -0,167).

The adjusted R square shows that if all independent variables are included, the model explains 11,4 percent variance of the latent variable. Of the explained variance of the entire model, 2,5 percent is explained by the general demographics. Furthermore, the coefficients of the partial F tests are significant in both steps. This means that adding a subgroup of variables in each step of this regression model leads to a significant improvement of the model.

Table 4.17: Results of multiple regression of face validity of PVE method ('points' experiment)

Independent variables	Step 1		Step 2	
	Coefficient	P-value	Coefficient	P-value
General demographics				
Dummy – 18-34 years	-0,125	0,037	-0,202	0,003
Dummy – high educated	0,169	0,006	0,151	0,012
Dummy – living together	-0,113	0,022	-0,104	0,030
Type of house – owner-occupied house	0,071	0,108	0,081	0,057
Case-specific characteristics				
Working for Schiphol – yes			-0,072	0,092
Spending free time near Schiphol – yes			0,073	0,082
Dummy – neutral satisfied with influence			-0,192	0,000
Dummy – (totally) unreliable information			-0,138	0,005
Dummy – neutral reliability of information			-0,152	0,001
Dummy – neutral independence of information			-0,167	0,005
Model information				
Constant	3,694	0,000	4,047	0,000
Partial F test	2,054	0,013	3,763	0,000
Adjusted R square	0,025		0,114	

4.6 Results explanatory personal characteristics of the face validity categories

Besides studying whether and which characteristics of the respondents influence the assessment of the latent variables, it is also researched whether and which personal characteristics influence the evaluation of the five face validity categories apart included in this research. Therefore, multinomial logistic regressions are performed for each category for each of the two experiments. This results in ten regressions of which the outcomes are presented below. To apply the multinomial logistic regression, the independent and dependent variables are recoded. An overview of this recoding is presented in Appendix J. The tables presented below are abbreviated. The variables with an significance level less than 0,050 are shown and the coefficient that are significant are shown in blue. The complete result tables are presented in Appendix K. Because hierarchical multinomial logistic regression is applied, the result table consist of a step one and a step two. The steps have been drawn up in the methodology (chapter 2). To answer the sub question, the focus is on the significant coefficients in step two. These coefficients or multinomial logits are indicated with a C and the significance with an S. Moreover, at the end of each result table the likelihood ratio test is stated. The significance of this test is shown. Furthermore, the coefficient presented are the standardized regression coefficients. In section 4.6.1 the multinomial logistic regressions of the 'sliders' experiment are discussed and in section 4.6.2 the regressions of the 'points' experiment are discussed.

4.6.1 Results multinomial logistic regressions of the 'sliders' experiment

First of all, a multinomial logistic regression is performed with the clarity category as dependent variable. The outcomes are presented in table 4.18. There are seven attributes in step two that load significantly on one or more comparisons when they are controlled for the other independent variables.

The multinomial logit for fulltime working respondent relative to non-working respondents is 1,348 unit lower for choosing neutral relative to (totally) disagree given all other predictor variables in the model are held constant. Furthermore, respondents living in the city center are more likely than respondents living out of the city to choose neutral or totally) agree in comparison with (totally) disagree. Respondents who have children living at home are more likely than respondents who don't have children living at home to choose neutral relative to (totally) agree (coefficient = 0,483). In addition, the multinomial logit for people living in the 'inner' area relative to respondents living in the 'outside' area is 0,912 unit lower for choosing neutral instead of (totally) disagree. In other words, respondents living in the 'inner' area are more likely than people in the 'outside' area to choose (totally) disagree. Moreover, respondents that are a member of a citizen organization are more likely to choose neutral or (totally) agree in comparison with (totally) disagree. Finally, participants that are (totally) dissatisfied with their influence on decisions about Schiphol are more likely to choose neutral instead of (totally) agree (coefficient = 0,618), while people who are neutral about their influence are more likely to choose neutral instead of (totally) disagree (coefficient = 1,015).

The significance of the likelihood ratio test is not significant in step two which means that the full model does not predict the clarity category better than the intercept-only model (significance = 0,141). In contrary, the likelihood ratio test of step one is significant (significance = 0,049). This means that the model with the general demographics predicts the clarity category better than the intercept-only model. The pseudo R square indicates that when all independent variables are included, the model explains 5,8 percent variance of the clarity.

Table 4.18: Results of multinomial logistic regression of clarity for 'sliders' experiment (where 'C' stands for coefficient and 'S' stands for significance)

Attributes	(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree		(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree	
	Step 1		Step 1		Step 1		Step 2		Step 2		Step 2	
	C	S	C	S	C	S	C	S	C	S	C	S
General demographics												
Working life												
Fulltime	-1,250	0,039	-0,242	0,323	-1,008	0,090	-1,348	0,029	-0,203	0,419	-1,144	0,059
Parttime	-0,831	0,220	-0,216	0,432	-0,615	0,354	-0,927	0,177	-0,179	0,521	-0,748	0,267
Not working	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Living environment												
Out of city	-0,874	0,047	-0,038	0,855	-0,835	0,051	-0,987	0,032	-0,053	0,806	-0,934	0,036
Outside city centre	-0,364	0,431	0,153	0,460	-0,516	0,255	-0,484	0,310	0,138	0,512	-0,622	0,184
Inside city centre	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Children living at home	-0,045	0,910	0,418	0,033	-0,463	0,230	-0,032	0,939	0,483	0,017	-0,515	0,203
Municipality	-0,824	0,033	-0,370	0,087	-0,454	0,215	-0,912	0,028	-0,416	0,065	-0,496	0,208
Case specific characteristics												
Member of citizen organization							18,034	0,000	-0,046	0,899	18,080	0,000
Satisfied with influence												
(Totally) dissatisfied							0,574	0,305	0,618	0,029	-0,044	0,939
Neutral							1,015	0,023	0,306	0,162	0,708	0,101
(Totally) satisfied							REF	REF	REF	REF	REF	REF
Model information												
Nr. Observations	648		648		648		648		648		648	
Likelihood ratio test	0,049		0,049		0,049		0,141		0,141		0,141	
Pseudo R^2	0,037		0,037		0,037		0,058		0,058		0,058	

Second, a multinomial logistic regression is performed with the unambiguity category as dependent variable. Table 4.19 presents the results. There are six attributes in step two that load significantly on one or more comparisons when they are controlled for the other independent variables.

A first significant multinomial logit is 18-34 years. When a respondent is between 18 and 34 years old, this respondent is relative to respondent who are 65 years or older more likely to rate unambiguity with neutral to (totally) disagree and to (totally) agree controlled by all other independent variables. Compared to woman, men are more likely to rate unambiguity with neutral instead of (totally) agree (coefficient = 0,456). Furthermore, people who work fulltime or parttime are less likely to rate unambiguity with neutral or (totally) agree relative to (totally) disagree compared to non-working people. It is notable that respondents who live in the 'inner' area are less likely to choose neutral regarding both (totally) agree and (totally) disagree. In the last comparison, it appears that residents of the 'inner' area are less likely to rate unambiguity with (totally) disagree compared to (totally) agree relative to residents in the 'outside' area. Finally, the multinomial logits for respondents who are (totally) dissatisfied with their influence on decisions about Schiphol and who are neutral about their influence relative to (totally) satisfied are 0,671 and 0,484 unit higher for choosing neutral over (totally) agree with respect to unambiguity.

The likelihood ratio test of step two significant which means that the full model does predict the unambiguity category better than the intercept-only model (significance = 0,022).

Following the pseudo R square, the model explains 6,6 percent variance of unambiguity when all independent variables are included. Of the explained variance of the entire model, 4,1 percent is explained by the general demographics.

Table 4.19: Results of multinomial logistic regression of unambiguity for 'sliders' experiment (where 'C' stands for coefficient and 'S' stands for significance)

	(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree		(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree	
	Step 1		Step 1		Step 1		Step 2		Step 2		Step 2	
Attributes	C	S	C	S	C	S	C	S	C	S	C	S
General demographics												
Age												
18-34 years	1,506	0,010	0,726	0,031	0,780	0,152	1,579	0,009	0,956	0,006	0,623	0,267
35-64 years	0,751	0,155	0,382	0,221	0,369	0,447	0,715	0,177	0,434	0,172	0,280	0,567
65 plus	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Gender	-0,053	0,868	0,402	0,039	-0,455	0,117	-0,056	0,864	0,456	0,024	-0,512	0,089
Working life												
Fulltime	-1,476	0,004	-0,136	0,600	-1,340	0,005	-1,448	0,005	-0,160	0,553	-1,288	0,008
Parttime	-1,460	0,007	-0,074	0,802	-1,386	0,006	-1,409	0,009	-0,055	0,855	-1,354	0,007
Not working	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Children living at home	-0,483	0,149	-0,413	0,052	-0,069	0,821	-0,407	0,239	-0,320	0,147	-0,086	0,785
Municipality	-1,133	0,001	-0,541	0,028	-0,592	0,042	-1,311	0,000	-0,667	0,010	-0,643	0,037
Case specific characteristics												
Satisfied with influence												
(Totally) dissatisfied							0,709	0,157	0,671	0,027	0,038	0,935
Neutral							0,617	0,106	0,484	0,042	0,134	0,702
(Totally) satisfied							REF	REF	REF	REF	REF	REF
Model information												
Nr. Observations	648		648		648		648		648		648	
Likelihood ratio test	0,012		0,012		0,012		0,022		0,022		0,022	
Pseudo R^2	0,041		0,041		0,041		0,066		0,066		0,066	

Third, a multinomial logistic regression is performed with the relevance category as dependent variable. Table 4.20 shows the results of this regression. There are five attributes in step two that load significantly on one or more comparisons when they are controlled for the other independent variables.

The significant attributes result in the following. A respondent between 18 and 34 years old is more likely than a respondent aged 65 or older to rate relevance with neutral relative to (totally) agree (coefficient = 0,973). Men are more likely to rate relevance with neutral instead of (totally) agree (coefficient = 0,625) compared to women. Furthermore, respondents who have children living at home are less likely than respondents who do not have children living at home to rate the relevance with neutral or (totally) agree relative to (totally) disagree. The multinomial logit for respondents spending their free time near Schiphol relative to respondents who do not spend their free time near Schiphol, is 1,168 unit lower for choosing (totally) agree over (totally) disagree. Respondents who are (totally) dissatisfied with their influence on decisions about Schiphol are relative to people who are (totally) satisfied less likely to rate relevance with neutral compared to (totally) disagree.

Both the likelihood ratio tests are not significant. This means that the model which includes the demographics does not predict the category relevance better than the intercept-only model. The pseudo R square has a value of 6,9 percent in step two. The model explains 6,9 percent variance of relevance when all independent variables are included. The general demographics themselves explain 4,2 percent variance.

Table 4.20: Results of multinomial logistic regression of relevance for 'sliders' experiment (where 'C' stands for coefficient and 'S' stands for significance)

	(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree		(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree	
	Step 1		Step 1		Step 1		Step 2		Step 2		Step 2	
Attributes	C	S	C	S	C	S	C	S	C	S	C	S
General demographics												
Age												
18-34 years	0,403	0,586	0,947	0,008	-0,544	0,436	0,380	0,626	0,973	0,009	-0,592	0,421
35-64 years	0,329	0,629	0,581	0,082	-0,253	0,692	0,434	0,535	0,616	0,072	-0,182	0,781
65 plus	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Gender	0,138	0,745	0,636	0,002	-0,498	0,215	0,090	0,839	0,625	0,003	-0,535	0,204
Children living at home	-0,968	0,033	-0,079	0,717	-0,889	0,039	-1,057	0,025	-0,059	0,794	-0,998	0,025
Case specific characteristics												
Spending free time near Schiphol							-0,893	0,140	0,275	0,460	-1,168	0,037
Satisfied with influence												
(Totally) dissatisfied							-1,432	0,028	-0,422	0,213	-1,010	0,095
Neutral							-0,275	0,621	0,250	0,290	-0,525	0,326
(Totally) satisfied							REF	REF	REF	REF	REF	REF
Model information												
Nr. Observations	648		648		648		648		648		648	
Likelihood ratio test	0,064		0,064		0,064		0,114		0,114		0,114	
Pseudo R^2	0,042		0,042		0,042		0,069		0,069		0,069	

In the fourth multinomial logistic regression readability is the dependent variable. The outcomes of this regression are shown in table 4.21. Four attributes have a significant value in step two given that all other independent variables are held constant in the model.

Compared to women, men are more likely to rate readability with neutral instead of (totally) agree (coefficient = 0,517). Furthermore, the multinomial logit for high educated participants is 0,852 unit higher for choosing neutral instead of (total) agree relative to participants that have a low education level. It is remarkable that respondents who spend their free time near Schiphol are more likely to choose neutral regarding readability instead of both (totally) disagree and (totally) agree. In the last comparison, it appears that residents who spend their free time near Schiphol are more likely to rate readability with (totally) agree compared to (totally) disagree.

Both the likelihood ratio tests are not significant. This means that the model with the demographics does not predict the category readability better than the intercept-only model. Following the pseudo R square, the model explains 6,3 percent variance of readability when all independent variables are included. Of the explained variance of the entire model, he general demographics themselves explain 3,5 percent variance.

Table 4.21: Results of multinomial logistic regression of readability for 'sliders' experiment (where 'C' stands for coefficient and 'S' stands for significance)

	(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree		(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree	
	Step 1		Step 1		Step 1		Step 2		Step 2		Step 2	
Attributes	C	S	C	S	C	S	C	S	C	S	C	S
General demographics												
Gender	0,818	0,060	0,513	0,015	0,305	0,455	0,764	0,091	0,517	0,017	0,248	0,558
Educational level												
High	0,568	0,392	0,924	0,004	-0,356	0,558	0,575	0,402	0,852	0,009	-0,277	0,661
Medium	0,283	0,672	0,545	0,094	-0,262	0,670	0,243	0,725	0,522	0,114	-0,279	0,660
Low	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Case specific characteristics												
Spending free time near Schiphol							18,547	0,000	0,692	0,042	17,855	0,000
Model information												
Nr. Observations	648		648		648		648		648		648	
Likelihood ratio test	0,262		0,262		0,262		0,354		0,354		0,354	
Pseudo R^2	0,035		0,035		0,035		0,063		0,063		0,063	

Finally, a multinomial logistic regression is performed with completeness as the dependent variable. Table 4.22 shows the results of this regression. Four attributes have a significant value in step two given that all other independent variables are held constant in the model. These four attributes are part of the general demographics.

The significant multinomial logits of the age group 18-34 year and educational level high can be interpreted in the same way as in the previous table 4.20. Compared to respondents who live in a rental house, respondents who live in an owner-occupied house are less likely to rate completeness with neutral relative to (totally) agree. Furthermore, both significant multinomial logits of the municipality variable are negative. It follows from this that people living in the 'inner' area are less likely to evaluate completeness with neutral or (totally) disagree compared to people living the 'outside' area.

The likelihood ratio test of step two is not significant with a significance value of 0,354. Therefore, the model with general demographics and the case-specific characteristics does not predict the category completeness better than the intercept-only model. Following the pseudo R square, the model explains 6,2 percent variance of completeness when all the characteristics are included. 3,9 Percent variance is explained by the general demographics.

Table 4.22: Results of multinomial logistic regression of completeness for 'sliders' experiment (where 'C' stands for coefficient and 'S' stands for significance)

	(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree		(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree	
	Step 1		Step 1		Step 1		Step 2		Step 2		Step 2	
Attributes	C	S	C	S	C	S	C	S	C	S	C	S
General demographics												
Age												
18-34 years	0,642	0,385	0,642	0,385	0,074	0,916	1,057	0,195	0,722	0,044	0,334	0,666
35-64 years	1,169	0,098	1,169	0,098	0,747	0,264	1,300	0,076	0,441	0,176	0,859	0,217
65 plus	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Educational level												
High	-0,099	0,889	-0,099	0,889	-0,767	0,259	-0,046	0,951	0,633	0,029	-0,679	0,335
Medium	-0,196	0,784	-0,196	0,784	-0,661	0,332	-0,187	0,799	0,427	0,140	-0,613	0,382
Low	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Type of house	-0,231	0,613	-0,231	0,613	0,251	0,566	-0,370	0,440	-0,487	0,019	0,117	0,798
Municipality	-1,280	0,004	-1,280	0,004	-0,957	0,019	-1,476	0,002	-0,390	0,117	-1,086	0,015
Model information												
Nr. Observations	648		648		648		648		648		648	
Likelihood ratio test	0,109		0,109		0,109		0,275		0,275		0,275	
Pseudo R^2	0,039		0,039		0,039		0,062		0,062		0,062	

4.6.2 Results multinomial logistic regressions of the 'points' experiment

This section follows the same structure as section 4.6.1, but focuses on the multinomial logistic regressions of the 'points' experiment. First, a multinomial logistic regression is performed in which clarity is defined as the dependent variable. The outcomes are presented in table 4.23. There are four attributes in step two that load significantly on one or more comparisons when controlled for the other independent variables.

The group of elderly people (65 plus) is less likely to evaluate clarity with neutral or (totally) agree relative to (totally) disagree compared to the age group of 18 to 34 years old. The multinomial logit for people that are inconvenienced by Schiphol relative to people that are not inconvenienced is 1,155 unit lower for choosing neutral relative to (totally) disagree given that all other predictor variables in the model are held constant. Respondents who regard their influence on decisions about Schiphol as neutral are more likely to evaluate clarity with neutral than (totally) agree relative to respondents who are (totally) satisfied with their influence on decisions about Schiphol. Finally, the positive multinomial logit of respondents who regard the reliability of information about Schiphol as neutral is significant as well. People who assess the reliability of information with neutral are more likely to assess the clarity with neutral than (totally) agree compared to people who regard information about Schiphol as (totally) reliable.

The likelihood ratio test in step two is significant (significance = 0,000). This means that the full model does predict the clarity better than the intercept-only model. The peudo R square in step two indicates that 10,3 percent variance is explained when all independent variables are included. Of the explained variance of the entire model, 2,9 percent is explained by the general demographics.

Table 4.23: Results of multinomial logistic regression of clarity for 'points' experiment (where 'C' stands for coefficient and 'S' stands for significance)

Attributes	(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree		(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree	
	Step 1		Step 1		Step 1		Step 2		Step 2		Step 2	
	C	S	C	S	C	S	C	S	C	S	C	S
General demographics												
Age												
18-34 years	-1,168	0,116	-0,053	0,871	-1,115	0,127	-1,452	0,040	0,268	0,450	-1,720	0,023
35-64 years	-0,825	0,253	0,275	0,356	-1,099	0,124	-1,094	0,138	0,225	0,482	-1,319	0,070
65 plus	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Case specific characteristics												
Inconvenienced by Schiphol							-1,155	0,018	-0,356	0,203	-0,799	0,088
Satisfied with influence												
(Totally) dissatisfied							-0,482	0,444	0,480	0,145	-0,962	0,114
Neutral							0,037	0,942	0,747	0,003	-0,710	0,154
(Totally) satisfied							REF	REF	REF	REF	REF	REF
Reliability of information												
(Totally) unreliable							-0,107	0,874	0,607	0,121	-0,714	0,269
Neutral							0,394	0,375	0,877	0,000	-0,482	0,263
(Totally) reliable							REF	REF	REF	REF	REF	REF
Model information												
Nr. Observations	528		828		528		528		528		528	
Likelihood ratio test	0,416		0,416		0,416		0,000		0,000		0,000	
Pseudo R^2	0,029		0,029		0,029		0,103		0,103		0,103	

Second, unambiguity is the dependent variable in a multinomial logistic regression. The results are presented in table 4.24. With regard to unambiguity, there are seven attributes in step two that are significant on one or more comparisons when they are controlled for the other predictor variables.

Regarding education level, it is remarkable that highly educated people are more likely to evaluate unambiguity with (totally) agree than with neutral or (totally) disagree relative to low-educated people. Moreover, people who live in an owner-occupied house are more likely to rate unambiguity with (totally) agree than with (totally) disagree compared to people who live in a rental house. It is remarkable that the multinomial logit for people who live near a fight path to people who do not live near to a flight path is 0,790 unit higher for choosing (totally) agree compared to (totally) disagree. Furthermore, people who work for Schiphol are less likely to evaluate unambiguity with (totally) disagree than neutral or (totally) agree relative to people who do not work at Schiphol. People who work for an organization that works closely with Schiphol are less likely to choose neutral instead of (totally) agree. Additionally, there is the reliability of information about Schiphol. People who regard the reliability of information as neutral are more likely to evaluate unambiguity with neutral or (totally) disagree compared to (totally) agree relative to people who think the information is (totally)reliable. Finally, the variable about the independence of information about Schiphol has significant attributes. People who regard the information as (totally) dependent are more likely to evaluate unambiguity with neutral instead of (totally) agree compared to people who regard the information as (totally) independent. In addition, people who have assessed the independence of information with neutral are more likely to assess unambiguity with neutral

than with (totally) disagree or (totally) agree compared to people who regard the information as (totally) independent.

The pseudo R square shows that when all independent variables are included, the model explains 11,4 percent variance of unambiguity. 3,8 Percent of the explained variance from the entire model can be explained by general demographics. Furthermore, the likelihood ratio test is significant in step two (significance = 0,000). The entire model predicts unambiguity better than the intercept-only model.

Table 4.24: Results of multinomial logistic regression of unambiguity for 'points' experiment (where 'C' stands for coefficient and 'S' stands for significance)

	(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree		(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree	
	Step 1		Step 1		Step 1		Step 2		Step 2		Step 2	
Attributes	C	S	C	S	C	S	C	S	C	S	C	S
General demographics												
Educational level												
High	0,163	0,757	-0,666	0,015	0,829	0,106	0,247	0,659	-0,656	0,025	0,903	0,039
Medium	-0,111	0,813	-0,257	0,329	0,146	0,751	0,099	0,843	-0,200	0,475	0,299	0,540
Low	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Type of house	0,762	0,043	-0,005	0,979	0,767	0,035	0,775	0,056	-0,086	0,690	0,861	0,028
Case specific characteristics												
Living near flight path							0,750	0,078	-0,040	0,866	0,790	0,043
Working for Schiphol							-1,659	0,032	-0,269	0,627	-1,390	0,044
Working closely with Schiphol							-1,293	0,084	-1,073	0,030	-0,220	0,738
Reliability of information												
(Totally) unreliable							-0,608	0,372	0,325	0,419	-0,933	0,148
Neutral							-0,322	0,497	0,432	0,042	-0,753	0,040
(Totally) reliable							REF	REF	REF	REF	REF	REF
Independence of information												
(Totally) dependent							0,196	0,772	1,240	0,002	-1,044	0,089
Neutral							1,367	0,026	1,567	0,000	-0,200	0,719
(Totally) independent							REF	REF	REF	REF	REF	REF
Model information												
Nr. Observations	582		582		582		582		582		582	
Likelihood ratio test	0,100		0,100		0,100		0,000		0,000		0,000	
Pseudo R^2	0,038		0,038		0,038		0,114		0,114		0,114	

Third, a multiple logistic regression is performed with the independent variable relevance. The results are shown in table 4.25. Regarding the relevance, there are six attributes in step two that are loading significant on one or more comparisons while controlled for the other predictor variables.

Men are more likely to evaluate relevance with (totally) agree than with neutral or (totally) disagree compared to women. The multinomial logit for high educated people relative to low educated people is 1,184 unit higher for choosing (totally) agree in comparison with (totally) disagree. Furthermore, the multinomial logit is also higher for people who live close to a flight path relative to people who do not live close to a flight path with a unit of 1,251 for choosing neutral instead of (totally) disagree. The variable reliability of information about Schiphol has two significant attributes as well. People who regard the reliability of information as (totally)

unreliable are less likely to evaluate relevance as (totally) agree than as (totally) disagree compared to people who regard the information as (totally) reliable. People who regard the reliability of information as neutral are more likely to evaluate relevance as neutral than as (totally) agree compares to people who regard the information as (totally) reliable.

The likelihood ratio test in step two is significant (significance = 0,022). The full model does predict the relevance better than the intercept-only model. The pseudo R square in step two indicates that 9,6 percent variance of relevance is explained when all independent variables are included. Of the explained variance of the entire model, 4,3 percent is explained by the general demographics.

Table 4.25: Results of multinomial logistic regression of relevance for 'points' experiment (where 'C' stands for coefficient and 'S' stands for significance)

	(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree		(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree	
	Step 1		Step 1		Step 1		Step 2		Step 2		Step 2	
Attributes	C	S	C	S	C	S	C	S	C	S	C	S
General demographics												
Gender	0,347	0,538	-0,572	0,007	0,920	0,040	0,566	0,351	-0,410	0,048	0,976	0,044
Educational level												
High	0,763	0,295	-0,419	0,158	1,182	0,042	0,848	0,284	-0,337	0,275	1,184	0,039
Medium	0,213	0,733	-0,141	0,620	0,354	0,553	0,342	0,622	-0,064	0,826	0,406	0,540
Low	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Case specific characteristics												
Living near flight path							1,251	0,040	-0,145	0,548	1,395	0,017
Reliability of information												
(Totally) unreliable							-1,224	0,169	0,474	0,277	-1,698	0,042
Neutral							0,659	0,339	0,650	0,009	0,008	0,990
(Totally) reliable							REF	REF	REF	REF	REF	REF
Independence of information												
(Totally) dependent							0,710	0,437	0,136	0,737	0,574	0,506
Neutral							1,295	0,047	0,252	0,451	1,043	0,152
(Totally) independent							REF	REF	REF	REF	REF	REF
Model information												
Nr. Observations	582		582		582		582		582		582	
Likelihood ratio test	0,196		0,196		0,196		0,022		0,022		0,022	
Pseudo R^2	0,043		0,043		0,043		0,096		0,096		0,096	

Fourth, a multinomial logistic regression is performed with readability as a dependent variable. Table 4.26 presents the outcomes. With regard to readability, there are nine attributes in step two that are loading significant on one or more comparisons.

The age group 18-34 years is more likely to evaluate readability with neutral or (totally) disagree relative to (totally) agree compared to the age group 65 plus. In terms of educational level, there are two significant attributes. High-educated people are more likely to rate readability with (totally) agree than with (totally) disagree in comparison with low-educated people. In addition, medium-educated people are less likely to assess readability with neutral than with (totally) agree relative to low-educated people. Moreover, the multinomial logit for people working for Schiphol is 1,061 unit higher for choosing relative to (totally) agree. It is notable that people who work for an organization that works closely with Schiphol are more

likely to choose (totally) disagree or (totally) agree instead of neutral. Furthermore, people who regard their influence on decisions about Schiphol as neutral are less likely to choose neutral or (totally) agree instead of (totally) disagree compared to people who are (totally) satisfied with their influence. The variable reliability of information about Schiphol contains two significant attributes. People who regard the information as (totally) unreliable are more likely to evaluate readability with (totally) disagree than with (totally) agree compared to people who regard the information as (totally) reliable. Additionally, people who regard the reliability of information as neutral are more likely to evaluate readability as neutral than as (totally) agree. Finally, people who regard the independence of information as neutral are more likely to also regard the readability as neutral than as (totally) agree compared to people who regard the information as (totally) independent.

The likelihood ratio tests in both steps are significant. Furthermore, the pseudo R square in step two indicates that 13,5 percent variance of readability is explained when all independent variables are included. The general demographics cause 6,7 percent variance of readability.

Table 4.26: Results of multinomial logistic regression of readability for 'points' experiment (where 'C' stands for coefficient and 'S' stands for significance)

	(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree		(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree	
	Step 1		Step 1		Step 1		Step 2		Step 2		Step 2	
Attributes	C	S	C	S	C	S	C	S	C	S	C	S
General demographics												
Age												
18-34 years	-1,811	0,119	0,750	0,047	-2,561	0,023	-2,076	0,094	0,942	0,027	-3,019	0,012
35-64 years	-1,000	0,392	0,128	0,734	-1,128	0,319	-1,153	0,345	0,010	0,980	-1,163	0,325
65 plus	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Educational level												
High	1,021	0,138	-0,692	0,023	1,712	0,010	1,301	0,085	-0,578	0,070	1,879	0,010
Medium	-0,317	0,561	-0,688	0,021	0,371	0,470	-0,137	0,819	-0,630	0,043	0,493	0,382
Low	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Case specific characteristics												
Working for Schiphol							-0,153	0,852	1,061	0,039	-1,213	0,113
Working closely with Schiphol							-2,214	0,006	-1,098	0,046	-1,117	0,080
Satisfied with influence												
(Totally) dissatisfied							-1,287	0,160	-0,601	0,136	-0,686	0,429
Neutral							-1,291	0,049	0,153	0,588	-1,444	0,022
(Totally) satisfied							REF	REF	REF	REF	REF	REF
Reliability of information												
(Totally) unreliable							-1,345	0,105	0,611	0,208	-1,956	0,010
Neutral							0,565	0,326	0,454	0,038	0,112	0,839
(Totally) reliable							REF	REF	REF	REF	REF	REF
Independence of information												
(Totally) dependent							0,681	0,438	0,371	0,401	0,310	0,703
Neutral							1,116	0,097	0,812	0,022	0,304	0,624
(Totally) independent							REF	REF	REF	REF	REF	REF
Model information												
Nr. Observations	582		582		582		582		582		582	
Likelihood ratio test	0,003		0,003		0,003		0,000		0,000		0,000	
Pseudo R^2	0,067		0,067		0,067		0,135		0,135		0,135	

Finally, a multinomial logistic regression is performed for the 'points' experiment with completeness as the dependent variable. The outcomes are presented in table 4.27. There are nine attributes in step two that all load significantly on one or more comparisons while they are controlled for the other predictor variables.

The age groups 18 to 34 years and 35 to 64 years are less likely to evaluate completeness with neutral or (totally) agree relative to (totally) disagree compared to the age group 65 plus. Men are more likely to rate completeness with neutral than with (totally) agree compared to women. Moreover, the multinomial logit for high-educated people relative to low-educated people is 1,369 unit higher for choosing (totally) agree compared to (totally) disagree. Respondents who live in an owner-occupied house are less likely to rate completeness with neutral relative to (totally) agree compared to respondents who live in a rental house. Furthermore, the multinomial logit for people living in the 'inner' area relative to people living in the 'outside' area is 2,217 unit higher for choosing neutral instead of (totally) disagree. Moreover, the multinomial logit for people working for Schiphol relative to people working not for Schiphol is 1,587 lower for choosing (totally) agree instead of (totally) disagree. Compared to people who are not customers of Schiphol, customers of Schiphol are more likely to rate completeness with neutral or (totally) agree relative to (totally) disagree. Respondents who regard their influence on decisions about Schiphol as neutral are more likely to evaluate completeness with neutral than (totally) agree relative to respondents who are (totally) satisfied with their influence on decisions about Schiphol. Finally, people who regard the reliability of information as neutral are more likely to evaluate completeness as neutral than as (totally) agree compared to people who consider information as (totally) reliable.

The likelihood ratio test is significant in step two (significance = 0,000). The full model does predict the relevance better than the intercept-only model. The pseudo R square shows that when all independent variables are included, the model explains 11,9 percent variance of completeness. 4,9 Percent of the explained variance from the entire model can be explained by general demographics.

Table 4.27: Results of multinomial logistic regression of completeness for 'points' experiment (where 'C' stands for coefficient and 'S' stands for significance)

Attributes	(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree		(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree	
	Step 1		Step 1		Step 1		Step 2		Step 2		Step 2	
	C	S	C	S	C	S	C	S	C	S	C	S
General demographics												
Age												
18-34 years	-16,133	0,000	0,054	0,878	-16,187	0,000	-17,220	0,000	0,211	0,580	-17,432	0,000
35-64 years	-15,249	0,000	-0,031	0,924	-15,217	0,000	-16,223	0,000	-0,064	0,856	-16,159	0,000
65 plus	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Gender	0,263	0,625	0,353	0,096	-0,090	0,861	0,103	0,868	0,581	0,011	-0,477	0,430
Educational level												
High	0,727	0,337	-0,299	0,314	1,026	0,159	1,220	0,152	-0,149	0,632	1,369	0,046
Medium	0,217	0,748	-0,092	0,746	0,309	0,634	0,307	0,677	0,045	0,879	0,262	0,712
Low	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Type of house	-0,772	0,158	-0,630	0,003	-0,142	0,788	-0,822	0,180	-0,631	0,005	-0,192	0,747
Municipality	1,738	0,103	0,251	0,292	1,486	0,158	2,217	0,044	0,291	0,245	1,926	0,102
Case specific characteristics												
Working for Schiphol							-0,881	0,295	0,707	0,153	-1,587	0,046
Schiphol customer							1,241	0,030	-0,276	0,213	1,517	0,023
Satisfied with influence												
(Totally) dissatisfied							-0,857	0,359	0,153	0,682	-1,010	0,259
Neutral							0,677	0,379	0,589	0,031	0,087	0,907
(Totally) satisfied							REF	REF	REF	REF	REF	REF
Reliability of information												
(Totally) unreliable							-1,129	0,234	0,287	0,528	-1,417	0,111
Neutral							0,863	0,236	0,664	0,008	0,200	0,778
(Totally) reliable							REF	REF	REF	REF	REF	REF
Model information												
Nr. Observations	582		582		582		582		582		582	
Likelihood ratio test	0,076		0,076		0,076		0,000		0,000		0,000	
Pseudo R^2	0,049		0,049		0,049		0,119		0,119		0,119	

4.7 Results identified clusters

The above multiple regressions and the multinomial logistic regressions show which characteristics of respondents influence the evaluation of face validity. In addition to these regressions, LCCA's are performed as well to identify groups with certain characteristics that collectively score high or low on certain categories of face validity. In total, two LCCA's are performed. This entails one LCCA for the 'sliders' experiment and one LCCA for the 'points' experiment. The result tables below are abbreviated. Only the covariates with a p-value less than 0,050 are shown. The complete result tables of the LCCA's are shown in Appendix L. The profile measures of the clusters are displayed in the result tables. Furthermore, the Wald statistics are presented in the result tables with the associated p-values. In section 4.7.1 the results of the LCCA of the 'sliders' experiment are shown. Section 4.7.2 presents the results of the LCCA of the 'points' experiment.

4.7.1 Identified clusters of the 'sliders' experiment

Table 4.28 presents the results of the LCCA of the 'sliders' experiment. The most optimal number of clusters is based on a BIC-value of 658,614. It follows from the BIC-value that three clusters can be identified. Cluster 1 consists of 55,64% of the sample, cluster 2 consists of 29,7% of the sample and cluster 3 contains 14,65% of the sample.

Table 4.28: Results of LCCA of the 'sliders' experiment

		Cluster 1	Cluster 2	Cluster 3	Wald	P-value
Cluster size		0,5564	0,2971	0,1465	-	-
Indicators						
Clarity	(Totally) disagree	0,0038	0,1237	0,1804	107,1037	5,5e-24
	Neutral	0,1829	0,6025	0,6205		
	(Totally) agree	0,8132	0,2738	0,1191		
Unambiguity	(Totally) disagree	0,0018	0,1571	0,4173	85,8031	2,3e-19
	Neutral	0,1023	0,5549	0,4914		
	(Totally) agree	0,8958	0,2980	0,0912		
Relevance	(Totally) disagree	0,0004	0,1712	0,0016	98,1450	4,9e-22
	Neutral	0,0718	0,6462	0,1364		
	(Totally) agree	0,9278	0,1826	0,8620		
Readability	(Totally) disagree	0,0003	0,1550	0,0030	97,5663	6,5e-22
	Neutral	0,0503	0,5869	0,1476		
	(Totally) agree	0,9494	0,2582	0,8494		
Completeness	(Totally) disagree	0,0017	0,1473	0,0025	90,7421	2,0e-20
	Neutral	0,1236	0,6064	0,1479		
	(Totally) agree	0,8746	0,2463	0,8496		
Covariates						
Age	18-34 years	0,2825	0,4102	0,2698	8,5402	0,014
	35-64 years	0,5178	0,4332	0,5627		
	65 years and older	0,1997	0,1566	0,1765		
Gender	Man	0,4923	0,5914	0,5617	7,0141	0,030
	Woman	0,5077	0,4086	0,4383		
Educational level	High	0,4424	0,5596	0,3801	8,3381	0,015
	Medium	0,3617	0,3202	0,4078		
	Low	0,1958	0,1201	0,2121		
Satisfied with influence	(Totally) unsatisfied	0,1602	0,1898	0,2805	6,5622	0,038
	Neutral	0,4822	0,4897	0,5195		
	(Totally) satisfied	0,3575	0,3205	0,2000		

All entered indicators that consists of the five face validity categories included in this study have a significant p-value. This indicates that the null hypothesis of the Wald test is rejected. Therefore, the coefficients of the indicators are not equal to zero. Regarding the covariates the general demographics and the case-specific characteristics are entered. Age, gender, educational level and satisfied with influence are significant. Therefore, the coefficients of these coefficients are not equal to zero.

Cluster 1 represents the approving raters of face validity. This cluster contains the highest percentage of (totally) agree on each indicator. Therefore, it the cluster that best assessed the face validity. For the answer option '(totally) agree', clarity has a share of 81,32%, unambiguity of 89,58%, relevance of 92,78%, readability of 94,94% and completeness a share of 87,46%. Cluster 1 has the largest share of the oldest age group which consists of people aged 65 years or older (19,97%) compared to the other two clusters. Moreover, cluster 1 contains the highest percentage of women compared to the other two clusters (50,77%). Furthermore, cluster 1 has relatively the most people who are (totally) satisfied with their influence on decisions taken around Schiphol (35,75%). In contrast, cluster 1 has also the least share of a number of characteristics compared to the other two clusters. Cluster 1 contains the fewest men (49,23%) and also the fewest people who are (totally) unsatisfied with their influence on decisions about Schiphol (16,02%).

The average raters of face validity are represented by cluster 2. This cluster has the highest share of the 'neutral' answer option for each indicator. For the answer option 'neutral', clarity has a share of 60,25%, unambiguity of 49,14%, relevance of 64,62%, readability of 58,69% and completeness a share of 60,64%. With regard to the characteristics of the persons in this cluster, the following is noticeable. Cluster 2 contains the largest share of people of 18-34 years old compared to the other two clusters (41,02%). Compared to the other two clusters, cluster 2 contains also the highest proportion of men (59,14%). Moreover, this cluster contains the highest proportion of highly educated people compared to the other clusters (55,96%). In contrast, cluster 2 has also the least share of a number of characteristics compared to the other two clusters. Cluster 2 contains out of the least people of 35-64 years old (43,32%) and of 65 years and older (15,66%). There are also the fewest women of this cluster (40,86%). Finally, this cluster also contains the fewest medium-educated (32,02%) and low-educated people (12,01%) compared to the other two clusters.

Cluster 3 represents the dependent raters of face validity. Within this cluster, the clarity and unambiguity categories are assessed most with (totally) disagreement with regard to the other two clusters. Clarity has a share of 18,04% in '(totally) disagree' and unambiguity 41,73%. These two categories were asked immediately after the Schiphol Social Council 'sliders' choice task. The categories relevance, readability and completeness have in contrast a high share of '(totally) agree'. Relevance has a share of 86,20%, readability of 84,94% and completeness 84,96% on the answer option '(totally) agree'. These three categories were questioned at the end of the PVE consultation. With regard to the characteristics of the persons in this cluster, the following is noticeable. Cluster 3 consists of the largest share of people of 35-64 years old (56,27%). Moreover, cluster 3 contains the highest share of medium-educated (40,78%) and low-educated people (21,21%) compared to the other two clusters. Finally, people who regard their influence on decisions about Schiphol as neutral (51,95%) or (totally) unsatisfied (28,05%) also have the highest share in cluster 3 compared to the other two clusters. In contrast, cluster 3 has also the least share of a number of characteristics compared to the other two clusters. Cluster 3 contains the lowest proportion of 18–34-year-old people (26,98%), the lowest proportion of highly educated people (38,01%) and the lowest proportion of people who are (totally) satisfied with their influence on decisions about Schiphol (20,00%).

4.7.2 Identified clusters of the 'points' experiment

Table 4.28 presents the results of the LCCA of the 'points' experiment. The most optimal number of clusters is based on a BIC-value of 491,903. It follows that three clusters are identified. Cluster 1 consists of 39,63% of the sample, cluster 2 consists of 38,96% of the sample and cluster 3 contains 21,41% of the sample.

Table 4.29: Results of LCCA of the 'points' experiment

		Cluster 1	Cluster 2	Cluster 3	Wald	P-value
Cluster size		0,3963	0,3896	0,2141	-	-
Indicators						
Clarity	(Totally) disagree	0,0007	0,1011	0,1356	27,1657	1,3e-6
	Neutral	0,0730	0,5369	0,5652		
	(Totally) agree	0,9263	0,3620	0,2991		
Unambiguity	(Totally) disagree	0,0001	0,0957	0,1548	8,1134	0,017
	Neutral	0,0247	0,5229	0,5675		
	(Totally) agree	0,9752	0,3814	0,2777		
Relevance	(Totally) disagree	0,0002	0,0092	0,1435	49,9491	1,4e-11
	Neutral	0,0395	0,2486	0,5899		
	(Totally) agree	0,9603	0,7422	0,2666		
Readability	(Totally) disagree	0,0001	0,0012	0,2222	31,4570	1,5e-7
	Neutral	0,0408	0,1258	0,6504		
	(Totally) agree	0,9591	0,8730	0,1274		
Completeness	(Totally) disagree	0,0003	0,0019	0,1485	64,7862	8,5e-15
	Neutral	0,0711	0,1671	0,6796		
	(Totally) agree	0,9286	0,8310	0,1719		
Covariates						
Age	18-34 years	0,2499	0,3404	0,4269	12,0621	0,002
	35-64 years	0,5404	0,4680	0,4433		
	65 years and older	0,2097	0,1917	0,1297		
Educational level	High	0,5259	0,3798	0,3578	9,3326	0,009
	Medium	0,3689	0,4190	0,4167		
	Low	0,1052	0,2012	0,2255		
Inconvenienced by Schiphol	Yes	0,2182	0,1997	0,1235	1,5379	0,046
	No	0,7818	0,8003	0,8765		
Reliability of information	(Totally) unreliable	0,0951	0,0998	0,1159	9,4972	0,009
	Neutral	0,3726	0,5641	0,5710		
	(Totally) reliable	0,5323	0,3361	0,3131		
Independence of information	(Totally) dependent	0,2216	0,3021	0,1716	6,4960	0,039
	Neutral	0,5025	0,5951	0,6354		
	(Totally) independent	0,2758	0,1028	0,1931		

All entered indicators that consists of the five face validity categories included in this study have a significant p-value. This indicates that the null hypothesis of the Wald test is rejected. Therefore, the coefficients of the indicators are not equal to zero. With regard to the covariates the general demographics and the case-specific characteristics are entered. Age, educational level, inconvenienced by Schiphol, reliability of information, and independence of

information are significant. Therefore, the coefficients of these coefficients are not equal to zero.

Cluster 1 represents the approving raters of face validity. This cluster has the highest percentage of '(totally) agree' on each indicator, like the LCCA of the 'sliders' experiment. This makes it the cluster that best assessed the face validity. For the answer option '(totally) agree', clarity has a share of 92,63%%, unambiguity of 97,52%%, relevance of 96,03%, readability of 95,91% and completeness a share of 92,86%. Compared to the other two clusters, cluster 1 has the largest share of the age group 35-64 year (54,04%) and of the age group which consists of people aged 65 years or older (20,97%). Moreover, cluster 1 contains the largest share of high-educated people compared to the other two clusters (52,59%). It is noticeable that cluster 1 also contains the largest share of people that are inconvenienced by Schiphol in comparison with the other clusters (21,82%). Furthermore, cluster 1 has the highest proportion of people who regard the reliability of information about Schiphol as (totally) reliable (53,23%). Finally, cluster 1 has the highest proportion of people who regard the independence of information as (totally) independent compared to the other clusters (27,58%).

Cluster 2 represents the dependent raters of face validity. The indicators of this cluster nowhere have the highest or lowest share per answer option and in that regard score average everywhere. However, it is noticeable that clarity and unambiguity are rated worse in this cluster than relevance, readability and completeness. Clarity and unambiguity are asked immediately after the Schiphol Social Council 'points' choice task and the other three categories are questioned at the end of the PVE consultation. With regard to the characteristics of the persons in this cluster, the following is noticeable. Cluster 2 contains the highest share of medium-educated (41,90%) compared to the other two clusters. With regard to the independence of information, cluster 2 contains the highest proportion of people who consider information as (totally) dependent (30,21%) compared to the other two clusters. In contrast, cluster 2 has the lowest share of people who regard the information about Schiphol as (totally) independent (10,28%).

Cluster 3 represents the average raters of face validity. This cluster has the highest share of the 'neutral' answer option for each indicator like cluster 3 in the LCCA of experiment 1. For the answer option 'neutral', clarity has a share of 56,62%, unambiguity of 56,75%, relevance of 58,99%, readability of 65,04% and completeness a share of 67,96%. With regard to the characteristics of the persons in this cluster, the following is noticeable. Cluster 3 contains the largest share of people of 18-34 years old compared to the other two clusters (42,69%). Compared to the other two clusters, cluster 3 contains the largest share of low-educated people (22,55%). It is noticeable that cluster 3 also consists of the largest share of people that are not inconvenienced by Schiphol compared to the other clusters (87,65%). With regard to the reliability of information, this LCCA results in the following. Compared to the other clusters, people who regard the reliability of the information about Schiphol as neutral (57,10%) or as (totally) unreliable (31,31%) are most represented in cluster 3. Finally, people who consider the independence of information about Schiphol as neutral are most represented in cluster 3 compared to the other two clusters (63,54%).

5. The benchmarks of the evaluation of face validity categories and the influence of case study properties

This chapter elaborates on the fundamentals of the benchmarks of the evaluation of face validity categories and the properties of a case study regarding the PVE method that may influence the evaluation of face validity. The third sub question is the focus of this chapter. As described in the methodology (section 2.6), there is a number of categories of face validity that have been included in this study which were also surveyed in previous PVE consultations. The previous consultations concern the categories clarity, relevance and completeness. The structure of this section is designed in such a way that these three categories are dealt with in turn. Previous PVE consultations are compared one by one with each other and the Schiphol Environmental Council consultation on the basis of one of the categories. This chapter starts with a description of the previous case studies on the basis of a typology. Of the face validity categories, the category clarity is discussed first. The cases about the climate consultation, the heat transition vision of Utrecht, the long-term Corona policy and the Schiphol Environmental Council consultation are compared with each other. This is followed by the relevance category in which the cases of the heat transition vision of Utrecht, the long-term Corona policy, Foodvalley and the Schiphol Environmental consultation are compared with each other. Finally, there is the completeness category. Here, the Foodvalley case is compared with the Schiphol Environmental Council consultation.

5.1 Description of previous PVE consultation case studies

This section contains the description of the previous case studies in which a PVE consultation took place and a face validity category is questioned. The section ends with a typology to compare the case studies.

5.1.1 The climate consultation case study

To achieve the national climate targets by 2030, measures must be applied to reduce emission of greenhouse gases such as CO_2. It is up to the government to make decisions about measures. Members of the Parliament have previously indicated that that they would like to involve the Dutch society. Therefore, the aim of this climate consultation case is to enable Dutch society to advice the government on climate policy. The PVE method has been applied for this (Mouter et al., 2021a). The old version of the online platform is used where respondents could advice the government.

In the climate consultation case, participants were given ten measures that reduce greenhouse gases. Information about the effects of the measures was shown to the citizens as well. The participants were asked to indicate to what extent the government must apply the measured to achieve the national climate targets. Thereafter, there was room for respondents to motivate their choices (Mouter et al., 2021a). Since the climate targets must be achieved by 2030, this consultation concerns measures for the long term. In this research, a ‘sliders’ choice task is applied. Therefore, this case is compared with the ‘sliders’ experiment of the Schiphol Environmental Council case which also consists of a ‘sliders’ choice task. These two case studies are the only two that contain a ‘sliders’ choice task. Furthermore, in the case study of the climate consultation the researchers are in charge of shaping the consultation.

The climate consultation case and the Schiphol case are compared based on the scores of the clarity statement. The clarity statement that is questioned in the climate consultation is as follows: I think that this research provides participants with sufficient information to be able to give useful advice to the government (Mouter et al., 2021a). This statement could be answered on a five-point Likert scale where 1 stands for totally disagree and 5 for totally agree, just like in the Schiphol case. In the climate consultation case study, anyone above eighteen years old living in the Netherlands was allowed to take part in the climate consultation.

5.1.2 The heat transition vision of Utrecht case study

A second case in which the category clarity is studied, is the heat transition vision Utrecht study. The Climate Agreement indicates that involving residents is crucial for a successful energy transition. That is why the participation process of the heat transition vision Utrecht includes various resources to involve different groups of stakeholders in the energy transition. One of the resources is a PVE performed by Delft University of Technology and the VU Amsterdam. In the case of the heat transition, residents were asked to give advice on achieving the objective to make 40.000 houses natural gas free by 2030. Therefore, this consultation focuses on the long term. The PVE consists of two steps. In the first step, residents could divide one hundred points between four different approaches. After residents divided their points, they were asked to motivate their choices. In the second step, residents could put together their own approach. This set-up was drawn up with stakeholders who were in charge. With the distribution of one hundred points, a 'points' choice task is used (Mouter et al., 2020). Therefore, the clarity statement of this case is compared with other consultations that also applied the 'points' choice task. Furthermore, any resident over the age of eighteen living in the city of Utrecht was allowed to participate in the consultation. So, this case makes use of an open consultation. However, this consultation still took place in the old version of the online platform.

The clarity statement included in the heat transition vision Utrecht case is presented to respondents at the end of the consultation. The clarity statement is formulated in the following way: "I received enough information to make a choice" (Mouter et al., 2020). This statement could be answered on a five-point Likert scale with 1 for totally disagree and 5 for totally agree. The heat transition vision Utrecht consultation also includes a statement about the relevance. The relevance statement in the Utrecht case formulated in the following way: "This is a good method for involving citizens in choosing between approaches to makes houses gas-free" (Mouter et al., 2020). The answer options for this statement consist of the same five-point Likert scale as the relevance statement in the Schiphol Environmental Council case. This statement is presented to the respondents at the end of the consultation.

5.1.3 The long-term Corona policy case study

A third case is the case of the long-term Corona policy. The aim of the study about Corona policy is that the Dutch government is thinking about designing a long-term strategy. This research is conducted by Populytics in collaboration with the National Institute for Public Health and the Environment (RIVM). The RIVM and the other stakeholders involved were in charge of setting up the PVE consultation. The aim of this research is to answer four questions related to the preferences of citizens about the implementations of Corona measures and the pursuit of goals with the Corona policy. Therefore, two different experiments, or two different PVE consultations, were carried out. In the first consultation, four scenarios are presented to

the respondents on how the Corona pandemic may develop. Citizens gave advice on the use of measures for each scenario. Since Corona was present in the Dutch population at the time of the consultation, the Corona measures had a direct impact on the people's personal lives. Therefore, this first experiment focuses on measures that can be deployed in the short term. In the second consultation, the preference of respondents regarding the goals of the Corona policy are examined. Respondents could assign points to various goals of this Corona policy (Geijsen et al., 2022). These goals are more tied to the values people have. Therefore, this second experiment focuses on the people's long-term values and the impacts on their personal life. These two different experiments are treated in this study as two different PVE consultations.

In the Corona policy case study, a data panel is used for a representative sample of respondents. Furthermore, there was an open consultation in which anyone could participate. Due to the availability of data, the results of the panel consultation are included in this research. Moreover, this consultation has been carried out in the new version of the online platform.

At the end of both experiments, a number of questions are asked about the respondents' experiences with this Corona policy research. One of these questions concerns the clarity category: "I believe that this study provided sufficient information to be able to provide advice to the government" (Geijsen et al., 2022). Like in the Schiphol Environmental Council case this question is answered on a five-point Likert scale where a score of 1 means totally disagree and 5 totally agree. Because both the experiments contain so-called 'points' choice tasks, the results of this clarity statement are compared with other case studies that also applied the 'points' choice task. In addition to clarity, the category relevance is also reflected in the long-term Corona policy case. The relevance is questioned in both experiments of this case at the end of the consultation where the respondents are asked a number of questions about their experiences with the Corona policy research. The statement about relevance contains five answer options on a Likert scale, which corresponds to the answer options in the Schiphol Environmental Council case. In the statement about relevance, respondents are asked whether this is a good method for involving citizens in choices that the government has to make regarding Corona policy (Geijsen et al., 2022).

5.1.4 The Foodvalley case study

The aim of the Foodvalley case is to involve residents of the Foodvalley Region in the future energy policy of their region. This PVE consultation is carried out as part of four steps that the Foodvalley Region is taking together with Populytics, Public Mediation and Platform Civic Participation in Government Policy (NPBO) to involve citizens in future energy policy. In the consultation of Foodvalley, five options were presented to the respondents about future energy. These five options are not concrete plans, but they are able to expose preferences and the advantages and disadvantages of the options. So, the focus of this consultation is on energy policy in the long term. The participants are asked to divide one hundred points among the five options (Spruit & Mouter, 2021). The Foodvalley case contains a 'points' choice task. Therefore, it is be compared with other consultations which also contains the 'points' choice task. During the setting up of this PVE consultation, the stakeholders were in charge.

In addition to the choice task, the respondents in the Foodvalley case were also asked whether the consultation they completed is a good way to involve citizens in making choices about the future (Spruit & Mouter, 2021). This statement is measured on a five-point Likert scale with 1 indicating totally disagree and 5 totally agree. Moreover, this statement is asked at the end of the consultation. The Foodvalley case contains a statement regarding completeness as well. The completeness statement is formulated in the following way: "I have had enough space to express my opinion" (Spruit & Mouter, 2021). The answer options for this statement consist of the same five-point Likert scale as the relevance statement. This statement was also presented to the respondents at the end of the consultation.

Moreover, any citizen over eighteen years old who lives in the Foodvalley region could participate in the consultation. So, there is an open consultation. In the Foodvalley case, the old version of the online platform was used where respondents could give their advice.

5.1.5 A typology of previous PVE consultation case studies

The case studies regarding the Schiphol Environmental Council, the climate consultation, the heat transition vision of Utrecht, the Corona policy and Foodvalley all have their own characteristics. The descriptions of these case studies above show that they differ in at least four factors. These are the impact on the respondent's personal life in the short or long term, whether it was an open or panel consultation, whether the new or old version of the online platform was used and who is in charge during the setup of the PVE consultation, the stakeholders or the researchers. Table 5.1 shows these four characteristics per case study.

Table 5.1: A typology of the characteristics of PVE consultation case studies

	Impact on personal life	Respondents	Platform	Who are in charge?
Schiphol	Short-term	Panel	New	Stakeholders
Climate consultation	Long-term	Open	Old	Researchers
Heat transition vision Utrecht	Long-term	Open	Old	Stakeholders
Long-term Corona policy exp. 1	Short-term	Panel	New	Stakeholders
Long-term Corona policy exp. 2	Long-term	Panel	New	Stakeholders
Foodvalley	Long-term	Open	Old	Stakeholders

5.2 The clarity category and the case study characteristics

Mann-Whitney U tests are carried out in order to make a comparison between the results of the clarity category in two different case studies with their characteristics. This entails that it is tested, for example, whether there is a significant difference in the scores that the respondents gave to the clarity statement between the Schiphol Environmental Council case study and the climate consultation. Therefore, the case studies which contain a statement about clarity are all compared with each other. However, the case studies with 'points' choice tasks are compared with 'points' choice tasks and 'sliders' choice tasks with 'sliders' choice tasks. The results of the tests of normality are presented in Appendix M. These tests of normality show that none of the variables is significant and therefore is not normally distributed. Table 5.2 presents the results of the Mann-Whitney U tests for the clarity category.

Table 5.2: Results of Mann-Whitney U tests of clarity between PVE consultation case studies

	PVE consultation	Number	Average score	Mann-Whitney U	P-value
Clarity	Schiphol	684	3,59	610191,00	0,003
	Climate consultation	2028	3,67		
Clarity	Schiphol	582	3,60	262681,00	0,913
	Heat transition vision Utrecht	321	3,53		
Clarity	Schiphol	582	3,60	559960,00	0,121
	Corona policy (exp. 1)	2005	3,62		
Clarity	Schiphol	582	3,60	776153,50	0,000
	Corona policy (exp. 2)	2005	3,84		
Clarity	Heat transition vision Utrecht	321	3,53	308221,00	0,006
	Corona policy (exp. 1)	2005	3,62		
Clarity	Heat transition vision Utrecht	321	3,53	275149,50	0,000
	Corona policy (exp. 2)	2005	3,84		
Clarity	Corona policy (exp. 1)	2005	3,62	1645817,50	0,000
	Corona policy (exp. 2)	2005	3,84		

From table 5.2 it becomes clear that that the scores on the clarity statement significantly differs between two cases in five of the seven Mann-Whitney U tests. First, the scores of the clarity statement in the Schiphol Environmental Council case differ significantly from the scores of the clarity statement in the climate consultation case (p-value = 0,003). The average score on the five-point Likert scale in the Schiphol case is 3,59 compared to an average of 3,67 in the climate consultation case. Moreover, these two cases differ on all four of the case study characteristics. Furthermore, the clarity from the perspective of the respondents is significantly higher in experiment two of the Corona policy case compared to the Schiphol Environmental Council case (p-value = 0,000). On the five-point Likert scale, experiment two of the Corona policy case has an average score of 3,84 while the Schiphol case has an average score of 3,60. It is remarkable that these two cases differ in one characteristic. The Schiphol case focuses on the short-term impact on the personal life of people and Corona experiment two on the long-term. Moreover, the scores of the clarity statement in the Utrecht case differ significantly from the scores of the clarity statement in the first experiment of the Corona policy case (p-value = 0,006). The average score of the Utrecht case is significantly lower than in the Utrecht case. The Utrecht case has an average score of 3,53 and the first Corona policy experiment has an average of 3,62. The only similarity between these two cases is that the stakeholders were in charge during the setup of these PVE consultations. From table 5.2 it becomes clear as well that the scores of the clarity statement in the Utrecht case differ significantly from the scores of the clarity statement in the second experiment of the Corona policy case (p-value = 0,000). On the five-point Likert scale, experiment two of the Corona policy case has an average score of 3,84 while the Utrecht case has an average score of 3,53. These two cases differ in two characteristics. The Utrecht case made use of the old version of the online platform while the second Corona experiment made use of the new version. Furthermore, the Utrecht case is an open consultation, while in Corona experiment two a

panel was used. Finally, in the first experiment of the Corona case study is the clarity significantly higher than in the second experiment (p-value = 0,000). The first experiment has an average rate of 3,62 and the second experiment has an average rate of 3,84. These two experiments only differ on the impact on the personal life, where the first experiment focuses on the short-term and the second experiment focuses on the long-term.

Additionally to the five Mann-Whitney U tests that show a significant difference, there are also two Mann Whitney U tests that show that there is no significant difference in the evaluation of the clarity category between two cases. On the one hand, it is remarkable that there is no significant difference in the evaluation of the clarity category between the Schiphol Environmental Council case and the Utrecht case (p-value = 0,913). These two cases differ in three of the four properties. The only similarity is that the stakeholders were in charge. On the other hand, there is no significant difference in the assessment of clarity between the Schiphol case and the first experiment of the Corona policy case study (p-value = 0,121). These two case studies have all four the characteristics of the typology in table 5.1 in common.

From the case studies above that include clarity it becomes clear that there is a bandwidth between the average scores 3,53 and 3,84 on a five-point Likert scale. On this scale, 1 stands for totally disagree and 5 stands for totally agree.

5.3 The relevance category and the case study characteristics

Mann-Whitney U tests are carried out again in order to make a comparison between the results of the relevance category in two different case studies with their characteristics. The results of the tests of normality are presented in Appendix M. These tests of normality show that none of the variables is significant and therefore is not normally distributed. Table 5.3 presents the results of the Mann-Whitney U tests for the clarity category.

Table 5.3: Results of Mann-Whitney U tests of relevance between PVE consultation case studies

	PVE consultation	Number	Average score	Mann-Whitney U	P-value
Relevance	Schiphol	582	3,88	372655,00	0,000
	Foodvalley	1556	3,53		
Relevance	Schiphol	582	3,88	71378,50	0,000
	Heat transition vision Utrecht	321	3,43		
Relevance	Schiphol	582	3,88	382369,00	0,801
	Corona policy (exp. 1)	2005	3,76		
Relevance	Schiphol	582	3,88	399378,00	0,014
	Corona policy (exp. 2)	2005	3,95		
Relevance	Foodvalley	1556	3,53	235861,00	0,097
	Heat transition vision Utrecht	321	3,43		
Relevance	Foodvalley	1556	3,53	1333072,00	0,000
	Corona policy (exp. 1)	2005	3,76		
Relevance	Foodvalley	1556	3,53	1207776,00	0,000
	Corona policy (exp. 2)	2005	3,95		
Relevance	Heat transition vision Utrecht	321	3,43	259551,50	0,000
	Corona policy (exp. 1)	2005	3,76		
Relevance	Heat transition vision Utrecht	321	3,43	231880,50	0,000
	Corona policy (exp. 2)	2005	3,95		
Relevance	Corona policy (exp. 1)	2005	3,76	1888733,50	0,000
	Corona policy (exp. 2)	2005	3,95		

Of the ten Mann-Whitney U tests performed in table 5.3, eight tests indicate that there is a significant difference between two case studies with regard to the assessment of the relevance category. All these ten tests have a p-value of 0,000. First, the Schiphol case has a significant higher average score of 3,88 for the relevance category on the five-point Likert scale than the Foodvalley case with an average score of 3,53. With regard to the properties in table 5.1, these two cases have nothing in common except that the stakeholders were in charge. Furthermore, the Schiphol case has also a significant higher evaluation of relevance than the Utrecht case. The Utrecht case has an average score of 3,43 on the five-point Likert scale. The only characteristic in common between the Schiphol and the Utrecht case is that the stakeholders were in charge. Next, the second experiment of the Corona case has a significant higher average score of 3,95 compared to the Schiphol case with an average score of 3,88 on the relevance category. Regarding the properties, these cases studies only differ in the impact on the personal life. Moreover, the scores of the clarity statement in the Foodvalley case differ significantly from the scores of the clarity statement in the first and second experiment of the Corona policy case. On the five-point Likert scale, experiment one of the Corona policy case has an average score of 3,76 and experiment two of 3,95 while the Foodvalley case has an average score of 3,53. Both Corona experiments made use of a panel of respondents and of the new platform, while Foodvalley was an open consultation and made use of the old platform. From table 5.3 it also becomes clear that the Foodvalley focuses on

the long-term impact on the personal life. Moreover, both the experiments of the Corona policy case have a significant higher average score on the relevance category than the Utrecht case. The average score of the Utrecht case on the five-point Likert scale is 3,43. The differences in properties between the Corona policy case and the Utrecht case are the same as the difference between the Corona policy case and the Foodvalley case. Finally, the second experiment of the Corona policy case has a significant higher average score of 3,95 for the relevance category on the five-point Likert scale than the first experiment of the same case with an average score of 3,76. These two experiments only differ in the impact on the personal life. Experiment one focuses on the short-term while experiment two focuses on the long-term.

Additionally to the eight Mann-Whitney U tests that show a significant difference, there are two Mann Whitney U tests that show that there is no significant difference in the evaluation of the relevance category between two cases. From table 5.3 it becomes clear that there is no significant difference between the evaluation of the relevance category of the Schiphol Environmental Council case and the first experiment of the Corona policy case (p-value = 0,801). Furthermore, there is no significant difference between the evaluation of the relevance category of the Foodvalley case and the heat transition vision Utrecht case which is depicted in table 5.3 as well (p-value = 0,097). It is remarkable that in both tests the cases show no differences in the four characteristics of table 5.1.

From the case studies above that include relevance it becomes clear that there is a bandwidth between the average scores 3,43 and 3,95 on a five-point Likert scale. On this scale, 1 stands for totally disagree and 5 stands for totally agree.

5.4 The completeness category and the case study characteristics

The Foodvalley case, as described in 5.1.4, contains a statement regarding completeness as well. Table 5.4 presents the results of the Mann-Whitney U test between the scores of the completeness statement of the Foodvalley case and the 'points' experiment of the Schiphol Environmental Council case.

Table 5.4: Results of Mann-Whitney U test of completeness PVE consultation case studies

	PVE consultation	Number	Average score	Mann-Whitney U	P-value
Completeness	Schiphol	582	3,89	530651,00	0,000
	Foodvalley	1556	4,11		

From table 5.4 it becomes clear that the scores on the completeness statement in the Foodvalley case significantly differ from the scores in the Schiphol Environmental Council case. The average score on the five-point Likert scale for this statement in the Foodvalley case is 4,11 compared to an average score of 3,89 in the Schiphol case. The respondents in the Foodvalley case evaluate the completeness of the consultation with a significantly higher score. That the two case studies differ significantly from each other in the evaluation of completes was to be expected since the two cases have only one of the four characteristics in common. In both cases, the stakeholders are in charge. It is noticeable regarding the completeness that the Foodvalley case consists of five options about future energy. However, these are not concrete plans but broad approaches for the long term. The Schiphol case, on

the other hand, contains concrete tasks of the Environmental House and the Schiphol Social Council.

Based on the Schiphol and Foodvalley cases, the bandwidth for the assessment of completeness is between 3,89 and 4,11 on a five-point Likert scale. On this scale, 1 stands for totally disagree and 5 stands for totally agree.

6. Face validity in practice

This chapter elaborates on the concerns of citizens and stakeholders regarding face validity in practice. This is done by addressing the fourth sub question of this research. In order to study the concerns in practice about face validity, four perspectives have been divided. In section 6.1, the concerns of local residents of Schiphol or respondents are discussed. Thereafter, the perspective of stakeholders who are involved in the process of designing the PVE consultation of the Schiphol Environmental Council is addressed. Third, the concerns from the perspective of a stakeholder who was not involved in the process of designing the PVE consultation of the Schiphol Environmental Council is discussed. Finally, the concerns about face validity in practice are explained from the perspective of the client.

6.1 The local residents and respondents

The perspective of the local residents and respondents has been collected in practice with regard to their concerns about face validity. Therefore, quotes from the open consultation are analysed, since face validity categories also emerge in these quotes that are not included in the consultation.

The statistical analyses performed in this research concern the panel consultation. In this panel consultation, an attempt is made to collect a representative sample of the population in the 54 municipalities around Schiphol. In addition to this panel, an open consultation is conducted. Every person aged eighteen or older is allowed to participate in this open consultation. Two additional questions related to face validity have been added to this open consultation. The questions are formulated in the following way: “please explain you answer to the previous questions”. The question was first asked after the statements about the clarity and unambiguity categories. This question was asked for the second time after the statements about the relevance, readability and completeness categories. From the analysis of the quotes from respondents to the open consultation, it is noticeable that respondents not only comment on the five included face validity categories but also other categories. That is why the answers on these questions are analysed to study the concerns about face validity in practice.

The face validity category that is most apparent in the quotes is completeness. First, one respondent indicated that the possible tasks of the Schiphol Social Council are not complete. Second, it appears that some respondents would have liked to have provided more information about a particular subject themselves. A third quote concerns the possibility of checking the option ‘other’ in a question.

“In addition to providing solicited and unsolicited advice to the minister and the House of Representatives, the Schiphol Social Council has to do many other things: 1. Coordination point, meeting place and support for all residents’ organisations in the area […] 13. Provide training for the members of the Schiphol Social Council, with the aim of ensuring that they are full discussion partners and that advice is given sufficient weight.”

“I would have liked to have been able to provide some more information about which ways/methods there are to involve citizens, and organisations that are specifically involved in this.”

“I sometimes missed the ‘other’ option.”

In addition to the comments of the respondents who indicate that the consultation is not totally complete, there are also respondents who indicate that the consultation is complete because of the sufficient opportunity for personal input or ‘overcomplete’.

“Plenty of room for personal input.”

“I think that the questions are broad enough and that I have had a good opportunity to clarify things.”

“Questions were here and there very ‘overcomplete’ with a lot of repetition, overlap.”

This ‘overcompleteness’ touches upon the next category. This is the feasibility category. This is also the most apparent category after completeness. The feasibility category is not included in the PVE consultation. A number of quotes follow that indicate that the length of the consultation constitutes a threshold to participate and therefore the feasibility is lower.

“I think such a long survey raises the threshold, I think many give up in advance or halfway through.”

“The questionnaire is too long and too detailed, so you lose the overview and there is a lot of repetition.”

“I found the questions too extensive and I think the focus is lacking. So, I worry that a support group will be set up with the Schiphol Social Council, which is certainly not my intention.”

“I think this takes a lot of time.”

“Bizarrely long survey aimed at discouraging participation.”

“This questionnaire is too long. In this way you scare people off and only the fanatics will participate.”

Two categories that appear significantly less in the quotes of the respondents of the open consultation, but which are included in the framework in this research, are clarity and unambiguity. The first quote below concerns clarity. The second quote also concerns clarity, since this respondent discusses the picture this respondent has obtained on the basis of information about the Schiphol Social Council. The insensitivity category emerges from this second quote as well. The distrust of the respondent could influence the assessment of face validity. The third and fourth quote below contain unambiguity.

“To be honest, I haven’t really gotten a picture of what the council exactly will do.”

“Didn’t get a clear picture or I mistrust the intent.”

“Questionnaire was long but clearly laid out.”

“There is not much to explain. Questions were clear.”

Furthermore, the face validity category aesthetics emerges in the quotes as well even though it was not included in the consultation. The quote below concerns the aesthetics of the mobile version of the online platform where respondents can complete the consultation.

“Good initiative. Curious about the outcome and what lessons can be learned from this. A shorter questionnaire would help. Also, that it is better to fill it in on a mobile phone.”

Finally, a summary quote is presented. This quote contains clarity, feasibility and completeness.

“I thought it was a fairly cumbersome survey with may duplications and not always clear questions. Also, too many open questions. Not a pleasant survey.”

6.2 Stakeholders involved in the process of designing the PVE consultation

The stakeholders involved in the process of designing the PVE consultation are listed in section 2.3. After analysing the data from the consultation, a draft report is drawn up. The stakeholders involved in the design of the consultation were given the opportunity to provide feedback on the draft report before it became public. Subsequently, the statements of the resident representatives on the ‘Regioforum’ are analysed.

6.2.1 Feedback on draft report

The stakeholders involved from Schiphol Group, Air Traffic Control the Netherlands, Natuur en Milieufederatie Noord-Holland, the municipality of Haarlemmermeer and Ouder-Amstel and four residents’ representatives of the Schiphol Social Environmental Council have communicated that they had no further comments on the draft report. The public affairs manager of VNO-NCW West had not responded at all to the request. However, one resident representative of the Schiphol Environmental Council submitted written comments. The comments regarding the face validity of the PVE consultation have been analysed. A total of three quotes can be traced back to face validity.

The first quote is as follows: “The participants of the Dynata panel are also used to this kind of complicated survey.” This first quote indicates the appropriateness of the difficulty level as well as the clarity. It is indicated that the survey is complex. In addition to a panel consultation, the consultation was also opened for everyone who wanted to participate. As a result, the quote can also be interpreted in such a way that the familiarity category is central in this quote. This is not included in the face validity framework that has been drawn up in this research.

The second quote is as follows: “Residents’ organisations have received many reactions that they [the citizens, red.] thought it was way too complicated.” This quote focuses in particular on the complexity of the consultation. Therefore, this quote can be linked to the appropriateness of the difficulty level which means how easy or difficult it is to make a choice

or can be linked to clarity. Clarity is about getting enough information to make a choice. It is not clear what kind of complexity is meant by this quote.

The third quote concerns the completeness category of face validity. The quote is as follows: “I have not come across a question about this.”

6.2.1 The ‘Regioforum’

The results of the PVE consultation were presented during the ‘Regioforum’ and the members of the Schiphol Environmental Council were allowed to comment on the presentation. During the comment round of the members of the council, no fundamental arguments were put forward about the PVE consultation with regard to face validity. However, one resident representative did consider face validity during the response round. The quote below states what the resident representative has claimed with regard to face validity.

“We have heard from many people that they thought it was so incredible complicated and long that they dropped out.”

The complexity mentioned by the resident representative can be linked to multiple categories of face validity. First, the complexity can be linked to the appropriateness of the difficulty level. This category focuses on how easy it is to make a choice. Second, complexity can also be linked to clarity, unambiguity or the readability since a lack of these three categories increase the complexity. Returning to section 3.2.2, an expert points out that the category of the difficulty level is a more inclusive category of complexity. On the other hand, clarity, unambiguity and readability are more demarcated categories. Both approaches measure the complexity.

The comment regarding the length of the PVE consultation can be linked to the feasibility category. In a further conversation with this representative, it turned out that the representative agrees that the PVE consultation is complete. This results in a trade-off between completeness and feasibility. Since the PVE consultation was longer, there was room to incorporate the ideas of every stakeholder in the consultation. If the consultation had been shorter, there would have been no room for the ideas of each stakeholder and the consultation would be regarded as incomplete.

6.3 Stakeholder who is not involved in the process of designing the PVE consultation

With regard to a stakeholder who is not involved in the design process of the PVE consultation, an interview is conducted with an employee of the mainport strategy department of the airline KLM. This stakeholder also received the draft report, as were the stakeholders who were involved in the design process. The reaction of this KLM employee was that this employee assumes that ‘the findings have been worked out neatly’, but that the questioning is somewhat guiding. This response forms the basis of an interview of which the protocol l can be found in Appendix N.

The answers given during the interview can be divided into three different categories of face validity. When asked why this employee thinks the question a bit guiding, the answer is as follows: "The question is somewhat closed. In particular, that you ask: what is important to you? Then you have six things that you can find important. You already are steering the answers here. You can only divide your points into a number of categories." The consequences

of a steering consultation touch on the face validity categories of completeness and insensitivity.

First, a consequence is that the consultation can be experienced as less complete if the consultation is guiding. "You start from a very limited scope of activities. That way you don't meet the people who want a free format. Then you also have to ask more open questions. If such a Schiphol Social Council exists, what are the most important tasks. And now all categories are already pre-sorted." Another quote that connects to the previous quote is as follows: "I think you then have a certain outcome which is already quite decisive, because people can't express their opinion outside of what you have written in the categories."

Another consequence of a consultative consultation is that the consultation evokes negative feelings. For example, the employee indicates: "You want to reach a goal, but then you still try to funnel and push even though the respondent does not want to click on the answers. However, the respondent must divide points to move a step further in the consultation." Therefore, the employee of KLM can well imagine that respondents would stop because of this. "If you have to divide points into categories that you don't feel comfortable with anyway, then you stop."

In addition to completeness and insensitivity, there is a third category of face validity that emerges in the interviewee's answers. It concerns the clarity of the consultation. The employee says about this: "It is quite difficult for people without context to fill this in. A number of colleagues in my department had also completed this. I had to explain some things to them about what exactly was the intention of such an Environmental House and what the difference was with the Schiphol Social Council." Moreover, it was not clear to the interviewee at first that the different tasks included in the consultation are based on the advice of van Geel (2020). For the interviewee, this does make a difference in the understanding of the tasks that have been included. If the tasks of the Schiphol Social Council and the Environmental House are really based on Van Geel's report, then this stakeholder would have worded it differently in the consultation: "That you first give a brief summary about van Geel. That a decision has already been taken and given that you decide to move on." Another suggestion from the KLM employee is to indicate the following in the consultation: "We do not start from scratch, because there is already an advice from van Geel and then you can illustratively indicate what he meant. Given this scope of activities, we want to weigh what is most important for the Schiphol Social Council." According to the interviewee, it is mainly about the context when it comes to clarity. It is important that the consultation is put in a certain context, such as in the context of van Geel's advice (2020). In this way it becomes clear on what information or knowledge the questions are based on.

6.4 The client

With regard to the client, two policy officers are interviewed who work at the ministry of Infrastructure and Water Management. These two policy officers who work for the client have been involved in the assignment of the PVE consultation about Schiphol. They have been closely involved in the design of the PVE consultation and the results of the consultation have also been presented to them. Moreover, a project manager of the ORS is interviewed. The purpose of these three interviews with both policy officers is to evaluate with the client what

went well and what could be improved regarding the phases that went through to set up the consultation.

The client interview protocol is presented in Appendix O. In this interview protocol, it is first asked what went well and what could be improved in each phase of the consultation set-up. A total of three phase are completed to set up the consultation. The first phase is about determining the goals and preconditions of the PVE consultation, phase two is about feedback on a more concrete version of the PVE and phase three is about feedback on the 99% version of the PVE consultation. Thereafter, the client is asked whether concessions have been made and what the client's role has been in this.

6.4.1 Evaluation of phase 1: the goals and preconditions of the PVE consultation

First, a policy officer indicated that he thought it was a good process in phase one. According to the first interviewee, this was mainly due to "going into it openly and collecting from residents and stakeholders what they want to know and how they can use the Schiphol Social Council and see the Environmental House. The ministry often argues that this is what the citizen means." He points out that the advantages of this open communication are that "on the one hand you are more transparent and on the other you are able to retrieve more objective information. You know better what is going on without filling this in for the other person." The second policy officer also points out the position of the ministry. "It is a problem for all of us. Especially in this specific example of Schiphol and the Environmental House, it is actually impossible for one party to solve it or to come up with the solution ourselves. We sometimes receive criticism within the ministry that the stakeholders are just informed. So, I thought it was really good that you were also so proactive about involving all stakeholders very early."

Furthermore, the first policy officer indicated that he was pleased with the meetings with stakeholders not being conducted by the ministry itself. "This gives stakeholders some degree of freedom to express themselves about the research design, without bringing things to the table such as: I speak with the ministry so I have to put forward my views well. So, it is an addition if an independent party collects the input from stakeholders without the client sitting directly at the table." Besides the fact that an independent design for shaping the PVE gives the stakeholders freedom, there are even more advantages, according to this policy officer. "If you speak to many stakeholders in advance that the set-up actually has more input or becomes better than if you only speak to a limited number of people. This also leads to a more supported design." The need for support for the research is also mentioned by the interviewee from the ORS. This ORS interviewee noticed the following: "At the final phase, I was very pleased with how little feedback was given from various parties, precisely because they were consulted in multiple stages." The policy officer continues his argument with: "If you speak to more people, you are more able to set up a better research method. With more people you have tested whether the method touches the core dilemma. You are able to obtain more perspectives. This does not mean that you have to give every perspective a place, you have to make trade-offs." The same argument is made by the second policy officer. "There are always parts or fields that you haven't thought of yet or that you think about as a researcher or ministry, we don't think that is that important. Then you pay less attention to it. If another party suddenly goes wild about it, that is a wake-up call. That can lead to different insights."

Another benefit of early stakeholder involvement is addressed by the second policy officer. “What is advantageous about it is that you can actually have a conversation about it. You sometimes notice that people get stuck in points of view and if you then open the conversation and enter into it, it eventually turns out that the points of view are a lot less strong than you first thought and that the parties also thought.” However, in phase one, this policy officer did not really have a good idea of who had been spoken to. This leads to the following improvement: “What could have been done better was to map out the entire process better. If in, then we're going to do this, then we're going to do that. In particular, what is expected of the various participants at a given moment.”

Two further suggestions for improvement were raised by the project manager. A first point from this interviewee is as follows: “I have the feeling that people are not always very aware of what is in the proposal. Precisely because it is communicated very shortly in advance. I can imagine that it will come out that way over time, but what I do remember about that proposal is that there was not a whole substantive discussion." In addition, this project manager also has a suggestion for a different angle that was not included in the design of the consultation: "If I look purely at residents who have built up expertise in the past period, they may have less need for an Environmental House for themselves since they have a position in the ORS. And as a result, the results may be different than when you asked blank residents, gosh what do you see in front of you? Perhaps it would have been even more complete if you had already included residents who are affected, i.e. from the Schiphol area, in your design beforehand.”

The above evaluations of phase one show that in this phase the client focuses on the completeness of the consultation. The client is pleased with the inclusions of stakeholders, with the open communication and with an independent party that has conversation with the stakeholders so they can speak freely. With the conversations with the stakeholders, multiple perspectives are also retrieved. In this way, all perspectives that are included in the consultation. Completeness is a face validity category that is included in the framework drawn up in section 3.3.2.

Complementary, the first policy officer indicates that involving stakeholders in several phases leads to more depth and therefore more completeness. "This allows residents and other stakeholders to think a layer deeper about what they think is really important.” This opinion is shared with the ORS project manager. “The sooner you do that [involving stakeholders, ed.], the more familiar people become with what is a PWE, how is it used, you name it. And I also think it is good to follow up on this by inserting different moments during the process so that people know better about this, it also contributes to sharpening and strengthening a vision or opinion about it. So, I think in that sense it benefits the content.”

6.4.2 Evaluation of phase 2: Feedback on the tightened PVE consultation

According to the first policy officer, what went well in this phase is that it was useful to not only give individual feedback via email, but also to have the online meetings. "This created more of a group conversation and you heard different opinions." What is also a positive element is that the client was given many moments in this phase to give feedback. "The feedback was also processed quickly." According to the second policy officer, the following went well: “I really liked how you took us through the storyboard and the readability test. That

is often forgotten. I thought it was very good that you pointed out why this is so important, because it also gives some insight into the research method."

According to the interviewee of the ORS, the feedback rounds to concretise the tasks of the Environmental House and the Schiphol Social Council were very helpful. "I think it works very well to just keep making translations on that. One round is really not enough. I think the more often you see something, the more you have the opportunity to consider whether this is what it is all about or are we still missing certain aspects in this. The number of moments has been very good." The project manager adds: "Although I do notice that you are quite capable of making things a little more understandable, while we often look for nuance from the ORS and ienw [ministry of Infrastructure and Water Management, red.]. It is a nice option because you can test again with your gosh do you mean by this one that you then say: yes guys this could be the case but we need to write it more understandable and we can maybe do that by using these and these words I think that's very good to make sure you keep talking about the same thing, but it is also understandable to an outsider." It can be deduced from this argumentation that the feedback rounds for concretising the tasks in the choice tasks have led to more clarity.

In addition to the positive elements there is also an area for improvement. The interviewee of the ORS suggests that an overview of the process must be created at the beginning of the process. "That can do so much with the feedback people give, because they can find more time for it or stay more in the research." In other words, this will improve the content of the feedback.

The first policy officer suggested another improvement. "Because it went so fast [with the feedback loops, red.], there were sometimes small errors or things that were not clearly expressed in the final list of questions. Sometimes that had to be removed again. It is extra important to formulate clearly since the members of the Schiphol Environmental Council could make commotion about this. This was mainly in the task descriptions of the Schiphol Social Council and the Environmental House. It would have been nice to create clarity earlier what those task descriptions exactly were." In this argumentation of the policy officer, concerns about the face validity categories unambiguity and clarity emerge.

According to the policy officer, a solution for this point of improvement lies in clearly stating what has changed and why. "It was less explicit in this second phase what input was given by us as a ministry or by the stakeholders. These flows of feedback ran parallel to each other. It would have been clear to hear in advance which tasks had changed due to the input from stakeholders so that we as a ministry could easily respond to that again." By communicating more openly about the changes, you create an understanding of why something has changed on the one hand, and you also create an overview so that it is easier to see what needs to be focused on when giving feedback. This could be achieved by not sending the whole consultation back and forth but only the parts that have been adjusted.

A concrete example of a communicated change is with regard to the order of the choice tasks of the Schiphol Social Council and the Environmental House. "It is good to state explicitly: we notice that there are all kinds of different opinions and this is what we are going to decide. This makes it easy to follow and is transparent."

In addition to the process of giving feedback, the first policy officer mentioned the language check in phase two. This policy officer experienced the language check as important in making the consultation accessible to everyone with the aim of completing it by as many respondents as possible. If there were no language checks, the consequences could be that people are more misunderstanding and will drop out sooner. This can also lead to a less representative study with more highly educated people. The policy officer cited a challenge to the ministry. "Policy texts are written at the Ministry, but this is not necessarily understandable for everyone. Can we use other words for this?"

The second policy officer also focuses on representativeness, but mentions an improvement. “Is the research representative? That is always a recurring question within the ministry, because officials want to be sure that they can use the research. That is to a large extent also covered by the panel, but you noticed that there is uncertainty about that. That might be something for further research. You may tell us quite at the beginning that we are going to do that panel soon, but if we also want to do a separate sample among a different group of people, that is possible, but keep this and this in mind.”

The above evaluations of phase two show that in this phase the client focuses mainly on the clarity and unambiguity of the consultation. The readability has also been named and the representativeness can be linked to the efficacy of the consultation.

6.4.3 Evaluation of phase 3: Feedback on the 99% version of the PVE consultation

When evaluating this third phase, no face validity categories were cited by the policy officers. However, one policy officer noted the following: "it is good to give stakeholders one last chance to think along since a lot of choices were made at the end of phase two. As far as confidence in the method is concerned, it is very good to let everyone have their last say." The ORS interviewee stated the following about phase three: “It's great that this phase has been included at all, because it's a plus if you communicate to citizens and stakeholders like: hey, that's the way we dealt with your comments. I think that also contributes greatly to the support of the draft report.”

6.4.4 Concessions

First, a policy officer refers to a concession on the length of the consultation. "If you see that people are not really willing to put in a lot of time, then it is a shame if not many people complete the open consultation. You saw for the paid panel that there were many reactions. It is good to improve accessibility to give more weight when you talk about the length of the open consultation. In this case of the ORS, support for the research has prevailed. That was a good choice for this case." This quote shows the concern between the balance of completeness and feasibility. As a possible solution, the policy officer cites here that the length trade-offs may differ between the panel and the open consultation. For example, a shorter version for the open consultation should be prepared.

This concession between completeness and feasibility is also reflected in the interview with the second policy officer. A number of his colleagues had completed the consultation. “A colleague thought the consultation was really good, but really long. I have also heard that from a number of other colleagues. They said: really good that you are doing this, but I can imagine that people give up because it really takes twenty to twenty-five minutes.” This policy officer

reflects on this by: “I think we may have made too few concessions due to the concession of the time it takes. If you really want a very brief survey, the concession is that you may not have enough information. But if you want to retrieve a lot of information, the concession is that it takes too long with the chance that people drop out.” In other words, in this study the choice was made to make a concession on the length of the consultation, so that no other concessions had to be made.

The length (or the feasibility) of the consultation is also discussed in the interview with the ORS project manager. “You still spend twenty minutes and society is now organised in such a way that everything has to be done quickly and quickly, and that has proven to be the case. That scares people off.” The interviewee continues the argument with: “A question that I ask myself is what you are going to shorten to. What are you going to work on? In doing so, you also fall short of all the preferences that exist in terms of motivation behind the subject the PWE is about.”

Moreover, a policy officer indicates that a concession has been made for the two different questions with a slider and a points choice task. "You really get two different results, and you therefore have to name everything twice in your research report. These two types of results are difficult to compare and interpret, because they are both measured on a different point scale." The ORS interviewee adds: “This provides broadening on the one hand, but also narrowing it on the other hand if you look at how many people have used the same tool. So, it does say something about representativeness. That is a concession.” Regarding these quotes, it can be inferred that this is a concession between the clarity of interpreting the results and reaching consensus on the method. With regard to this concession, the client would have liked to have had a more leading role. "As a client, you ultimately want to shape the research. As a ministry, we think it is important to collect all input from stakeholders and respondents. That is why it might be good to make a harder choice instead of the choice with the sliders and points task."

Another concession that has been made is not to collect a sample that is representative throughout the Netherlands, but only for the municipalities around Schiphol. This concession was made because a Netherlands-wide sample led to a lot of commotion among the stakeholders. “This was clearly discussed with each other and in that sense, it felt as if we could really make a choice and therefore had a leading role as client. The opinion of stakeholders is of course something to take seriously.”

Finally, during the discussion of phase three, the second policy officer also referred to the draft report: “Ultimately, the investigation resulted in a very extensive report. Actually, I couldn't really find the time to read a very large report. I found it difficult to think of what should I prioritise now? Or where is feedback still needed?” This policy officer confirms that this is a concession between a report that is as complete as possible versus a short report with priorities. This concession can be interpreted as the concession between completeness and feasibility, but in this case from the perspective of the client who wants to use the results of the consultation for policy formation.

6.5 Overview of the concerns of face validity in practice

In sections 6.1 to 6.4 the concerns about face validity are examined in practice. This section presents a table with an overview of the face validity categories that lead to concerns in practice. This includes the concerns of the local residents and respondents, stakeholders involved in the process designing the PVE consultation, a stakeholder not involved in the process of designing the PVE consultation and the client. Table 6.1 shows this overview of face validity concerns in practice.

Table 6.1: Overview of the concerns of face validity in practice

	Completeness	Clarity	Feasibility	Unambiguity	Insensitivity	Aesthetics	Familiarity	Appropriateness of difficulty level	Readability	Efficacy
Local residents and respondents	X	X	X	X	X	X				
Stakeholders involved in the process of designing the PVE consultation	X	X	X	X			X	X	X	
Stakeholder not involved in the process of designing the PVE consultation	X	X			X					
Client	X	X	X	X						X
Total times mentioned	4	4	3	3	2	1	1	1	1	1

Table 6.1 shows that the categories completeness and clarity emerge as a concern of each of the four perspectives. This is followed by feasibility and unambiguity. This is apparent since feasibility is not included in the face validity framework established in this research (Chapter 3.3). Other categories of face validity emerge less often in the four perspectives.

7. Discussion

This chapter presents a discussion based on the results of this study. Within this discussion, the limitations and implications of this research are addressed. After face validity is placed in context with regard to the concept of validity, discussion points are first presented about the design of the face validity framework with regard to the first sub question. This is followed by discussions points about the panel of respondents and the data used to perform analyses that are part of the second sub question. Subsequently, discussion points regarding the third and fourth sub questions are presented. Finally, two more comprehensive points of discussion are raised.

A first point of discussion concerns an implication about face validity per se. As mentioned in the introduction (section 1.1.3), the concept of face validity is a umbrella term. Within validity there are different types of validity. Among others, Drost (2011) describes six types of validity that fall under the principle of construct validity. Examples are content validity, predictive validity, concurrency validity and face validity. So, face validity is just a part of the umbrella term validity. From this it follows that when a consultation is face-valid, this is only a condition of validity. Face validity by itself does not provide a sufficient conclusion regarding the overall validity of a PVE consultation.

Subsequently, there are a number of points of discussion that can be traced back to the design of the face validity framework. First of all, this study included five categories in the face validity framework that are identified from literature. However, practice showed that feasibility is also an important category (section 6.1 to 6.4). If there was even more room to measure face validity, the unsensitivity category should also be included as shown by an in-depth interview in section 6.3. However, face validity is measured without feasibility in this study. In retrospect, this category should have been included in the framework if it is up to the practical perspective.

Furthermore, a face validity category can be queried in several ways as is also apparent from section 3.1.6.1. For example, the clarity category can concern whether an item or instrument is understandable, but clarity may also concern the adequacy of the instruction (Desai and Patel, 2020). Another example is readability. Oluwatayo (2012) distinguishes between two criteria that both relate to legibility. One criterion is about the correct spelling of difficult words and the other criterion is about the readability in general or the understandability. However, it is a limitation of this study that each face validity category is questioned on the basis of one statement and thus in one way. When multiple statements could be questioned per category, a factor analysis could also be performed per category to test whether each statement actually measures the same category. In this way it is possible to test the categorization of face validity as applied in this research.

A third limitation is that the five categories that eventually ended up in the face validity framework depend on the experts who were interviewed. With other experts you might end up with a different combination of categories. This would also provide a different framework. Moreover, it may be stated that the concept of efficacy is rather vaguely defined in the literature. Therefore, it is doubtful whether the experts have properly understood this category. This vagueness may have resulted in this category being selected less often. In

addition, the five categories selected in this study are specific to the PVE method. The validity experts were given an explanation about the PVE method in the interview (Appendix C) and PVE experts were also interviewed who were allowed to select the five most important face validity categories for the PVE method. However, this means that the results from this study cannot be generalized for methods other than the PVE method.

Another point of discussion that refers to the first sub question is about the legitimacy as presented in section 3.1.2. In this section the legitimacy of the PVE method is discussed with the perspective of the PVE as isolated method. The choice of this approach coincides with the choice to focus in this study on measuring preferences of respondents, which is part of the isolated PVE method. As a result, legitimacy from the perspective of a broader repertoire of methods that can help a policymaker gain insight into what people think has not been addressed. There are different types of validity. It might be the case that, for example, the PVE method scores less on a certain type of validity. However, that does not mean that it is a lesser method, but that the PVE method is an addition to the flaws of another method.

Additionally to the discussion about the design of the face validity framework, there are points for discussion regarding the second sub question. More specifically these points deal with the use of a panel of respondents and the use of the data from the consultation. By using a panel of respondents, these respondents are already selected who can use an online web tool. This is because a panel contains people who more often participate in online surveys. In such a panel there are no people who have no idea at all how a PVE consultation should be carried out online. So, a pre-selection is already made on the basis of the use of a panel. A hypothesis is that people in a panel have a higher face validity compared to a consultation than people who are more digitally literate. As a result, the results regarding the assessment of face validity may have been framed which may have resulted in a higher assessment of face validity. However, this would not be the case if, for example, letters about participating in this consultation are sent to randomly selected citizens from the surrounding municipalities of Schiphol.

Another limitation is that the respondents who dropped out might have the opinion that the consultation has a low face validity, but there are no results of these respondents because they dropped out. Moreover, the analyses of this study only included the answers of respondents who assessed all face validity categories and filled in their demographic characteristics in order to minimize the chance of bias. If the face validity ratings of these drop outs could still be measured, the average ratings of the assessment of face validity might have been lower.

Regarding the third sub question about properties that influence the evaluation of face validity between different case studies and the benchmarks of face validity, there are also a couple of points of discussion. First, the evaluation of completeness is compared between only two case studies. It follows that the set bandwidth for completeness consists of two results. As a result, the set bandwidth contains uncertainty. For the categories clarity and unambiguity, only five case studies have been compared on which the bandwidth is based. Due to this low number of case studies, there is an uncertainty margin in the bandwidth and therefore also in the benchmark as target value for the evaluation of a face validity category. Due to its uncertainty, it might be the case that the target values that will be given in section 8.1 do not lead to an

improvement in face validity in the future because the benchmark has been set too low. On the other hand, these case studies may have exceptionally high evaluations of face validity, as a results of which the benchmark is set unrealistic and too ambitious.

Furthermore, the case studies included in this research have all taken place in the Netherlands. This is a limitation, as face validity can be assessed differently in other countries. It is arguable that a good assessment differs per county or culture. An example are the grades given for the central final exam in the Netherlands and the United States of America. Where in the Netherlands there is more of a sixes culture, in the USA you will be assessed with a high mark if you score slightly above average. This is reflected in the results of the final exam. In 2022, the average grade for the final exam in the Netherlands was 6,42 (Examenoverzicht, 2022). In the USA this was a grade point average of 3,0 (Roberts, 2021). This is converted a 7,4 in Dutch grades. The people in the Netherlands are more likely to give a lower average score. This grading culture could influence the way in which face validity is assessed.

A similar explanation can be given to the question when face validity is high enough. What constitutes a good assessment of face validity depends first of all on the culture, as argued above. Secondly, a good assessment of face validity varies by gender. In this research, men more often give extreme answers (totally agree or totally disagree) to the face validity statements than women (see section 4.7). To determine when face validity is high or good enough, applying rules of thumb would offer a solution. This research provides a first design for rules of thumb for face validity assessments by means of benchmarks and bandwidths for a face validity category (section 5.2 to 5.4). For further drawing up rules of thumb for when face validity is high or good enough, an example can be taken of construct validity. Following Del Greco et al. (1987), “construct validity refers to the extent to which the new questionnaire conforms to existing ideas or hypotheses concerning the concepts (construct) that are being measured.” A way of stating that there is construct validity is if there is a high correlation between an established measure and the questionnaire. The article by Del Greco (1987) gives an example on appetite. Construct validity can be established by measuring a variable related to appetite, such as health status. In this case, the health status is the established measure. When it comes to correlation, there are clear rules of thumb. A coefficient between 0,3 and 0,5 corresponds to a low correlation, between 0,5 and 0,7 corresponds to a moderately correlation and correlation is considered strong with a coefficient above 0,7. Possibly similar rules can also be applied for face validity. In practice, the aim is to achieve that 70% of the participants agree with the validity statements about face validity. Concluding, for face validity reliable rules of thumb still need to be drawn up to determine whether face validity is high/good enough. This research provides a first step op this design.

Moreover, there is an outlier in the comparisons between different case studies with regard to the assessment of clarity in section 5.2. The assessment of clarity in the Schiphol case and the heat transition vision in Utrecht are in agreement, while the four properties of the established typology do not match. This suggests that other factors also play a role in the assessment of face validity. However, these additional factors are not covered in this research. A possibility for a factor could be the scale of the project.

Concerning the fourth sub question about the practical perspective of concerns about face validity, there are also two points of discussion. The results of this sub question show that the

respondents found the questionnaire to be very long (section 6.1). The residents' representatives also expressed themselves about its length after hearing the opinions of their followers (section 6.2). As a time investment has been requested from people of which some of them perceive as too long, the question can be asked what the effect of a long consultation is on the answers of the respondents. During a long consultation, people may become impatient and may become less concentrated. This may affect the accuracy of their answers. Of the respondents whose answers have been analysed in this study, it appears that clarity and unambiguity have a lower average score. These two categories were questioned early in the consultation. However, the other three categories of relevance, readability and completeness score higher. It can be argued that these results are influenced by the impatience or concentration capacity of the respondent. The respondent may have thought about the end of the consultation out of complacency or lack of concentration: it will be all right. This is where a trade-off arises between the completeness and the feasibility of a consultation.

Furthermore, the strategic game of the stakeholders and in particular the residents' representative must be taken into account. Their feedback on the draft report depends on when the report is sent and on the results of the report. For example, for one resident's representative the results were in line with expectations. This representative has not provided further feedback. Another resident representative criticised a number of points related to face validity. In the 'Regioforum' it therefore emerged that this representative was less satisfied with the results.

Finally, there are two more general implications. The first general implication is about the normalization of statistics. In chapter 5 of this study, bandwidths are drawn up on the basis of average scores from different case studies on face validity categories. On the basis of an average score, a general picture can be created about how a certain face validity category scores and thus whether most respondents have understood the consultation. Therefore, the average score is often applied in statistics. However, with the results of a PVE consultation, a policymaker can draw up policy. When it comes to policy, it can be stated that inclusiveness must be a priority. The question then is whether the average score provides a complete picture to aim for inclusiveness. An alternative is to look at the spread and strive for no respondent to assess the face validity with (totally) disagree. In this way, the aim is that everyone understands the consultation and that inclusiveness is ensured.

The last implication is that in this study the main focus is on investigating how the face validity currently scores in the PVE method. A follow-up question is how these scores can be improved. The descriptive results (section 4.2) show that clarity and unambiguity have the lowest average face validity score in comparison with readability, relevance and completeness. With regard to clarity, it is difficult from this study to state what information may have been missing about the possible tasks of the Schiphol Social Council and the explanation of the consultation. For the unambiguity it is difficult to state what was not clear about each task. To find out and thereby increase the value of the face validity assessment, a possible suggestion is to do a test before all respondents are able to complete the consultation. During this test, which somewhat resembles an interview, there is a test person who completes the consultation. While filling in the questionnaire, the test taker is asked which information is missing and what he or she understands by each option in a choice task.

The most ideal would be to test this with several people to conclude whether they understand the same thing under options of a choice task. If there is time pressure for the researcher to go live with the consultation or it is difficult to find test subjects, co-researchers may be involved. Importantly, these researchers did not participate in this consultation beforehand. Such an interview may also help improve readability. The Schiphol case showed that two options from a choice task were not completely clear in the panel consultation. The wording has been changed in the open consultation and this has resulted in a change in the ranking of the options of the choice task. In addition, clarity and unambiguity are surveyed immediately after the Schiphol Social Council choice task. It would be interesting to investigate whether the scores of these two categories would differ if they were questioned at the very end of the consultation like the other categories.

Regarding the relevance, the respondents were asked whether they consider the PVE method a good method for expressing their opinion. Not included in the consultation about Schiphol is why this method is used or what the added value of this method is. Perhaps if respondents receive information about this, they will be more convinced of this method. It could also be an added value to interview a number of respondents about why they think the PVE method is or is not a good method for expressing their opinion. Based on their motivations, further steps can be taken to improve the face validity assessment of this category. In addition, it must also be clear what happens to the results, just as is done in the Schiphol case.

Regarding completeness, there are two suggestions that could potentially increase the value of the evaluation of this category. A first suggestion is that an open question could be asked at the end of each consultation whether the respondent would like to comment anything about his or her participation. In this way a respondent can express all his or her opinions that could not be given during the consultation. A second suggestion is to investigate whether completeness is increased if not all points have to be divided in a choice task. This suggestion stems from the interview with the stakeholder who was not involved in the design of the consultation (section 6.3). If a respondent does not agree with the options in the choice task, he or she does have the option to continue the consultation. This respondent can be asked why not all points were distributed and which options were missing. Finally, a trade-off between completeness and feasibility is identified in this research. When it is not possible to shorten the consultation because of the completeness, but this would lead to lower feasibility, it is an option to consider splitting the consultation in different parts. Each respondent will then complete a part of the entire consultation. In this case, for example, the parts of the Schiphol Social Council and the Environmental House could be separated. This consideration plays a particular role in the open consultation, as is apparent from this study.

8. Conclusions and recommendations

The Participatory Value Evaluation method is a relatively new method that serves multiple purposes as an evaluation method to measure preferences of respondents. To this moment, little empirical research has yet been conducted into the validity of the PVE method. The concept of validity is a umbrella concept in which different types of validity can be identified. One of these types is face validity. In an earlier PVE consultation Amsterdam-Wind has shown that a lack of face validity has led to a lot of commotion and resistance. Therefore, this research has gained insight in how to measure face validity within the PVE method and how involved citizens and stakeholders evaluate the face validity within this method. In this chapter the main conclusions of this research are presented. After the four sub questions have been answered, the main research question is answered. This is followed by the recommendations for future research in section 8.2. Finally, this chapter presents the recommendations for practice. These are recommendations for those designing a PVE consultation.

8.1 Conclusions

During the first part of this research the following sub question is addressed: *how to design a framework that is able to measure face validity regarding the PVE method?* The results of the literature review and the expert interviews lead to the following three conclusions.

The first conclusion is about the importance of a face validity assessment. If the content of the PVE consultation is inappropriate or irrelevant, there is a likelihood that the results obtained from this instrument provide false information and decisions of respondents that are misleading for policymakers (Anastasi & Urbina, 2007). It follows that if the PVE consultation is not face-valid, there is a chance that citizens will not agree with the decisions that the politicians make based on the results of a PVE consultation. On the other hand, misleading information in the consultation can also cause respondents to make choices that they would not make otherwise. Reasoned from political legitimacy, a non-face-valid PVE may lead to infectiveness of policy decisions. Reasoned from social legitimacy, if respondents and stakeholders consider an instrument to be face-valid, this ensures acceptance of the instrument and therefore increases the usefulness of the instrument. This shows the importance of a face validity test.

The next conclusion is about the setup of a face validity framework. When setting up an assessment of face validity, it is important to take the following four framework items into account as stated in sections 3.1.3 to 3.1.6. This is based on the assumption that it has already been determined what would be assessed, for example a test item or an instrument. First it needs to be determined if the raters are experts or laypeople. Second, it must be determined if the method of the assessment is a questionnaire, a ‘think aloud’ interview or a focus group. Third, it must be determined in which stage of the development the assessment will take place. Fourth, it must be considered whether it is an absolute or a relative assessment.

Finally, it can be concluded that when evaluating the face validity of the PVE as an evaluation method, these are the five most recommended categories: clarity, unambiguity, relevance, readability and completeness. In sections 3.1.6 and 3.3.2 these categories are further elaborated.

The second sub question is defined in the following way: *what are the similarities and differences of two different types of a choice task regarding a PVE consultation on the evaluation of face validity?* The statistical analyses performed to answer this second sub question lead to the following conclusions.

Following from the results of the descriptive statistics, clarity and unambiguity are rated the lowest with an average of 3,60 and 3,61 on a five-point Likert scale where 1 stands for totally disagree and 5 stands for totally agree across the total sample. This is followed by the relevance and completeness categories with an average of 3,85. Readability has the highest rating with an average of 3,89. It follows that the respondents consider the PVE consultation face valid rather than neutral or not face valid. This leads to the conclusion that based on the Schiphol consultation, it is recommended to focus in the following PVE consultations in particular on clarity and unambiguity if the goal is to increase face validity.

Furthermore, the factor analyses show that several categories of the established framework of sub question one load onto a latent variable that measures face validity. This result leads to the following conclusion. If the aim is to measure the assessment of face validity, more than one category of face validity should be questioned.

Moreover, it can be concluded that there is no significant difference in the evaluation of face validity between the two experiments as described in section 4.3. The first experiment contains a 'sliders' choice task and the second experiment a 'points' choice task.

Despite the fact that there is no difference in the evaluation of face validity between the two experiments, it can be concluded that respondents are more indifferent to a 'sliders' choice task than to a 'points' choice task. First, this is apparent from the multiple regression analyses where the 'points' experiment with the 'points' choice task has more demographic characteristics that have a significant influence on the evaluation of face validity (section 4.5). Second, this is apparent from the multinomial logistic regressions because of the significance of the likelihood ratio test (section 4.6). In the 'sliders' experiment, almost all of these tests are not significant. Therefore, the regression model does not explain the evaluation of the face validity categories better than no model. However, in the 'points' experiment these tests are significant in each of the five categories. This means that the model with the demographic characteristics can better predict the evaluation of the face validity categories than no model.

A similar conclusion can be drawn from the latent class cluster analyses. From these analyses, it is remarkable that in the 'sliders' experiment the majority of the respondents (55,64%) rated the face validity highly. In the 'points' experiment, the clusters are more equally distributed. Here, 39,63% of the respondents rate the face validity highly, 38,96% rated clarity and unambiguity as neutral and the other three categories as high and 21,41% rated face validity as neutral or low. Concluding, the opinions with regard to face validity in the experiment with the 'points' choice task differ more widely.

The third sub question is defined in the following way: *which properties influence the differences in the evaluation of face validity between different case studies and what are the benchmarks of those differences?*

A first conclusion is that the four properties impact on personal life, the respondents, the platform and who is in charge influence the differences in the evaluation of face validity between different case studies. These characteristics are further elaborated in section 5.1. If one of these four properties differ between two consultations, this leads to a significant difference in the assessment of face validity. If these four properties all match between two consultations, this does not lead to a significant difference in the assessment of face validity.

The evaluations of the face validity categories clarity, relevance and completeness are compared between different PVE consultation case studies. Those results lead to the following conclusions. First, for the case studies in which clarity is included, it appears that the bandwidth regarding the assessment of clarity lies between the average scores 3,53 and 3,84 on a five-point Likert scale. On this scale, 1 stands for totally disagree and 5 stands for totally agree. If the goal is to score high on the clarity category, the aim should be to achieve an average evaluation score above 3,67. Only a third of the case studies achieve this score. Be aware that if the impact on personal life is in the short term, this will result in a lower average evaluation score regarding clarity than impact on the long term.

Second, for the case studies in which relevance is included, it appears that the bandwidth regarding the assessment of relevance lies between the average scores 3,53 and 3,95 on a five-point Likert scale. On this scale, 1 stands for totally disagree and 5 stands for totally agree. The aim should be to achieve an average evaluation score above 3,88 if the goals is to score high on the relevance category. Only a quarter of the case studies achieve this score. Be aware that if the impact on personal life is in the short term, this will result in a lower average evaluation score regarding clarity than impact on the long term.

Third, for the case studies in which completeness is included, the benchmark regarding the evaluation of completeness lies between the average scores 3,89 and 4,11 on a five-point Likert scale. On this scale, 1 stands for totally disagree and 5 stands for totally agree. Based on two case studies, if the goal is to score high on the completeness category the aim should be to achieve an average evaluation score around 4,11.

Although the first three sub questions focus on identifying and analysing face validity categories that follow from the literature, the last sub question focuses on the face validity categories that evoke in practice. During the last part of this research the following sub question is addressed: *to what extent do the concerns of citizens and stakeholders with regard to face validity correspond in practice with the established framework?*

From practice, a conclusion is that most concerns of the local residents and respondents, stakeholders involved and not involved in the process of designing of the PVE consultation and the client can be traced back to four face validity categories. These are the categories completeness, clarity, unambiguity and feasibility. These three are concerned by all four of the above groups. Only the stakeholder who is not involved in the design of the consultation did not recognize the concern regarding feasibility. Completeness, clarity and unambiguity are all included in the face validity framework of this research (section 3.3.2). Therefore, it is concluded that from a practical point of view the feasibility category should also be included in the face validity framework.

Finally, the main research question is answered. The main research question is defined in the following way: *how to measure face validity regarding the PVE method and how do respondents evaluate the face validity of a PVE consultation?*

The main research question consists of two parts. The first part concerns the measurement of face validity regarding the PVE method. From the literature review and the expert interviews it can be concluded that five categories that fall under face validity are most suitable for measuring the face validity of the PVE method. The five most recommended categories are: clarity, unambiguity, relevance, readability and completeness. These categories are assessed on the basis of the following framework items in this study. First, the face validity is assessed by laypeople. Second, the evaluation of these categories is in the form of a questionnaire. Furthermore, the evaluation takes place after the implementation of the PVE consultation. Finally, an absolute assessment is applied. The face validity of the PVE consultation of the ORS case study has been evaluated on the basis of this framework.

However, if it had been up to the concerns of face validity in practice, the feasibility category would also have been included. This can be concluded based on the feedback from and interviews with the local residents and respondents, stakeholders involved and not involved in the process of designing the PVE consultation and the client.

The second part of the main research question consists of the evaluation of face validity of a PVE consultation by respondents. For the evaluation, the framework established in the first part of the main research question is applied to the PVE consultation of the ORS case study.

The respondents consider the PVE consultation face valid rather than neutral or not face valid. Clarity and unambiguity are rated the lowest and readability the highest by the respondents. Furthermore, it makes no difference to the evaluation of face validity which type of choice task respondents evaluate. In this study, the assessments of the ‘points’ choice task and the ‘sliders’ choice task are compared with each other. Despite the fact that there is no difference in the evaluation of face validity between the two types of choice tasks, it can be concluded that respondents are more indifferent to a ‘sliders’ choice task than to a ‘points’ choice task.

Besides the evaluation of face validity in the ORS case study, categories of face validity were also questioned in some previous PVE consultation. On the basis of these assessments, benchmarks are set as a target for the evaluation of face validity for subsequent consultations. If the goal is to score high on the clarity category, the aim should be to achieve an average evaluation score above 3,67 on a five-point Likert scale. The aim should be to achieve an average evaluation score above 3,88 if the goals is to score high on the relevance category. For completeness, the aim should be to achieve an average evaluation score around 4,11.

8.2 Recommendations for future research

This section provides recommendations for further research. The recommendations arise from the results of this study which give rise to further research.

First of all, the results of the latent class cluster analyses in section 4.7 give rise to further research. These results show that with a ‘sliders’ choice task the majority of the respondents

is part of the same cluster that rates face validity high. With a 'points' choice task it is noticeable that three clusters are identified in which the respondents are more equally distributed. However, this research is the first study in which LCCAs have been applied to identify clusters of respondents who collectively rate face validity high or low. This method makes it possible to identify clusters with the corresponding (demographic) characteristics of the respondents. Since this is the first study to include LCCAs to analyse the assessment of face validity, there is uncertainty about the reliability of these results. Further research is needed to reduce this uncertainty. In future research, performing LCCAs in other case studies about the assessment of face validity can be applied. This makes it possible to perform reliability tests and to compare the results of the LCCAs around face validity in multiple studies.

Furthermore, the LCCAs performed in section 4.7 show that the greatest improvement regarding face validity can be made for several respondents with certain characteristics. For both the 'points' and the 'sliders' choice task there is room for improvement with regard to face validity with 18- to 34-year-olds, the low-educated people and people who are most suspicious of the subject of the consultation. The multinomial regression analyses also show in both experiments that young people between the ages of 18 and 34 years rated all five face validity categories lower. With regard to the low educated, it is remarkable that in the 'sliders' experiment they rated readability lower. Low-educated people also scored lower in completeness and relevance. Therefore, a suggestion for further research is to investigate how the face validity of these groups can possibly be increased and what their needs are regarding face validity.

Another recommendation also arises from remarkable results of the multinomial logistic regressions. These remarkable results are related to the respondents who live in the 'inner' area of Schiphol or who are inconvenienced by Schiphol. Respondents who live in a municipality in the 'inner' area rate clarity, unambiguity and completeness lower in the 'sliders' experiment than people who live in the 'outside' area. Respondents who are inconvenienced by Schiphol rate the clarity lower in the 'points' experiment than people who are not inconvenienced. These two results seem to imply a proximity effect. A hypothesis that follows from this is that people who live close to a problem situation, such as at Schiphol, will rate the face validity lower. A recommendation is to further investigate this proximity effect and to test this hypothesis. Related questions are how big is this effect and what are the consequences of a proximity effect?

In addition, based on the multiple regression analyses and the multinomial logistic regressions, it appears that more demographic characteristics influence face validity in a 'points' choice task compared to a 'sliders' choice task. This led to the conclusion that respondents are more indifferent to a 'sliders' choice task. From these results and conclusions follows the question of why more demographic characteristics influence the assessment of face validity in a 'points' choice task. An argument could be that a 'points' PVE is more like a survey than a 'sliders' PVE. As a result, the 'points' PVE is probably more familiar among the respondents and they know better what they think of this way of questioning. This may lead to more extreme opinions. 'Sliders' PVEs are a relatively new type of questioning. So far, the 'sliders' choice task has only been applied in the climate consultation (Mouter et al., 2021a) and in the Schiphol Environmental case. It is possible that respondents are less familiar with

this and therefore have less extreme opinions. However, this argument is a hypothesis. A recommendation for future research is to test this hypothesis.

Furthermore, based on the results of the third sub question on the benchmarks of the evaluation of face validity categories and the influence of case study properties, three suggestions for future research follow. First, in this study the assessment of a face validity category is compared between different case studies with their own properties. Included properties in this study are the impact on personal life, the respondents, the platform and who is in charge. However, this study did not investigate how much effect an individual property has on the assessment of face validity. In future research this could possibly be researched by applying a regression model. From this research, new benchmarks can be set that apply to a specific property.

A second suggestion for future research that follows from the third sub question is that only a comparison is made between the old version and the new version of the online platform in this research regarding the platform property of a case study. However, there is also a mobile version of the platform. Therefore, it is a recommendation for future research to investigate the difference in influence on face validity between the new version and the mobile version of the platform. The hypothesis is that the mobile version has a lower face validity because a lot of scrolling has to be done to complete a consultation.

A third recommendation from this sub question follows from an outlier. In the face validity clarity, it appears that there is no difference in the assessment of face validity between the Schiphol case and the Utrecht case. However, these cases differ on each of the four properties included in this study (section 5.1). Therefore, this result is not in line with the expectations and with the other results. The question that arises is how this outlier can be explained. In addition, there is also the question which other properties have influence on face validity besides the four properties included in this study. This outlier gives rise to future research.

A fourth recommendation linked to the third sub question is about the spreading. In section 8.1 conclusions are drawn about the benchmarks for the assessment for face validity categories based on the average evaluation score. However, when the goals of the PVE method is to achieve full inclusiveness since the results will be used for policymaking, it may be important to investigate the spread as well. The average score shows when most people are able to participate in a PVE consultation while the spread shows when everyone is able to join. Therefore, it is a recommendation for follow-up research to study how it can be ensured that respondents do not judge the face validity categories with (totally) disagree. This research has shown that especially 18- to 34-year-olds and low-educated people assess the face validity low.

Finally, the results of sub question four about the concerns of face validity in practice shows that there is a field of tension between the completeness and feasibility of a consultation. In the case of the Schiphol Environmental Council, there is consensus from the stakeholders about the structure of the consultation. However, this resulted in comments from respondents that the consultation took too much time. In this case, a less complete consultation from the perspective of stakeholders could have led to a higher feasibility from the perspective of respondents. Future research can focus on what this field of tension exactly

looks like, how the trade-offs can be set up in this field of tension and what the effect of the trade-offs is on the consultation and the perspectives of respondents, stakeholders and/or researchers.

8.3 Recommendations for practice

In addition to the recommendations for future research, this section provides recommendations for practice. These recommendations are intended for researchers who make the final decisions about the design of a PVE consultation.

A first recommendation is not to make the trade-off between a 'points' or a 'sliders' choice task based on the assessment of face validity. Regarding the evaluation of face validity, the two types of choice tasks do not differ from each other. The fact that these evaluations do not differ can serve as an argument to convince the involved stakeholders to apply a 'sliders' choice task. In contract, the recommendation is to base the trade-off between the 'sliders' and the 'points' choice task on the indifference. When indifference with regard to the starting point, it is recommended to apply the 'sliders' choice task. A reason why indifference to a choice task is pursued is to exclude that the method influences the results. When there are more extreme opinions, these opinions can evoke in emotions that ultimately influence the answers that the respondents fill in. This reason can also be used to convince the involved stakeholders to apply a 'sliders' choice task.

The following recommendation follows from the results of the LCCAs which show that a number of demographic characteristics can be focused on to increase face validity. For both the 'points' and the 'sliders' choice task there is room for improvement regarding face validity with 18- to 34-year-olds, average and low-educated people and people who are most suspicious of the subject of the consultation. More specifically for the 'sliders' choice task, improvements regarding the evaluation of face validity are possible regarding women. It is advisable to focus in particular on the evaluation of face validity on these demographics characteristics when testing a consultation. Respondents with these characteristics tend to rate face validity the lowest compared to all other respondents.

A third recommendation is to include the face validity categories in the conversations with the involved stakeholders when designing the PVE consultation. In this way expectations can be managed. An example from the Schiphol case where face validity categories are used in conversations with stakeholders is with the concretization of the definition of the tasks in the choice tasks. The tasks of the Environmental House and the Schiphol Social Council by van Geel (2020) have been specified in a number of steps based on the categories clarity and unambiguity. These two categories formed the criteria for shaping the tasks. Another example from this case where communicating about the categories could have provided added value is the trade-off between completeness and feasibility. It emerged from the discussions with stakeholders that they wanted to add more questions to the design. However, they did not realize up front that the consultation for respondents was considered to be too long, as the results show in section 6.2. By informing the stakeholders in advance about the expected feasibility, the feedback about the feasibility could possibly have been prevented.

A final recommendation for practice is to include statements to evaluate face validity categories in upcoming PVE consultations. On the one hand, it is recommended because a

high face validity is important for, among other things, the reliability of the results from a PVE as argued in section 3.1.2. On the other hand, statements about face validity have only been included in a few case studies. Adding these statements in subsequent case studies provide a more reliable picture with regard to face validity and the possible benchmarks to set as a goal.

Literature

Alderson, J. C., Clapham, C. & Wall, D. (1995). Language Test Construction and Evaluation. Cambridge: University Press.

Anastasi, A. & Urbina, S. (2007). Psychological testing (2nd impression). Pearson. NJ: Prentice-Hall.

Arnstein, S. R. (1969). A Ladder Of Citizen Participation. *Journal of the American Institute of Planners*, *35*(4), 216–224. https://doi.org/10.1080/01944366908977225

Austin, J., & Delaney, P. F. (1998). Protocol Analysis as a Tool for Behavior Analysis. *The Analysis of Verbal Behavior*, *15*(1), 41–56. https://doi.org/10.1007/bf03392922

Banna, J. C., Vera Becerra, L. E., Kaiser, L. L., & Townsend, M. S. (2010). Using Qualitative Methods to Improve Questionnaires for Spanish Speakers: Assessing Face Validity of a Food Behavior Checklist. *Journal of the American Dietetic Association*, *110*(1), 80–90. https://doi.org/10.1016/j.jada.2009.10.002

Bannigan, K., & Watson, R. (2009). Reliability and validity in a nutshell. *Journal of Clinical Nursing*, *18*(23), 3237–3243. https://doi.org/10.1111/j.1365-2702.2009.02939.x

Bakker, M. (2022). Opinie: 'Stroomopwek met wind is in de stad onhaalbaar, maar die mening is taboe.' *Het Parool*. https://www.parool.nl/columns-opinie/opinie-stroomopwek-met-wind-is-in-de-stad-onhaalbaar-maar-die-mening-is-taboe~b9fe1064/

Barnes, M., Newman, J., Knops, A., & Sullivan, H. (2003). Constituting "the public" in public participation. *Public Administration*, *81*(2), 379–399. https://doi.org/10.1111/1467-9299.00352

Belone, L., Lucero, J. E., Duran, B., Tafoya, G., Baker, E. A., Chan, D., Chang, C., Greene-Moton, E., Kelley, M. A., & Wallerstein, N. (2016). Community-Based Participatory Research Conceptual Model. *Qualitative Health Research*, *26*(1), 117–135. https://doi.org/10.1177/1049732314557084

Berenschot (2020). *Deelrapportages Berenschot advies Schiphol vernieuwd verbinden.* Retrieved from https://www.omgevingsraadschiphol.nl/wp-content/uploads/2020/12/Deelrapportages-Berenschot-ihkv-advies-Schiphol-Vernieuwd-verbinden.pdf

Bogner, A., Littig, B. & Menz, W. (2009). Introduction: Expert Interviews – An Introduction to a New Methodological Debate. In: A. Bogner, B. Littig & W. Menz (eds.), Interviewing Experts (pp. 1-13). Palgrave Macmillan.

Bokhorst, A. M. (2014). Bronnen van legitimiteit: Over de zoektocht van de wetgever naar zeggenschap en gezag. Boom juridische uitgever.

Bouwhuis, S. (2011). Legitieme besluiten over infrastructurele projecten. Een kritische discoursanalyse van de participatiepraktijk bij infrastructurele projecten in Nederland. Radbout Universiteit Nijmegen.

Bouwmeester, M. (2021). Effects of goal-dependent implementation choices on the achievement of goals in Participatory Value Evaluation processes. Delft University of Technology.

Bowen, G.A. (2009). Document Analysis as a Qualitative Research Method. *Qualitative Research Journal, 9*(2), 27-40. https://doi.org/10.3316/QRJ0902027

Boynton, P. M., & Greenhalgh, T. (2004). Selecting, designing, and developing your questionnaire. *BMJ*, *328*(7451), 1312–1315. https://doi.org/10.1136/bmj.328.7451.1312

Broder, H. L., McGrath, C., & Cisneros, G. J. (2007). Questionnaire development: face validity and item impact testing of the Child Oral Health Impact Profile. *Community Dentistry and Oral Epidemiology*, *35*(s1), 8–19. https://doi.org/10.1111/j.1600-0528.2007.00401.x

Bruijn, J. A. de, Heuvelhof, E., ten & Veld, R. J. In 't (2002). *Procesmanagement. Over procesmanagement en besluitvorming*. Schoonhoven: Academic Service.

Buijs, A., & Boonstra, F. (2020). Natuurbeleid betwist. Visies op legitimiteit en het natuurbeleid. KNNV Uitgeverij.

Burgess, R.G. (2002). *In the field: An introduction to field research.* Routledge.

Burton, P. (2009). Conceptual, theoretical and practical issues in measuring the benefits of public participation. *Evaluation, 15*(3), 263-284. https://doi.org/10.1177/1356389009105881

CBS (2021). *Bevolking; onderwijsniveau en migratieachtergrond 2003-2021.* Retrieved from https://opendata.cbs.nl/statline/#/CBS/nl/dataset/82275NED/table?fromstatweb

CBS (2021). *Regionale kerncijfers Nederland.* Retrieved from https://opendata.cbs.nl/statline/#/CBS/nl/dataset/70072NED/table?fromstatweb

Chabrol, H., Rousseau, A., Rodgers, R., Callahan, S., Pirlot, G., & Sztulman, H. (2005). A Study of the Face Validity of the 40 Item Experiment of the Defense Style Questionnaire (DSQ-40). *The Journal of Nervous and Mental Disease*, *193*(11), 756–758. https://doi.org/10.1097/01.nmd.0000185869.07322.ed

Christiano, T. (1997). The Significance of Public Deliberation. In. J. Bohman en W. Rehg (eds.) Deliberative Democracy: Essays on Reason and Politics. Cambridge: The MIT Press, pp.243-277.

Clark, M.D., Determann, D., Petrou, S., et al (2014). Discrete choice experiments in health economics: a review of the literature. *Pharmaoeconomics, 32*(9), 883-901. https://doi.org/10.1002/hec.1697

Connell, J., Carlton, J., Grundy, A., Taylor Buck, E., Keetharuth, A. D., Ricketts, T., Barkham, M., Robotham, D., Rose, D., & Brazier, J. (2018). The importance of content and face validity in instrument development: lessons learnt from service users when developing the Recovering Quality of Life measure (ReQoL). *Quality of Life Research*, *27*(7), 1893–1902. https://doi.org/10.1007/s11136-018-1847-y

Creswell, J.W. & Plano Clark, V.L. (2017). *Designing and Conducting Mixed Methods Research*. 3rd edition. London: SAGE Publication Ltd.

Damme, van A., Noppe J. & Verhage, A. (2017). Police legitimacy: an introduction. *Policing: An International Journal*, *40*(3), 474–479. https://doi.org/10.1108/pijpsm-05-2017-0058

Dang, W. (2020). How culture shapes environmental public participation: case studies of China, the Netherlands, and Italy. *Journal of Chinese Governance, 5*(3), 390-412. https://doi.org/10.1080/23812346.2018.1443758

Dartée, K. (2018). Practicing Participatory Value Evaluation. Assessing the applicability of the participatory value evaluation method for public decision-making on urban storm water management in a The Hague case study. Delft University of Technology.

Dawes, J. (2008). Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. *International Journal of Market Research, 50*(1), 61-104. https://doi.org/10.1177/147078530805000106

Dekker, T., Koster, P., & Mouter, N. (2019). The Economics of Participatory Value Evaluation. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3323645

Del Greco, L., Walop, W., & McCarthy, R. H. (1987). Questionnaire development: 2. Validity and reliability. *CMAJ: Canadian Medical Association Journal*, *136*(7), 699-700.

Dempsey, P.A. & Dempsey, A.D. (1992). Nursing research and basic statistical applications, 3rd edn. Jones and Barlett, Boston.

Desai, S. & Patel, N. (2020). ABC of Face Validity for Questionnaire. *International Journal of Pharmaceutical Sciences Review and Research*, *65*(1), 164–168. https://doi.org/10.47583/ijpsrr.2020.v65i01.025

Drost, E.A. (2011). Validity and reliability in social science research. *Education Research and perspectives, 38*(1), 105-123.

Edelenbos, J., Domingo, A., Klok, P.-J., & van Tatenhove, J. (2006). Burgers als beleidsadviseurs.een vergelijkend onderzoek naar acht projecten van interactieve beleidsvorming bij driedepartementen.

Engbersen, G., Bochve, M. van, Boom, J. de, Bussemaker, J., Farisi, B. el, Krouwel, A., Lindert, J. van, Rusinovic, K., Snel, E., Heck, L. van, Veen, H. van der & Wensveen, P. van (2021). *De laag-vertrouwen samenleving.* https://www.eur.nl/essb/media/99176

Engström, S.M., Leksell, J., Johansson, U. B., Eeg-Olofsson, K., Borg, S., Palaszewski, B., & Gudbjörnsdottir, S. (2018). A disease-specific questionnaire for measuring patient-reported outcomes and experiences in the Swedish National Diabetes Register: Development and evaluation of content validity, face validity, and test-retest reliability. *Patient Education and Counseling*, *101*(1), 139–146. https://doi.org/10.1016/j.pec.2017.07.016

Ericsson, K., & Simon, H. (1993). Protocol analysis: Verbal reports as data. Cambridge: MIT Press.

Examenoverzicht. (2022). *Statistieken over het Eindexamen: het Ultieme Overzicht (2022)*. https://www.examenoverzicht.nl/examen-informatie/algemeen/statistieken#:%7E:text=Op%20het%20vwo%20wordt%20zowel,Engels%20(7%2C0).

Few, R., Brown, K., & Tompkins, E. L. (2007). Public participation and climate change adaptation: avoiding the illusion of inclusion. *Climate Policy*, *7*(1), 46–59. https://doi.org/10.1080/14693062.2007.9685637

Frantz, A., & Holmgren, K. (2019). The Work Stress Questionnaire (WSQ) – reliability and face validity among male workers. *BMC Public Health*, *19*(1). https://doi.org/10.1186/s12889-019-7940-5

Freeman, F. (1963). Theory and practice of psychological testing (third ed.). New York, NY: Holt, Rinehart and Winston.

Gaber, J., & Gaber, S. L. (2010). Using face validity to recognize empirical community observations. *Evaluation and Program Planning*, *33*(2), 138–146. https://doi.org/10.1016/j.evalprogplan.2009.08.001

Geel, P. van (2020). *Schiphol vernieuwd verbinden.* Retrieved from https://open.overheid.nl/repository/ronl-da36fb08-45f1-46a6-a5c7-ec054ccdb920/1/pdf/bijlage-2-eindadvies-schiphol-vernieuwd-verbinden.pdf

Geijsen, T., Vries, M. de, Maas, W., Tuit, C., Fillerup, L. & Mouter, N. (2022). *Coronabeleid op de lange termijn: Welke doelen en maatregelen vinden Nederlanders belangrijk?* Retrieved from https://populytics.nl/wp-content/uploads/2022/04/Rapport-Coronabeleid-op-de-lange-termijn2.pdf

Geus, T.F. de (2019). Decision-making in Participatory Value Evaluation. Delft University of Technology.

Hardesty, D. M., & Bearden, W. O. (2004). The use of expert judges in scale development. *Journal of Business Research*, *57*(2), 98–107.

https://doi.org/10.1016/s0148-2963(01)00295-8

Hojat, M., & Gonnella, J. S. (2010). An instrument for measuring pharmacist and physician attitudes towards collaboration: Preliminary psychometric data. *Journal of Interprofessional Care*, *25*(1), 66–72. https://doi.org/10.3109/13561820.2010.483368

Holloway, L., Humphrey, L., Heron, L., Pilling, C., Kitchen, H., Højbjerre, L., Strandberg-Larsen, M., & Hansen, B. B. (2014). Patient-reported outcome measures for systemic lupus erythematosus clinical trials: a review of content validity, face validity and psychometric performance. *Health and Quality of Life Outcomes*, *12*(1), 1-14. https://doi.org/10.1186/s12955-014-0116-1

Horwood, J., Sutton, E., & Coast, J. (2013). Evaluating the Face Validity of the ICECAP-O Capabilities Measure: A "Think Aloud" Study with Hip and Knee Arthroplasty Patients. *Applied Research in Quality of Life*, *9*(3), 667–682. https://doi.org/10.1007/s11482-013-9264-4

Ianniello, M., Iacuzzi, S., Fedele, P., & Brusati, L. (2018). Obstacles and solutions on the ladder of citizen participation: a systematic review. *Public Management Review*, *21*(1), 21–46. https://doi.org/10.1080/14719037.2018.1438499

Ito, P. (1980). 7 Robustness of ANOVA and MANOVA test procedures. *Handbook of Statistics*, 199–236. https://doi.org/10.1016/s0169-7161(80)01009-7

Jacobs, L. R., Cook, F.L. & Delli Carpini, M.X. (2009), Talking Together: Public Deliberation and Political Participation in America, Chicago, IL: University of Chicago Press.

Janssen, E.M., Marshall, D.A., Hauber, A.B. & Bridges, J.F.P. (2017). Improving the quality of discrete-choice experiments in health: how can we assess validity and reliability? *Pharmacoeconomics, 17*(6), 531-542. https://doi.org/10.1080/14737167.2017.1389648

Kaklamanou, D., Armitage, C. J., & Jones, C. R. (2012). A further look into compensatory health beliefs: A think aloud study. *British Journal of Health Psychology*, *18*(1), 139–154. https://doi.org/10.1111/j.2044-8287.2012.02097.x

Kennedy, L. G., Kichler, E. J., Seabrook, J. A., Matthews, J. I., & Dworatzek, P. D. (2019). Validity and Reliability of a Food Skills Questionnaire. *Journal of Nutrition Education and Behavior*, *51*(7), 857–864. https://doi.org/10.1016/j.jneb.2019.02.003

Knofczynski, G.T. & Mundfrom, D. (2007). Sample sizes when using multiple linear regression for prediction. *Educational and Psychological Measurement, 68*(3), 431-442. https://doi.org/10.1177/0013164407310131

Liang, J., Bi, G., & Zhan, C. (2020). Multinomial and ordinal Logistic regression analyses with multi-categorical variables using R. *Annals of Translational Medicine*, *8*(16), 982. https://doi.org/10.21037/atm-2020-57

Maginn, P. (2006). *Urban policy analysis through a qualitative lens: Overview to special*

issue. Urban Policy and Research, 24(1), 1–15. https://doi.org/10.1080/08111140600590650

Maithel, S., Sierra, R., Korndorffer, J., Neumann, P., Dawson, S., Callery, M., Jones, D., & Scott, D. (2005). Construct and face validity of MIST-VR, Endotower, and CELTS. *Surgical Endoscopy*, *20*(1), 104–112. https://doi.org/10.1007/s00464-005-0054-4

Marshall, G. (2005). The purpose, design and administration of a questionnaire for data collection. *Radiography*, *11*(2), 131–136. https://doi.org/10.1016/j.radi.2004.09.002

Matell, M.S. & Jacoby, J. (1972). Is there an optimal number of alternatives for Likert-scale items? Effects of testing time and scale properties. *Journal of Applies Psychology, 56*(6), 506-509. https://doi.org/10.1037/h0033601

Mazepus, H. (2017) What makes political authorities legitimate? Students' ideas about legitimacy in five European democracies and hybrid regimes. *Contemporary Politics, 23*(3), 306-327. https://doi.org/10.1080/13569775.2017.1306762

Meredith, J. (1998). Building operations management theory through case and field research. *Journal of Operations Management*, *16*(4), 441–454. https://doi.org/10.1016/s0272-6963(98)00023-0

Meuser, M. & Nagel, U. (2009). The Expert Interview and Changes in Knowledge Production. In: A. Bogner, B. Littig & W. Menz. (2009) (Eds.). *Interviewing Experts*. Palgrave Macmillan.

Moiser, C. (1947). A critical examination of the concepts of face validity. *Educational and Psychological Measurement, 7*(2), 191–205. https://doi.org/10.1177/001316444700700201

Moores, K. L., Jones, G. L., & Radley, S. C. (2012). Development of an instrument to measure face validity, feasibility and utility of patient questionnaire use during health care: the QQ-10. *International Journal for Quality in Health Care*, *24*(5), 517–524. https://doi.org/10.1093/intqhc/mzs051

Morgan, D. L. (1996). Focus Groups. *Annual Review of Sociology*, *22*(1), 129–152. https://doi.org/10.1146/annurev.soc.22.1.129

Mouffe, C. (2005) On the Political. London: Routledge

Mousazadeh, S., Rakhshan, M., & Mohammadi, F. (2017). Investigation of content and face validity and reliability of sociocultural attitude towards appearance questionnaire-3 (SATAQ-3) among female adolescents. *Iranian Journal of Psychiatry*, *12*(1), 15-20.

Mouter, N., Beek, L. van, Ruijter, A. de, Hernandez, J.I., Schouten, S., Noord, L. Van & Spruit, S. (2021a). *Brede steun voor ambitieus klimaatbeleid als aan vier voorwaarden is voldaan.* Retrieved from https://d2k0ddhflgrk1i.cloudfront.net/TBM/PWE/Cases/Eindrapport%20Klimaatraadpleging%2017062021.pdf

Mouter, N., Koster, P., & Dekker, T. (2021b). Contrasting the recommendations of participatory value evaluation and cost-benefit analysis in the context of urban mobility investments. *Transportation Research Part A: Policy and Practice*, *144*, 54–73. https://doi.org/10.1016/j.tra.2020.12.008

Mouter, N., Shortall, R. M., Spruit, S. L., & Itten, A. V. (2021c). Including young people, cutting time and producing useful outcomes: Participatory value evaluation as a new practice of public participation in the Dutch energy transition. *Energy Research & Social Science*, *75*, 101965. https://doi.org/10.1016/j.erss.2021.101965

Mouter, N., Spruit, S., Itten, A., Shortall, R., Herhandez, J.I., Collewet, M., Koster, P. & Borst, P. (2020). *Bewoners kiezen aardgasvrije wijken: eindrapport en achtergronden.* Commissioned by the municipality of Utrecht. Retrieved from https://d2k0ddhflgrk1i.cloudfront.net/TBM/PWE/Warmtetransitie%20Utrecht/PWE%20Utrecht%20aardgasvrij%20warmtetransitie%20rapport%20FINAL.pdf

Nevo, B. (1985). Face validity revisited. *Journal of Educational Measurement*, *22*(4), 287–293. https://doi.org/10.1111/j.1745-3984.1985.tb01065.x

Newfields, T. (2002). Challenging the notion of face validity. *SHIKEN: The JALT Testing & Evaluation SIG Newsletter*, *6*(3), 14.

Nouws, S. J. J. (2020). Finding a balance between meaningful and useful participation by improving information provision: Assessing the effectiveness of information provision approaches in participatory value evaluation on empowering participants to give informed input on urban climate adaptation projects. Delft University of Technology.

Nunnally, J. C. & Bernstein, I.H. (1994). Psychometric Theory. 3rd edn. New York: McGrawHill.

Oluwatayo, J.A. (2012). Validity and reliability issues in educational research. *Journal of educational and social research, 2*(2), 391-391. https://doi.org/10.5901/jesr.2012.v2n2.391

Omgevingsraad Schiphol (2015). *Instellingsdocument Omgevingsraad Schiphol.* Retrieved from https://www.omgevingsraadschiphol.nl/wp-content/uploads/2015/01/Instellingsdocument-Omgevingsraad-Schiphol-1.pdf

Omgevingsraad Schiphol (2017). *Rapportage hinderbelevingsonderzoek 2017.* Retrieved at March 29, 2022, from https://www.omgevingsraadschiphol.nl/wp-content/uploads/2017/06/Rapportage-Hinderbelevingsonderzoek-2017.pdf

Ozawa, C.P. (2012), Planning resilient communities: Insights from experiences with risky technologies. In Bruce Evan Goldstein (ed.), Collaborative Resilience: Moving through Crisis to Opportunity, Cambridge, MA: MIT Press, pp.19–38.

Peeters, T. (2020). Studying participant decision-making processes in Participatory Value Evaluation. Delft University of Technology.

Pelet, J., Khan, J., & Papadopoulou, P. (2012). Towards a scale for perceptions of mobile interaction: Establishing content and face validity.

Pfadenhauer, M. (2009). At Eye Level: The Expert Interview. A Talk between Expert and Quasi-expert. In: A. Bogner, B. Littig & W. Menz (eds.), Interviewing Experts (pp. 81-97). Palgrave Macmillan.

Plano Clark, V. L., & Creswell, J. (2008). The Mixed Method Reader. Los Angeles: CA: Sage Publication.

Populytics (2021). *5449 Amsterdammers denken mee over windenergie in Amsterdam.* Retrieved from https://assets.amsterdam.nl/publish/pages/1002875/bijlage_3_rapport_online_raadpleging_pwe_amsterdammers_denken_mee.pfd

Quick, K.S. & Bryson, J. (2016). Theories of public participation in governance. https://www.researchgate.net/publication/282733927_Theories_of_public_participation_in_governance

Raad voor het Openbaar Bestuur. (2005). Burgers betrekken—Publicatie—Raad voor het Openbaar Bestuur [Publicatie]. Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. https://www.raadopenbaarbestuur.nl/documenten/publicaties/2005/05/01/burgers-betrekken

Randall, D.M. & Fernandes, M.F. (1991). The social desirability response bias in ethics research. *Journal of Business Ethics, 10*(11), 805-817. https://doi.org/10.1007/bf00383696

Roberts, D. M. (2000). Face validity: is there a place for this in measurement? *SHIKEN: The JALT Testing and Evaluation SIG Newsletter, 4*(2), 6-7.

Roberts, D. (2021). *Average GPA in High School 2022 and Past Years - ThinkImpact*. https://www.thinkimpact.com/average-gpa-in-high-school/#:%7E:text=The%20average%20GPA%20in%20US,average%20male%20GPA%20is%202.9.

Rosener, J.B. (1978). Citizen participation: can we measure its effectiveness? *Public Administration Review, 38*(5), 457. https://doi.org/10.2307/975505

Rotteveel, A. H., Lambooij, M. S., Over, E. A. B., Hernández, J. I., Suijkerbuijk, A. W. M., de Blaeij, A. T., de Wit, G. A., & Mouter, N. (2022). If you were a policymaker, which treatment would you disinvest? A participatory value evaluation on public preferences for active disinvestment of health care interventions in the Netherlands. *Health Economics, Policy and Law*, 1–16. https://doi.org/10.1017/s174413312200010x

Rowe, G., & Frewer, L. J. (2004). Evaluating Public-Participation Exercises: A Research Agenda. *Science, Technology, & Human Values*, *29*(4), 512–556. https://doi.org/10.1177/0162243903259197

Sarantopoulos, A., Spagnol, G. S., Newbold, D., & Li, L. M. (2017). Establishing face validity of the EPLIT questionnaire. *British Journal of Healthcare Management*, *23*(5), 221–227. https://doi.org/10.12968/bjhc.2017.23.5.221

Sartori, R. (2010). Face validity in personality tests: psychometric instruments and projective techniques in comparison. *Quality & Quantity*, *44*(4), 749–759. https://doi.org/10.1007/s11135-009-9224-0

Sartori, R., & Pasini, M. (2006). Quality and Quantity in Test Validity: How can we be Sure that Psychological Tests Measure what they have to? *Quality & Quantity*, *41*(3), 359–374. https://doi.org/10.1007/s11135-006-9006-x

Sato, T., & Ikeda, N. (2015). Test-taker perception of what test items measure: a potential impact of face validity on student learning. *Language Testing in Asia*, *5*(1), 124-130. https://doi.org/10.1186/s40468-015-0019-z

Scharpf, F.W. (1997). Economic integration, democracy and the welfare state. *Journal of European Public Policy, 4*(1), 18-36.

Secolsky, C. (1987). On the Direct Measurement of Face Validity: A Comment on Nevo. *Journal of Educational Measurement*, *24*(1), 82–83. https://doi.org/10.1111/j.1745-3984.1987.tb00265.x

Shorten, A. & Smith, J. (2017). Mixed methods research: expanding the evidence base. *Evidence Based Nursing, 20*(3), 74-75. https://doi.org/10.1136/eb-2017-102699

Shotland, A., Alliger, G. M., & Sales, T. (1998). Face Validity in the Context of Personnel Selection: A multimedia Approach. *International Journal of Selection and Assessment*, *6*(2), 124–130. https://doi.org/10.1111/1468-2389.00081

Spruit, S., & Mouter, N. (2020). *1376 inwoners van Súdwest-Fryslân over het toekomstige energiebeleid van hun gemeente: de uitkomsten van een raadpleging*. https://d2k0ddhflgrk1i.cloudfront.net/TBM/PWE/Cases/Energie%20in%20SWF/20200624%20Resultaten%20Raadpleging%20Toekomst%20van%20Energie%20SWF.pdf

Spruit, S. & Mouter, N. (2021). *1795 Inwoners over het toekomstige energiebeleid van regio Foodvalley.* Retrieved from https://populytics.nl/wp-content/uploads/2022/04/Rapport-Regio-Foodvalley.pdf

Stallard, P., & Rayner, H. (2005). The Development and Preliminary Evaluation of a Schema Questionnaire for Children (SQC). *Behavioural and Cognitive Psychotherapy*, *33*(2), 217–224. https://doi.org/10.1017/s1352465804001912

Taherdoost, H. (2016). Validity and Reliability of the Research Instrument; How to Test the Validation of a Questionnaire/Survey in a Research. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3205040

Tsang, S., Royse, C., & Terkawi, A. (2017). Guidelines for developing, translating, and validating a questionnaire in perioperative and pain medicine. *Saudi Journal of Anaesthesia*, *11*(5), 80. https://doi.org/10.4103/sja.sja_203_17

Tweed, M., & Cookson, J. (2001). The face validity of a final professional clinical examination. *Medical Education*, *35*(5), 465–473. https://doi.org/10.1046/j.1365-2923.2001.00895.x

UCLA (2021). *One-way MANOVA; SPSS Data Analysis Examples.* Retrieved from https://stats.oarc.ucla.edu/spss/dae/one-way-manova/

Zainal, Z. (2007). Case study as a research method. *Jurnal Kemanusiaan, (9),* 1-6.

Zheng, Y., & Schachter, H. L. (2016). Explaining Citizens' E-Participation Use: the Role of Perceived Advantages. *Public Organization Review*, *17*(3), 409–428. https://doi.org/10.1007/s11115-016-0346-2

Appendix A Research flow diagram

Figure A.1 presents the detailed research flow diagram of this research.

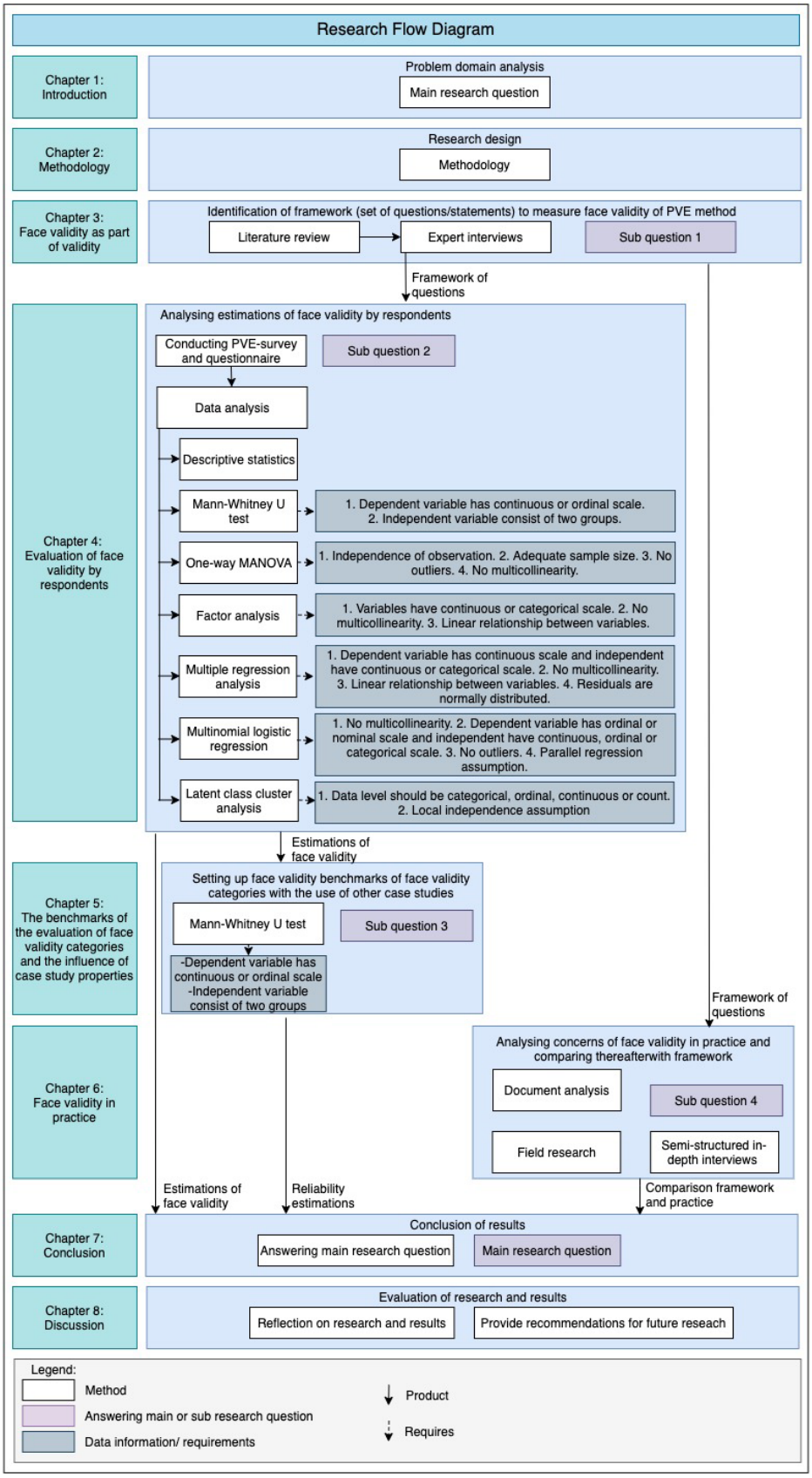


Figure A.1: Detailed research flow diagram

Appendix B Literature review

Selection of articles for literature review for knowledge gap and main research question

A literature review is conducted to define a knowledge gap and a main research question (section 1.1). Scopus and Google Scholar were the databases used to search for references. Combinations of the following keywords were used for this search: Participatory Value Evaluation, PVE, public participation, citizen participation and (face) validity. Thereby, the master thesis reports of Nouws (2020) and Bouwmeester (2021) were used for backward snowballing, since these reports provide a clear overview of the current status of the PVE method. Articles were selected on their relevance with regard to the PVE method, the practice of public participation related to this method and the influence of information provision within the PVE.

Selection of articles for literature review for the first sub research question

Search words have been used to find literature for the literature review. The following search used have been used:

"Face validity" OR "Content validity" OR "Validity" OR "Validation" AND "Assessment" OR "Measurement" OR "Questionnaire" OR "Process" OR "Main themes" OR "Categories" OR "Criteria" OR "Research instrument" OR "Psychometric test" OR ("Citizen" OR "Public) "Participation" OR "Importance" OR "Legitimacy" OR "Raters" OR "Expert"

Furthermore, four articles have been used for backward and forward snowballing. These four articles are: Nevo (1985), Gaber & Gaber (2010), Hardesty & Bearden (2004) and Connell et al. (2018).

Appendix C Expert interview protocol

In this appendix, the protocol used for the expert interviews is presented. Since two experts have been interviewed in the field of psychology under which face validity falls, two experts in the field of the PVE method and one expert with knowledge in both fields, the protocol differs between the types of experts. For example, PVE method experts need more explanations about the concept of face validity. This is indicated in the protocol. The structure of this protocol is as follows. First of all, there is an introduction in which the consequences of participation were made known and the experts were asked for their consent. Thereafter, three interview topics are discussed. First, there is a topic about the completeness of categories identified in the literature review. Second, there is a topic about prioritizing five categories. Finally, there is a topic about the specific questioning of the categories in terms of statements.

Expert interview protocol

Introduction

- Thank the interviewee for his or her time and participation to this interview.
- Indicate that the interviewee may stop at any time during the interview without giving a reason.
- The interview data is securely stored in the data centre of Delft University of Technology.
- Explain that this interview is part of a master thesis. The topic concerns the face validity of the PVE method.
- The purpose of this interview is to prioritize face validity categories for the PVE method.
- Does the participant have any questions?
- Ask if the audio of the interview may be recorded. This question is first asked without the audio on and repeated once the audio is on.

Completeness of categories

For the PVE method experts: Face validity is a type of validity that deals with the operationalization of an instrument, such as the PVE. Authors state that it is also about the 'looks and the feel'. In other words, a test has face validity when individuals agree that the test appears valid with regard to the type of measurement. A question that is often asked in the literature is: is this research meaningful at first sight?

Various categories of face validity are distinguished in the literature. The categories that I have distinguished are shown in table C.1.

To what extent do you think these categories are complete or are you missing one or more categories of face validity?

Table C.1: Overview of face validity categories

Category
Clarity
Relevance
Readability
Appropriateness of difficulty level
Unambiguity
Aestethicy
Completeness
Feasibility
Efficacy
Insensitivity
Familiarity

Prioritizing of categories

In my master thesis I will present statements in a PVE consultation to the respondents who will assess face validity.

For psychology experts: In short, the PVE method is a relatively new research method to evaluate policy options. This method offers the possibility to facilitate participation for large groups of citizens. Citizens can provide advice on a policymaker's policy issue in a low-threshold manner. Citizens are put in the shoes of the decisionmaker. An important part of a PVE consultation is the choice task. An example of what a choice task looks like, is shown in figure C.1.

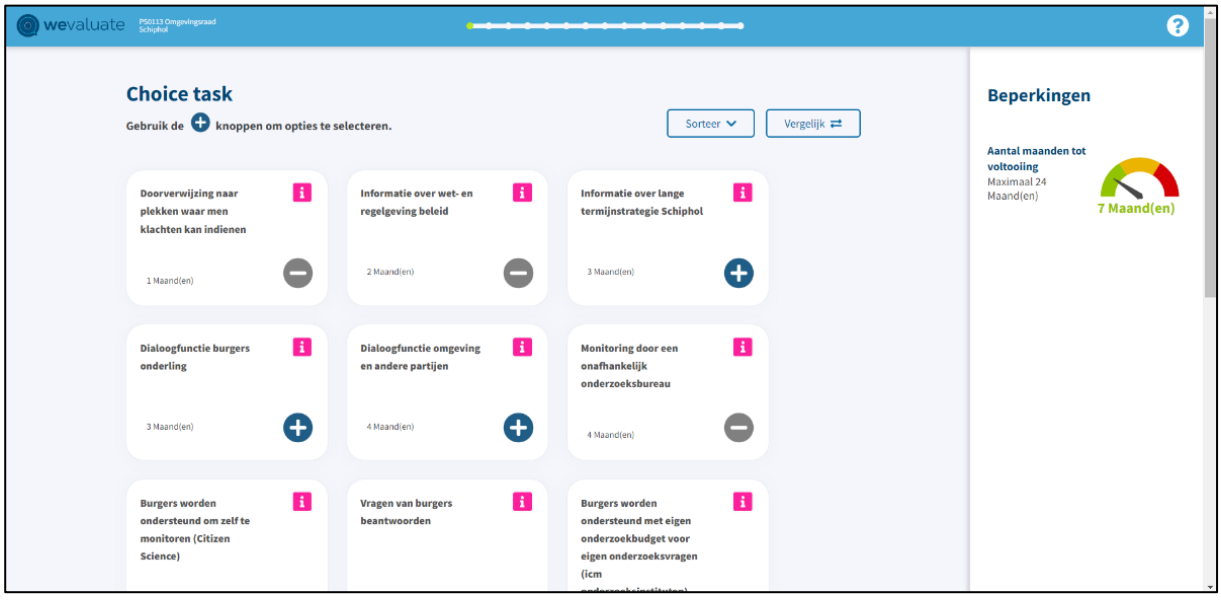


Figure C.1: Example of a choice task

The PVE consultation that I will use is in my master thesis is about the case of the Schiphol Environmental Council (ORS). The ORS is a council that local residents can turn to with questions, comments and requests about the public interests of Schiphol Airport. However, the van Geel committee concluded that the decision-making model in the ORS no longer

works. That is why van Geel has proposed two new entities that contribute to more intensive and broader participation and improved information provision. These two entities are called the Schiphol Social Council and the Environmental House. To measure face validity in this case study, the following statements have been linked to the categories (see table C.2). However, there is place in this PVE consultation to measure five categories. From this follows the questions:

Which five categories (from table C.2 + any additions) do you think are the most important for the PVE method and can you substantiate this?
Could you rank your five chosen categories?

Table C.2: Statements linked to categories

Category of face validity	Statement
Clarity	I have received sufficient information to make a choice.
Relevance	I think this is a good method for expressing my opinion on how citizens should be involved in decision-making about Schiphol.
Readability	I found the choice task understandable.
Appropriateness of difficulty level	I found it easy to make a choice.
Unambiguity	I found it clear with the choice task what was meant by each task.
Aesthetics	I thought the platform in which I made the choice task looked attractive.
Completeness	I felt that I could give all my opinions on how citizens are involved in decision-making about Schiphol and how information should be provided.
Feasibility	It was doable to complete the survey within twenty minutes.
Efficacy	I thought that the choice task is of a scientific level.
Insensitivity	I felt that the choice task evoked too many negative feelings in me.
Familiarity	Before I filled in this choice tasks, I was already familiar with the idea of setting up a Schiphol Social Council and an Environmental House.

Suggestions for statements

Based on the five categories selected by the expert, the following question is addressed:
Do you have any comments or remarks on the statements as they are drawn up per category?

Code list

The following code list is used to transcribe the expert interview. The expert interviews were held in Dutch, so the codes are listed in English as well as in Dutch below.

Dutch	**English**
Duidelijkheid	Clarity
Relevantie	Relevance
Leesbaarheid	Readability
Geschiktheid van de moelijkheidsgraad	Appropriateness of difficulty level
Eenduidigheid	Umambiguity
Esthetiek	Aesthetics
Volledigheid	Completeness
Haalbaarheid	Feasibility
Doeltreffendheid	Efficacy
Ongevoeligheid	Insensitivity
Bekendheid	Familiarity
Ergernis	Annoyance
Capabel	Capable
Mogelijkheden	Capabilities
Belangrijk	Important
Ervaring	Experience

Appendix D Design of the PVE consultation

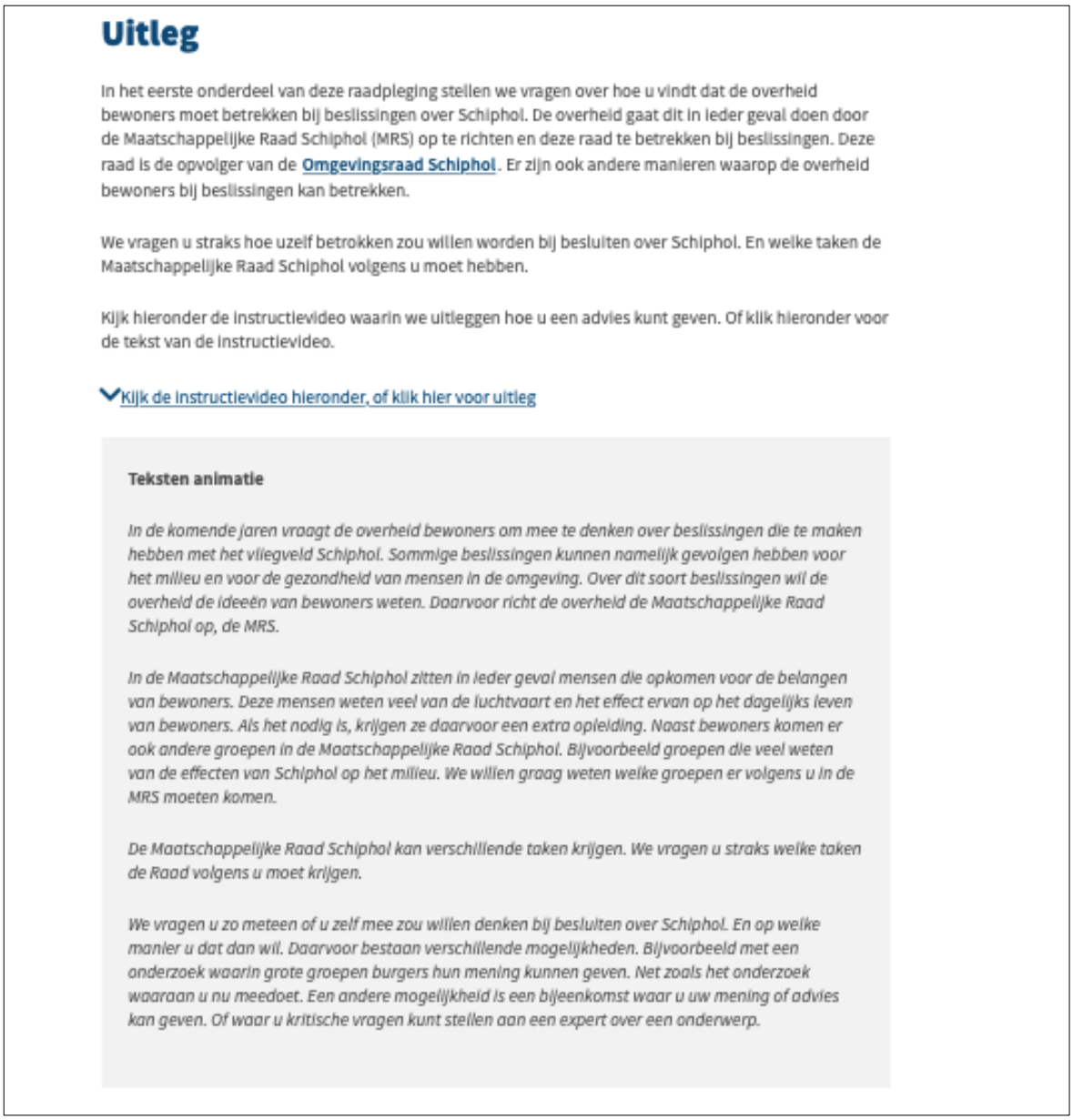
This Appendix presents the design of whole the PVE consultation. The whole PVE design consists of an introduction, an instruction, a part about how respondents want to be involved in in decisions about Schiphol, a part about the Schiphol Social Council which includes a choice task and in-depth questions, a part about the Environmental House which included a choice task and in-depth questions and finally there are the last general questions. The whole PVE consultation is presented in the figures below.

Introduction

Figure D.1: Introduction of the PVE consultation

Figure D.2: Information agreement of the PVE consultation

Instruction



Uitleg

In het eerste onderdeel van deze raadpleging stellen we vragen over hoe u vindt dat de overheid bewoners moet betrekken bij beslissingen over Schiphol. De overheid gaat dit in ieder geval doen door de Maatschappelijke Raad Schiphol (MRS) op te richten en deze raad te betrekken bij beslissingen. Deze raad is de opvolger van de **Omgevingsraad Schiphol**. Er zijn ook andere manieren waarop de overheid bewoners bij beslissingen kan betrekken.

We vragen u straks hoe uzelf betrokken zou willen worden bij besluiten over Schiphol. En welke taken de Maatschappelijke Raad Schiphol volgens u moet hebben.

Kijk hieronder de instructievideo waarin we uitleggen hoe u een advies kunt geven. Of klik hieronder voor de tekst van de instructievideo.

Kijk de instructievideo hieronder, of klik hier voor uitleg

Teksten animatie

In de komende jaren vraagt de overheid bewoners om mee te denken over beslissingen die te maken hebben met het vliegveld Schiphol. Sommige beslissingen kunnen namelijk gevolgen hebben voor het milieu en voor de gezondheid van mensen in de omgeving. Over dit soort beslissingen wil de overheid de ideeën van bewoners weten. Daarvoor richt de overheid de Maatschappelijke Raad Schiphol op, de MRS.

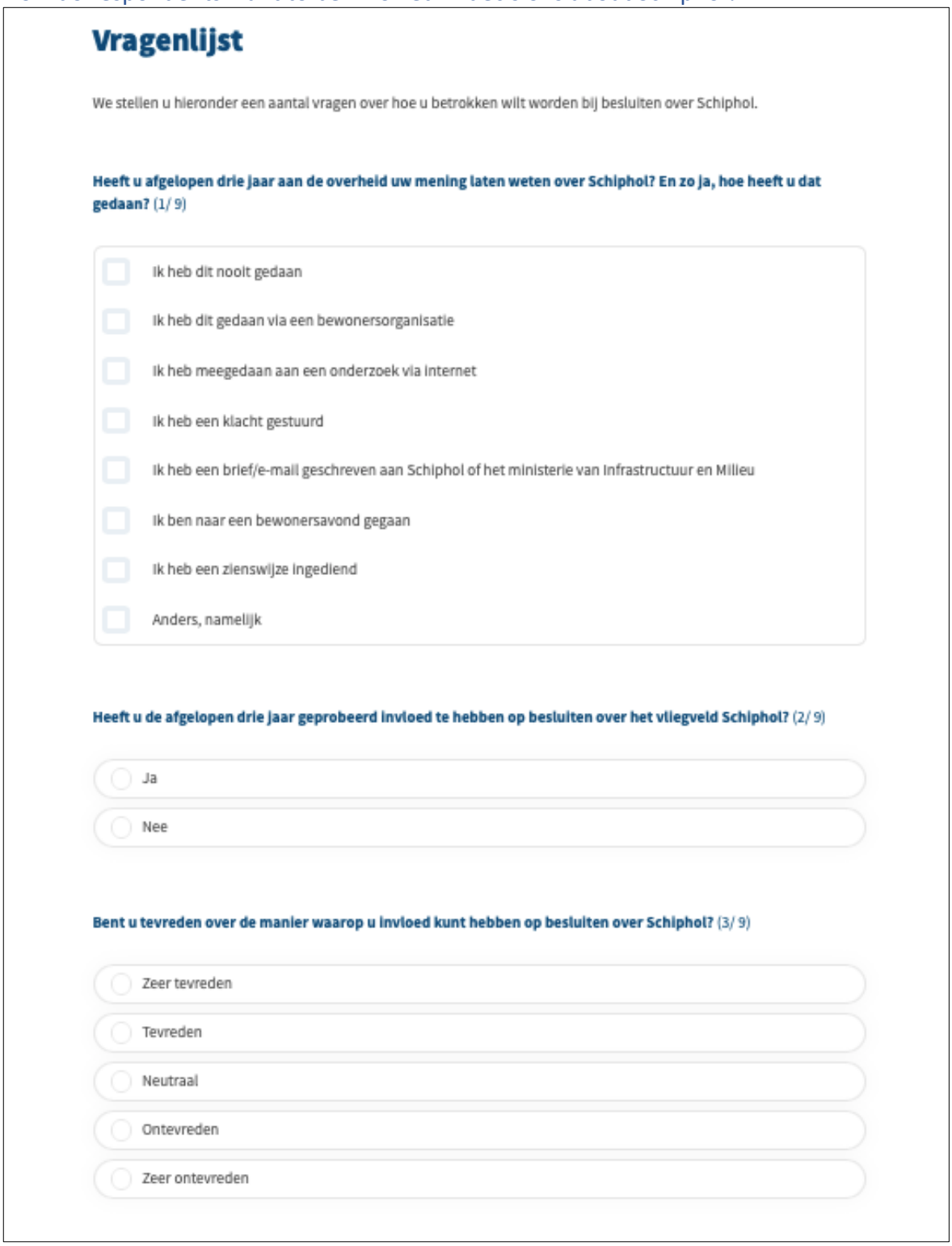
In de Maatschappelijke Raad Schiphol zitten in ieder geval mensen die opkomen voor de belangen van bewoners. Deze mensen weten veel van de luchtvaart en het effect ervan op het dagelijks leven van bewoners. Als het nodig is, krijgen ze daarvoor een extra opleiding. Naast bewoners komen er ook andere groepen in de Maatschappelijke Raad Schiphol. Bijvoorbeeld groepen die veel weten van de effecten van Schiphol op het milieu. We willen graag weten welke groepen er volgens u in de MRS moeten komen.

De Maatschappelijke Raad Schiphol kan verschillende taken krijgen. We vragen u straks welke taken de Raad volgens u moet krijgen.

We vragen u zo meteen of u zelf mee zou willen denken bij besluiten over Schiphol. En op welke manier u dat dan wil. Daarvoor bestaan verschillende mogelijkheden. Bijvoorbeeld met een onderzoek waarin grote groepen burgers hun mening kunnen geven. Net zoals het onderzoek waaraan u nu meedoet. Een andere mogelijkheid is een bijeenkomst waar u uw mening of advies kan geven. Of waar u kritische vragen kunt stellen aan een expert over een onderwerp.

Figure D.3: Instruction of the PVE consultation

How do respondents want to be involved in decisions about Schiphol?



Vragenlijst

We stellen u hieronder een aantal vragen over hoe u betrokken wilt worden bij besluiten over Schiphol.

Heeft u afgelopen drie jaar aan de overheid uw mening laten weten over Schiphol? En zo ja, hoe heeft u dat gedaan? (1/ 9)

- [ ] Ik heb dit nooit gedaan
- [ ] Ik heb dit gedaan via een bewonersorganisatie
- [ ] Ik heb meegedaan aan een onderzoek via internet
- [ ] Ik heb een klacht gestuurd
- [ ] Ik heb een brief/e-mail geschreven aan Schiphol of het ministerie van Infrastructuur en Milieu
- [ ] Ik ben naar een bewonersavond gegaan
- [ ] Ik heb een zienswijze ingediend
- [ ] Anders, namelijk

Heeft u de afgelopen drie jaar geprobeerd invloed te hebben op besluiten over het vliegveld Schiphol? (2/ 9)

- ○ Ja
- ○ Nee

Bent u tevreden over de manier waarop u invloed kunt hebben op besluiten over Schiphol? (3/ 9)

- ○ Zeer tevreden
- ○ Tevreden
- ○ Neutraal
- ○ Ontevreden
- ○ Zeer ontevreden

Figure D.4: Questions about how do respondents want to be involved – part 1

Kunt u uitleggen waarom u dit vindt? (4/ 9)

We laten verschillende manieren zien waarop u betrokken zou kunnen worden bij de besluitvorming over Schiphol. Aan welke manieren zou u meedoen? (5/ 9)

	Ik zou hier nooit aan meedoen.	Ik zou hier soms aan meedoen	Ik zou hier regelmatig aan meedoen.	Ik zou hier vaak aan meedoen.	Ik weet niet of ik hier aan mee zou doen.
Een onderzoek via internet, waarin veel bewoners hun mening kunnen geven. Op dit moment doet u aan zo'n onderzoek mee.	○	○	○	○	○
Een bijeenkomst voor bewoners waar u uw mening kunt geven aan mensen van Schiphol.	○	○	○	○	○
Een bijeenkomst voor bewoners waar u uw mening kunt geven aan mensen van de overheid.	○	○	○	○	○
Een bijeenkomst voor bewoners waar een expert vertelt over een onderwerp. En waar u kritische vragen kunt stellen.	○	○	○	○	○
Een bijeenkomst voor bewoners waar u hoort hoe het gaat met plannen die met Schiphol te maken hebben. En waar u vragen kunt stellen.	○	○	○	○	○

Kunt u uitleggen waarom u deze antwoorden hebt gegeven? (6/ 9)

Figure D.5: Questions about how do respondents want to be involved – part 2

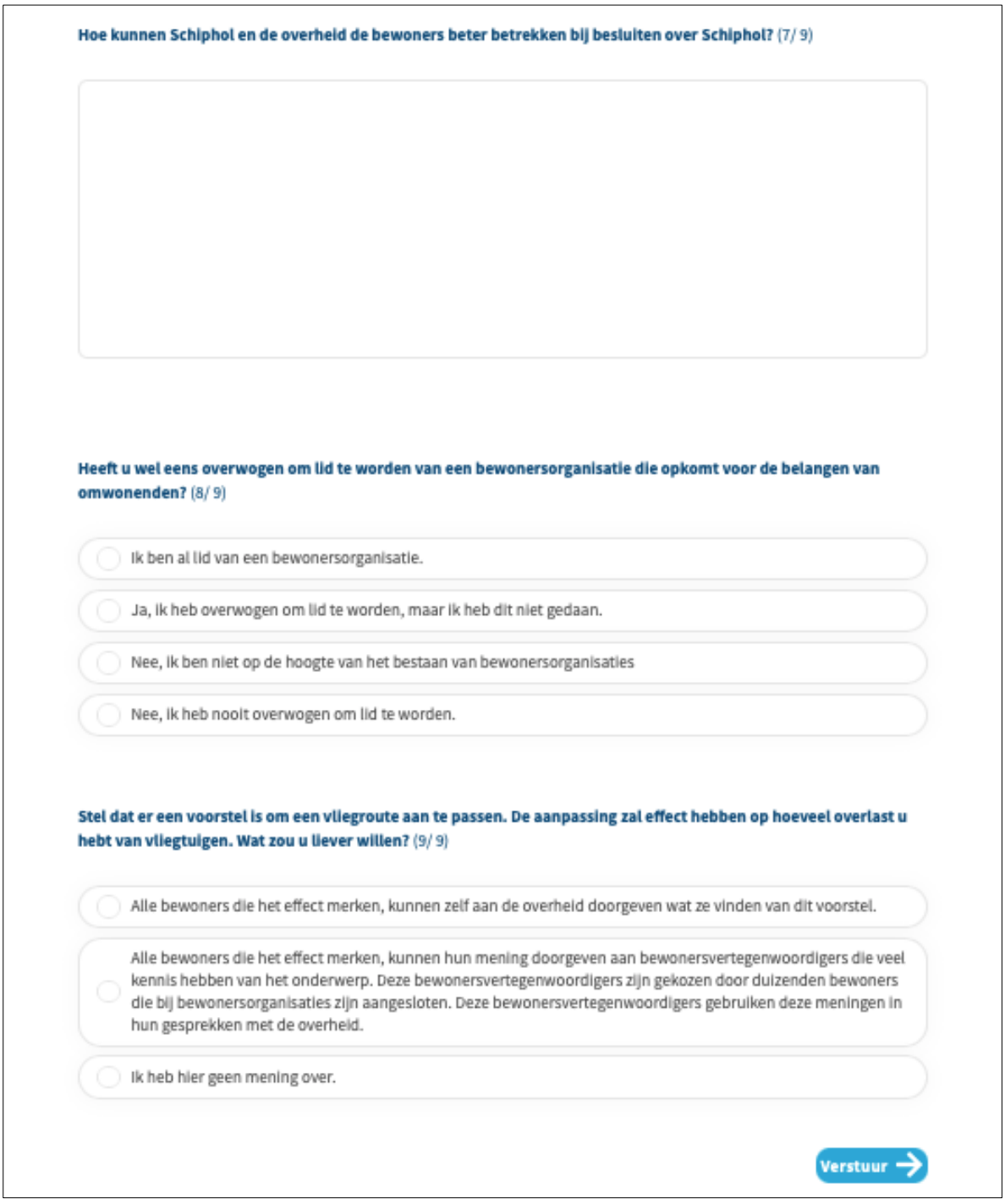


Figure D.6: Questions about how do respondents want to be involved – part 3

The Schiphol Social Council

After the introduction of the choice task (figure D.7), the respondent is shown one of the two following figures D.8 or D.9. These two figures show the two different choice tasks which make the difference between the two experiments. Figure D.8 presents the 'sliders' choice task and figure D.9 the 'points' choice task.

Choice task of the Schiphol Social Council

Introductie keuzetaak MRS

We vragen u zo meteen welke taken de Maatschappelijke Raad Schiphol volgens u moet hebben.

Kijk hieronder de instructievideo waarin we uitleggen hoe u een advies kunt geven. Of klik hieronder voor de tekst van de instructievideo.

Kijk de animatie hieronder, of klik hier voor uitleg

Zo meteen gaat u uw eerste advies geven. U ziet acht taken die de Maatschappelijke Raad Schiphol zou kunnen krijgen. We willen weten welke taken de raad volgens u moet hebben.

Bij elke taak ziet u een schuifje dat u kunt verplaatsen. Vindt u dat een taak hoort bij het werk van de Maatschappelijke Raad Schiphol? Dan zet u het schuifje naar rechts. Vindt u dat iets bij het werk van de Raad hoort én dat de Raad er veel aandacht aan moet besteden? Dan zet u het schuifje helemaal naar rechts.

Met dit onderzoek willen we graag leren wat u de belangrijkste taken vindt van de Maatschappelijke Raad Schiphol. Daarom kunt u er niet voor kiezen dat de Raad aan alle taken veel aandacht moet besteden.

Rechtsboven in het scherm ziet u een metertje. Dit laat zien hoe druk de Raad het krijgt met de taken die u heeft gekozen. Als het metertje in het rood staat kunt u niet nog meer taken kiezen of meer aandacht voor taken vragen. U mag er ook voor kiezen dat de Raad geen of weinig taken uitvoert.

Wilt u weten wat er gebeurt als de Maatschappelijke Raad een taak gaat uitvoeren? Klik dan op i. U ziet dan bijvoorbeeld hoeveel inspanning een taak kost.

Als u klaar bent met uw advies, klik dan op 'volgende'.

Figure D.7: Introduction of the Schiphol Social Council choice task

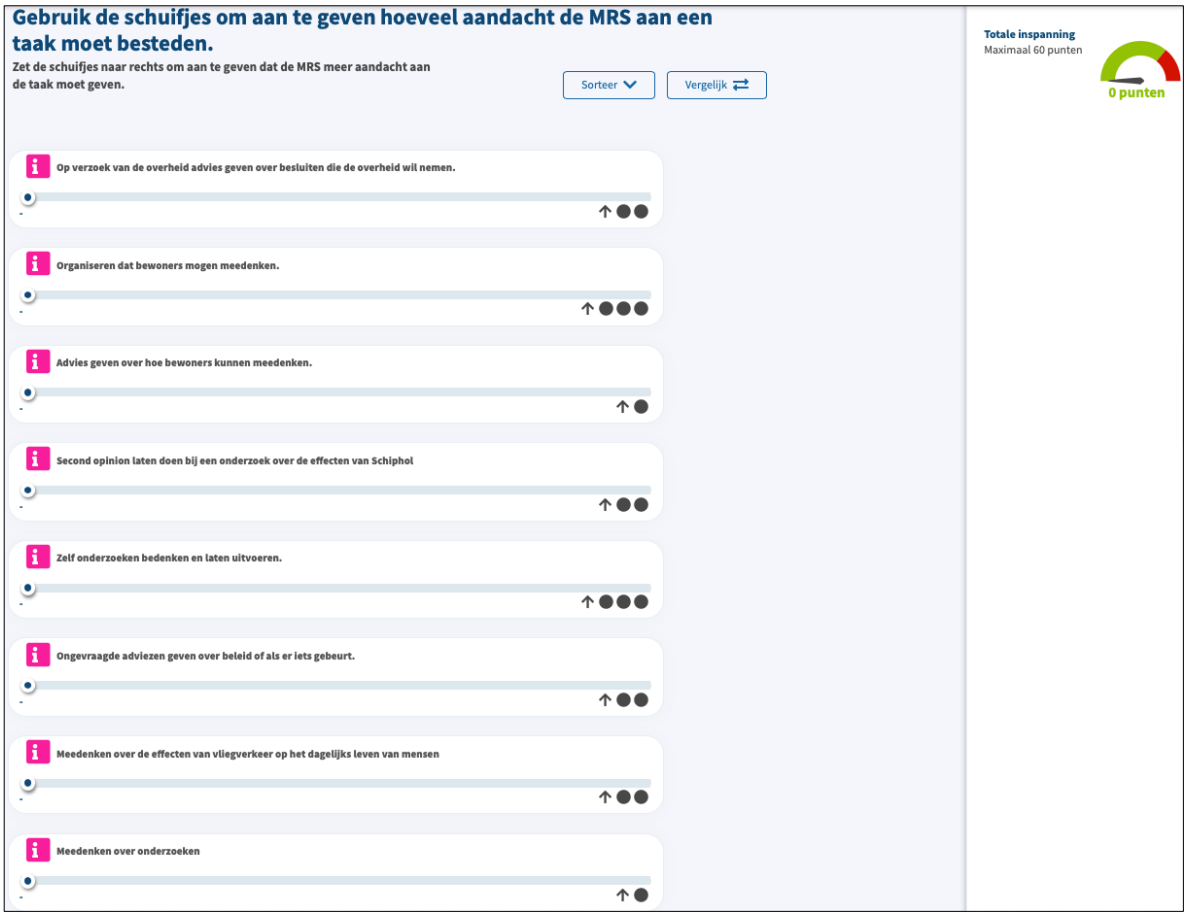


Figure D.8: Experiment 1 of the Schiphol Social Council choice task ('sliders' choice task)

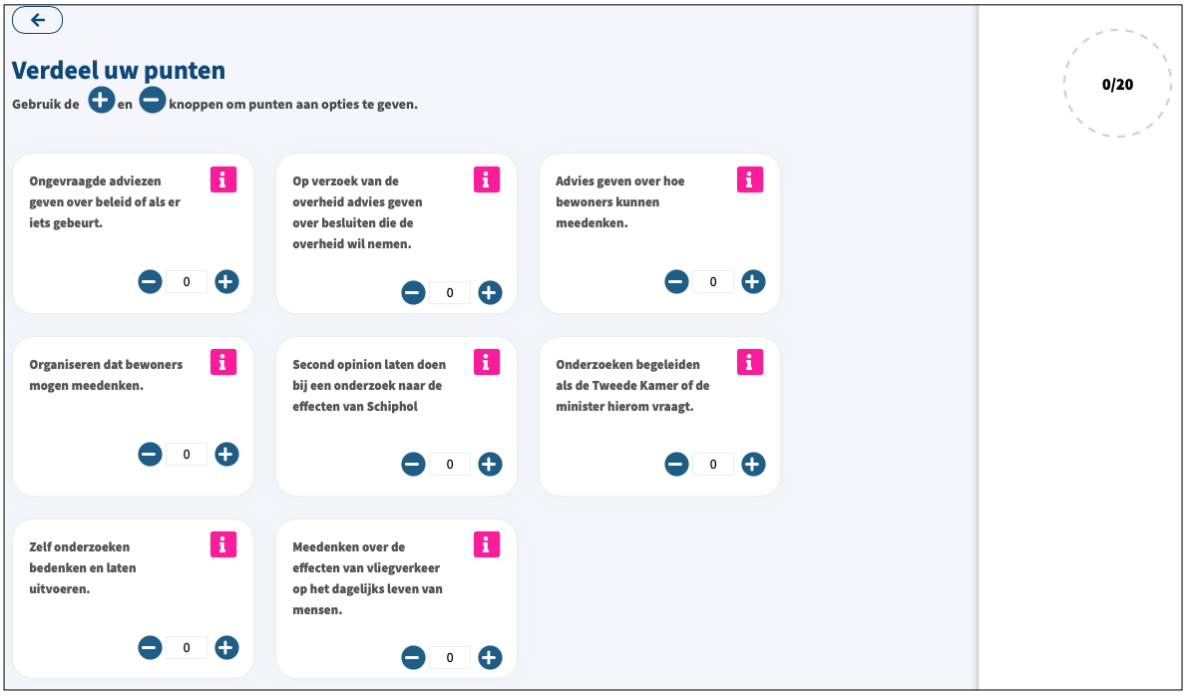


Figure D.9: Experiment 2 of the Schiphol Social Council choice task ('points' choice task)

Figure D.10: Motivation for each choice in the Schiphol Social Council choice task

In-depth question about the Schiphol Social Council

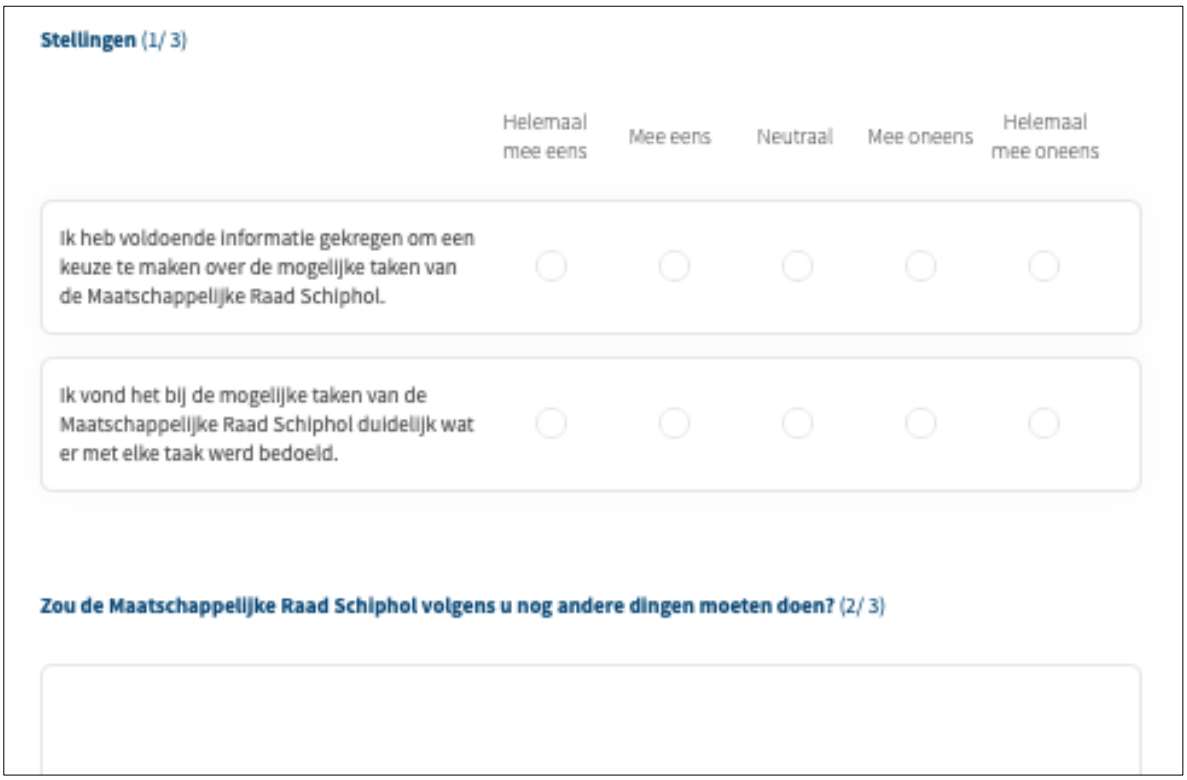


Figure D.11: In-depth questions about the Schiphol Social Council – part 1

Kunt u voor de volgende groepen aangeven hoe belangrijk u het vindt dat ze onderdeel uitmaken van de Maatschappelijke Raad Schiphol?* (3/ 5)

	Zeer belangrijk	Belangrijk	Neutraal	Onbelangrijk	Zeer onbelangrijk	Geen mening
Bewoners die de belangen vertegenwoordigen van omwonenden die overlast ervaren van Schiphol.	○	○	○	○	○	○
Bewoners die de belangen vertegenwoordigen van mensen die voordelen ervaren van Schiphol (bijvoorbeeld omdat ze er werken).	○	○	○	○	○	○
Milieuorganisaties	○	○	○	○	○	○
Werkgeversvereniging en Werknemersvereniging (vakbonden)	○	○	○	○	○	○
Het ministerie van Infrastructuur en Waterstaat	○	○	○	○	○	○
Wetenschappers die expert zijn op het gebied van de effecten van Schiphol op gezondheid, geluid en milieu.	○	○	○	○	○	○

Kunt u uw antwoord toelichten? (4/ 5)

Zijn er groepen of organisaties die volgens u een rol moeten krijgen in de Maatschappelijke Raad Schiphol, maar die niet in het rijtje stonden (5/ 5)

Figure D.12: In-depth questions about the Schiphol Social Council – part 2

Algemene vragen

We willen graag weten wat u belangrijk vindt wanneer de overheid bewoners betrekt bij besluiten over Schiphol. U krijgt zo meteen een aantal stellingen te zien. We vragen u wat u van deze stellingen vindt.

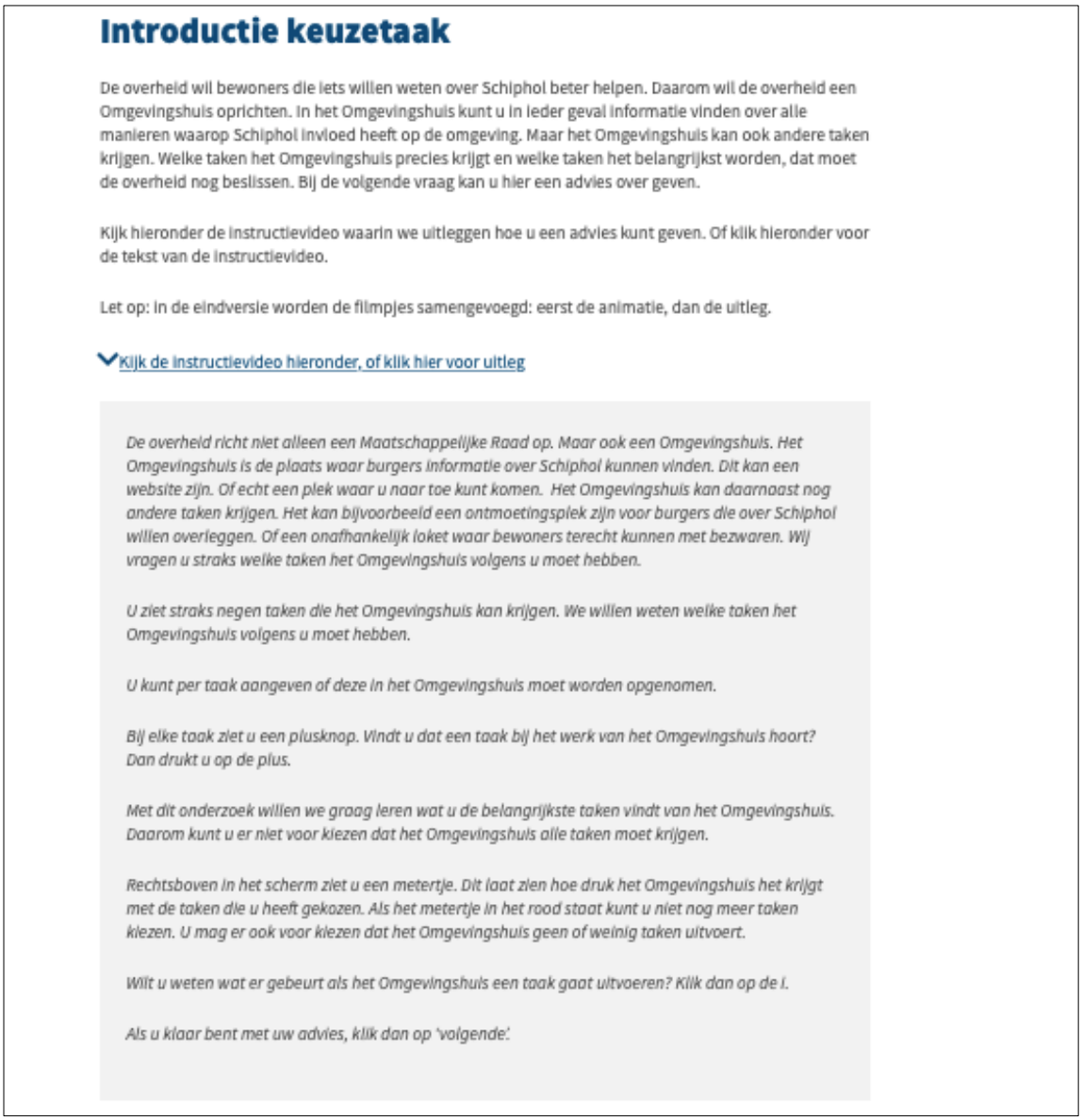
Stellingen (1/ 1)

	Zeer mee eens	Mee eens	Neutraal	Mee oneens	Zeer mee oneens
De groep bewoners die mag meedenken bij besluiten over Schiphol moet een goede afspiegeling vormen van de mensen die voordelen hebben van Schiphol en de mensen die overlast hebben van Schiphol.	○	○	○	○	○
De groep bewoners die mag meedenken bij besluiten over Schiphol moet een goede afspiegeling vormen van de mensen die overlast hebben van Schiphol.	○	○	○	○	○
Bewoners die heel veel overlast hebben van Schiphol (omdat zij dichtbij de baan wonen) moeten meer te zeggen hebben bij besluiten over Schiphol dan bewoners die weinig of geen overlast hebben.	○	○	○	○	○
Bewoners die overlast hebben van Schiphol moeten meer te zeggen hebben bij besluiten over Schiphol dan bewoners die voordelen ervaren van Schiphol.	○	○	○	○	○
De groep bewoners die mag meedenken bij besluiten over Schiphol moet een goede afspiegeling vormen van de hele Nederlandse bevolking.	○	○	○	○	○
Gaat een besluit over een verandering van een vliegroute? Dan mogen daar alleen bewoners over meedenken voor wie de overlast gaat veranderen door de nieuwe vliegroute.	○	○	○	○	○
Gaat een besluit over hoeveel vliegtuigen 's nachts van en naar Schiphol mogen vliegen? Dan mogen daar alleen bewoners over meedenken voor wie de overlast hierdoor gaat veranderen.	○	○	○	○	○
Laat de overheid bewoners meedenken over besluiten? Dan moeten ze bij hun besluit laten zien wat ze met deze adviezen hebben gedaan.	○	○	○	○	○
Laat de overheid bewoners meedenken over besluiten? Dan moeten ze de adviezen van bewoners zoveel mogelijk overnemen.	○	○	○	○	○

Figure D.13: In-depth questions about the Schiphol Social Council and participation principles – part 3

The Environmental House

Choice task about the Environmental House



Introductie keuzetaak

De overheid wil bewoners die iets willen weten over Schiphol beter helpen. Daarom wil de overheid een Omgevingshuis oprichten. In het Omgevingshuis kunt u in ieder geval informatie vinden over alle manieren waarop Schiphol invloed heeft op de omgeving. Maar het Omgevingshuis kan ook andere taken krijgen. Welke taken het Omgevingshuis precies krijgt en welke taken het belangrijkst worden, dat moet de overheid nog beslissen. Bij de volgende vraag kan u hier een advies over geven.

Kijk hieronder de instructievideo waarin we uitleggen hoe u een advies kunt geven. Of klik hieronder voor de tekst van de instructievideo.

Let op: in de eindversie worden de filmpjes samengevoegd: eerst de animatie, dan de uitleg.

Kijk de instructievideo hieronder, of klik hier voor uitleg

De overheid richt niet alleen een Maatschappelijke Raad op. Maar ook een Omgevingshuis. Het Omgevingshuis is de plaats waar burgers informatie over Schiphol kunnen vinden. Dit kan een website zijn. Of echt een plek waar u naar toe kunt komen. Het Omgevingshuis kan daarnaast nog andere taken krijgen. Het kan bijvoorbeeld een ontmoetingsplek zijn voor burgers die over Schiphol willen overleggen. Of een onafhankelijk loket waar bewoners terecht kunnen met bezwaren. Wij vragen u straks welke taken het Omgevingshuis volgens u moet hebben.

U ziet straks negen taken die het Omgevingshuis kan krijgen. We willen weten welke taken het Omgevingshuis volgens u moet hebben.

U kunt per taak aangeven of deze in het Omgevingshuis moet worden opgenomen.

Bij elke taak ziet u een plusknop. Vindt u dat een taak bij het werk van het Omgevingshuis hoort? Dan drukt u op de plus.

Met dit onderzoek willen we graag leren wat u de belangrijkste taken vindt van het Omgevingshuis. Daarom kunt u er niet voor kiezen dat het Omgevingshuis alle taken moet krijgen.

Rechtsboven in het scherm ziet u een metertje. Dit laat zien hoe druk het Omgevingshuis het krijgt met de taken die u heeft gekozen. Als het metertje in het rood staat kunt u niet nog meer taken kiezen. U mag er ook voor kiezen dat het Omgevingshuis geen of weinig taken uitvoert.

Wilt u weten wat er gebeurt als het Omgevingshuis een taak gaat uitvoeren? Klik dan op de i.

Als u klaar bent met uw advies, klik dan op 'volgende'.

Figure D.14: Introduction of the Environmental House choice task

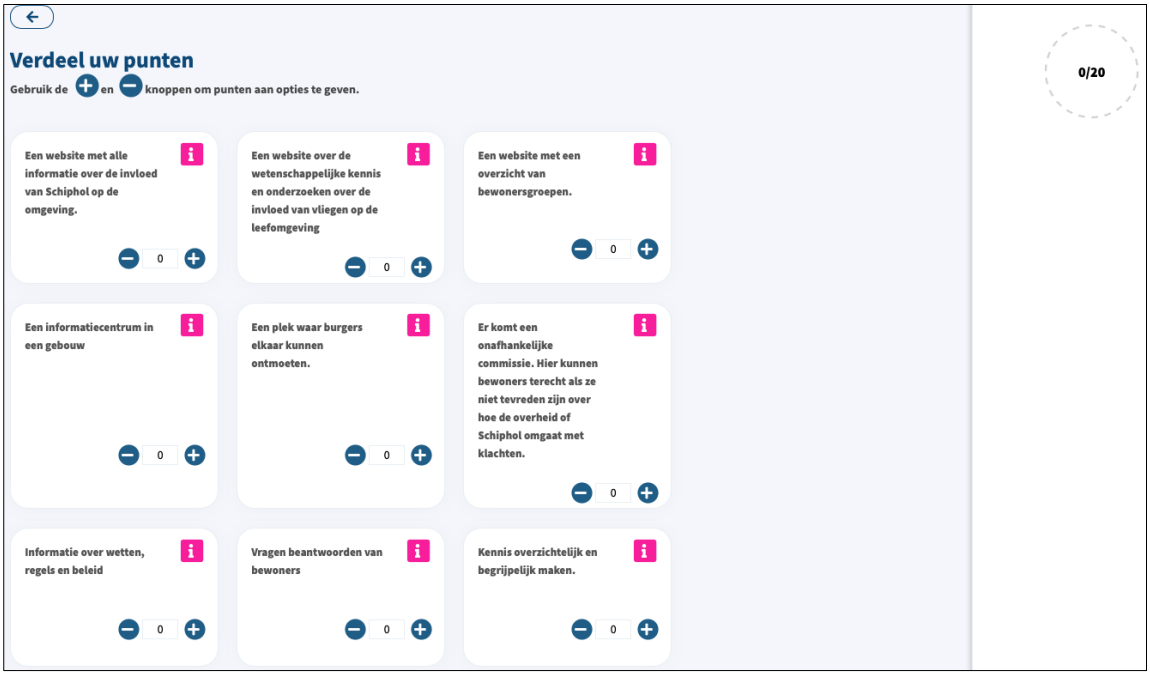


Figure D.15: The Environmental House choice task

Figure D.16: Motivation for each choice in the Environmental House choice task

In-depth questions about the Environmental House

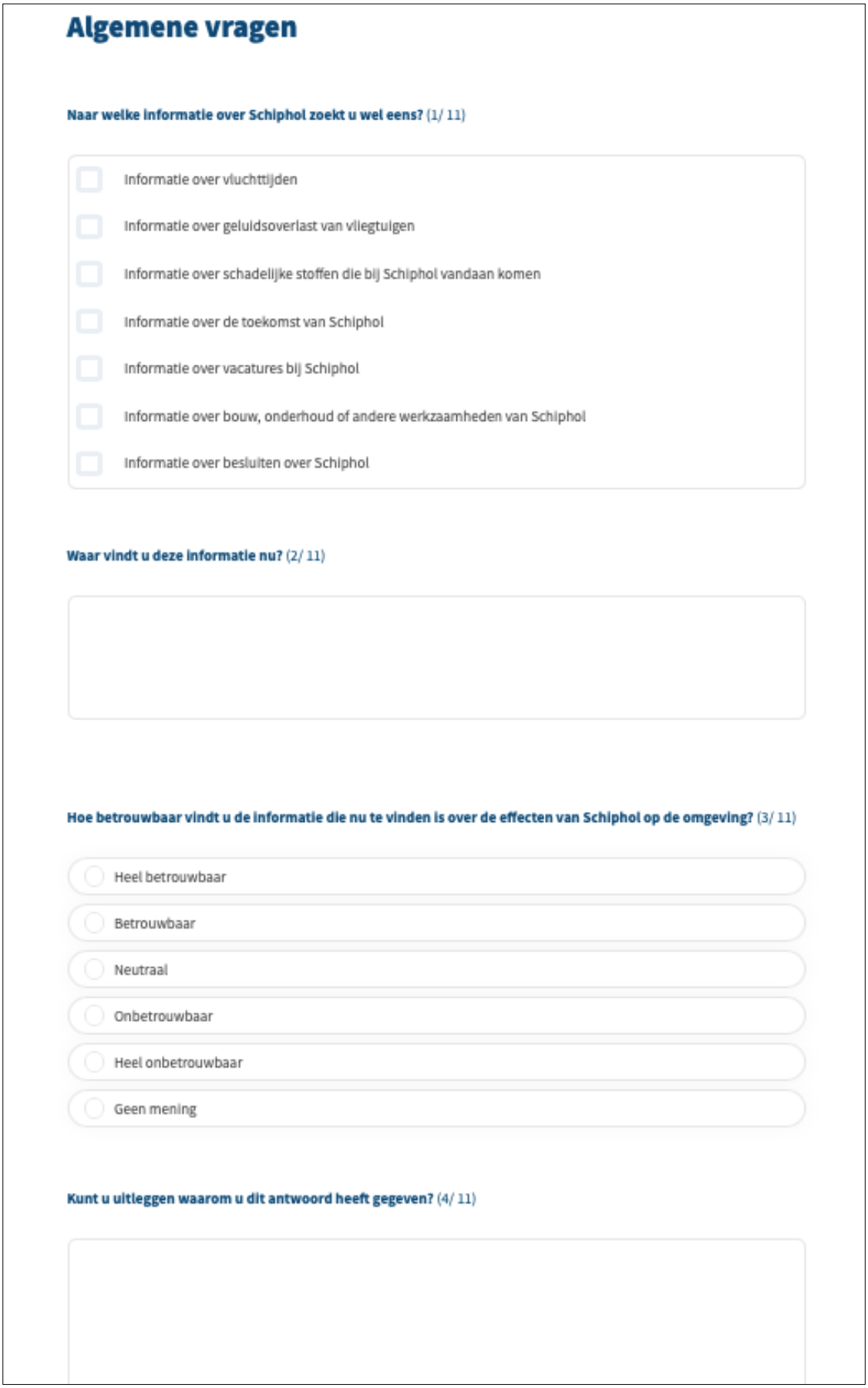


Figure D.17: In-depth questions about the Environmental House – part 1

Algemene vragen

Hoe onafhankelijk vindt u de informatie die nu te vinden is over de effecten van Schiphol op de omgeving? (5/ 11)

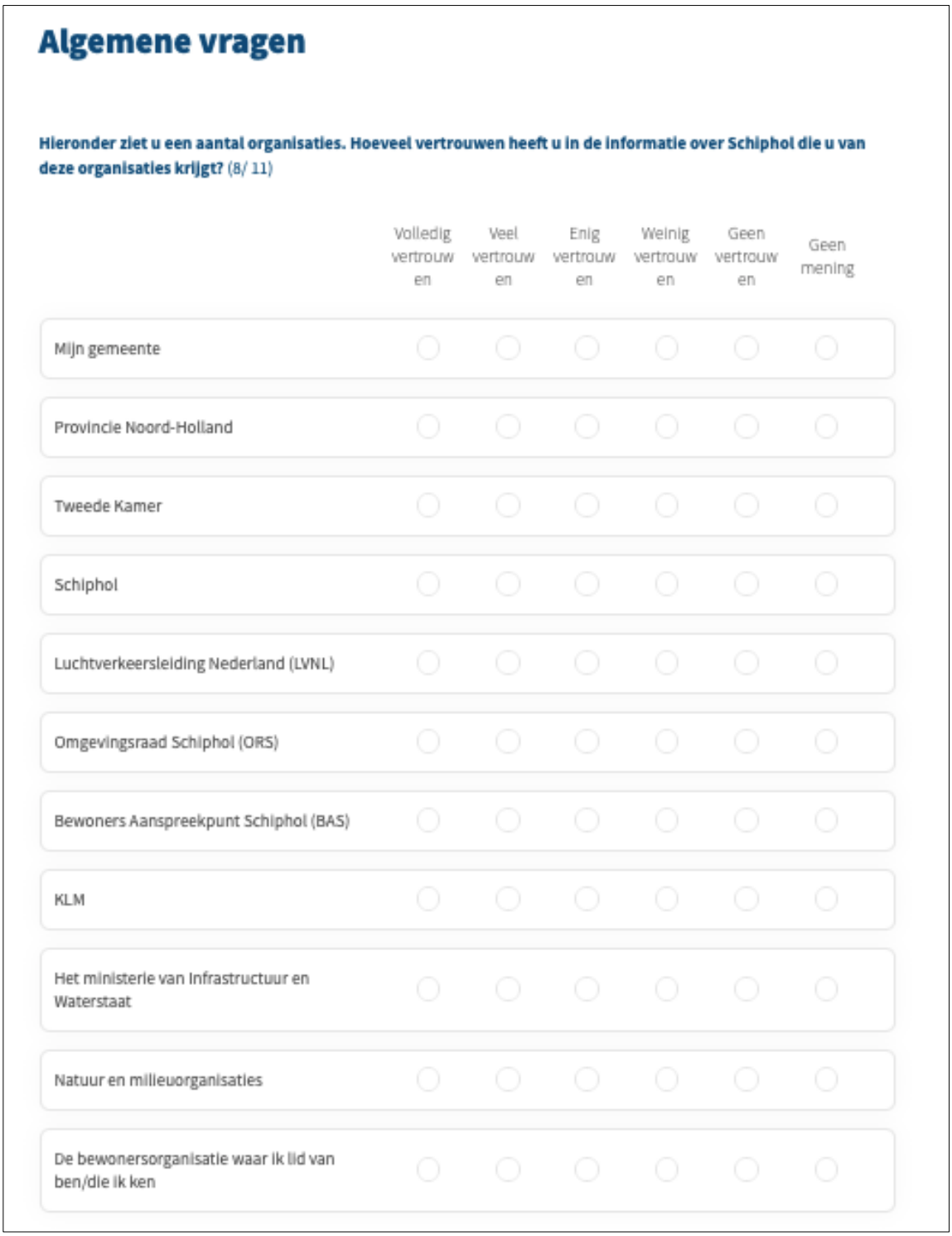
- ○ Heel onafhankelijk
- ○ Onafhankelijk
- ○ Neutraal
- ○ Niet onafhankelijk
- ○ Helemaal niet onafhankelijk
- ○ Geen mening

Kunt u uitleggen waarom u dit antwoord heeft gegeven? (6/ 11)

Waar let u op, om te beoordelen of onderzoeksinformatie betrouwbaar is? (7/ 11)

- ☐ Zijn de onderzoeksbureaus onafhankelijk?
- ☐ Heeft het onderzoeksbureau als doel winst te maken of niet?
- ☐ Wie heeft de opdracht gegeven voor het onderzoek?
- ☐ Mag iedereen de onderzoeksrapporten lezen?
- ☐ Mag iedereen de gegevens zien die de onderzoekers hebben verzameld?
- ☐ Is er een second opinion uitgevoerd en openbaar beschikbaar?
- ☐ Mochten bewoners meedenken of advies geven over het onderzoek?
- ☐ Anders, namelijk

Figure D.18: In-depth questions about the Environmental House – part 2



Algemene vragen

Hieronder ziet u een aantal organisaties. Hoeveel vertrouwen heeft u in de informatie over Schiphol die u van deze organisaties krijgt? (8/ 11)

	Volledig vertrouwen	Veel vertrouwen	Enig vertrouwen	Weinig vertrouwen	Geen vertrouwen	Geen mening
Mijn gemeente	○	○	○	○	○	○
Provincie Noord-Holland	○	○	○	○	○	○
Tweede Kamer	○	○	○	○	○	○
Schiphol	○	○	○	○	○	○
Luchtverkeersleiding Nederland (LVNL)	○	○	○	○	○	○
Omgevingsraad Schiphol (ORS)	○	○	○	○	○	○
Bewoners Aanspreekpunt Schiphol (BAS)	○	○	○	○	○	○
KLM	○	○	○	○	○	○
Het ministerie van Infrastructuur en Waterstaat	○	○	○	○	○	○
Natuur en milieuorganisaties	○	○	○	○	○	○
De bewonersorganisatie waar ik lid van ben/die ik ken	○	○	○	○	○	○

Figure D.19: In-depth questions about the Environmental House – part 3

Over welke onderwerpen zou u graag gemakkelijk meer informatie willen kunnen vinden? (9/ 11)

Wat vindt u voorbeelden van goede informatie over de effecten van Schiphol? (10/ 11)

Hoe kunnen de overheid en Schiphol zorgen dat burgers betere informatie krijgen over Schiphol? (11/ 11)

Figure D.20: In-depth questions about the Environmental House – part 4

Last general questions

Laatste algemene vragen

U gaat nu naar het laatste onderdeel van deze raadpleging. In het kader van wetenschappelijk onderzoek willen we graag weten hoe verschillende typen bewoners aankijken tegen het betrekken van burgers en informatievoorziening en daarom stellen we een aantal vragen over uzelf.

Hoe heeft u te maken met Schiphol? (1/ 13)

- [ ] Ik werk voor Schiphol
- [ ] Ik ben klant of reiziger van Schiphol
- [ ] De organisatie waar ik werk, werkt nauw samen met Schiphol
- [ ] Ik woon in de buurt van Schiphol
- [ ] Ik ben lid van een bewonersorganisatie
- [ ] Ik heb overlast van Schiphol
- [ ] Ik besteed mijn vrije tijd in de buurt van Schiphol
- [ ] Geen van deze antwoorden past bij mij

Figure D.21: Last general questions – part 1

Stellingen (2/ 13)

	Nooit	1x	2x	3x	4x	5x - 7x	8x - 10x	Meer dan 10x
Hoe vaak vloog u in 2021 vanaf Schiphol voor een zakelijke reis?	○	○	○	○	○	○	○	○
Hoe vaak vloog u in 2021 vanaf Schiphol voor een privéreis zoals vakantie of bezoek van familie/vrienden?	○	○	○	○	○	○	○	○
Hoe vaak vloog u in 2019 (voor de coronacrisis) vanaf Schiphol voor een zakelijke reis?	○	○	○	○	○	○	○	○
Hoe vaak vloog u in 2019 (voor de coronacrisis) vanaf Schiphol voor een privéreis zoals vakantie of bezoek van familie/vrienden?	○	○	○	○	○	○	○	○

Woont u bij een vliegroute van Schiphol? (3/ 13)

- ○ Ja
- ○ Nee
- ○ Weet ik niet

Stellingen* (4/ 13)

	Helemaal mee eens	Mee eens	Neutraal	Mee oneens	Helemaal mee oneens	Geen mening
Ik vind dit onderzoek een goede manier om mijn mening te kunnen geven over de Maatschappelijke Raad Schiphol en het Omgevingshuis.	○	○	○	○	○	○
Ik vond de vragen die aan mij werden gesteld in dit onderzoek begrijpelijk.	○	○	○	○	○	○
Ik vond dat ik al mijn meningen kon geven over hoe burgers betrokken moeten worden bij besluitvorming over Schiphol en hoe informatie moet worden gegeven.	○	○	○	○	○	○

Figure D.22: Last general questions – part 2

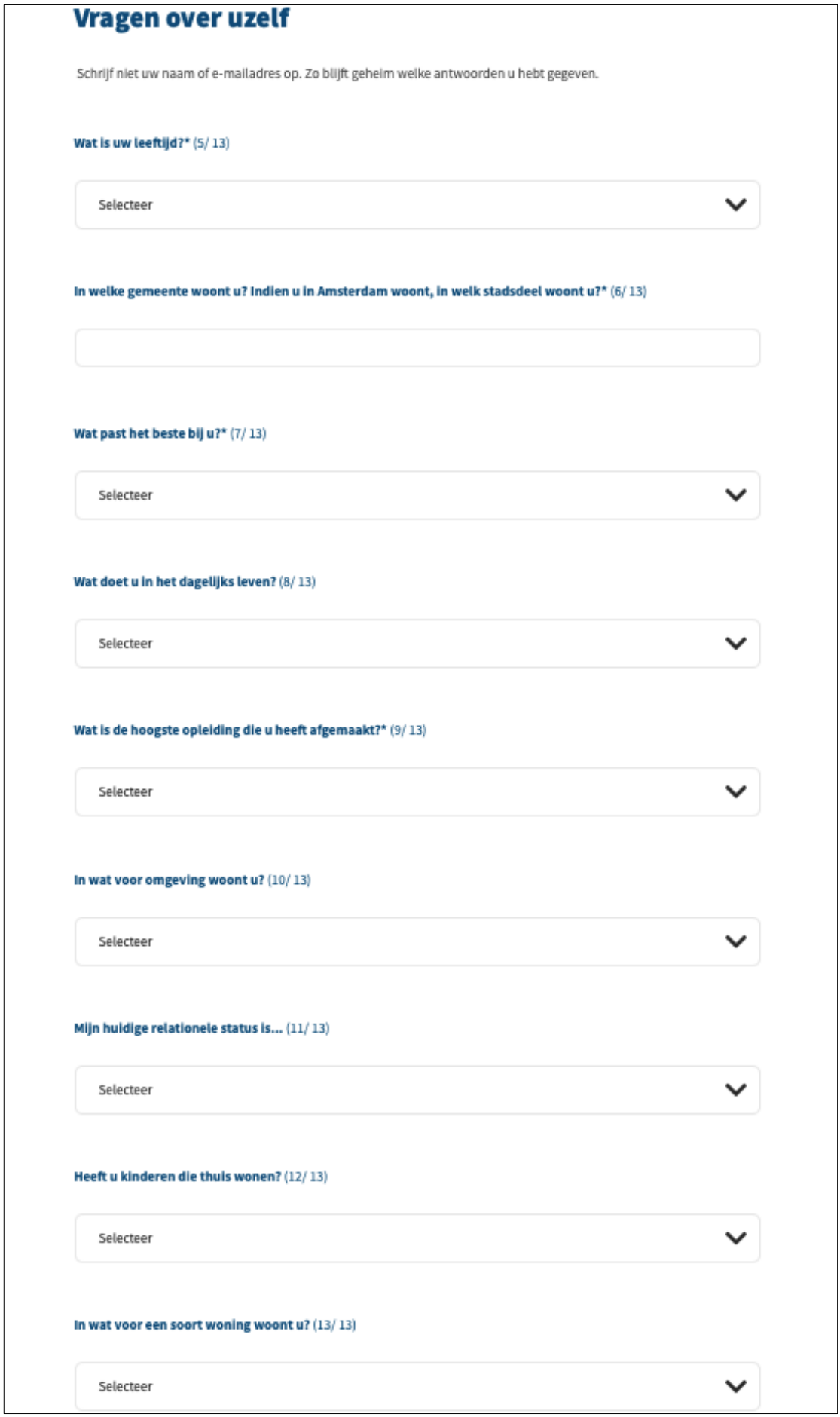


Figure D.23: Last general questions – part 3

Appendix E Tests of parallel lines

Tests of parallel line are performed per experiment for all five face validity categories included in this research. The results are presented in table E.1. The proportional odds assumption in SPSS is commonly referred to as the test of parallel lines (UCLA, 2021).

Table E.1: Result tests of parallel lines

	Step 1		Step 2	
Dependent variable	Chi-square	p-value	Chi-square	p-value
'Sliders' experiment				
Clarity	74,968	0,002	113,628	0,010
Unambiguity	94,999	0,000	144,840	0,000
Relevance	75,122	0,001	164,366	0,000
Readability	73,810	0,002	92,500	0,180
Completeness	85,091	0,000	114,800	0,008
'Points' experiment				
Clarity	80,342	0,000	126,925	0,001
Unambiguity	53,886	0,103	135,487	0,000
Relevance	60,973	0,044	87,075	0,302
Readability	56,505	0,067	126,345	0,001
Completeness	74,311	0,002	102,357	0,045

Appendix F Characteristics of the sample

Table F.1 presents the general demographic characteristics and their categorization. Moreover, table F.1 present the percentages per category in the 'sliders' experiment, the 'points' experiment and the total sample.

Table F.1: Overview of general demographic characteristics

Operationalised characteristics		Experiment 1	Experiment 2	Total sample
Age	18-34 years	31,8%	32,3%	32,1%
	35-64 years	49,8%	49,1%	49,5%
	65 years and older	18,3%	18,6%	18,4%
Gender	Man	53,3%	50,2%	51,8%
	Woman	46,7%	49,8%	48,2%
Educational level	High	46,7%	43,3%	45,1%
	Medium	35,7%	39,9%	37,7%
	Low	17,5%	16,8%	17,2%
Working life	Fulltime	51,1%	48,1%	49,7%
	Parttime	18,6%	21,1%	19,8%
	Not working	30,3%	30,8%	30,5%
Living environment	Out of city	30,9%	33,0%	31,9%
	Outside city centre	30,4%	30,1%	30,3%
	Inside city centre	38,7%	36,9%	37,8%
Relational status	Married	41,8%	43,0%	42,3%
	Living together	24,1%	23,0%	23,6%
	Single or not cohabiting, widow/widower or divorced	34,2%	34,0%	34,1%
Children living at home	Children living at home	38,4%	39,2%	38,7%
	No children living at home	61,6%	60,8%	61,3%
Type of house	Owner-occupied house	60,4%	61,9%	61,1%
	Rental house	39,6%	38,1%	38,9%
Municipality	'Inner' area 58 dB(A)	27,8%	23,4%	22,4%
	'Outside' area 48 dB(A)	72,2%	76,6%	77,6%

With regard to the municipality characteristic in the table above, the 54 municipalities around Schiphol are divided in two groups. A first group of municipalities is experiencing 58 dB(A) noise from Schiphol and is also referred to as the inner area. The second group of municipalities is experiencing 48 dB(A) noise from Schiphol and is also referred to as the outside area. The municipalities that have even only a small part in the inner area, are grouped under the inner area. This leads to the following division of the inner and outside area (Omgevingsraad Schiphol, 2015):

- Inside area: Uithoorn, Amstelveen, Aalsmeer, Kaag en Braassem, Haarlemmermeer, Uitgeest, Zaanstad & Velsen.
- Outside area: Alkmaar, Almere, Alphen aan den Rijn, Amsterdam, Bergen, Beverwijk, Blaricum, Bloemendaal, Bodegraven-Reeuwijk, Castricum, de Ronde Venen, Diemen, Dijk en Waard, Edam-Volendam, Eemnes, Gooise Meren, Haarlem, Heemskerk, Heemstede, Heiloo, Hillegom, Hilversum, Huizen, Katwijk, Landsmeer, Laren, Leiden, Leiderdorp, Lelystad, Lisse, Nieuwkoop, Noordwijk, Oegstgeest, Oostzaan, Ouder-Amstel, Purmerend, Stichtse Vecht, Teylingen, Voorschoten, Waterland, Wijdemeren, Woerden, Wormerland, Zaanstad, Zandvoort, Zoeterwoude.

Table F.2 presents the case-specific characteristics and their categorization.

Table F.2: Overview of case-specific characteristics

Operationalised characteristics		Experiment 1	Experiment 2	Total sample
Living near a flight path	Yes	72,2%	74,4%	73,2%
	No	27,8%	25,6%	26,8%
Working for Schiphol	Yes	4,7%	5,3%	5,0%
	No	95,4%	94,7%	95,0%
Schiphol customer or traveler	Yes	53,1%	51,0%	52,1%
	No	46,9%	49,0%	47,9%
The organization where I work, works closely with Schiphol	Yes	7,0%	6,9%	6,9%
	No	93,0%	93,1%	93,1%
Member of a citizen organization	Yes	6,4%	5,3%	5,9%
	No	93,6%	94,7%	94,1%
Inconvenienced by Schiphol	Yes	18,3%	19,1%	18,7%
	No	81,7%	80,9%	81,3%
Spending free time near Schiphol	Yes	7,1%	8,8%	7,9%
	No	92,9%	91,2%	92,1%
Satisfied with the way in which decisions about Schiphol can be influenced	(Totally) unsatisfied	18,6%	20,8%	19,7%
	Neutral	48,9%	48,6%	48,8%
	(Totally) satisfied	32,5%	30,6%	31,6%
Reliability of the information that can be found about the effects of Schiphol on the environment	(Totally) unreliable	9,5%	10,1%	9,8%
	Neutral	41,3%	49,0%	44,9%
	(Totally) reliable	49,2%	40,9%	45,9%
Independence of the information that can be found about the effects of Schiphol on the environment	(Totally) dependent	21,4%	24,2%	22,8%
	Neutral	56,4%	56,7%	56,5%
	(Totally) independent	22,2%	19,1%	20,7%

Appendix G Tests of normality (for section 4.3)

Table G.1 presents of the results of the tests of normality of the five face validity categories included in this research per experiment. It is observed that none of the variables has a p-value greater then 0,000. It follows that for each variable the null hypothesis must be rejected. The null hypothesis indicates that the Likert scale scores of each category statement are normally distributed. Therefore, Mann-Whitney U tests are performed since this test does not assume a normal distribution of the data in contrast with the independent samples t-test.

Table G.1: Results of tests of normality of five face validity categories per experiment

		Kolmogorov-Smirnov			Shapiro-Wilk		
	Experiment	statistic	df	P-value	statistic	df	P-value
Clarity	1	0,259	648	0,000	0,856	648	0,000
	2	0,270	582	0,000	0,855	582	0,000
Unambiguity	1	0,283	648	0,000	0,63	648	0,000
	2	0,278	582	0,000	0,853	582	0,000
Relevance	1	0,279	648	0,000	0,847	648	0,000
	2	0,287	582	0,000	0,838	582	0,000
Readability	1	0,289	648	0,000	0,841	648	0,000
	2	0,299	582	0,000	0,829	582	0,000
Completeness	1	0,272	648	0,000	0,847	648	0,000
	2	0,282	582	0,000	0,836	582	0,000

Appendix H Coding of multiple regression analysis

In table H.1 the dummy coding applied in the multiple regression analysis is presented. In the dummy coding, the attribute level that has always the value zero is the reference category.

Table H.1: Overview of dummy coding applied in the multiple regression analyses

Variable	Categories	Coding	
General demographics			
Age	18-34 years	1	0
	35-64 years	0	1
	65 years and older	0	0
Gender	Man	1	
	Woman	0	
Educational level	High	1	0
	Medium	0	1
	Low	0	0
Working life	Fulltime	1	0
	Parttime	0	1
	Not working	0	0
Living environment	Out of city	1	0
	Outside city centre	0	1
	Inside city centre	0	0
Relational status	Married	1	0
	Living together	0	1
	Single or not cohabiting, widow(er) or divorced	0	0
Children living at home	Yes	1	
	No	0	
Type of house	Owner-occupied house	1	
	Rental house	0	
Municipality	'Inner' area 58 dB(A)	1	
	'Outside' area 48 dB(A)	0	
Case-specific characteristics			
Living near a flight path	Yes	1	
	No	0	
Working for Schiphol	Yes	1	
	No	0	
Schiphol customer (or traveller)	Yes	1	
	No	0	
The organization where I work, works closely with Schiphol	Yes	1	
	No	0	
Member of citizen organization	Yes	1	
	No	0	
Inconvenienced by Schiphol	Yes	1	
	No	0	
Spending free time near Schiphol	Yes	1	

	No	0	
Satisfied with the way in which decisions about Schiphol can be influenced (satisfied with influence)	(Totally) unsatisfied	1	0
	Neutral	0	1
	(Totally) satisfied	0	0
Reliability of the information that can be found about the effects of Schiphol on the environment (reliability of information)	(Totally) unreliable	1	0
	Neutral	0	1
	(Totally) reliable	0	0
Independence of the information that can be found about the effects of Schiphol on the environment (Independence of information)	(Totally) dependent	1	0
	Neutral	0	1
	(Totally) independent	0	0

Appendix I Complete results of the multiple regression analyses

This appendix contains of six tables showing the complete results of the multiple regression analyses. The variables with a p-value less than 0,050 are shown in blue. Because hierarchical multiple regression analyses are applied, the result tables consist of a step one and a step two. These steps have been drawn up in the methodology (chapter 2). First of all, the result tables of the 'sliders' experiment are shown with first the results of the latent variable which deals with the face validity of the Schiphol Social Council 'sliders' choice task. Thereafter follow the results of the multiple regression analysis of the latent variable which deals with the face validity of the PVE consultation in general. Finally, the results of the latent variable which deals with a more comprehensive view on the face validity of the PVE method. After the tables of the 'sliders' experiment, the results of the 'points' experiment are presented, which follow the same structure.

Table I.1: Complete results of multiple regression of face validity of Schiphol Social Council 'sliders' choice task ('sliders' experiment)

Independent variables	Step 1		Step 2	
	Coefficient	P-value	Coefficient	P-value
General demographics				
Dummy – 18-34 years	0,022	0,745	-0,009	0,899
Dummy – 35-65 years	0,024	0,719	0,022	0,746
Gender – man	-0,067	0,114	-0,076	0,080
Dummy – high educated	-0,047	0,418	-0,039	0,509
Dummy – medium educated	0,009	0,875	0,012	0,835
Dummy – fulltime	-0,066	0,251	-0,073	0,218
Dummy – parttime	-0,067	0,185	-0,070	0,171
Dummy – out of city	-0,041	0,363	-0,039	0,395
Dummy – out of city centre	-0,035	0,439	-0,033	0,475
Dummy – married	0,084	0,103	0,077	0,138
Dummy – living together	0,036	0,451	0,036	0,463
Children living at home – yes	-0,050	0,265	-0,065	0,158
Type of house – owner-occupied house	0,013	0,756	0,010	0,828
Municipality – 'inner' area 58 dB(A)	-0,013	0,743	-0,008	0,856
Case-specific characteristics				
Living near a flight path – yes			-0,019	0,655
Working for Schiphol – yes			-0,029	0,493
Schiphol customer/traveller – yes			-0,010	0,820
The organization where I work, works closely with Schiphol – yes			-0,006	0,886
Member of a citizen organization – yes			0,088	0,039
Inconvenienced by Schiphol – yes			-0,058	0,197
Spending free time near Schiphol – yes			0,011	0,790
Dummy – (totally) unsatisfied with influence			-0,043	0,403
Dummy – neutral satisfied with influence			-0,008	0,866
Dummy – (totally) unreliable information			0,000	0,993
Dummy – neutral reliability of information			-0,032	0,500
Dummy – (totally) dependent information			0,010	0,857
Dummy – neutral independence of information			0,003	0,956
Model information				
Constant	3,707	0,000	3,813	0,000
Partial F test	0,844	0,620	0,769	0,794
Adjusted R square	-0,003		-0,010	

Table I.2: Complete results of multiple regression of face validity of PVE consultation in general ('sliders' experiment)

Independent variables	Step 1		Step 2	
	Coefficient	P-value	Coefficient	P-value
General demographics				
Dummy – 18-34 years	-0,093	0,171	-0,112	0,112
Dummy – 35-65 years	0,004	0,948	-0,005	0,945
Gender – man	-0,053	0,211	-0,051	0,234
Dummy – high educated	-0,115	0,048	-0,099	0,094
Dummy – medium educated	-0,037	0,505	-0,035	0,539
Dummy – fulltime	0,052	0,367	0,059	0,315
Dummy – parttime	-0,006	0,909	-0,004	0,943
Dummy – out of city	-0,006	0,902	0,007	0,879
Dummy – out of city centre	0,013	0,778	0,018	0,698
Dummy – married	-0,006	0,900	-0,011	0,836
Dummy – living together	-0,007	0,886	-0,002	0,970
Children living at home – yes	-0,039	0,375	-0,040	0,381
Type of house – owner-occupied house	0,073	0,091	0,057	0,201
Municipality – 'inner' area 58 dB(A)	-0,016	0,691	-0,008	0,845
Case-specific characteristics				
Living near a flight path – yes			-0,023	0,588
Working for Schiphol – yes			-0,062	0,141
Schiphol customer/traveller – yes			0,003	0,941
The organization where I work, works closely with Schiphol – yes			-0,016	0,706
Member of a citizen organization – yes			0,006	0,885
Inconvenienced by Schiphol – yes			-0,028	0,532
Spending free time near Schiphol – yes			-0,039	0,327
Dummy – (totally) unsatisfied with influence			-0,035	0,493
Dummy – neutral satisfied with influence			-0,089	0,075
Dummy – (totally) unreliable information			-0,018	0,715
Dummy – neutral reliability of information			-0,009	0,856
Dummy – (totally) dependent information			0,010	0,866
Dummy – neutral independence of information			0,076	0,181
Model information				
Constant	3,947	0,000	4,016	0,000
Partial F test	0,025	0,313	0,963	0,520
Adjusted R square	0,003		-0,002	

Table I.3: Complete results of multiple regression of face validity of PVE method ('sliders' experiment)

Independent variables	Step 1		Step 2	
	Coefficient	P-value	Coefficient	P-value
General demographics				
Dummy – 18-34 years	-0,053	0,437	-0,081	0,252
Dummy – 35-65 years	0,015	0,826	0,008	0,912
Gender – man	-0,069	0,105	-0,072	0,097
Dummy – high educated	-0,101	0,081	-0,087	0,144
Dummy – medium educated	-0,021	0,705	-0,018	0,751
Dummy – fulltime	0,003	0,954	0,005	0,931
Dummy – parttime	-0,036	0,470	-0,036	0,476
Dummy – out of city	-0,024	0,599	-0,014	0,757
Dummy – out of city centre	-0,008	0,854	-0,004	0,935
Dummy – married	0,036	0,480	0,030	0,563
Dummy – living together	0,013	0,789	0,016	0,741
Children living at home – yes	-0,051	0,252	-0,059	0,200
Type of house – owner-occupied house	0,056	0,193	0,044	0,329
Municipality – 'inner' area 58 dB(A)	-0,017	0,667	-0,009	0,825
Case-specific characteristics				
Living near a flight path – yes			-0,025	0,558
Working for Schiphol – yes			-0,057	0,182
Schiphol customer/traveller – yes			-0,003	0,952
The organization where I work, works closely with Schiphol – yes			-0,014	0,744
Member of a citizen organization – yes			0,047	0,272
Inconvenienced by Schiphol – yes			-0,047	0,293
Spending free time near Schiphol – yes			-0,022	0,590
Dummy – (totally) unsatisfied with influence			-0,045	0,383
Dummy – neutral satisfied with influence			-0,065	0,195
Dummy – (totally) unreliable information			-0,012	0,800
Dummy – neutral reliability of information			-0,021	0,652
Dummy – (totally) dependent information			0,012	0,839
Dummy – neutral independence of information			0,053	0,348
Model information				
Constant	3,851	0,000	3,395	0,000
Partial F test	1,010	0,441	0,832	0,711
Adjusted R square	0,000		-0,007	

Table I.4: Complete results of multiple regression of face validity of Schiphol Social Council 'points' choice task ('points' experiment)

Independent variables	Step 1		Step 2	
	Coefficient	P-value	Coefficient	P-value
General demographics				
Dummy – 18-34 years	-0,011	0,876	-0,108	0,115
Dummy – 35-65 years	-0,036	0,610	-0,028	0,677
Gender – man	0,060	0,174	-0,018	0,673
Dummy – high educated	0,100	0,108	0,103	0,087
Dummy – medium educated	0,046	0,447	0,044	0,441
Dummy – fulltime	0,056	0,294	0,047	0,362
Dummy – parttime	0,070	0,255	0,042	0,473
Dummy – out of city	0,011	0,824	0,029	0,533
Dummy – out of city centre	0,030	0,529	0,061	0,174
Dummy – married	-0,022	0,686	-0,024	0,639
Dummy – living together	-0,094	0,057	-0,094	0,048
Children living at home – yes	0,030	0,535	-0,035	0,457
Type of house – owner-occupied house	0,068	0,125	0,087	0,039
Municipality – 'inner' area 58 dB(A)	-0,095	0,026	-0,075	0,068
Case-specific characteristics				
Living near a flight path – yes			0,054	0,191
Working for Schiphol – yes			-0,012	0,782
Schiphol customer/traveller – yes			0,036	0,390
The organization where I work, works closely with Schiphol – yes			0,042	0,320
Member of a citizen organization – yes			0,023	0,587
Inconvenienced by Schiphol – yes			-0,027	0,539
Spending free time near Schiphol – yes			0,073	0,079
Dummy – (totally) unsatisfied with influence			-0,129	0,019
Dummy – neutral satisfied with influence			-0,197	0,000
Dummy – (totally) unreliable information			-0,072	0,137
Dummy – neutral reliability of information			-0,150	0,002
Dummy – (totally) dependent information			-0,133	0,032
Dummy – neutral independence of information			-0,205	0,000
Model information				
Constant	3,433	0,000	3,943	0,000
Partial F test	1,523	0,098	3,886	0,000
Adjusted R square	0,012		0,118	

Table I.5: Complete results of multiple regression of face validity of PVE consultation in general ('points' experiment)

Independent variables	Step 1		Step 2	
	Coefficient	P-value	Coefficient	P-value
General demographics				
Dummy – 18-34 years	-0,173	0,013	-0,216	0,002
Dummy – 35-65 years	-0,065	0,347	-0,059	0,386
Gender – man	0,023	0,595	-0,032	0,464
Dummy – high educated	0,174	0,005	0,146	0,017
Dummy – medium educated	0,057	0,337	0,044	0,450
Dummy – fulltime	-0,028	0,595	-0,025	0,637
Dummy – parttime	0,013	0,830	0,014	0,814
Dummy – out of city	0,016	0,743	0,025	0,587
Dummy – out of city centre	0,017	0,717	0,035	0,446
Dummy – married	-0,069	0,191	-0,064	0,220
Dummy – living together	-0,097	0,047	-0,084	0,084
Children living at home – yes	0,045	0,344	0,007	0,880
Type of house – owner-occupied house	0,055	0,211	0,056	0,198
Municipality – 'inner' area 58 dB(A)	-0,032	0,452	-0,034	0,426
Case-specific characteristics				
Living near a flight path – yes			0,060	0,158
Working for Schiphol – yes			-0,096	0,028
Schiphol customer/traveller – yes			0,056	0,191
The organization where I work, works closely with Schiphol – yes			-0,005	0,911
Member of a citizen organization – yes			0,002	0,957
Inconvenienced by Schiphol – yes			0,077	0,090
Spending free time near Schiphol – yes			0,054	0,204
Dummy – (totally) unsatisfied with influence			-0,026	0,642
Dummy – neutral satisfied with influence			-0,139	0,009
Dummy – (totally) unreliable information			-0,149	0,003
Dummy – neutral reliability of information			-0,115	0,017
Dummy – (totally) dependent information			-0,052	0,413
Dummy – neutral independence of information			-0,098	0,101
Model information				
Constant	3,868	0,000	4,116	0,000
Partial F test	2,077	0,012	2,748	0,000
Adjusted R square	0,025		0,075	

Table I.6: Complete results of multiple regression of face validity of PVE method ('points' experiment)

Independent variables	Step 1		Step 2	
	Coefficient	P-value	Coefficient	P-value
General demographics				
Dummy – 18-34 years	-0,125	0,037	-0,202	0,003
Dummy – 35-65 years	-0,063	0,368	-0,055	0,415
Gender – man	0,045	0,304	-0,031	0,471
Dummy – high educated	0,169	0,006	0,151	0,012
Dummy – medium educated	0,062	0,301	0,052	0,365
Dummy – fulltime	0,008	0,887	0,005	0,917
Dummy – parttime	0,043	0,483	0,030	0,609
Dummy – out of city	0,016	0,738	0,031	0,494
Dummy – out of city centre	0,026	0,577	0,053	0,233
Dummy – married	-0,059	0,271	-0,056	0,274
Dummy – living together	-0,113	0,022	-0,104	0,030
Children living at home – yes	0,046	0,339	-0,012	0,802
Type of house – owner-occupied house	0,071	0,108	0,081	0,057
Municipality – 'inner' area 58 dB(A)	-0,068	0,108	-0,060	0,150
Case-specific characteristics				
Living near a flight path – yes			0,068	0,104
Working for Schiphol – yes			-0,072	0,092
Schiphol customer/traveller – yes			0,056	0,181
The organization where I work, works closely with Schiphol – yes			0,017	0,689
Member of a citizen organization – yes			0,013	0,765
Inconvenienced by Schiphol – yes			0,040	0,365
Spending free time near Schiphol – yes			0,073	0,082
Dummy – (totally) unsatisfied with influence			-0,081	0,146
Dummy – neutral satisfied with influence			-0,192	0,000
Dummy – (totally) unreliable information			-0,138	0,005
Dummy – neutral reliability of information			-0,152	0,001
Dummy – (totally) dependent information			-0,100	0,107
Dummy – neutral independence of information			-0,167	0,005
Model information				
Constant	3,694	0,000	4,047	0,000
Partial F test	2,054	0,013	3,763	0,000
Adjusted R square	0,025		0,114	

Appendix J Coding of multinomial logistic regression

In table J.1 the coding applied in the multinomial logistic regression is presented. For each variable it is indicated which attribute level form the reference category.

Table J.1: Overview of coding of variables applied in the multinomial logistic regression

Variable	Categories	Coding
General demographics		
Age	18-34 years	1
	35-64 years	2
	65 years and older	3 (reference category)
Gender	Man	0
	Woman	1 (reference category)
Educational level	High	1
	Medium	2
	Low	3 (reference category)
Working life	Fulltime	1
	Parttime	2
	Not working	3 (reference category)
Living environment	Out of city	1
	Outside city centre	2
	Inside city centre	3 (reference category)
Relational status	Married	1
	Living together	2
	Single or not cohabiting, widow(er) or divorced	3 (reference category)
Children living at home	Yes	0
	No	1 (reference category
Type of house	Owner-occupied house	0
	Rental house	1 (reference category)
Municipality	'Inner' area 58 dB(A)	0
	'Outside' area 48 dB(A)	1 (reference category)
Case specific characteristics		
Living near a flight path	Yes	0
	No	1 (reference category)
Working for Schiphol	Yes	0
	No	1 (reference category)
Schiphol customer (or traveller)	Yes	0
	No	1 (reference category)
The organization where I work, works closely with Schiphol	Yes	0
	No	1 (reference category)
Member of citizen organization	Yes	0
	No	1 (reference category)
Inconvenienced by Schiphol	Yes	0
	No	1 (reference category)
Spending free time near Schiphol	Yes	0

	No	1 (reference category)
Satisfied with the way in which decisions about Schiphol can be influenced (satisfied with influence)	(Totally) unsatisfied	1
	Neutral	2
	(Totally) satisfied	3 (reference category)
Reliability of the information that can be found about the effects of Schiphol on the environment (reliability of information)	(Totally) unreliable	1
	Neutral	2
	(Totally) reliable	3 (reference category)
Independence of the information that can be found about the effects of Schiphol on the environment (Independence of information)	(Totally) dependent	1
	Neutral	2
	(Totally) independent	3 (reference category)
Dependent variable		
Clarity/unambiguity/relevance /readability/completeness	(Totally) disagree	1
	Neutral	2
	(Totally) agree	3 (reference category)

Appendix K Complete results of the multinomial logistic regressions

This appendix contains of ten tables showing the complete results of the multinomial logistic regressions. The attributes with a p-value less than 0,050 are shown in blue. Because hierarchical multinomial logistic regressions are applied, the result tables consist of a step one and a step two. These steps have been drawn up in the methodology (chapter 2). In the tables below, the standardized coefficient is indicated with a C and the significance with an S. First of all, the complete result tables of the five categories of the 'sliders' experiment are shown. The complete results of the 'points' experiment are presented after the tables of the 'sliders' experiment.

Table K.1: Complete results of the multinomial logistic regression of clarity for the 'sliders' experiment (where 'C' stands for coefficient and 'S' stands for significance)

	(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree		(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree	
	Step 1		Step 1		Step 1		Step 2		Step 2		Step 2	
Attributes	C	S	C	S	C	S	C	S	C	S	C	S
General demographics												
Age												
18-34 years	-0,433	0,583	0,072	0,817	-0,506	0,514	-0,342	0,674	0,199	0,537	-0,541	0,497
35-64 years	-0,445	0,555	0,042	0,884	-0,487	0,509	-0,436	0,566	0,039	0,894	-0,475	0,525
65 plus	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Gender	-0,494	0,189	0,077	0,670	-0,571	0,119	-0,519	0,179	0,110	0,555	-0,628	0,094
Educational level												
High	-0,132	0,814	0,175	0,474	-0,308	0,578	-0,122	0,833	0,140	0,577	-0,262	0,645
Medium	-0,498	0,377	-0,235	0,349	-0,264	0,632	-0,456	0,428	-0,250	0,326	-0,206	0,714
Low	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Working life												
Fulltime	-1,250	0,039	-0,242	0,323	-1,008	0,090	-1,348	0,029	-0,203	0,419	-1,144	0,059
Parttime	-0,831	0,220	-0,216	0,432	-0,615	0,354	-0,927	0,177	-0,179	0,521	-0,748	0,267
Not working	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Living environment												
Out of city	-0,874	0,047	-0,038	0,855	-0,835	0,051	-0,987	0,032	-0,053	0,806	-0,934	0,036
Outside city centre	-0,364	0,431	0,153	0,460	-0,516	0,255	-0,484	0,310	0,138	0,512	-0,622	0,184
Inside city centre	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Relational status												
Married	0,465	0,313	-0,162	0,470	0,627	0,161	0,412	0,400	-0,167	0,465	0,579	0,224
Living together	0,849	0,088	0,159	0,503	0,690	0,156	0,837	0,103	0,164	0,500	0,673	0,180
Single, not cohabiting, widow(er) or divorced	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Children living at home	-0,045	0,910	0,418	0,033	-0,463	0,230	-0,032	0,939	0,483	0,017	-0,515	0,203
Type of house	0,173	0,649	0,004	0,983	0,169	0,648	0,198	0,629	-0,007	0,973	0,204	0,608
Municipality	-0,824	0,033	-0,370	0,087	-0,454	0,215	-0,912	0,028	-0,416	0,065	-0,496	0,208
Case specific characteristics												
Living near flight path							0,191	0,650	-0,045	0,828	0,236	0,564
Working for Schiphol							0,214	0,817	-0,440	0,322	0,654	0,459
Schiphol customer							0,274	0,468	0,013	0,944	0,261	0,476
Working closely with Schiphol							0,472	0,510	-0,072	0,844	0,544	0,431
Member of citizen organization							18,034	0,000	-0,046	0,899	18,080	0,000

Inconvenienced by Schiphol							0,215	0,696	0,110	0,652	0,105	0,846
Spending free time near Schiphol							0,855	0,291	0,177	0,593	0,678	0,394
Satisfied with influence												
(Totally) dissatisfied							0,574	0,305	0,618	0,029	-0,044	0,939
Neutral							1,015	0,023	0,306	0,162	0,708	0,101
(Totally) satisfied							REF	REF	REF	REF	REF	REF
Reliability of information												
(Totally) unreliable							-0,176	0,802	0,027	0,939	-0,203	0,765
Neutral							-0,393	0,375	0,081	0,695	-0,474	0,273
(Totally) reliable							REF	REF	REF	REF	REF	REF
Independence of information												
(Totally) dependent							-0,062	0,912	-0,211	0,494	0,148	0,785
Neutral							0,379	0,435	-0,077	0,755	0,456	0,333
(Totally) independent							REF	REF		REF	REF	REF
Model information												
Nr. Observations	648		648		648		648		648		648	
Likelihood ratio test	0,049		0,049		0,049		0,141		0,141		0,141	
Pseudo R^2	0,037		0,037		0,037		0,058		0,058		0,058	

Table K.2: Complete results of the multinomial logistic regression of unambiguity for the 'sliders' experiment (where 'C' stands for coefficient and 'S' stands for significance)

	(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree		(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree	
	Step 1		Step 1		Step 1		Step 2		Step 2		Step 2	
Attributes	C	S	C	S	C	S	C	S	C	S	C	S
General demographics												
Age												
18-34 years	1,506	0,010	0,726	0,031	0,780	0,152	1,579	0,009	0,956	0,006	0,623	0,267
35-64 years	0,751	0,155	0,382	0,221	0,369	0,447	0,715	0,177	0,434	0,172	0,280	0,567
65 plus	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Gender	-0,053	0,868	0,402	0,039	-0,455	0,117	-0,056	0,864	0,456	0,024	-0,512	0,089
Educational level												
High	-0,333	0,463	-0,053	0,840	-0,279	0,510	-0,393	0,399	-0,165	0,546	-0,227	0,600
Medium	-0,367	0,424	-0,068	0,798	-0,299	0,485	-0,381	0,416	-0,111	0,685	-0,270	0,535
Low	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Working life												
Fulltime	-1,476	0,004	-0,136	0,600	-1,340	0,005	-1,448	0,005	-0,160	0,553	-1,288	0,008
Parttime	-1,460	0,007	-0,074	0,802	-1,386	0,006	-1,409	0,009	-0,055	0,855	-1,354	0,007
Not working	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Living environment												
Out of city	-0,300	0,406	-0,044	0,847	-0,257	0,439	-0,391	0,296	-0,109	0,642	-0,282	0,411
Outside city centre	-0,217	0,549	0,046	0,837	-0,263	0,433	-0,327	0,381	-0,045	0,845	-0,282	0,414
Inside city centre	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Relational status												
Married	0,116	0,762	-0,072	0,763	0,188	0,592	0,029	0,940	-0,088	0,718	0,118	0,744
Living together	0,172	0,679	-0,060	0,814	0,231	0,549	0,197	0,647	-0,069	0,791	0,267	0,504
Single, not cohabiting, widow(er) or divorced	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Children living at home	-0,483	0,149	-0,413	0,052	-0,069	0,821	-0,407	0,239	-0,320	0,147	-0,086	0,785
Type of house	0,452	0,162	0,259	0,205	0,193	0,515	0,473	0,169	0,301	0,157	0,172	0,585
Municipality	-1,133	0,001	-0,541	0,028	-0,592	0,042	-1,311	0,000	-0,667	0,010	-0,643	0,037
Case specific characteristics												

Living near flight path							0,165	0,657	0,288	0,204	-0,123	0,719
Working for Schiphol							-0,952	0,161	0,021	0,966	-0,973	0,093
Schiphol customer							-0,156	0,621	0,214	0,277	-0,370	0,203
Working closely with Schiphol							0,647	0,278	0,631	0,098	0,016	0,977
Member of citizen organization							0,743	0,376	-0,380	0,358	1,123	0,156
Inconvenienced by Schiphol							-0,184	0,648	0,219	0,401	-0,403	0,282
Spending free time near Schiphol							1,081	0,105	0,521	0,128	0,560	0,385
Satisfied with influence												
(Totally) dissatisfied							0,709	0,157	0,671	0,027	0,038	0,935
Neutral							0,617	0,106	0,484	0,042	0,134	0,702
(Totally) satisfied							REF	REF	REF	REF	REF	REF
Reliability of information												
(Totally) unreliable							0,105	0,865	-0,096	0,799	0,201	0,726
Neutral							-0,040	0,911	0,134	0,548	-0,174	0,602
(Totally) reliable							REF	REF	REF	REF	REF	REF
Independence of information												
(Totally) dependent							0,023	0,964	0,276	0,401	-0,253	0,596
Neutral							-0,015	0,972	0,141	0,603	-0,156	0,691
(Totally) independent							REF	REF	REF	REF	REF	REF
Model information												
Nr. Observations	648		648		648		648		648		648	
Likelihood ratio test	0,012		0,012		0,012		0,022		0,022		0,022	
Pseudo R^2	0,041		0,041		0,041		0,066		0,066		0,066	

Table K.3: Complete results of the multinomial logistic regression of relevance for the 'sliders' experiment (where 'C' stands for coefficient and 'S' stands for significance)

	(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree		(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree	
	Step 1		Step 1		Step 1		Step 2		Step 2		Step 2	
Attributes	C	S	C	S	C	S	C	S	C	S	C	S
General demographics												
Age												
18-34 years	0,403	0,586	0,947	0,008	-0,544	0,436	0,380	0,626	0,973	0,009	-0,592	0,421
35-64 years	0,329	0,629	0,581	0,082	-0,253	0,692	0,434	0,535	0,616	0,072	-0,182	0,781
65 plus	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Gender	0,138	0,745	0,636	0,002	-0,498	0,215	0,090	0,839	0,625	0,003	-0,535	0,204
Educational level												
High	-0,363	0,569	0,537	0,059	-0,900	0,137	-0,276	0,675	0,452	0,121	-0,728	0,242
Medium	-0,230	0,727	0,110	0,705	-0,340	0,585	-0,176	0,796	0,040	0,892	-0,216	0,737
Low	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Working life												
Fulltime	0,644	0,259	-0,394	0,151	1,038	0,055	0,597	0,312	-0,471	0,099	1,068	0,055
Parttime	0,218	0,719	-0,038	0,900	0,256	0,654	0,158	0,805	-0,093	0,766	0,251	0,677
Not working	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Living environment												
Out of city	-0,012	0,980	-0,275	0,246	0,263	0,554	-0,016	0,975	-0,326	0,179	0,310	0,509
Outside city centre	0,482	0,324	-0,044	0,847	0,525	0,260	0,500	0,320	-0,088	0,706	0,588	0,221
Inside city centre	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Relational status												
Married	-0,936	0,099	-0,158	0,519	-0,778	0,153	-0,928	0,117	-0,148	0,557	-0,780	0,170

Living together	-0,935	0,121	-0,296	0,264	-0,640	0,268	-0,851	0,175	-0,266	0,331	-0,586	0,330
Single, not cohabiting, widow(er) or divorced	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Children living at home	-0,968	0,033	-0,079	0,717	-0,889	0,039	-1,057	0,025	-0,059	0,794	-0,998	0,025
Type of house	0,208	0,632	-0,065	0,755	0,273	0,507	0,083	0,858	-0,103	0,636	0,186	0,670
Municipality	-0,400	0,379	-0,100	0,675	-0,300	0,478	-0,478	0,423	-0,130	0,598	-0,348	0,444
Case specific characteristics												
Living near flight path							0,596	0,222	0,390	0,101	0,206	0,652
Working for Schiphol							-0,294	0,747	-0,027	0,956	-0,268	0,758
Schiphol customer							-0,177	0,682	0,281	0,168	-0,458	0,262
Working closely with Schiphol							-0,008	0,992	0,194	0,617	-0,202	0,783
Member of citizen organization							-0,234	0,749	0,114	0,782	-0,347	0,613
Inconvenienced by Schiphol							0,276	0,633	-0,023	0,932	0,299	0,585
Spending free time near Schiphol							-0,893	0,140	0,275	0,460	-1,168	0,037
Satisfied with influence												
(Totally) dissatisfied							-1,432	0,028	-0,422	0,213	-1,010	0,095
Neutral							-0,275	0,621	0,250	0,290	-0,525	0,326
(Totally) satisfied							REF	REF	REF	REF	REF	REF
Reliability of information												
(Totally) unreliable							-0,485	0,542	0,547	0,138	-1,033	0,180
Neutral							-0,555	0,272	-0,021	0,929	-0,534	0,263
(Totally) reliable							REF	REF	REF	REF	REF	REF
Independence of information												
(Totally) dependent							0,923	0,227	0,198	0,546	0,726	0,323
Neutral							0,216	0,711	-0,126	0,642	0,342	0,537
(Totally) independent							REF	REF	REF	REF	REF	REF
Model information												
Nr. Observations	648		648		648		648		648		648	
Likelihood ratio test	0,064		0,064		0,064		0,114		0,114		0,114	
Pseudo R^2	0,042		0,042		0,042		0,069		0,069		0,069	

Table K.4: Complete results of the multinomial logistic regression of readability for the 'sliders' experiment (where 'C' stands for coefficient and 'S' stands for significance)

	(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree		(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree	
	Step 1		Step 1		Step 1		Step 2		Step 2		Step 2	
Attributes	C	S	C	S	C	S	C	S	C	S	C	S
General demographics												
Age												
18-34 years	0,553	0,465	0,642	0,085	-0,088	0,900	0,299	0,706	0,712	0,065	-0,413	0,574
35-64 years	0,999	0,180	0,360	0,300	0,639	0,357	0,888	0,244	0,390	0,271	0,499	0,482
65 plus	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Gender	0,818	0,060	0,513	0,015	0,305	0,455	0,764	0,091	0,517	0,017	0,248	0,558
Educational level												
High	0,568	0,392	0,924	0,004	-0,356	0,558	0,575	0,402	0,852	0,009	-0,277	0,661
Medium	0,283	0,672	0,545	0,094	-0,262	0,670	0,243	0,725	0,522	0,114	-0,279	0,660
Low	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Working life												
Fulltime	-0,812	0,198	-0,241	0,402	-0,571	0,336	-0,785	0,234	-0,286	0,333	-0,499	0,421
Parttime	-0,241	0,736	-0,036	0,909	-0,205	0,762	-0,197	0,790	-0,034	0,917	-0,163	0,816

Not working	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Living environment												
Out of city	-0,426	0,401	0,006	0,981	-0,432	0,366	-0,354	0,503	-0,035	0,886	-0,319	0,522
Outside city centre	-0,556	0,284	-0,160	0,508	-0,395	0,418	-0,584	0,273	-0,236	0,343	-0,348	0,487
Inside city centre	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Relational status												
Married	0,268	0,625	-0,252	0,331	0,520	0,312	0,145	0,796	-0,298	0,260	0,443	0,398
Living together	0,226	0,680	-0,095	0,726	0,321	0,531	0,302	0,597	-0,102	0,715	0,403	0,450
Single, not cohabiting, widow(er) or divorced	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Children living at home	-0,109	0,816	0,044	0,843	-0,154	0,727	-0,047	0,924	0,091	0,695	-0,138	0,766
Type of house	-0,452	0,321	-0,291	0,175	-0,161	0,708	-0,511	0,295	-0,252	0,258	-0,260	0,573
Municipality	0,479	0,343	0,414	0,072	0,065	0,892	0,521	0,331	0,354	0,141	0,166	0,743
Case specific characteristics												
Living near flight path							-0,383	0,480	0,144	0,559	-0,526	0,301
Working for Schiphol							-1,207	0,166	-0,022	0,965	-1,185	0,134
Schiphol customer							0,142	0,746	0,214	0,313	-0,072	0,860
Working closely with Schiphol							-0,393	0,610	-0,428	0,278	-0,821	0,259
Member of citizen organization							0,939	0,409	0,185	0,648	0,755	0,496
Inconvenienced by Schiphol							0,980	0,159	0,310	0,253	0,670	0,318
Spending free time near Schiphol							18,547	0,000	0,692	0,042	17,855	0,000
Satisfied with influence												
(Totally) dissatisfied							-0,558	0,431	0,187	0,570	-0,745	0,262
Neutral							-0,379	0,509	0,223	0,379	-0,601	0,267
(Totally) satisfied							REF	REF	REF	REF	REF	REF
Reliability of information												
(Totally) unreliable							0,169	0,856	-0,177	0,664	0,346	0,692
Neutral							-0,234	0,640	0,132	0,584	-0,365	0,434
(Totally) reliable							REF	REF	REF	REF	REF	REF
Independence of information												
(Totally) dependent							-0,330	0,648	0,188	0,590	-0,518	0,444
Neutral							-0,039	0,950	0,063	0,827	-0,102	0,863
(Totally) independent							REF	REF	REF	REF	REF	REF
Model information												
Nr. Observations	648		648		648		648		648		648	
Likelihood ratio test	0,262		0,262		0,262		0,354		0,354		0,354	
Pseudo R^2	0,035		0,035		0,035		0,063		0,063		0,063	

Table K.5: Complete results of the multinomial logistic regression of completeness for the 'sliders' experiment (where 'C' stands for coefficient and 'S' stands for significance)

	(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree		(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree	
	Step 1		Step 1		Step 1		Step 2		Step 2		Step 2	
Attributes	C	S	C	S	C	S	C	S	C	S	C	S
General demographics												
Age												
18-34 years	0,642	0,385	0,642	0,385	0,074	0,916	1,057	0,195	0,722	0,044	0,334	0,666
35-64 years	1,169	0,098	1,169	0,098	0,747	0,264	1,300	0,076	0,441	0,176	0,859	0,217
65 plus	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Gender	0,559	0,206	0,559	0,206	0,214	0,612	0,740	0,108	0,379	0,060	0,360	0,411

Educational level												
High	-0,099	0,889	-0,099	0,889	-0,767	0,259	-0,046	0,951	0,633	0,029	-0,679	0,335
Medium	-0,196	0,784	-0,196	0,784	-0,661	0,332	-0,187	0,799	0,427	0,140	-0,613	0,382
Low	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Working life												
Fulltime	-0,445	0,480	-0,445	0,480	-0,044	0,942	-0,401	0,547	-0,346	0,201	-0,055	0,932
Parttime	-0,626	0,353	-0,626	0,353	-0,171	0,790	-0,835	0,237	-0,421	0,170	-0,414	0,539
Not working	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Living environment												
Out of city	0,717	0,167	0,717	0,167	0,735	0,137	0,857	0,133	-0,017	0,940	0,875	0,090
Outside city centre	0,603	0,239	0,603	0,239	0,571	0,243	0,816	0,134	0,047	0,837	0,768	0,140
Inside city centre	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Relational status												
Married	-0,254	0,633	-0,254	0,633	-0,135	0,789	-0,248	0,655	-0,072	0,773	-0,176	0,738
Living together	0,444	0,476	0,444	0,476	0,358	0,552	0,328	0,612	0,083	0,751	0,246	0,693
Single, not cohabiting, widow(er) or divorced	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Children living at home	-0,567	0,234	-0,567	0,234	-0,421	0,355	-0,269	0,585	-0,083	0,706	-0,187	0,691
Type of house	-0,231	0,613	-0,231	0,613	0,251	0,566	-0,370	0,440	-0,487	0,019	0,117	0,798
Municipality	-1,280	0,004	-1,280	0,004	-0,957	0,019	-1,476	0,002	-0,390	0,117	-1,086	0,015
Case specific characteristics												
Living near flight path							0,125	0,807	-0,020	0,929	0,145	0,767
Working for Schiphol							-0,007	0,994	0,278	0,538	-0,285	0,739
Schiphol customer							0,431	0,352	-0,067	0,734	0,498	0,261
Working closely with Schiphol							-1,366	0,052	-0,442	0,302	-0,924	0,143
Member of citizen organization							-0,109	0,894	0,055	0,892	-0,164	0,831
Inconvenienced by Schiphol							0,745	0,260	0,073	0,782	0,671	0,291
Spending free time near Schiphol							-0,105	0,886	0,517	0,138	-0,621	0,373
Satisfied with influence												
(Totally) dissatisfied							-0,605	0,357	0,101	0,746	-0,706	0,253
Neutral							0,408	0,466	0,361	0,126	0,047	0,930
(Totally) satisfied							REF	REF	REF	REF	REF	REF
Reliability of information												
(Totally) unreliable							-0,313	0,690	0,286	0,435	-0,599	0,426
Neutral							-0,046	0,933	0,142	0,527	-0,187	0,718
(Totally) reliable							REF	REF	REF	REF	REF	REF
Independence of information												
(Totally) dependent							1,003	0,165	0,276	0,396	0,727	0,291
Neutral							0,985	0,099	0,047	0,864	0,938	0,097
(Totally) independent							REF	REF	REF	REF	REF	REF
Model information												
Nr. Observations	648		648		648		648		648		648	
Likelihood ratio test	0,109		0,109		0,109		0,275		0,275		0,275	
Pseudo R^2	0,039		0,039		0,039		0,062		0,062		0,062	

Table K.6: Complete results of the multinomial logistic regression of clarity for the 'points' experiment (where 'C' stands for coefficient and 'S' stands for significance)

Attributes	(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree		(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree	
	Step 1		Step 1		Step 1		Step 2		Step 2		Step 2	
	C	S	C	S	C	S	C	S	C	S	C	S
General demographics												
Age												
18-34 years	-1,168	0,116	-0,053	0,871	-1,115	0,127	-1,452	0,040	0,268	0,450	-1,720	0,023
35-64 years	-0,825	0,253	0,275	0,356	-1,099	0,124	-1,094	0,138	0,225	0,482	-1,319	0,070
65 plus	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Gender	0,212	0,572	-0,165	0,391	0,377	0,301	0,239	0,551	0,133	0,525	0,106	0,785
Educational level												
High	-0,090	0,868	-0,254	0,341	0,164	0,756	0,144	0,797	-0,190	0,511	0,334	0,544
Medium	-0,225	0,659	-0,234	0,371	0,009	0,985	-0,084	0,873	-0,204	0,468	0,120	0,817
Low	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Working life												
Fulltime	0,169	0,731	-0,201	0,448	0,369	0,442	0,159	0,754	-0,135	0,633	0,294	0,554
Parttime	-0,153	0,773	-0,448	0,118	0,295	0,566	-0,242	0,660	-0,468	0,128	0,226	0,670
Not working	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Living environment												
Out of city	0,117	0,779	-0,035	0,873	0,152	0,707	0,002	0,996	-0,089	0,705	0,092	0,831
Outside city centre	0,334	0,460	-0,193	0,385	0,526	0,230	0,288	0,544	-0,356	0,132	0,645	0,164
Inside city centre	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Relational status												
Married	0,402	0,407	0,140	0,545	0,261	0,581	0,400	0,434	0,136	0,578	0,264	0,600
Living together	-0,465	0,305	0,002	0,994	-0,467	0,284	-0,534	0,270	-0,023	0,933	-0,510	0,276
Single, not cohabiting, widow(er) or divorced	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Children living at home	-0,282	0,485	-0,388	0,071	0,106	0,787	-0,333	0,450	-0,182	0,432	-0,151	0,724
Type of house	0,192	0,608	-0,200	0,310	0,392	0,280	0,314	0,431	-0,260	0,216	0,575	0,142
Municipality	-0,175	0,680	0,252	0,241	-0,427	0,302	0,000	1,000	0,225	0,329	-0,225	0,617
Case specific characteristic												
Living near flight path							0,692	0,105	-0,105	0,648	0,797	0,057
Working for Schiphol							0,159	0,866	0,261	0,601	-0,102	0,911
Schiphol customer							0,247	0,525	-0,082	0,683	0,329	0,383
Working closely with Schiphol							1,067	0,338	-0,119	0,780	1,186	0,276
Member of citizen organization							-1,033	0,238	-0,675	0,251	-0,358	0,636
Inconvenienced by Schiphol							-1,155	0,018	-0,356	0,203	-0,799	0,088
Spending free time near Schiphol							-0,220	0,763	-0,711	0,068	0,491	0,472
Satisfied with influence												
(Totally) dissatisfied							-0,482	0,444	0,480	0,145	-0,962	0,114
Neutral							0,037	0,942	0,747	0,003	-0,710	0,154
(Totally) satisfied							REF	REF	REF	REF	REF	REF
Reliability of information												
(Totally) unreliable							-0,107	0,874	0,607	0,121	-0,714	0,269
Neutral							0,394	0,375	0,877	0,000	-0,482	0,263
(Totally) reliable							REF	REF	REF	REF	REF	REF
Independence of information												

(Totally) dependent							-0,323	0,672	0,239	0,514	-0,562	0,442
Neutral							-0,265	0,690	0,497	0,102	-0,762	0,231
(Totally) independent							REF	REF	REF	REF	REF	REF
Model information												
Nr. Observations	528		528		528		528		528		528	
Likelihood ratio test	0,416		0,416		0,416		0,000		0,000		0,000	
Pseudo R^2	0,029		0,029		0,029		0,103		0,103		0,103	

Table K.7: Complete results of the multinomial logistic regression of unambiguity for the 'points' experiment (where 'C' stands for coefficient and 'S' stands for significance)

	(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree		(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree	
	Step 1		Step 1		Step 1		Step 2		Step 2		Step 2	
Attributes	C	S	C	S	C	S	C	S	C	S	C	S
General demographics												
Age												
18-34 years	-0,391	0,538	-0,050	0,880	-0,342	0,581	-0,343	0,609	0,240	0,500	-0,583	0,371
35-64 years	-0,083	0,892	-0,207	0,498	0,124	0,835	-0,219	0,730	-0,345	0,291	0,126	0,837
65 plus	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Gender	-0,149	0,685	-0,199	0,310	0,050	0,888	-0,141	0,722	0,018	0,932	-0,159	0,675
Educational level												
High	0,163	0,757	-0,666	0,015	0,829	0,106	0,247	0,659	-0,656	0,025	0,903	0,039
Medium	-0,111	0,813	-0,257	0,329	0,146	0,751	0,099	0,843	-0,200	0,475	0,299	0,540
Low	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Working life												
Fulltime	0,218	0,667	0,064	0,812	0,154	0,755	0,368	0,492	0,168	0,555	0,200	0,700
Parttime	-0,332	0,528	-0,225	0,438	-0,107	0,833	-0,113	0,839	-0,134	0,666	0,021	0,969
Not working	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Living environment												
Out of city	0,620	0,155	0,221	0,323	0,399	0,346	0,495	0,287	0,135	0,573	0,359	0,422
Outside city centre	0,266	0,532	-0,015	0,948	0,281	0,492	0,288	0,522	-0,101	0,677	0,389	0,368
Inside city centre	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Relational status												
Married	-0,404	0,393	-0,253	0,286	-0,150	0,743	-0,370	0,460	-0,213	0,399	-0,157	0,746
Living together	-0,587	0,213	0,249	0,330	-0,836	0,069	-0,602	0,241	0,306	0,264	-0,907	0,070
Single, not cohabiting, widow(er) or divorced	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Children living at home	-0,562	0,165	-0,126	0,563	-0,436	0,265	-0,455	0,307	0,058	0,807	-0,513	0,231
Type of house	0,762	0,043	-0,005	0,979	0,767	0,035	0,775	0,056	-0,086	0,690	0,861	0,028
Municipality	-0,248	0,550	0,390	0,076	-0,638	0,114	-0,204	0,648	0,402	0,087	-0,606	0,163
Case specific characteristics												
Living near flight path							0,750	0,078	-0,040	0,866	0,790	0,043
Working for Schiphol							-1,659	0,032	-0,269	0,627	-1,390	0,044
Schiphol customer							-0,219	0,577	-0,249	0,224	0,030	0,937
Working closely with Schiphol							-1,293	0,084	-1,073	0,030	-0,220	0,738
Member of citizen organization							1,518	0,133	0,671	0,168	0,848	0,380
Inconvenienced by Schiphol							-0,351	0,492	-0,115	0,682	-0,236	0,630
Spending free time near Schiphol							-0,181	0,791	-0,274	0,469	0,093	0,885
Satisfied with influence												
(Totally) dissatisfied							-0,526	0,410	-0,239	0,477	-0,287	0,639

Neutral							-0,054	0,915	0,316	0,211	-0,370	0,454
(Totally) satisfied							REF	REF	REF	REF	REF	REF
Reliability of information												
(Totally) unreliable							-0,608	0,372	0,325	0,419	-0,933	0,148
Neutral							-0,322	0,497	0,432	0,042	-0,753	0,040
(Totally) reliable							REF	REF	REF	REF	REF	REF
Independence of information												
(Totally) dependent							0,196	0,772	1,240	0,002	-1,044	0,089
Neutral							1,367	0,026	1,567	0,000	-0,200	0,719
(Totally) independent							REF	REF	REF	REF		REF
Model information												
Nr. Observations	582		582		582		582		582		582	
Likelihood ratio test	0,100		0,100		0,100		0,000		0,000		0,000	
Pseudo R^2	0,038		0,038		0,038		0,114		0,114		0,114	

Table K.8: Complete results of the multinomial logistic regression of relevance for the 'points' experiment (where 'C' stands for coefficient and 'S' stands for significance)

	(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree		(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree	
	Step 1		Step 1		Step 1		Step 2		Step 2		Step 2	
Attributes	C	S	C	S	C	S	C	S	C	S	C	S
General demographics												
Age												
18-34 years	-0,357	0,722	0,099	0,781	-0,456	0,639	0,013	0,990	0,310	0,409	-0,298	0,769
35-64 years	-0,591	0,527	-0,038	0,911	-0,554	0,541	-0,676	0,486	-0,046	0,896	-0,630	0,501
65 plus	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Gender	0,347	0,538	-0,572	0,007	0,920	0,040	0,566	0,351	-0,410	0,048	0,976	0,044
Educational level												
High	0,763	0,295	-0,419	0,158	1,182	0,042	0,848	0,284	-0,337	0,275	1,184	0,039
Medium	0,213	0,733	-0,141	0,620	0,354	0,553	0,342	0,622	-0,064	0,826	0,406	0,540
Low	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Working life												
Fulltime	0,231	0,749	0,184	0,527	0,047	0,946	0,441	0,573	0,259	0,392	0,182	0,809
Parttime	-0,264	0,710	0,014	0,965	-0,278	0,683	-0,401	0,603	0,017	0,959	-0,418	0,570
Not working	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Living environment												
Out of city	-0,588	0,314	-0,126	0,608	-0,463	0,409	-0,779	0,210	-0,161	0,527	-0,618	0,298
Outside city centre	-0,159	0,808	-0,144	0,557	-0,015	0,981	0,111	0,879	-0,207	0,412	0,318	0,654
Inside city centre	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Relational status												
Married	-0,732	0,300	-0,159	0,539	-0,573	0,402	-0,819	0,283	-0,157	0,554	-0,662	0,371
Living together	-0,830	0,247	0,122	0,655	-0,952	0,171	-0,747	0,338	0,084	0,767	-0,832	0,272
Single, not cohabiting, widow(er) or divorced	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Children living at home	-0,406	0,247	-0,317	0,184	-0,089	0,865	-0,263	0,664	-0,238	0,347	-0,025	0,966
Type of house	-0,516	0,459	-0,212	0,324	-0,303	0,569	-0,690	0,264	-0,226	0,309	-0,463	0,439
Municipality	0,083	0,350	0,242	0,314	-0,159	0,776	-0,166	0,800	0,222	0,374	-0,387	0,540
Case specific characteristics												
Living near flight path							1,251	0,040	-0,145	0,548	1,395	0,017
Working for Schiphol							-1,572	0,079	-0,278	0,629	-1,294	0,098
Schiphol customer							-0,107	0,851	-0,190	0,386	0,082	0,881
Working closely with Schiphol							-0,777	0,452	-0,807	0,123	0,030	0,974

Member of citizen organization							-0,268	0,765	0,568	0,266	-0,836	0,311
Inconvenienced by Schiphol							0,459	0,601	-0,161	0,597	0,620	0,466
Spending free time near Schiphol							-1,058	0,259	-0,479	0,282	-0,578	0,504
Satisfied with influence												
(Totally) dissatisfied							-0,146	0,879	-0,101	0,779	-0,045	0,961
Neutral							0,068	0,922	0,275	0,307	-0,208	0,756
(Totally) satisfied							REF	REF	REF	REF	REF	REF
Reliability of information												
(Totally) unreliable							-1,224	0,169	0,474	0,277	-1,698	0,042
Neutral							0,659	0,339	0,650	0,009	0,008	0,990
(Totally) reliable							REF	REF	REF	REF	REF	REF
Independence of information												
(Totally) dependent							0,710	0,437	0,136	0,737	0,574	0,506
Neutral							1,295	0,047	0,252	0,451	1,043	0,152
(Totally) independent							REF	REF	REF	REF	REF	REF
Model information												
Nr. Observations	582		582		582		582		582		582	
Likelihood ratio test	0,196		0,196		0,196		0,022		0,022		0,022	
Pseudo R^2	0,043		0,043		0,043		0,096		0,096		0,096	

Table K.9: Complete results of the multinomial logistic regression of readability for the 'points' experiment (where 'C' stands for coefficient and 'S' stands for significance)

	(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree		(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree	
	Step 1		Step 1		Step 1		Step 2		Step 2		Step 2	
Attributes	C	S	C	S	C	S	C	S	C	S	C	S
General demographics												
Age												
18-34 years	-1,811	0,119	0,750	0,047	-2,561	0,023	-2,076	0,094	0,942	0,027	-3,019	0,012
35-64 years	-1,000	0,392	0,128	0,734	-1,128	0,319	-1,153	0,345	0,010	0,980	-1,163	0,325
65 plus	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Gender	-0,090	0,841	0,151	0,500	-0,241	0,572	-0,260	0,610	0,355	0,142	-0,614	0,205
Educational level												
High	1,021	0,138	-0,692	0,023	1,712	0,010	1,301	0,085	-0,578	0,070	1,879	0,010
Medium	-0,317	0,561	-0,688	0,021	0,371	0,470	-0,137	0,819	-0,630	0,043	0,493	0,382
Low	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Working life												
Fulltime	0,124	0,840	0,196	0,529	-0,072	0,902	0,167	0,808	0,246	0,455	-0,079	0,903
Parttime	-0,587	0,373	-0,113	0,746	-0,474	0,438	-0,839	0,241	-0,093	0,804	-0,746	0,261
Not working	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Living environment												
Out of city	-0,456	0,389	-0,380	0,151	-0,076	0,879	-0,482	0,387	-0,473	0,091	-0,008	0,987
Outside city centre	-0,098	0,855	-0,086	0,735	-0,012	0,981	0,126	0,831	-0,131	0,621	0,257	0,648
Inside city centre	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Relational status												
Married	-0,180	0,773	0,025	0,930	-0,205	0,728	-0,275	0,686	0,019	0,948	-0,294	0,648
Living together	-0,416	0,455	0,217	0,455	-0,633	0,228	-0,427	0,490	0,183	0,549	-0,610	0,298
Single, not cohabiting, widow(er) or divorced	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Children living at home	-0,095	0,850	-0,032	0,898	-0,062	0,895	0,091	0,874	0,146	0,590	-0,056	0,918
Type of house	0,149	0,743	-0239	0,297	0,388	0,362	0,186	0,704	-0,279	0,243	0,465	0,315

Municipality	0,147	0,814	-0,094	0,728	0,241	0,682	0,301	0,659	-0,094	0,740	0,395	0,542
Case specific characteristics												
Living near flight path							0,078	0,883	-0,201	0,429	0,280	0,581
Working for Schiphol							-0,153	0,852	1,061	0,039	-1,213	0,113
Schiphol customer							0,375	0,474	-0,202	0,393	0,577	0,246
Working closely with Schiphol							-2,214	0,006	-1,098	0,046	-1,117	0,080
Member of citizen organization							0,269	0,790	-0,624	0,320	0,893	0,316
Inconvenienced by Schiphol							0,050	0,948	-0,055	0,872	0,105	0,885
Spending free time near Schiphol							0,373	0,748	-0,453	0,347	0,826	0,448
Satisfied with influence												
(Totally) dissatisfied							-1,287	0,160	-0,601	0,136	-0,686	0,429
Neutral							-1,291	0,049	0,153	0,588	-1,444	0,022
(Totally) satisfied							REF	REF	REF	REF	REF	REF
Reliability of information												
(Totally) unreliable							-1,345	0,105	0,611	0,208	-1,956	0,010
Neutral							0,565	0,326	0,454	0,038	0,112	0,839
(Totally) reliable							REF	REF	REF	REF	REF	REF
Independence of information												
(Totally) dependent							0,681	0,438	0,371	0,401	0,310	0,703
Neutral							1,116	0,097	0,812	0,022	0,304	0,624
(Totally) independent							REF	REF	REF	REF	REF	REF
Model information												
Nr. Observations	582		582		582		582		582		582	
Likelihood ratio test	0,003		0,003		0,003		0,000		0,000		0,000	
Pseudo R^2	0,067		0,067		0,067		0,135		0,135		0,135	

Table K.10: Complete results of the multinomial logistic regression of completeness for the 'points' experiment (where 'C' stands for coefficient and 'S' stands for significance)

	(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree		(Totally) disagree vs. neutral		(Totally) agree vs. neutral		(Totally) disagree vs. (totally) agree	
	Step 1		Step 1		Step 1		Step 2		Step 2		Step 2	
Attributes	C	S	C	S	C	S	C	S	C	S	C	S
General demographics												
Age												
18-34 years	-16,133	0,000	0,054	0,878	-16,187	0,000	-17,220	0,000	0,211	0,580	-17,432	0,000
35-64 years	-15,249	0,000	-0,031	0,924	-15,217	0,000	-16,223	0,000	-0,064	0,856	-16,159	0,000
65 plus	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Gender	0,263	0,625	0,353	0,096	-0,090	0,861	0,103	0,868	0,581	0,011	-0,477	0,430
Educational level												
High	0,727	0,337	-0,299	0,314	1,026	0,159	1,220	0,152	-0,149	0,632	1,369	0,046
Medium	0,217	0,748	-0,092	0,746	0,309	0,634	0,307	0,677	0,045	0,879	0,262	0,712
Low	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Working life												
Fulltime	-0,324	0,709	0,042	0,885	-0,366	0,665	-0,121	0,898	0,053	0,863	-0,174	0,850
Parttime	-1,197	0,171	0,089	0,779	-1,286	0,129	-1,608	0,088	0,050	0,880	-1,659	0,069
Not working	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Living environment												
Out of city	0,092	0,879	-0,127	0,603	0,219	0,706	0,106	0,872	-0,190	0,456	0,296	0,639

Outside city centre	0,051	0,935	-0,180	0,465	0,231	0,701	0,273	0,699	-0,303	0,237	0,576	0,399
Inside city centre	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Relational status												
Married	-0,537	0,462	0,013	0,960	-0,550	0,436	-0,836	0,318	-0,022	0,936	-0,814	0,317
Living together	0,045	0,948	0,421	0,124	-0,375	0,577	-0,277	0,731	0,329	0,256	-0,606	0,438
Single, not cohabiting, widow(er) or divorced	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF	REF
Children living at home	-0,234	0,687	-0,085	0,720	-0,149	0,789	-0,046	0,945	0,013	0,960	-0,059	0,927
Type of house	-0,772	0,158	-0,630	0,003	-0,142	0,788	-0,822	0,180	-0,631	0,005	-0,192	0,747
Municipality	1,738	0,103	0,251	0,292	1,486	0,158	2,217	0,044	0,291	0,245	1,926	0,102
Case specific characteristics												
Living near flight path							-0,057	0,932	-0,285	0,236	0,228	0,730
Working for Schiphol							-0,881	0,295	0,707	0,153	-1,587	0,046
Schiphol customer							1,241	0,030	-0,276	0,213	1,517	0,023
Working closely with Schiphol							-0,787	0,380	-0,301	0,512	-0,486	0,560
Member of citizen organization							1,063	0,304	0,396	0,427	0,667	0,499
Inconvenienced by Schiphol							0,079	0,931	-0,380	0,225	0,459	0,603
Spending free time near Schiphol							-1,096	0,277	-0,793	0,099	-0,303	0,742
Satisfied with influence												
(Totally) dissatisfied							-0,857	0,359	0,153	0,682	-1,010	0,259
Neutral							0,677	0,379	0,589	0,031	0,087	0,907
(Totally) satisfied							REF	REF	REF	REF	REF	REF
Reliability of information												
(Totally) unreliable							-1,129	0,234	0,287	0,528	-1,417	0,111
Neutral							0,863	0,236	0,664	0,008	0,200	0,778
(Totally) reliable							REF	REF	REF	REF	REF	REF
Independence of information												
(Totally) dependent							-0,846	0,416	-0,133	0,735	-0,713	0,476
Neutral							-0,686	0,419	0,015	0,963	-0,701	0,392
(Totally) independent							REF	REF	REF	REF	REF	REF
Model information												
Nr. Observations	582		582		582		582		582		582	
Likelihood ratio test	0,076		0,076		0,076		0,000		0,000		0,000	
Pseudo R^2	0,049		0,049		0,049		0,119		0,119		0,119	

Appendix L Complete results latent class cluster analyses

This appendix contains of two tables showing the complete results of the LCCA's of both experiments. Of the clusters, the profile measures are displayed. Furthermore, the Wald statistics are presented in the result tables with the associated p-values. For the 'sliders' experiment, the most optimal BIC-value is 658,614 which entails three clusters. For the 'points' experiment, the most optimal BIC-value is 491,903 which entails three clusters.

Table L.1: Complete results of latent class cluster analysis of the 'sliders' experiment

		Cluster 1	Cluster 2	Cluster 3	Wald	P-value
Cluster size		0,5564	0,2971	0,1465	-	-
Indicators						
Clarity	(Totally) disagree	0,0038	0,1237	0,1804	107,1037	5.5e-24
	Neutral	0,1829	0,6025	0,6205		
	(Totally) agree	0,8132	0,2738	0,1191		
Unambiguity	(Totally) disagree	0,0018	0,1571	0,4173	85,8031	2.3e-19
	Neutral	0,1023	0,5549	0,4914		
	(Totally) agree	0,8958	0,2980	0,0912		
Relevance	(Totally) disagree	0,0004	0,1712	0,0016	98,1450	4.9e-22
	Neutral	0,0718	0,6462	0,1364		
	(Totally) agree	0,9278	0,1826	0,8620		
Readability	(Totally) disagree	0,0003	0,1550	0,0030	97,5663	6.5e-22
	Neutral	0,0503	0,5869	0,1476		
	(Totally) agree	0,9494	0,2582	0,8494		
Completeness	(Totally) disagree	0,0017	0,1473	0,0025	90,7421	2.0e-20
	Neutral	0,1236	0,6064	0,1479		
	(Totally) agree	0,8746	0,2463	0,8496		
Covariates						
Age	18-34 years	0,2825	0,4102	0,2698	8,5402	0,014
	35-64 years	0,5178	0,4332	0,5627		
	65 years and older	0,1997	0,1566	0,1765		
Gender	Man	0,4923	0,5914	0,5617	7,0141	0,030
	Woman	0,5077	0,4086	0,4383		
Educational level	High	0,4424	0,5596	0,3801	8,3381	0,015
	Medium	0,3617	0,3202	0,4078		
	Low	0,1958	0,1201	0,2121		
Working life	Fulltime	0,4863	0,5150	0,5908	5,5857	0,061
	Parttime	0,1856	0,1836	0,1965		
	Not working	0,3280	0,3014	0,2126		
Living environment	Out of city	0,3128	0,2708	0,3755	2,9983	0,22
	Outside city centre	0,2978	0,2970	0,3368		
	Inside city centre	0,3894	0,4323	0,2877		
Relational status	Married	0,4351	0,4021	0,3877	0,6152	0,74
	Living together	0,2338	0,2509	0,2381		

	Single or not cohabiting, widow(er) or divorced	0,3311	0,3470	0,3743		
Children living at home	Yes	0,3846	0,3858	0,3685	0,0023	1,00
	No	0,6154	0,6142	0,6315		
Type of house	Owner-occupied house	0,6300	0,5578	0,5948	1,3523	0,51
	Rental house	0,3700	0,4422	0,4052		
Municipality	'Inner' area 58 dB(A)	0,2194	0,2107	0,2256	0,1211	0,94
	'Outside' area 48 dB(A)	0,7806	0,7893	0,7744		
Living near flight path	Yes	0,7175	0,7416	0,7073	0,7651	0,68
	No	0,2825	0,2584	0,2927		
Working for Schiphol	Yes	0,0416	0,0603	0,0380	0,3537	0,84
	No	0,9584	0,9397	0,9620		
Schiphol customer	Yes	0,5097	0,5558	0,5676	1,3267	0,52
	No	0,4903	0,4442	0,4324		
Working closely with Schiphol	Yes	0,0635	0,1021	0,0189	1,2810	0,53
	No	0,9365	0,8979	0,9811		
Member of citizen organization	Yes	0,0591	0,0945	0,0189	1,8839	0,39
	No	0,9409	0,9055	0,9811		
Inconvenienced by Schiphol	Yes	0,1755	0,1952	0,1904	0,0511	0,97
	No	0,8245	0,8048	0,8096		
Spending free time near Schiphol	Yes	0,0649	0,0996	0,0398	2,0479	0,36
	No	0,9351	0,9004	0,9602		
Satisfied with influence	(Totally) unsatisfied	0,1602	0,1898	0,2805	6,5622	0,038
	Neutral	0,4822	0,4897	0,5195		
	(Totally) satisfied	0,3575	0,3205	0,2000		
Reliability of information	(Totally) unreliable	0,0823	0,1177	0,0964	1,7671	0,41
	Neutral	0,3961	0,4195	0,4689		
	(Totally) reliable	0,5216	0,4628	0,4347		
Independence of information	(Totally) dependent	0,2048	0,2384	0,2036	0,1606	0,92
	Neutral	0,5636	0,5190	0,6499		
	(Totally) independent	0,2316	0,2426	0,1465		

Table L.2: Complete results of latent class cluster analysis of the 'points' experiment

		Cluster 1	Cluster 2	Cluster 3	Wald	P-value
Cluster size		0,3963	0,3896	0,2141	-	-
Indicators						
Clarity	(Totally) disagree	0,0007	0,1011	0,1356	27,1657	1,3e-6
	Neutral	0,0730	0,0,5369	0,5652		
	(Totally) agree	0,9263	0,3620	0,2991		
Unambiguity	(Totally) disagree	0,0001	0,0957	01548	8,1134	0,017
	Neutral	0,0247	0,5229	0,5675		
	(Totally) agree	0,9752	0,3814	0,2777		
Relevance	(Totally) disagree	0,0002	0,0092	0,1435	49,9491	1,4e-11
	Neutral	0,0395	0,2486	0,5899		
	(Totally) agree	0,9603	0,7422	0,2666		
Readability	(Totally) disagree	0,0001	0,0012	0,2222	31,4570	1,5e-7
	Neutral	0,0408	0,1258	0,6504		
	(Totally) agree	0,9591	0,8730	0,1274		
Completeness	(Totally) disagree	0,0003	0,0019	0,1485	64,7862	8,5e-15
	Neutral	0,0711	0,1671	0,6796		
	(Totally) agree	0,9286	0,8310	0,1719		
Covariates						
Age	18-34 years	0,2499	0,3404	0,4269	12,0621	0,0024
	35-64 years	0,5404	0,4680	0,4433		
	65 years and older	0,2097	0,1917	0,1297		
Gender	Man	0,5446	0,4622	0,4942	1,3774	0,50
	Woman	0,4554	0,5378	0,5058		
Educational level	High	0,5259	0,3798	0,3578	9,3326	0,0094
	Medium	0,3689	0,4190	0,4167		
	Low	0,1052	0,2012	0,2255		
Working life	Fulltime	0,5203	0,4203	0,5191	2,8241	0,24
	Parttime	0,2166	0,2045	0,2140		
	Not working	0,2630	0,3752	0,2669		
Living environment	Out of city	0,3390	0,3312	0,3105	0,2477	0,88
	Outside city centre	0,3107	0,2872	0,3066		
	Inside city centre	0,3503	0,3815	0,3828		
Relational status	Married	0,5147	0,3949	0,3349	0,4844	0,78
	Living together	0,1854	0,2313	0,3113		
	Single or not cohabiting, widow(er) or divorced	0,2998	0,3738	0,3538		
Children living at home	Yes	0,4322	0,3486	0,3953	0,0929	0,95
	No	0,5678	0,6514	0,6047		
Type of house	Owner-occupied house	0,6607	0,6298	0,5201	2,1842	0,34
	Rental house	0,3393	0,3702	0,4799		
Municipality	'Inner' area 58 dB(A)	0,2096	0,2801	0,1936	2,0853	0,35
	'Outside' area 48 dB(A)	0,7904	0,7199	0,8064		
	Yes	0,7807	0,7580	0,6503	3,0584	0,22

Living near flight path	No	0,2193	0,2420	0,3497		
Working for Schiphol	Yes	0,0580	0,0186	0,1077	3,1429	0,21
	No	0,9420	0,9814	0,8923		
Schiphol customer	Yes	0,5274	0,5538	0,3995	5,3096	0,070
	No	0,4726	0,4462	0,6005		
Working closely with Schiphol	Yes	0,1056	0,0306	0,0700	3,1644	0,21
	No	0,8944	0,9694	0,9300		
Member of citizen organization	Yes	0,0763	0,0275	0,0576	0,7101	0,70
	No	0,9237	0,9725	0,9424		
Inconvenienced by Schiphol	Yes	0,2182	0,1997	0,1235	1,5379	0,046
	No	0,7818	0,8003	0,8765		
Spending free time near Schiphol	Yes	0,1244	0,0818	0,0299	5,3564	0,069
	No	0,8756	0,9182	0,9701		
Satisfied with influence	(Totally) unsatisfied	0,2055	0,2496	0,1365	0,3893	0,82
	Neutral	0,4189	0,5143	0,5600		
	(Totally) satisfied	0,3757	0,2361	0,3035		
Reliability of information	(Totally) unreliable	0,0951	0,0998	0,1159	9,4972	0,087
	Neutral	0,3726	0,5641	0,5710		
	(Totally) reliable	0,5323	0,3361	0,3131		
Independence of information	(Totally) dependent	0,2216	0,3021	0,1716	6,4960	0,039
	Neutral	0,5025	0,5951	0,6354		
	(Totally) independent	0,2758	0,1028	0,1931		

Appendix M Tests of normality (for section 5)

Table M.1 presents the results of the tests of normality of the variables related to answering the third sub question. It is remarkable that none of the variables has a p-value greater than 0,000. It follows that for each variable the null hypothesis must be rejected. The null hypothesis indicates that the Likert scale scores of each statement are normally distributed. Therefore, Mann-Whitney U tests are performed to compare different cases. In contrary to the independent samples t-test, the Mann-Whitney U test does not assume a normal distribution of the data. Moreover, the Kolmogorov-Smirnov test and the Shapiro-Wilk test are both variants of the test of normality.

Table M.1: Results of tests of normality of face validity categories in previous PVE consultations

	Kolmogorov-Smirnov			Shapiro-Wilk		
	statistic	df	P-value	statistic	df	P-value
Clarity Schiphol (exp. 1)	0,270	582	0,000	0,855	582	0,000
Clarity climate consultation	0,290	2028	0,000	0,859	2028	0,000
Clarity heat transition vision Utrecht	0,270	321	0,000	0,880	321	0,000
Clarity corona policy (exp. 1)	0,229	2005	0,000	0,879	2005	0,000
Clarity corona policy (exp. 2)	0,276	2005	0,000	0,854	2005	0,000
Relevance Schiphol (exp. 1)	0,287	582	0,000	0,838	582	0,000
Relevance heat transition vision Utrecht	0,243	321	0,000	0,892	321	0,000
Relevance corona policy (exp. 1)	0,242	2005	0,000	0,848	2005	0,000
Relevance corona policy (exp. 2)	0,263	2005	0,000	0,846	2005	0,000
Relevance Foodvalley	0,267	1556	0,000	0,877	1556	0,000
Completeness Schiphol (exp. 1)	0,282	582	0,000	0,836	582	0,000
Completeness Foodvalley	0,305	1556	0,000	0,774	1556	0,000

Appendix N Stakeholder who is not involved in the process of designing the PVE consultation interview protocol

In this appendix, the protocol used for the interview with the stakeholder who is not involved in the process of designing the PVE consultation is presented. The draft report of the customer's end product was sent to this stakeholder. This stakeholder responded that the questioning in the consultation was somewhat steering. For example, the stakeholder would have preferred an open question and not already formulated tasks of the Schiphol Social Council or the Environmental House. This stakeholder indicates that this form of questioning has influence on the findings. This interview takes a closer look at this response steering. In this interview, in-depth questions are asked about this comment.

Expert interview protocol

Introduction

- Thank the interviewee for his or her time and participation to this interview.
- Indicate that the interviewee may stop at any time during the interview without giving a reason.
- The interview data is securely stored in the data centre of Delft University of Technology.
- Explain that this interview is part of a master thesis. The topic concerns the face validity of the PVE method.
- The purpose of this interview is to research the perspective of a stakeholder who is not involved in the process of designing the PVE consultation about the validity of this consultation.
- Does the participant have any questions?
- Ask if the audio of the interview may be recorded. This question is first asked without the audio on and repeated once the audio is on.

Questions

You have been sent the draft report of the PVE consultation on the Schiphol Social Council and the Environmental House. In response, you sent an email. The following questions further elaborate on your response.

Why do you think the questioning a bit steering?

What are the consequences of a steering consultation?

What are the advantages and disadvantages of the results of open questions about, for example, the preferences for possible tasks of the Schiphol Social Council compared to the way the questions are asked in the consultation?

What are other examples of questions that you found too directing besides formulated possible tasks of the Schiphol Social Council?

What are other points from the consultation that you would have liked to see different regarding measuring citizens' preferences in a good way?

Appendix O Client interview protocol

In this appendix, the protocol used for the interview with the client is presented. The structure of this protocol is as follows. The structure of this protocol is as follows. First of all, there is an introduction in which the consequences of participation were made known and the experts were asked for their consent. Thereafter, the three phases in which the PVE consultation is drawn up, are discussed. For each phase is discusses what went well and what could be improved. Finally, questions are asked about the concessions that have been made and the role of the client.

Expert interview protocol

Introduction

- Thank the interviewee for his or her time and participation to this interview.
- Indicate that the interviewee may stop at any time during the interview without giving a reason.
- The interview data is securely stored in the data centre of Delft University of Technology.
- Explain that this interview is part of a master thesis. The topic concerns the face validity of the PVE method.
- The purpose of this interview is to evaluate with the client what went well and what could be improved regarding to the different phases in setting up the PVE consultation.
- Does the participant have any questions?
- Ask if the audio of the interview may be recorded. This question is first asked without the audio on and repeated once the audio is on.

Evaluation per phase

During the determination of the preconditions and frameworks and the setting up of the PVE consultation regarding the participation and information facilities around Schiphol for local residents, three phases can be identified that have been completed. In this interview will be discussed for each phase what went well and what could be improved.

Phase 1: Determining the goals and preconditions of the PVE

During this preliminary research, the preconditions, scope and goals of the PVE are established. This is done in three sessions with stakeholders, policymakers and residents. With regard to the stakeholders, employees of Schiphol Group, Air Traffic Control the Netherlands and VNO-NCW West are included. Furthermore, policymakers of the municipality of Haarlemmermeer and Ouder-Amstel are included. During these exploratory discussions, questions were asked such as: what do you think is the core dilemma, what would you like to know from citizens, what choices must be made, where can citizens still inspire or provide ideas and what do you think of the idea of the Schiphol Social Council or an Environmental House?

Subsequently, a concept PVE consultation was drawn up. In this concept, the consultation started with a choice task about the Environmental House and its concrete tasks. With regard to the Schiphol Social Council, three issues are presented at different levels: local, regional and national. Finally, there are a number of questions about what information local residents would like to have and what information they are currently looking for. Therefore, the end

product of phase one is a concrete proposal for a PVE. This proposal is discussed during a meeting with the client.

What went well in this phase and what could be done better in your opinion?

Phase 2: Feedback on the tightened PVE consultation

After phase one was completed, phase 2 started with sharpening the PVE consultation and developing it more concretely. For example, the possible tasks of the Environmental House and the Schiphol Social Council are determined. This version of the PVE consultation is also implemented in the online platform with an instruction video in order to give the involved stakeholder and idea of the final consultation.

After this concretization of the PVE, another round is done with conversations with the involved stakeholders and the client. They are allowed to provide feedback. By involving various stakeholders in the design process, the aim was to achieve broader support for the final recommendations.

With regard to the content of the consultation, a number of changes in the design have been made at the end of this phase. For example, it is decided that the Schiphol Social Council should be the first choice task of the consultation and thereafter the choice task of the Environmental House, because the discussions about the Council caused more commotion. In addition, it is decided to leave out the issues at different scale levels but instead to formulate more concrete tasks for the Council. Participation principles are questioned after the choice task of the Schiphol Social Council. After the selection of the tasks for the Environmental House, in-depth questions are asked about the information needs of local residents. Another change that is made is about the demographic characteristics of the respondents being surveyed. For example, it emerged that both respondents had to be asked whether respondents live under a flight route and whether they experience any nuisance. Finally, it is decided to move the questions about how respondents would like to participate to the beginning of the consultation.

Finally, in this phase it is also decided that a representative group of citizens living in the BRS municipalities could participate and not a representative group of citizens for the Netherlands as a whole.

What went well in this phase and what could be done better in your opinion?

Phase 3: Feedback on the 99% version of the PVE consultation

The elaboration of the above changes took place in phase 3. After this elaboration, a final check is made by those involved and they were also allowed to provide their final comments. It is communicated to each of the stakeholders why their comments were or were not included in the consultation.

What went well in this phase and what could be done better in your opinion?

Concessions

During the phases it became clear that the stakeholders had different needs regarding the questions in the consultation. Ultimately, an attempt is made to find a middle ground that led

to consensus. This middle ground for the stakeholder may not meet the respondents' requirements when it comes to means of participation.

Have concessions been made between the wishes of the stakeholders and the participating citizens? What concessions are involved?

The role of the client

How do you see your role as a client?

Code list

The following code list is used to transcribe the expert interview. The expert interviews were held in Dutch, so the codes are listed in English as well as in Dutch below.

Dutch	**English**
Duidelijkheid	Clarity
Relevantie	Relevance
Leesbaarheid	Readability
Volledigheid	Completeness
Haalbaarheid	Feasibility
Commotie	Commotion
Consensus	Consensus
Omgevingshuis	Environmental House
MRS	Schiphol Social Council
Transparantie	Transparency
Concretisering	Concretization
ORS	ORS