Mapping Leisure Shopping Trip Decision Making: Validation of the CNET Interview Protocol

Tim De Ceunynck (corresponding author) [a], Diana Kusumastuti [b], Els Hannes [c], Davy Janssens [a], Geert Wets [a]

[a] Transportation Research Institute, Hasselt University
Wetenschapspark 5/6, 3590 Diepenbeek, Belgium
Fax: +32(0)11 26 91 99
Tel.: +32(0)11 26 -- -- {91 18; 91 28; 91 58}
Email: {tim.deceunynck; davy.janssens; geert.wets}@uhasselt.be

[b] Centre for Transport Studies, University of Twente
Drienerlolaan 5, 7522 NB Enschede, The Netherlands
Fax: +31 (0) 534 89 4040
Tel.: +31 (0) 534 89 4004
Email: D.Kusumastuti@utwente.nl

[c] PHL University College, dpt. PHL-Architecture
University campus, building E 3590, Diepenbeek, Belgium
Tel.: +32 11 24 92 06
E-mail: els.hannes@phl.be

Submitted and accepted for publication in Quality & Quantity

Submission date: June 22nd, 2011
Acceptance date: October 20th, 2011
Publication date: 2011

Abstract

Qualitative research methods can provide an in-depth understanding of how people come to certain decisions, providing valuable input to ground behavioural assumptions in activity-based travel demand models and to implement high impact policy measures to change travel behaviour. The CNET interview protocol is a semi-structured personal interview method to elicit the mental representation of individuals' decision making. There is a risk of bias caused by the interviewer's interpretation of the respondents' answers. Therefore, the quality of the CNET interview protocol is assessed by evaluating its trustworthiness using intercoder reliability tests. Krippendorff's alpha is identified as the most appropriate measure. The intercoder reliability is sufficiently high. Consequently, the CNET interview protocol can be considered a valid method to measure and map individuals' considerations in complex spatio-temporal decision problems.

Keywords: intercoder reliability, Krippendorff's alpha, CNET interview protocol, leisure shopping, decision making

1 Introduction

Research about travel behaviour has yielded critical insights into choices that individuals and households make about their daily travel. To study travel behaviour, most researchers rely on quantitative methods to explore travel patterns. They collect a limited amount of information about a research topic from a large number of entities, and perform statistical analysis to be able to draw conclusions that can be generalized to a certain population group (Clifton & Handy, 2003).

Another less frequently used way to study travel behaviour, are qualitative research methods. As opposed to quantitative research methods, qualitative methods gather very rich and detailed information from a small number of entities. The aim of these studies is an in-depth exploration of selected issues. The relatively limited usage of qualitative research methods in the field of transportation might be attributable to the fact that qualitative research has often been criticized. A first issue is that the small sample sizes usually do not allow to draw generalized conclusions, because formal statistical testing cannot be applied and the samples used are usually drawn randomly (Niaz, 2006; Strauss & Corbin, 1998). A second issue is that some researchers believe that qualitative methods often suffer from a lack of scientific rigour. One of their arguments is that they believe that conclusions in qualitative research often depend on subjective interpretations from the researcher (Leiva, Ríos, & Martínez, 2006). However, when qualitative methods are given the same attention to rigour in the research design, data selection, data analysis and interpretation as traditional quantitative studies, they can complement quantitative approaches or stand as a legitimate research method in their own right (Clifton & Handy, 2003).

While quantitative research methods mainly capture observed outcomes of travel decisions, qualitative research methods are able to explore how people come to a certain decision, and why they reach a particular decision outcome. Quantitative studies usually do not provide detailed answers about these “why” and “how” questions. Therefore, qualitative methods can help to fill these knowledge gaps that are left by quantitative techniques (Clifton & Handy, 2003).

Understanding why and how certain travel patterns arise is highly important because of various reasons. First of all, this knowledge can be used to ground behavioural assumptions that underlie disaggregated activity-based travel demand models (Kusumastuti, Hannes, D. Janssens, Wets, Dellaert, & Arentze, 2009a). These models simulate and predict activity-travel patterns by modelling travel behaviour at a disaggregated level, as the outcome of individual travel decisions based on a personal activity schedule, constraints, preferences, household interactions, etc. instead of modelling trips at an aggregated level (D. Janssens, Wets, Timmermans, & Arentze, 2007). Therefore, these activity-based models need a fundamental understanding not only about travel decision outcomes, but also about the decision process to be able to produce reliable predictions and analyses. And secondly, this underlying knowledge can help policy makers to implement high impact and effective policies to change travel behaviour (De Ceunynck, Kusumastuti, Hannes, D. Janssens, & Wets, 2011).

Quantitative methods often rely on surveys, which have some limitations. The most important problem is that surveys are often used in circumstances where the issues under study are defined very clearly, and the responses of participants are anticipated (Clifton & Handy, 2003). This way, the survey instruments not only narrowly frame the questions, but they also limit the possible range of answers. Therefore, the possibilities of surveys are bounded by the perspectives and the goals of the survey developers (Poulenez-Donovan & Ulberg, 1994). In other words, surveys are not suited to reveal results that are not initially (at least partly) anticipated by the researcher. Qualitative research methods, however, do not suffer from this issue because of their broader approach (Clifton & Handy, 2003). Research methods that are most often used in qualitative research are face-to-face interviews, participant observations and focus groups.

The Causal Network Elicitation Technique (CNET) interview protocol, which is used in this study, is a qualitative semi-structured personal interview technique to elicit individuals' constructs and beliefs and their interconnections when making leisure shopping travel decisions, in a structured mental representation of the decision problem (Kusumastuti, Hannes, D. Janssens, Wets, & Dellaert, 2009a).

In the CNET interview protocol, an interviewer asks a series of open-ended probing questions, asking for considerations that have an influence on the respondents' leisure shopping travel behaviour. The respondent can mention each consideration that comes to his mind, and the respondents' answers are coded by means of an extensive pre-defined list of variables. However, this final step implies a risk of subjective interpretation by the interviewer, because it is not always straightforward to filter the variables of interest from respondents' open answers.

The CNET interview protocol is still a relatively new research method. So far, it has been applied successfully to assess individuals' travel decision making processes in a hypothetical situation (den Hartog, Arentze, Dellaert, & Timmermans, 2005), and the technique has recently been adopted to assess individuals' leisure-shopping trip decisions in a real world setting (Kusumastuti, Hannes, D. Janssens, Wets, & Dellaert, 2009b). However, the reliability of the results of this protocol have not been formally assessed before. Especially the risk of subjective interpretation by the interviewer needs to be studied. The focus of the paper is to assess the intercoder reliability of the CNET interview protocol. To this end, a sample of the voice recorded interviews coded in the study by Kusumastuti et al. (2009b) is recoded using the same method by a second coder, and differences and similarities are analyzed using intercoder reliability measures.

The paper is structured as follows: section 2 provides the theoretical background about decision making, mental representations, the CNET interview protocol, qualitative research and the assessment of the quality of qualitative research; section 3 provides the calculation of the intercoder reliability of the CNET interview protocol; section 4 analyzes and discusses the results; section 5 summarizes the most important conclusions.

2 Theoretical background

In the first part of this section, a short introduction to human decision making and mental representation is presented. In the second and third part, the CNET

interview protocol in general and the experiment of interest are briefly described. Qualitative research is discussed in the fourth part. Fifth, the assessment of the quality of qualitative research is presented. In the sixth part, intercoder reliability is discussed. And finally, different intercoder reliability measures are provided.

2.1 Background: decision making and mental representation

One of the most influential theories about human decision making is the rational choice theory, which assumes that people calculate the likely costs and benefits of an action before deciding what to do, and choose the alternative that yields the highest expected utility (Henrich et al., 2001; J. Scott, 2000). Despite its importance, the theory has received some critique because of its unrealistic assumptions (e.g. fully informed decision maker, fully rational decision process), for example by Kroneberg (2006). Indeed, people's decision making process can also be seen as a process relying on a number of simplifying heuristics, rather than extensive algorithmic processing (Gilovich, Griffin, & Kahneman, 2002). These heuristics are efficient rules of thumb of the type if-then(-else) to get to a decision relatively easily.

However, for new or infrequent decisions, people do not always have ready-made solutions for all possible contexts of the decision environment. In that case, a complex and deliberative cognitive process is activated in which different considerations are linked by means of causal relations (Kusumastuti, Hannes, D. Janssens, Wets, & Dellaert, 2010). This is called a mental representation. It is activated when decision makers face complex decision problems. A mental representation consists of various contexts in the decision environment, the decision maker's benefit requirements, instruments of the decision alternatives, and the causal relationships between these variables (Dellaert, Arentze, & Timmermans, 2008). Thus, four types of variables can be distinguished in the mental representation: decision, contextual, instrumental and benefit variables.

Decision variables represent the decision alternatives available to the decision maker. For each decision variable, there is a set of pre-defined choice alternatives

(Arentze, Dellaert, & Timmermans, 2008). For instance, a decision variable "transport mode choice" could have the alternatives car, bus and bike.

Different choice alternatives have various characteristics, leading to different consequences for the decision maker. The characteristics of the choice alternatives can be considerations in the decision making process, and they are called instrumental variables. Instrumental aspects can be observed and operated by the decision maker (Kusumastuti, Hannes, D. Janssens, Wets, & Dellaert, 2009a). In a transport mode choice, instruments of the different transport modes can be for instance shelter provision, travel time, flexibility, easiness for parking, etc.

Contextual variables refer to situations, circumstances, and constraints in the decision environment that can have an influence on the outcome of the decision making process, but that cannot be controlled by the decision maker (Arentze et al., 2008). These can be natural forces like weather conditions, or a number of constraints, like for instance capability, coupling and authority constraints (Hägerstrand, 1970).

Benefit variables are directly related to utility. They describe the impact of the state of the contextual and instrumental variables on the fundamental needs and the well-being of the decision maker (Dellaert et al., 2008). Examples of benefit variables can be the desire to gain efficiency, physical comfort, etc. from the decision.

The final element of the mental representation are the causal links between these types of variables to obtain a network representation of the decision problem. This causal network indicates the individual's beliefs about the way the decision variables activate the consideration of other decision variables (Kusumastuti et al., 2009b).

The smallest building block of a mental representation is called a "cognitive subset", which consists of a context, a benefit and an instrument. An example that is shown in Figure 1 is for instance the cognitive subset {weather, shelter provision, comfort} for the transport mode choice. However, a cognitive subset

can be irrespective of the context of the decision. This means it always applies, in any circumstances. For instance, it is possible that someone always considers the type of stores in a zone in the shopping location choice, because it always influences the person's shopping efficiency, irrespective of the context of the decision. In that case, the cognitive subset {normally, type of stores, efficiency} has no contextual variable (Kusumastuti et al., 2011).

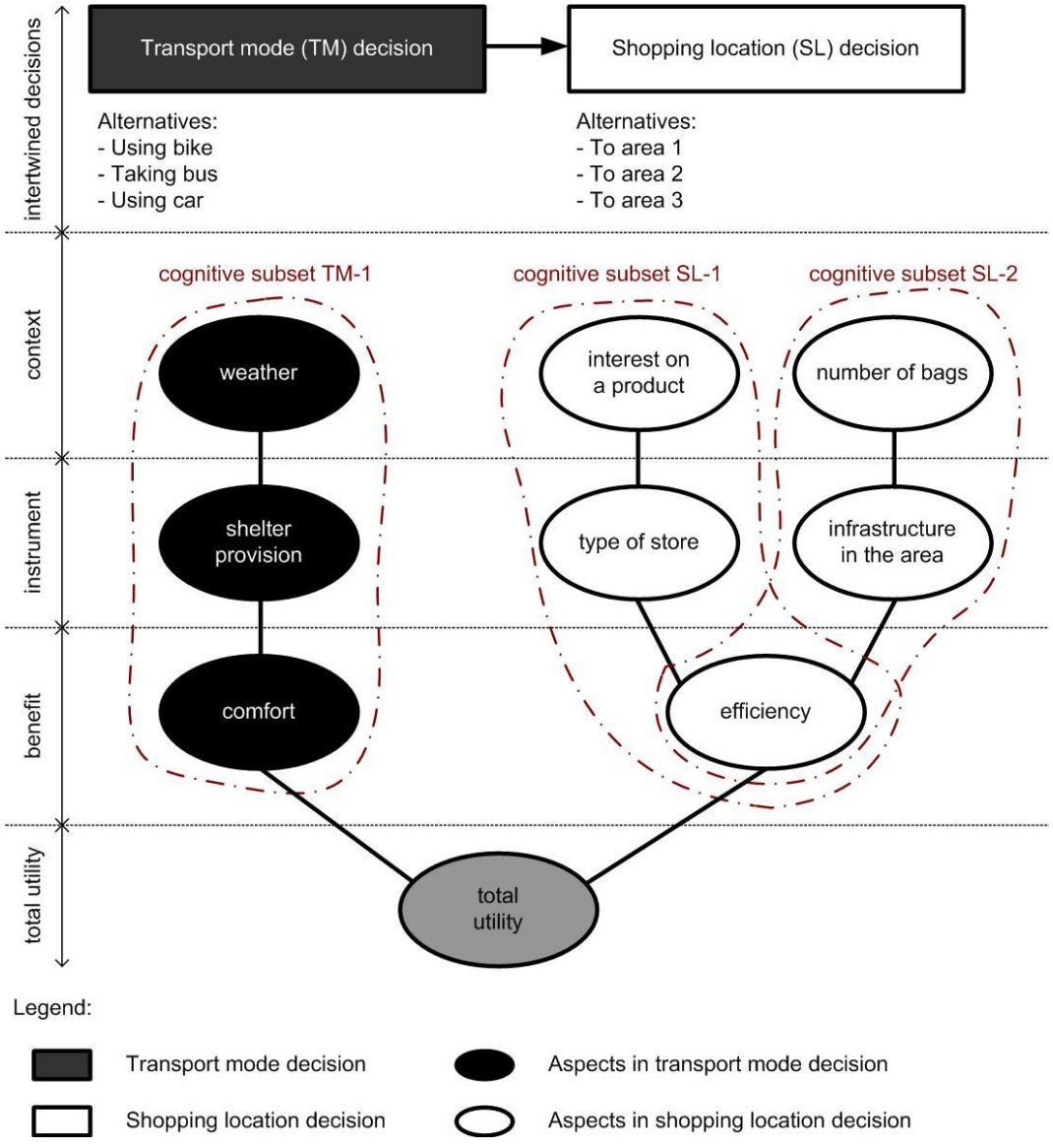

Figure 1: An example of the different variable types and their links in a mental representation.

2.2 The CNET interview protocol

The CNET interview protocol is structured along the lines of mental problem representation in human decision making. Respondents' considerations when

making decisions are systematically questioned and coded by means of a predefined list of variables that is defined prior to conducting the interviews. The aim of the list of variables is to convert primarily unstructured information (i.e. freely expressed thoughts of respondents) into a structured format. The list of variables does not have to be exhaustive (Arentze et al., 2008). If a respondent mentions a variable that does not correspond to any item in the list, the variable is added to the list.

For each consideration that is mentioned, a standardized continuation question is available, based on the type of variable it concerns. This way, the process results in the generation of an individual, context specific mental representation that visualizes the relevant contextual properties, the instruments of the decision variables and the benefits that respondents want to obtain from making the decision (Dellaert et al., 2008). More details about the protocol are presented in the next subsection, where the CNET interview protocol is applied to a leisure shopping context.

The process of coding the open answers by means of the predefined list of variables implies a risk of subjective interpretation by the interviewer. There is a risk that the interviewer interprets the respondents' answer incorrectly, resulting in indicating a wrong variable, overlooking a variable or including an abundant variable. In the CNET interview protocol, this problem is partially overcome by verifying the interviewer's selected code with the respondent. Verification is a process of checking and confirming the interpretation to make sure that an answer is interpreted correctly. This way, the analysis becomes self-correcting (Morse, Barrett, Mayan, Olson, & Spiers, 2002). In the CNET interview protocol, verification is done after each question. The respondent then states whether he agrees with the interviewer's interpretation or not.

2.3 An experiment: using the CNET interview protocol to elicit individuals' leisure shopping decisions

In this paper, the quality of the interview protocol to derive the elements and structure of a mental representation is assessed, based on its application by

Kusumastuti et al. (2009b) to planning a leisure shopping activity in the city centre of Hasselt in Belgium. Before the start of the interview, the research setting is explained to the respondents. For the purpose of that study, the following scenario is presented to a sample of 26 respondents, all students at Hasselt University: *"Suppose you have a vague plan in mind to go leisure shopping in the city centre of Hasselt in the near future. Leisure shopping is related to collecting shopping information (e.g. available stores, products that are sold, price, etc.). In addition, you have to buy a small gift for a friend".* The respondents are supposed to make the following decisions: (1) where to go shopping in terms of three distinct zones: the main shopping street, the boutique area and the gallery area (shown to them on a map). (2) the transport mode to go to the city centre, with a choice between car, bus and bike; and (3) when to go shopping, i.e. next Saturday, another Saturday or on a weekday. Moreover, respondents are asked to presume that they live about 5 km from Hasselt, that they know how to ride a car and a bike and that a bus stop is located within walking distance from their home.

For each decision, a similar procedure is used. The interviewer indicates the choice alternatives and asks the respondent which considerations come to his mind when making a choice between these alternatives. The respondent is supposed to mention all considerations that are of influence to make the decision one by one. The responses are verified to the predefined lists of variables. The variable lists of the transport mode choice exist of 17 contextual variables, 28 instrumental variables and 17 benefit variables. For the shopping location choice, 15 contextual, 26 instrumental and 17 benefit variables are defined. For the decision when to execute leisure shopping, the lists exist of 9 contextual, 13 instrumental and 17 benefit variables. The list of benefit variables is identical for all three decisions. For each type of variable that is mentioned, there is a subsequent standard question in the interview protocol. These questions make it possible to yield all cognitive subsets entirely (Arentze et al., 2008).

Suppose that for the transport mode choice a contextual variable is mentioned first, for instance *weather conditions*, which is a contextual variable. The interviewer has to complete the respondent's cognitive subset by eliciting the aspects that are related to *weather conditions*. This is done by asking *why* weather

influences the decision. The respondent may indicate that weather is considered because of the need to *have comfort*, which is a benefit variable. The interviewer now has to reveal the instrumental aspect related to the variables *weather conditions* and *having comfort* by asking: *"How can your choice of transport mode influence your comfort in different weather conditions?"*. The respondent may indicate then that car and bus provide shelter and bike does not. In this case, the interviewer adds *shelter provision* as the instrumental aspect. This way, one cognitive subset related to the transport mode decision is completed (Kusumastuti, Hannes, D. Janssens, Wets, & Dellaert, 2009a).

The interviewer then repeats the first question of the elicitation process: *"What other considerations come to your mind relating to the transport mode decision when you go leisure shopping in Hasselt?"*. As a second case, suppose a benefit aspect is now mentioned first. For instance, the respondent indicates that *efficiency* is important. Then, the interviewer has to elicit a related instrument or context by asking *how* the benefit aspect *efficiency* is influenced. The respondent may indicate that having *efficiency* is important depending on his *time availability*, which is a contextual aspect. The interviewer then continues by asking a question to elicit a related instrumental aspect (Kusumastuti, Hannes, D. Janssens, Wets, & Dellaert, 2009a).

Eventually, in case an instrumental aspect is mentioned first, the interviewer has to elicit related benefit(s) or context(s). Suppose the respondent mentions *vehicle speed* first (instrumental aspect). To elicit benefits or contexts related to *vehicle speed*, the interviewer has to ask a *why* question. When the respondent mentions a benefit variable, e.g. *efficiency*, one cognitive subset can be considered as complete, i.e. {normally, vehicle speed, efficiency}. However, when the answer to this first question is a contextual aspect (e.g. *time availability*), the interviewer has to ask another *why* question to reveal the benefit aspect to complete the cognitive subset (Kusumastuti, Hannes, D. Janssens, Wets, & Dellaert, 2009a).

2.4 Methodology

2.4.1 Quality of qualitative research

The quality of qualitative research is covered by the concept "trustworthiness". There are four criteria of importance to ensure trustworthiness in qualitative research: credibility, transferability, dependability and confirmability (Morse et al., 2002).

Credibility is an evaluation of whether the findings of the research are a "credible" conceptual interpretation of the information that is obtained from the respondents or not (Fenton & Mazulewicz, 2008). The question is whether the results are credible from the participant's perspective, since the goal of qualitative research is to describe or understand the phenomenon from the viewpoint of the respondent. That is why the respondents themselves are the only ones who can legitimately judge about the credibility of the results (Trochim, 2006).

Transferability is the degree to which the method can be applied or transferred beyond the borders of the project (Trochim, 2006). Note that this transferability is not the same as the concept of generalizability in quantitative research. In quantitative research, a statistically founded sample drawing with a sufficiently large sample size is chosen from the population. Therefore, the results can be generalized to the full population. In qualitative research however, this is usually not the case. Here, the question is rather: "Is it possible to apply this research method to a broader population and/or other circumstances?" (Fenton & Mazulewicz, 2008). The person who wishes to transfer the research method is responsible for assessing whether it is wise to make the transfer or not (Trochim, 2006).

Dependability is the assessment of the quality of the integrated process of data gathering, data analysis and theory development (Fenton & Mazulewicz, 2008). Dependability emphasizes the necessity to bear the continuously changing context in mind, in which the research takes place (Trochim, 2006). In quantitative research, the aim of the concept of reliability is that the research is repeatable. However, this is impossible in qualitative research because of the ever-changing characteristics of the context. Knowledge generated by means of qualitative research is not absolute, but it is restrained to time, context and culture.

Dependability must be considered as the agreement between the documented data and what actually happened during the research (Goodson, 2004).

Confirmability refers to the degree to which results can be confirmed or supported by others. Confirmability is a measure of how well the findings of the researcher are supported by the gathered data. It refers to the objectivity with which the research is carried out. The researcher can never be fully objective though. Qualitative research assumes that each researcher brings a unique point of view into the research (Trochim, 2006). However, the data analysis process is made objective by searching for all sorts of possible explanations for the phenomenon under investigation, reporting theoretically meaningful variables and granting others access to the data to judge the way important interpretations are deduced from the empirical material (Goodson, 2004).

Besides these criteria, specific tools are available to check to what extent the quality requirements of qualitative research are fulfilled. Of these tools, "intercoder reliability" is most relevant to this study. That is why other tools (e.g. thick description, theoretical sampling, reflexivity of the researcher) are not discussed in this paper. Intercoder reliability will be explained in the following subsection.

2.4.2 Tools for assessing qualitative research: intercoder reliability

Intercoder reliability is the general term for the degree to which different coders who judge an aspect of a message or object get to the same conclusion. This is a crucial component of certain types of analysis, such as analyzing interviews using open-ended questions. Interpretations of data can never be valid when they are not intercoder reliable (Hak & Bernts, 1996; Lombard, Snyder-Duch, & Bracken, 2002).

The quality of assigned codes for answers to open-ended questions, as is the case in this study, depends on two things: on the one hand the validity of the coding process, which is the degree to which the codes represent the true meaning, and on

the other hand the (intercoder) reliability, which is the degree to which the interpretations of different coders coincide (Hak & Bernts, 1996).

The validity of the coding process is the degree to which the theoretically relevant aspects of the answers are truly represented by the codes. This refers to the quality of the coding list. The purpose of coding is to decide if, and how, a theoretically relevant aspect is observable in a response. The validity of the coding process is most often considered to be independent of the quality of the coding instructions. The quality of the relation between the responses and the codes cannot be considered to be independent of the coding itself. Validity is a matter of argumentation (Hak & Bernts, 1996). So, in brief, the validity of the coding process is depending on the quality of the coding list, and the learning process of the coders.

Unlike validity, the intercoder reliability can be assessed without referring to the main point of the coding process. Reliability is a matter of calculation. The intercoder reliability is considered to depend upon the implementation of the instructions by the coders (Hak & Bernts, 1996).

So, acquiring the same coding results by different coders is considered to be a sign of theoretical solidity (Hak & Bernts, 1996). This way, intercoder reliability implicitly functions as a measure of validity: when the coding instructions are valid (so when the instructions are theoretically warranted, and the coders are well-trained), and the coding process produces reliable codes, it is assumed that this is the result of a valid implementation of the instructions. The fact that the interpretations are intercoder reliable, is nevertheless not sufficient to assume that the used methods are valid, but it is a necessary condition. It is not a sufficient condition because the degree of agreement could be the result of a so-called "training artifact". This implies that the research design and the implementation themselves lead to certain measurement results and relations, and not the reality that is investigated (Hak & Bernts, 1996). A training artifact can be caused, for example, by errors in the coding list, or by too pointed instructions to the coders. More simply stated: they all do the same, and get the same results, but that is because they all do something wrong because of an error in the training process.

An important practical reason to realize intercoder reliability is that it enables a division of the coding work among multiple coders. A high level of difference in interpretation amongst coders suggests that there are weaknesses in the research methods, like poor operational definitions, categories or education of the coders (Lombard et al., 2002).

To calculate the intercoder reliability, the following steps are to be taken (Lombard et al., 2002):

1. Select one or more suitable measures of intercoder reliability. There are several measures of intercoder reliability, e.g. percent agreement (Lombard et al., 2002), Cohen's kappa (Cohen, 1960), Scott's pi (W. A. Scott, 1955), Krippendorff's alpha (Krippendorff, 2003),… The choice depends on the characteristics of the variables, like the measurement level, the number of coders and the expected division into different categories. If percent agreement is selected, it is recommended that a second indicator is used that accounts for agreement by chance. More information about the different measures of intercoder reliability is provided in the next subsection.
2. Acquire the correct tools to calculate the selected indicator(s) (e.g. statistical software packages).
3. Select an appropriate minimum level of acceptation for the reliability of the indicator(s). This depends on the nature of the study and the measure. Coefficients of 0,90 or higher are nearly always acceptable, 0,70 can be appropriate for explorative studies. Higher criteria should be used for liberal indices (e.g. percent agreement) and lower criteria can be used for more conservative indices (e.g. Krippendorff's alpha).
4. Assess the intercoder reliability. This occurs in several steps.
    A. Informal assessment of the reliability during training of the coders. This is an informal test with a small number of units. In this research, one cognitive subset is considered as one unit. The informal reliability assessment in this research took place by recoding one short interview (about 30 minutes) from voice recordings.

B. Formal assessment of the reliability by means of a try-out test. Here, as a rule of thumb, a sample of about 30 units should be taken. If the level of reliability is sufficiently high, one can move on to the full sample. If this is not the case, extra training should be done, the coding instruments and procedures have to be refined, or (in exceptional cases) one or more coders should be replaced. In this research, formal try-out assessment consisted of recoding two long interviews (about 2 hours).

C. Formal assessment of the reliability of the full sample. This step is the focus of this paper. The appropriate size of the sample depends on many factors, but should be no less than 50 units or 10% of the total research (Lombard et al., 2002). For the formal assessment of this study, three respondents are randomly selected for each of the three decisions. The selected sample meets these conditions.

D. Use a procedure to include the sample of the validity assessment in the research. Unless the reliability is perfect, there will be disagreement for some units in the sample. Depending on the characteristics of data and coders, it can for instance be decided to accept the interpretation of the majority, to let the researcher or another expert cut the knot, or discuss about the disagreements. Since this validity assessment study has been disconnected from the analysis of the research results, this step is not relevant for this study.

In literature, there is some disagreement about which quality criteria can be assessed with the intercoder reliability method (Goodson, 2004; Hak & Bernts, 1996). In addition, it also depends on how criteria of credibility, transferability, dependability and confirmability are delimited exactly. According to the way the criteria are delimited in this study, the intercoder reliability can mainly be used to test the dependability, since it concerns a second opinion about what elements are actually present in the respondents' answers (Goodson, 2004). As stated before, it is also an indirect measure for the credibility and the transferability of the research, because acquiring the same coding results by different coders is seen as a sign of theoretical solidity. Finally, an intercoder reliability assessment also allows to judge the confirmability of the research, because this refers to the

neutrality of the research. So, it can be tested how neutral, or "free of subjective interpretation" the technique is (Hak & Bernts, 1996). This implies that the intercoder reliability allows to evaluate all four quality criteria of qualitative research.

2.4.3 Intercoder reliability measures

To select the most appropriate measure of intercoder reliability, the characteristics of the data are analyzed first. The intercoder reliability measure has to be suited for data involving only two coders. Furthermore, the measure has to be able to deal with a small sample size and missing data. Moreover, the measure should preferably be sophisticated enough to account for, for instance, agreement by chance. The rest of this section describes the reasoning behind the chosen measure, based on a literature review.

Several measures of intercoder reliability were identified. The most commonly used measures are percent agreement (Lombard et al., 2002), Cohen's kappa (Cohen, 1960), Scott's pi (W. A. Scott, 1955) and Krippendorff's alpha (Krippendorff, 2003). The data characteristics in this research influence the choice of the measure: the list of variables consists of a large number of possible codings, which implies that the number of codings for each possible coding is small, the overall sample size is relatively small, and the codings are nominal in nature.

Percent agreement is the most simple measure of intercoder reliability. It is calculated by counting the number of cases for which there is agreement between the coders, and dividing the outcome by the total number of cases considered. This value is often criticized in literature because of its simplicity: for instance, it does not take the possibility of agreement by chance into account (Leiva et al., 2006; Lombard et al., 2002). However, because of the large number of variables used in this research, the possibility for agreement purely by chance is very small. Therefore, the measure seems actually suitable for this study. Literature suggests that a researcher should not use only percent agreement to calculate intercoder reliability (Lombard et al., 2002). Hence, a more sophisticated measure is selected in addition.

Cohen's kappa (Cohen, 1960) is another measure of intercoder reliability. It is generally thought to be a more robust measure of intercoder reliability than percent agreement since it takes agreement by chance into account (Leiva et al., 2006). However this correction for agreement by chance will be very small, since the large number of possible codings makes agreement by random chance very unlikely. Hence, the results of an analysis using Cohen's kappa will be very close to the value that is obtained with percent agreement. So, although Cohen's kappa is a good measure of intercoder reliability, the added value in this research is limited, and the measure is not selected for this research.

Scott's pi (W. A. Scott, 1955) also takes agreement by chance into account, but the measure has some restrictions. The measure can only deal with research in which there are two coders, nominal data and large sample sizes. The small number of codings for each variable results in a violation of the last restriction for this measure. Therefore, the measure is not suitable for this study.

Krippendorff's alpha (Krippendorff, 2003) is a rather sophisticated measure to assess the intercoder reliability. It is not a mere correction of percent agreement for chance like for instance Scott's pi and Cohen's kappa. It also takes the tendency of the coders to choose certain codes more often than others into account. In other words: if coding "x" is used more often by coder 1, and coding "y" is used more often by coder 2, the measure accounts for this preference, because coders tend to stick to codes they have already used before. This is called "proclivity". Furthermore, Krippendorff's alpha is applicable to any number of coders and acknowledges metrics other than nominal as well. It accepts missing data, and can deal with small sample sizes (Krippendorff, 2003). Hence, the measure meets all postulated requirements. Hence, Krippendorff's alpha is chosen as the second measure of intercoder reliability for this study because of its sophisticated nature and because of the fact that it can deal with small sample sizes and missing data.

One small adjustment had to be made to allow the software package "R" to include missing values in the calculation of Krippendorff's alpha. Missing values

occur when one coder codes a variable where the other coder does not. The problem is that, when there are missing values for one or more coders, the software checks whether there are other coders who coded the value too. So, in order to be included, a variable has to be coded by at least two coders. However, since this research only involves two coders, all cases where one coder codes a variable while the other does not, would be omitted by this procedure. This is highly undesirable, since variables that are coded by one coder, but not by the other, appear to be an important part of the total disagreement among both coders. To counter this, it is decided to include a variable called "missing value". This way missing values are considered as a separate variable by the software instead of missing. Including it this way does not influence the calculated value, while allowing the missing values to be omitted, would result in a strong overestimation of the agreement.

Because of the extensive list of variables, the corrections for chance and proclivity are expected to be small. This means that it is expected that the sophisticated Krippendorff's alpha measure will not differ strongly from the simple percent agreement measure. If this assumption is correct, it will be satisfactory for researchers using the CNET interview protocol to simply assess intercoder reliability by calculating the percent agreement instead of using more complicated measures, which are more cumbersome to calculate.

3 Analysis

3.1 Numerical example of percent agreement and Krippendorff's alpha calculation

A simple example is included to clarify the calculation of Krippendorff's alpha for two coders. For a more detailed explanation to calculate Krippendorff's alpha, the reader is referred to Krippendorff (2003). First, the dataset has to be gathered. For this example, the data is presented in Table 1.

	Case 1	Case 2	Case 3	Case 4	Case 5
Coder 1	Coding "A"	Coding "B"	Coding "B"	Coding "C"	Coding "A"
Coder 2	Coding "A"	Coding "B"	Coding "B"	Missing	Coding "B"

Table 1: Dataset example.

From the table, it appears that the coders agree on three codings, disagree on one coding, and one coding is only noted by one observer. This results in a percent agreement of 0,60 (3 agreements out of 5).

To calculate Krippendorff's alpha, the so-called coincidence matrix has to be formulated. This means that the pairs of coded variables have to be added to a symmetric matrix. The cells on the diagonal represent the codings for which there is agreement among the coders. Notice that for each pair with agreement among the coders, a value of two is added in the diagonal cell. The coincidence matrix for this example is shown in Table 2.

	A	B	C	Missing	**Total**
A	2	1	-	-	**3**
B	1	4	-	-	**5**
C	-	-	-	1	**1**
Missing	-	-	1	-	**1**
Total	**3**	**5**	**1**	**1**	**10**

Table 2: Coincidence matrix example.

The next step is to calculate the alpha-value, using the following formula:

Where O_{cc} = value of diagonal cell “cc”

N_c = row total

n = sample size

For this example, the calculation is the following:

As can be seen, the value for Krippendorff’s alpha (0,4375) is substantially lower than the percent agreement value (0,60). This is because the number of different codings was quite low in the example, making it very likely that some of the agreement is due to chance.

3.2 Intercoder reliability of the CNET interview protocol

After the selection of the measure, the assessment of the intercoder reliability is started. From previous research (Kusumastuti et al., 2009b), digital voice recordings of 26 interviews are available in which respondents are asked for their considerations for the *when*, *transport mode* and *shopping location* in a leisure shopping setting in digital format. These interviews have already been coded during the interview by the interviewer. The length of the interviews differs considerably, ranging from a length of 30 minutes to 1,5 hour. The calculation of Krippendorff’s alpha is done by means of the statistical software package R. Percent agreement is calculated in Excel.

First, a minimum acceptable level of agreement is selected. The percent agreement is added to have an idea about the difference between both measures, and because its interpretation is more straightforward. Furthermore, the percent agreement can be calculated for the cognitive subsets at large, which is not possible for Krippendorff’s alpha. Since this is an explorative study, and

Krippendorff's alpha is a conservative measure, a value of 0,7 or higher is considered satisfactory (Lombard et al., 2002).

Krippendorff's alpha is calculated for each variable type (contextual, instrumental, evaluative) in each decision (when, TM, SL). So this results in 9 alpha-values. Percent agreement is also calculated for each of them. Furthermore, percent agreement is calculated for each variable type in total (i.e., all contextual variables put together for each of the three decisions). Because the variable lists for the contextual and the instrumental variables have different variables for each decision, it is not possible to calculate Krippendorff's alpha for the contextual and instrumental variables, taken together over all decisions. Only the list of benefit variables is equal for all three decisions, so the alpha-value can be calculated for it. The fact that Krippendorff's alpha is not calculable for the contextual and the instrumental variables, is due to the fact that the coincidence matrix requires labels for the rows and columns. If these labels are not equal for each decision, it is not possible to draw up the coincidence matrix, and thus it is not possible to calculate Krippendorff's alpha. For the same reason, it is not possible to calculate Krippendorff's alpha for the cognitive subsets at large (i.e., instead of calculating the intercoder reliability for each component of the cognitive subset separately, calculating the agreement for the cognitive subset as a whole). The results can be seen in the following figure.

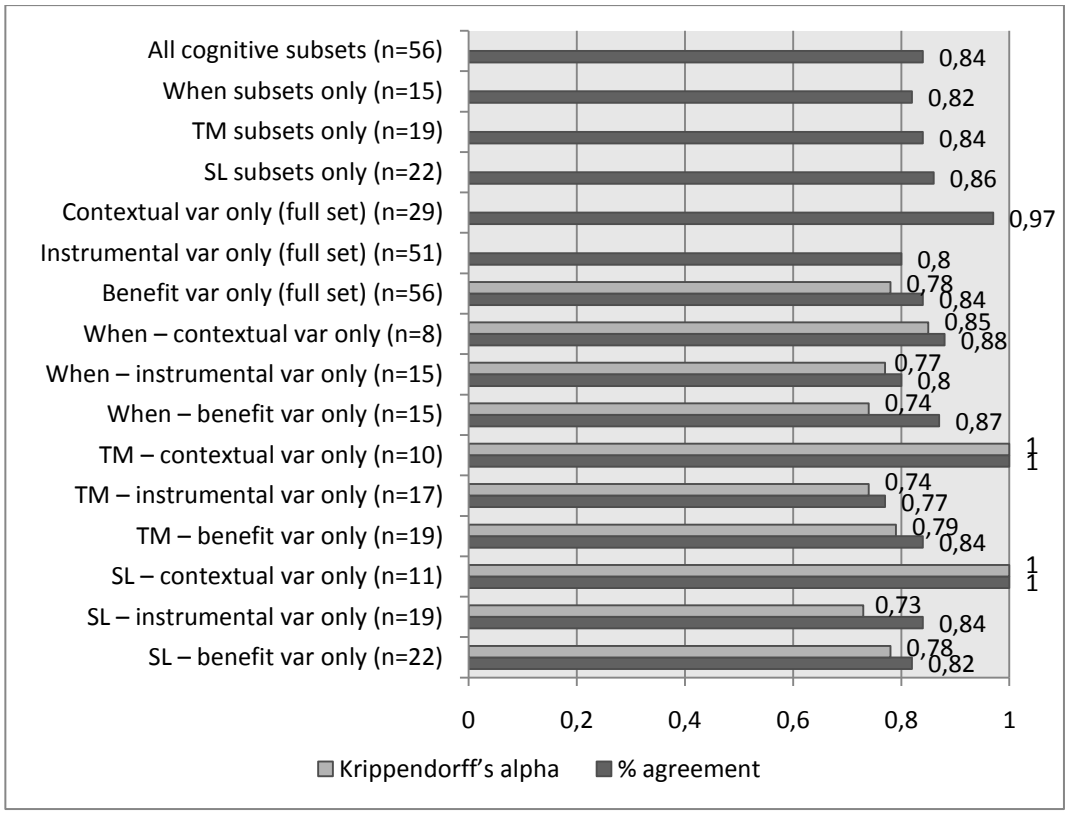

Figure 2: Percent agreement and Krippendorff's alpha values.

Most importantly, it can be seen immediately that all Krippendorff's alpha values pass the minimum acceptable level criterion of 0,70. Therefore, it can be concluded that the level of agreement between both coders is acceptable. The percent agreement values are 0 to 12,4% higher than the Krippendorff's alpha values. This confirms the premise that the differences between both values are relatively small.

Note that the alpha-values for the contextual variables for each decision are very high (85,1%, 100% and 100% respectively). Since all other alpha-values are in the range of 70-80%, it can be concluded that the agreement about the contextual variables is substantially higher than the agreement about the instrumental and the evaluative variables.

For the contextual and instrumental variables over all cognitive subsets, only percent agreement can be calculated. For the benefit variables, both measures can be calculated because the list of variables is the same for every decision variable. Not surprisingly, the value of the contextual variable over all cognitive subsets is

the highest one, since the values of the contextual variables for the separate decisions are all very high too.

There are no substantial differences in percent agreement on the level of the cognitive subsets at large for the different decisions. Their values are within a range of a few percent from each other. So the intercoder reliability is more or less equal for each single decision. For the cognitive subsets of all three decisions put together, a percent agreement of 84,2% was calculated.

So, in general, it can be concluded that the intercoder reliability for the CNET interview protocol is very well.

4 Results and discussion

In this section, the results that are presented in the previous section are discussed first. Next, the trustworthiness of the method is assessed based on the results presented in the previous section, and experiences that are obtained from measuring the intercoder reliability of the method.

4.1 Results

As shown in the previous chapter, the coefficient for the contextual variables is highest. A possible explanation is that contexts are "clearly observable states of the world", which makes them easier to explain for the respondents, and more straightforward to recognize for the interviewer. Instrumental variables relate to characteristics of the choice options. This is somewhat more abstract. The benefit aspects are the most abstract level of reasoning in the mental representation, and they refer to personal experiences or values that might even be pursued unconsciously. Therefore, they are generally the hardest variables to come up with for the respondents. Respondents have to recognize and express these "soft values", while interviewers have to categorize this abstract matter, which appears to be far from easy. That is why initially, it was expected that the percent agreement for the benefit variables would be substantially lower than for the

instrumental variables. Since this is not the case, it can be concluded that the coding list is well defined, and the definitions can be communicated sufficiently.

A possible explanation for present differences, is the fact that the interviewer has to make quick decisions during the interview itself, while the second coder who re-listens the interview has more time to reason about the answer. He can also listen to parts of the interview multiple times, and carefully weigh every word of the respondent.

Interview problems that may occur during the interview, but that are not observable in the intercoder reliability coefficients are the problems of proclivity of the respondent, and the problem of suggestion by the interviewer. Suggestion by the interviewer means that the interviewer accidentally raises a stimulus for the respondent to mention a certain value. This should be avoided because the consideration might not have been on the respondent's mind initially. Therefore, producing cues can have a significant influence on the obtained results. Suggestion is noted for 2 instrumental variables and 6 benefit variables in the set of re-listened interviews. So, the problem of suggestion by the interviewer seems to be largest for the benefit variables. The impact of suggestion seems relatively limited in this research, but researchers who wish to adopt the CNET interview protocol should remember to pay attention to this issue, and try to estimate the impact on the results.

Proclivity of the respondent is in fact comparable to proclivity by the interviewer: it means that a respondent tends to mention a particular variable more often because he "learned" that the interviewer is satisfied with this answer. This problem is particularly relevant for the benefit variables, since contextual or instrumental variables are hardly ever repeated by respondents. It is noted in 5 cases, all of which concerned two particular benefit variables: "Efficiency, saving time & effort" and "Having fun". This is probably because these are categories that match well with quite a few instrumental variables. However, it is very difficult to judge whether the repetition of a variable is in fact proclivity, or actually a real benefit that the respondent wants to obtain from the previously mentioned instrumental variable. Researchers using the CNET interview protocol

should try to estimate the impact of proclivity of the respondent on the research results.

An advantage of the face-to-face interview technique is the possibility to clarify the questions, and ask for more explanation if the respondent's answer is not clear to the interviewer. The verification technique is an example of this. The disadvantage of this direct contact between interviewer and respondent is that it involves a risk of bias by the interviewer, as respondents may react to the personal characteristics of the interviewer. They might for instance give socially desirable responses. Another disadvantage of personal interviews is the high marginal cost per interview (Boardman, Greenberg, Vining, & Weimer, 2005).

One final point is that respondents are asked what they would do in a specific situation. So, the respondent's intention is measured. Although intention is of direct influence to actual behaviour, there are still other influences that directly affect behaviour, like for instance habit, barriers and skills (Armitage & Conner, 2000; Pelsmacker & W. Janssens, 2007). Habit, for instance, is a variable from the variable list that is rarely stated by respondents, but when it comes down to it, will have a large impact. For instance, a respondent can state that he will go leisure shopping by bike when the weather is nice, but when the situation occurs in real life, he could go by car anyway. This intention-behaviour gap is important to keep in mind in behavioural research.

4.2 Trustworthiness evaluation of the CNET interview protocol

Concerning the credibility of the interview, it has to be stressed that the process of verification is part of the interview protocol. This is a good way to improve the credibility of the method, since the only one who is able to legitimately judge the credibility of the researcher's interpretation is the respondent himself. A problem with verification is that there is a need to verify systematically each response because, even if the researcher is very sure about an interpretation, he might still be incorrect. However, systematically verifying each answer could annoy both respondent and interviewer, because a lot of questions seem quite obvious. For

instance, the respondent could state “I consider the weather in my TM decision”. Bearing the verification technique in mind, the interviewer has to ask then: “So you consider ‘weather’?”. This increases the burden of the interview on both interviewer and respondent. Since the high burden for participants is a limitation of the method to start with, this is undesirable. Furthermore, verification does not guarantee that the researcher’s interpretation is correct. This is because the variable that is mentioned by the researcher (to ask whether it is the right interpretation), could have a different interpretation to the respondent.

There is still another problem with verification. Since the interviewer is only human, he is bound to make mistakes. An incorrect interpretation may lead to suggestion by the interviewer. It could cause a reaction like: “No, that is not exactly what I mean, but now that you mention it, I do consider that as well”. It is very important that the interviewer avoids to mention considerations that are initially not in the respondent’s mind, because this can influence the results. Another possibility is that the respondent is having some trouble to explain what he or she means exactly. The interviewer could ask then, for verification: “So you mean …”, and the respondent might agree. This could indeed be a good interpretation of the answer, but it might also be that the respondent wants to get himself out of this difficult situation, and therefore agrees with the explanation. In some way, this is also a form of suggestion.

However, despite these minor remarks, it is concluded that the credibility of the method is fine. Making use of verification in a well-considered way leads to a credible representation of the respondents’ decision making problem.

Transferability in qualitative research is not the same as generalizability of the results in quantitative research. Generalizing the results of a qualitative research is always risky, and exceeds the aim of most of these studies. Qualitative research most often aims to obtain an in depth understanding of a phenomenon, as is the case with the CNET interview protocol, or an exploration of a new and complex decision problem. Therefore, the possibility to transfer the method is more important than the possibility to transfer the results. It can be assumed that context or coincidence play a role. If a respondent would take the interview again, he

could mention some other variables than he did in the first interview. This means that the results of a process of spontaneously mentioning variables is influenced by contextual or coincidental influences that trigger the recollection of certain variables at a certain moment (Kusumastuti et al., 2009b). This might be partially attributable to the hypothetical situation from which the respondents in this research have to reason. Because of the hypothetical situation, respondents only have basic information about the circumstances in which the decision making process takes place. That is why they might not think of some aspects that are of influence on their shopping behaviour in real life, like for instance the possibility to combine the trip in a tour with other activities. This issue could be of less importance when people can reason from their real life situation. However, it is also possible that the opposite is true. Since it is quite probable that at least part of the decision making process occurs unconsciously, people could be having more difficulties to mention their considerations, or they might not be aware of some facts about the environment in which their decision making process takes place. This is an interesting topic for further research.

If the method is transferable, this means that other researchers are able to collect situation-specific information about variables that are included in the decision making process in the particular circumstances in which they are interested. A high level of intercoder reliability hints to the fact that the research method is valid if the list of variables is formulated well, and there are no training artifacts. A lot of time and effort is put in the development of the list of variables. Also, the list was updated continuously during the interviews, and afterwards during the assessment of the intercoder reliability. It is therefore unlikely that there are any biases because of the list of variables. Since the second coder was trained by the initial interviewer herself, it is impossible to exclude the presence of a training artifact. However, since the interviewer is competent in the coding technique, and because the second coder is an independent outsider who had no further involvement in the research, there is no reason to suspect training artifacts.

Recall that dependability should be considered as the correspondence between the documented data and what actually happened during the research. It is a judgment of the integrated process of data gathering, data-analysis and the development of

theories (Fenton & Mazulewicz, 2008). During the data gathering, there have been no changes in the design of the interview. The group of respondents has been selected at one certain point in time, and they are all interviewed in a short time period. As mentioned before, the only changes that occurred during the interview, is the addition of new variables to the list. Since respondents are not allowed to see the list of variables, this does not affect the respondents' answers.

It is important to keep in mind that measuring the cognitive subsets of the respondents is no "hard science". So it is difficult to determine whether the subsets that the interviewer registers during the interview measure exactly what is going on in the respondent's mind. However, the fact that the intercoder reliability is high gives more confidence that this is indeed the case, since the second coder has come to similar results based on the same data. This indicates a good dependability of the method.

Confirmability refers to the extent the results can be confirmed or supported by others. Here, intercoder reliability is an important tool. In the interview protocol, interviewers have to try to remain as objective as possible in their interview technique. They have to ask the same questions to each respondent in the same situation. While recoding the interviews, it has been noticed there have been no problems with this. Since the findings of the second coder are similar to the findings of the interviewer, her findings are confirmed. Also, the process has been made objective by registering the reasons why there is disagreement about certain codings. This makes it possible to discuss the results to make sure there are no misunderstandings. Hence, it can be concluded that there are no problems regarding the confirmability of the method.

5 Conclusions

Qualitative research methods are able to explore in detail how people come to a certain decision. This knowledge can be used to ground behavioural assumptions that underlie activity-based travel demand models, and it can help policy makers to implement high impact policy measures to influence travel behaviour. The Causal Network Elicitation Technique (CNET) interview protocol is a method to

map the mental representation of an individual's decision making process by making use of a semi-structured interview. The interview protocol must be seen as a method to obtain the mental representation of the respondent's decision problem, to code it by means of a predefined list of variables, and to transform it into a decision network. There are four types of variables in the decision network: decision variables, contextual variables, instrumental variables and benefit variables.

There is a risk of bias caused by the interviewer's interpretation of the respondents' answers. The purpose of this paper is to measure this bias by calculating the intercoder reliability. These results, and the personal experience of the researcher during the study, make it possible to assess the quality of this qualitative research method. This can be indicated with the general term "trustworthiness".

The results of the intercoder reliability assessment are very good. All calculated values for Krippendorff's alpha exceed the postulated value of 0,70. The results of the percent agreement calculation are 0 – 12,4% higher, which confirms the expectation that percent agreement is a relatively well-suited measure of intercoder reliability for this research method, because the corrections for proclivity and chance are small. Furthermore, the specific nature of the research technique results in some limitations for the more sophisticated intercoder reliability measures. For instance, calculating the intercoder reliability of the cognitive subsets at large is not possible for the more complicated measures. Therefore, researchers who want to assess the intercoder reliability of the CNET interview protocol for a specific research setting do not have to calculate a sophisticated measure, but can settle for the straightforward percentage agreement measure.

Over the cognitive subsets at large, a percent agreement of 84,2% has been found for all decisions put together. There are no substantial differences among the different decisions.

The calculation of the intercoder reliability results in a judgment of the four criteria of trustworthiness. To ensure credibility of the research results, verification is part of the interview protocol. Despite some minor remarks, it can be concluded that this provides a good credibility. The high level of intercoder reliability found in this case study hints to a good transferability of the method, since there are no reasons to suspect validity problems with the coding list, nor training artifacts. Furthermore, the fact that the Krippendorff's alpha values are similar for all decisions is an indication that different types of decisions can be investigated with the method. Although it is difficult to determine whether the subsets the interviewer codes measure exactly what is going on in the respondent's mind, the high level of intercoder reliability indicates that this is indeed the case, so the dependability is satisfactory. Objectivity is ensured in a number of ways, which means that the confirmability of the method is good as well.

The conclusion of this study is that the CNET interview protocol has a high level of thrustwortiness. This means it can be concluded that the CNET interview protocol is a good method to reveal the decision making process of individuals regarding leisure shopping. The extent to which the research method can be transferred (to other target groups, to other types of decisions, …) is a subject of further research. Another topic for further research is scaling up the research method to a quantitative application. A first large scale application of the CNET methodology is presented in De Ceunynck et al. (2011) and in Kusumastuti et al. (n.d.).

6 References

Arentze, T. A., Dellaert, B. G. C., & Timmermans, H. J. P. (2008). Modeling and Measuring Individuals' Mental Representations of Complex Spatio-Temporal Decision Problems. *Environment and Behavior*, *40*(6), 843-869. doi:10.1177/0013916507309994

Armitage, C. J., & Conner, M. (2000). Social cognition models and health behavior: a structured review. *Psychology and Health*, *15*, 173-189.

Boardman, A., Greenberg, D., Vining, A., & Weimer, D. (2005). *Cost Benefit Analysis: Concepts and Practice* (3rd ed.). Prentice Hall.

De Ceunynck, T., Kusumastuti, D., Hannes, E., Janssens, D., & Wets, G. (2011). What Drives People? Analyzing Leisure-shopping Trip Decision Making. *Proceedings of the 90th Annual Meeting of the Transportation Research Board*. Washington D.C., USA.

Clifton, K. J., & Handy, S. L. (2003). Qualitative Methods in Travel Behaviour Research. *Stopher, P., Jones, P. (eds.) Transport Survey Quality and Innovation*. Oxford, UK: Elsevier Science Ltd.

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

Dellaert, B. G. C., Arentze, T. A., & Timmermans, H. J. P. (2008). Shopping context and consumers' mental representation of complex shopping trip decision problems. *Journal of Retailing*, *84*(2), 219-232. doi:10.1016/j.jretai.2008.02.001

Fenton, B., & Mazulewicz, J. (2008). Trustworthiness. Retrieved October 8, 2009, from http://www.omnivise.com/research/trustworthiness.htm

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and Biases: The Psychology of Intuitive Judgment* (1st ed.). Cambridge, United Kingdom: Cambridge University Press.

Goodson, L. (2004). *Qualitative Research in Tourism: Ontologies, Epistemologies and Methodologies* (1st ed.). Routledge.

Hägerstrand, T. (1970). What about people in regional science? *Papers in Regional Science*, *24*(1), 7-21.

Hak, T., & Bernts, T. (1996). Coder training: Theoretical training or practical socialization? *Qualitative Sociology*, *19*(2), 235-257. doi:10.1007/BF02393420

den Hartog, A., Arentze, T. A., Dellaert, B. G. C., & Timmermans, H. J. P. (2005). Eliciting Causal Reasoning Mechanisms Underlying Activity-Travel Choice: An Interview Protocol Based on Bayesian Belief Networks. *Proceedings of the 84th Annual Meeting of the Transportation Research Board*. Washington D.C., USA.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies. *The American Economic Review*, *91*(2), 73-78.

Janssens, D., Wets, G., Timmermans, H. J. P., & Arentze, T. A. (2007). Modelling shortterm dynamics in activity-travel patterns: the FEATHERS model. *Proceedings of the 11th World Conference on Transportation Research*. Presented at the WCTRS, Berkeley, USA.

Krippendorff, K. H. (2003). *Content Analysis: An Introduction to Its Methodology* (2nd ed.). Sage Publications, Inc.

Kroneberg, C. (2006). *The Definition of the Situation and Variable Rationality: The Model of Frame Selection as a General Theory of Action* (No. No.

06-05). SonderForschungsBereich 504 (p. 44). Mannheim: Universität Mannheim.

Kusumastuti, D., De Ceunynck, T., Hannes, E., Janssens, D., Wets, G., & Dellaert, B. G. C. (n.d.). Assessing Travel Demand Management Measures Based on Individuals' Mental Representations. *Paper is prepared for Transportation Research Part A.*

Kusumastuti, D., Hannes, E., Depaire, B., Vanhoof, K., Janssens, D., Wets, G., & Dellaert, B. G. C. (2011). An interactive computer-based interface to support the discovery of individuals' mental representations and preferences in decisions problems: An application to travel behavior. *Computers in Human Behavior*, *27*(2), 997-1011. doi:10.1016/j.chb.2010.12.004

Kusumastuti, D., Hannes, E., Janssens, D., Wets, G., Dellaert, B. G. C. & Arentze, T.A. (2009a). Qualitative and quantitative comparisons of the CNET interview and the CNET card game to explore contextual, instrumental and evaluative aspects in individuals' fun shopping travel decisions. *Proceedings of the 16th International EIRASS Conference on Retailing and Consumer Services*. Niagara Falls, Canada.

Kusumastuti, D., Hannes, E., Janssens, D., Wets, G., & Dellaert, B. G. C. (2009b). Qualitative exploration of contextual, instrumental and evaluative aspects in individuals' fun shopping travel decisions. *Proceedings of the 12th International Conference on Travel Behavior Research*. Jaipur, India.

Kusumastuti, D., Hannes, E., Janssens, D., Wets, G., & Dellaert, B. G. C. (2010). Scrutinizing Individuals' Leisure-Shopping Travel Decisions to Apprise Activity-Based Models of Travel Demand. *Transportation, 37(4),647-661*.

Leiva, F. M., Ríos, F. J. M., & Martínez, T. L. (2006). Assessment of Interjudge Reliability in the Open-Ended Questions Coding Process. *Quality & Quantity, 40*(4), 519-537. doi:10.1007/s11135-005-1093-6

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability. *Human Communication Research*, *28*(4), 587-604. doi:10.1111/j.1468-2958.2002.tb00826.x

Morse, J. M., Barrett, M., Mayan, M., Olson, K., & Spiers, J. (2002). Verification Strategies for Establishing Reliability and Validity in Qualitative Research. *International Journal of Qualitative Methods*, *1*(2), 13-22.

Niaz, M. (2006). Can Findings of Qualitative Research in Education be Generalized? *Quality & Quantity, 41*(3), 429-445. doi:10.1007/s11135-006-9015-9

Pelsmacker, P. D., & Janssens, W. (2007). The effect of norms, attitudes and habits on speeding behavior: Scale development and model building and estimation. *Accident Analysis & Prevention*, *39*(1), 6-15. doi:10.1016/j.aap.2006.05.011

Poulenez-Donovan, C. J., & Ulberg, C. (1994). Seeing the Trees and Missing the Forest: Qualitative Versus Quantitative Research Findings in a Model Transportation Demand Management Program Evaluation. *Transportation Research Record: Journal of the Transportation Research Board*, *1459*, 1-6.

Scott, J. (2000). Rational choice theory. *Browning, G., Halcli, A., & Webster, F. (eds.). Understanding contemporary society*. London, UK: SAGE publications.

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

Strauss, A. L., & Corbin, J. M. (1998). *Basics of qualitative research: techniques and procedures for developing grounded theory*. SAGE.

Trochim, W. M. K. (2006). Qualitative Validity. *Research Methods Knowledge Base*. Retrieved October 8, 2009, from http://www.socialresearchmethods.net/kb/qualval.php