UNIVERSITY OF TWENTE.

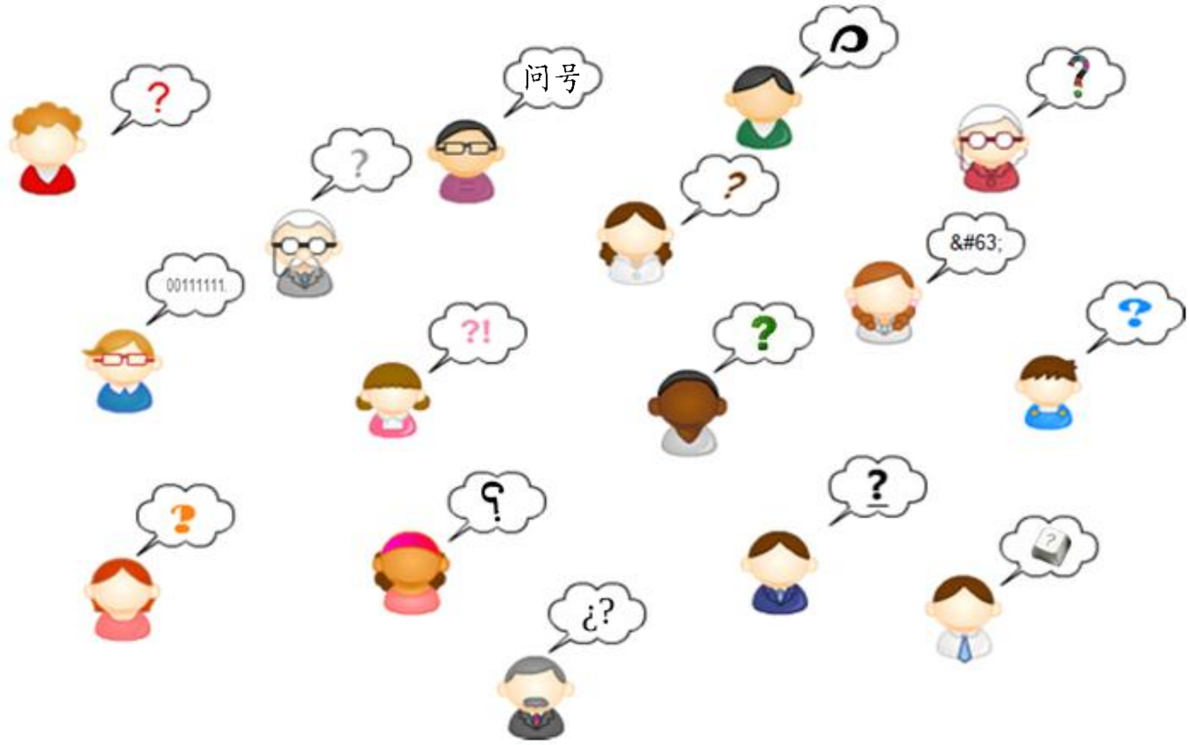
# DIFFERENCES BETWEEN USERS AND THEIR USER PROBLEMS

## A USABILITY TEST FOR WEHKAMP.NL

Inge Pluijlaar

University of Twente, The Netherlands - Enschede

11-2011



## Abstract

If it appears that different kinds of users experience different kinds of user problems, improving a website may depend on contradictory principles. Within this paper we report on a small scale study, which aimed to investigate (1) whether different kinds of users -concerning age, gender, education, internet experience and product experience- differ in the types of problems they experience and (2) whether it is enough to test only 5-7 users to find 85% of the different, existing usability problems. It appeared that at least some significant differences can be found on the experience of knowledge based problems. We also estimated that with 19 participants we had only uncovered around 81% of the existing problems. Implications and research challenges are included.

# 1. Introduction

The designs of websites should support the goals and needs of the end users. More and more professional creators and owners of websites are aware that when, in the development and evaluation process,  sufficient attention is given to usability, a number of benefits arise. For example, in e-commerce websites, better usability may lead to increased conversion, less costs for customer support and less development and redevelopment costs (Burnett & Ditsikas, 2006) (Bias & Mayhew, 2005).

Website usability refers to the amount of effectiveness, efficiency and satisfaction with which users achieve their specific goals, within a given context, using a website (ISO9241-11, 1998). There exist different definitions of usability, of which the one just mentioned, provided by the International Standardization Organization, may be the most commonly cited and generally accepted. Other definitions vary in whether they put the focus on usability as a quality attribute of a product or on usability as a profit arising from the use of the product (Bevan, 2001).  Although this is important to consider, more interesting than how we define usability, is how we determine the usability of a specific product, of which the final aim would be to discover how we can improve the product.

There are several methods for determining usability. In some of these methods predictions are made without the involvement of users. For example, in heuristic evaluation an expert on human-computer interaction reviews the product under study. However, involving end users in evaluation is recognized to be of great importance to really elicit and understand the actual problems that customers experience using a website (Burnett & Ditsikas, 2006).

Usability testing is one of many techniques that contribute to a good User-Centred Design (UCD). The basis of UCD lies in the principle that the user is positioned in the centre of the process, methods and procedures for designing usable websites (Rubin & Chisnell, 1994).

In a customary usability test an experimenter observes users that perform a number of representative tasks using the product (or website) under study. Various data is collected. Some common performance measures include (Rogers, Sharp & Preece, 2007):

- Time on task
- Number and type of user problem occurrences
- Number of users experiencing a particular problem
- Number of successful task completions

Experts and practitioners in the field of usability often recognise that there is no such thing as a ‘general user’ to which usability principles ‘generally’ apply. For example, Jacob Nielsen (2006) stresses:

“Anyone who's done user testing knows that there are tremendous individual differences among users.”

Another quote that supports this assumption derives from Steve Krug (2006):

“There is no Average User.”

This study aims to shed light on the individual differences between users in relation to their individual user problems. Different user characteristics might lead to different kinds of problems, which would also require different solutions.

## 2. Diversity in internet use

Diversity in Internet use has implications in several fields, such as governmental information and communications technology (ICT) deployment, e-commerce, (functional) web design, and usability testing methodology.

Many studies exist in which gender-, age-, cultural-, personality- and / or experience differences in specific internet uses are assessed. Several studies focus on different uses and usage motives of internet applications, like e-mail, entertainment, interpersonal communication through chat rooms and other social media, educational assistance, etc. They are often conducted through surveys. Other studies relate user characteristics to online performance. Although not complete, the next paragraphs  provide an overview of previous research on diversity in internet use.

### 2.1 Implications of diversity in Internet use

Since the rise of ICT, governments have seen opportunities to use the internet to share and spread information among citizens and offer accessible services to all society (Selwyn, 2004). There has been considerable debate on the inequalities of people’s access to the internet and their prowess in its use. Differences between users imply differences between the individual profits they may experience with regard to governmental ICT services. A phenomenon that is referred to as ‘the digital divide’(Helsper, 2010). Governments are concerned about discriminating underprivileged people, like elderly, disabled people, and people with lower socio-economic status. By understanding differences between users we would try to tackle the challenge to not disadvantage certain groups of the society in making use of digital governmental services.

Another field in which differences between patterns of use have important implications is e-commerce (Weiser, 2000). The internet as a medium has the advantage that users can be tracked through cookies. Many businesses already make use of this by matching relevant ad campaigns to users, according to the internet pages they visited and the keywords they searched for. This marketing strategy is referred to as ‘behavioural targeting’, or ‘personalized advertising’. When individual differences between users are better understood it becomes easier for publishers and advertisers to provide users with personally attractive deals.

Further challenges lie in conversion rate optimization (CRO) with respect to individual differences between users. For example, in a webshop different users might respond to different persuasive stimuli designed to elicit a purchase. When for every user stimuli are provided that answer to his individual preferences and desires, the overall revenue by a certain number of visits to the website may increase.

If it turns out that different users experience different types of user problems, interesting design implications emerge. There might be no ‘optimal’ design solution that supports all users. This would imply that functional and graphic designers should search for solutions that take different, perhaps even contradictory usability principles into account.

If user diversity influences usability test-results, and if the intended user group for a product is broad, the test-sample should also contain a great variety of users. There is an on-going debate on what number of test participants is enough to elicit the greatest amount of usability problems of a product. Opinions vary from five participants to 12, to the conclusion that the percentage of found problems can only be estimated after a specific test has been conducted (Schmettow, 2011). Examining the influence of diversity is an interesting addition to this discussion, because, if many differences exist between the types of problems encountered by different users, generally a larger amount of  participants would be required to uncover all problems.

## 2.2 Gender differences and internet use

Men and women traditionally have different attitudes towards the internet and use it for different purposes (Li & Kirkup, 2007). Although there seem to be some shifts, the trends consistently show that women's attitudes towards computers are more negative than those of men and that women use them less often than men. Women are also less self-assured in finding information on the Internet, even though their performance is no worse than that of men. Recent studies suggest that gender differences in the amount of internet usage are fading, but differences in motives and types of usage still exist.

For example, Weiser (2000) found that men used internet primarily for entertainment and leisure purposes. Women used it mainly for interpersonal communication and educational assistance. However, the findings of this particular study may be out-dated, for there have been considerable shifts in the popularity of internet and it's different applications. Weiser (2000) acknowledged that we should keep an eye on the future trends of differences in internet use patterns by different users.

More recently Helsper (2010) found that one of the few types of internet usage that is dominated by women were health-related activities. Unlike what has been found in earlier research, he did not find that women were more likely to use the internet for communication purposes. Helsper (2010) argues that even though the gender divide may decrease, differences between the internet use of men and women will continue to exist:

"Offline gender roles influence online behaviour like they do other behaviour, and this is likely to continue even when the current tech-savvy generation grows older."

He reinforces this statement with the finding that gender differences vary not only between generations, but also between different life stages in terms of employment and marital status.

## 2.3 Age differences and internet use

Together with gender and education, age is one of the most studied variables in digital divide research (Van Deursen & Van Dijk, 2009). Researchers try to uncover possible (cognitive) skill related differences between younger and older internet users. Freudenthal (2004) looked at age differences in relation to the capacity to retrieve information. Participants were asked to answer a couple of questions using a hierarchical menu structure. Several underlying cognitive psychological constructs were measured: movement speed, spatial ability, spatial memory, working memory capacity and reasoning speed. Elderly appeared to be slower than younger people on the overall task. Each step in the menu structure seemed to go with increasing differences in speed. Both movement speed, reasoning speed and spatial ability appeared to be of influence. Freudenthal (2004) concludes that the navigation of websites or other applications should not be designed as deep menu structures. Applying other methods to arrange various categories will help to avoid disadvantaging older people in using the application.

In a large-scale study Van Deursen and Van Dijk (2009) examined individual skill related problems that users experience while navigating the internet. They focused on four levels of internet skills: operational, formal, informational and strategic. Findings include that in particular, people of higher age experience more problems related to operational- and formal skill. They did not perform worse on the other two types of skills. Older participants even appeared to be better than younger participants at selecting relevant pages from search results. Therefore Van Deursen and Van Dijk (2009) recommend to look at differences in performance in a detailed way. The different characteristics of users may have both drawbacks and advantages that should be accounted for in the different fields of interest.

## 2.4 Cultural differences and internet use

People with different cultural backgrounds, living in different societies, might have different attitudes towards computers and the internet and use them differently (Li & Kirkup, 2007). Different studies have confirmed this hypthesis. For example, Li and Kirkup (2007) found significant differences between Chinese and British students in their attitudes towards and use of computers and internet. Chinese students had less prior experience with computers and were less likely to use computers for educational assistance than British students. However, the Chinese were more confident about their advanced computer skills. In both countries  men used computers more for e-mail and for playing computer games than women, and men were more self-assured in their computer use. Gender differences were greater among British students than for Chinese students.

## 2.5 Personality differences and internet use

Amiel and Sargent (2004) examined internet use and usage motives in relation to the personality types described by Eysenck and Eysenck (1985): psychoticism, extraversion and neuroticism. It turned out that neurotic participants used the internet for information purposes and to feel a sense of belonging. Extravert participants didn’t see the internet as communication medium and used it primarily as a tool for achieving certain goals. Participants scoring high on psychoticism were interested in ‘deviant, defiant and sophisticated’ internet usage.

In another study Burnett and Ditsikas (2006) conducted a usability test in which performance was compared over differences in personality. In this study they tried to ascertain whether extravert people undergoing a usability test reveal more usability problems than introvert participants do. The aim was for experimenters to be able to establish usability tests more efficiently. If it was shown that extravert people elicit more problems, using them as participants would help find more usability problems with less participants. Results showed that extravert participants revealed 40% more usability problems than introvert participants did.

Burnett and Ditsikas (2006) bring up that we then should consider not only what type of personality reveals most problems, but also if those problems qualitatively cover all problems experienced by the different types of users.

### 2.6 Differences in education as well as cognitive skills and internet use

People with lower levels of education and lower cognitive abilities generally show less proficiency in using the internet than highly educated people and people with more efficient perceptual skills and style do (Van Deursen & Van Dijk, 2009) (Kim, 2001) (Johnson, 2008) (Al-maskari & Sanderson, 2011). However, it appears that differences may decrease as people have more experience with the use of the internet (Kim, 2001). Johnson (2008) even put forward that the more people use the internet, the better their cognitive capacity. He argued that the internet functions as a tool that extends the cognitive processing abilities of people, and that by gaining more experience with this tool, the overall cognitive performance may be improved.

### 2.7 Differences in product experience and internet use

The characteristics, that determine who is a novice user of a specific website and who is an expert, are twofold. First, users differ in their experience and skills in general computer and internet use. The second factor that influences their level of expertise is the duration of, and frequency in use of a specific website. Faulkner and Wick (2005) acknowledge that categorising participants of a usability test on both these characteristics helps to uncover more, and more diverse usability problems. Dividing users on the basis of their expertise contributes to a good understanding of what these problems comprise of. As a result, better choices can be made on how and with what priority to improve elements of a website.

In general, experienced internet users have shown to outperform novice users (Van Deursen & Van Dijk, 2009) (Kim, 2001). In order to decrease this difference, in product design, sufficient attention should be directed at the learnability of the application. As it appears that experience may also influence the effects of other distinguishing characteristics, like cognitive skills, this further calls for proper intervention of computer courses in educational programs.

# 3. Research questions

## 3.2 Research question 1: diversity in users

This study is specifically interested in how differences between individual users result in differences between the types of problems they experience. To classify types of user problems we  relied on the Skill, Rule, Knowledge (SRK) based approach, provided by Rasmussen (1979). This framework helps to distinguish and understand the types of errors that occur in the interaction between human and (computer) system (Embrey & Lane, 1990). The different levels of information processing, ‘skill’, ‘rule’ and ‘knowledge’ based, vary in the degree conscious or automatic behaviour is applied. Skill based behaviour requires routine and little conscious awareness. In rule based behaviour people apply units of solutions from previous experiences to deal with new situations. Knowledge based information processing is required when no routines or rules are available. In this case a person’s interactions take place in a very conscious manner. Reason (1990) extended the SRK-approach in a detailed model describing how the different types of information processing are characterised and related. While in progress, people switch between the different levels of conscious behaviour. This model is known as the Generic Error Modeling System (GEMS).

The different types of information processing are each associated with certain types of human failure (Reason,1990). Errors that occur when an operator has the right intentions, but fails to deliver the right execution, are referred to as slips. Slips typically indicate skill-based problems. For example, a person intends to send an e-mail with attachment. He prepares the message, but then forgets to attach the document before sending the e-mail. In this case, although the operator knows very well how to attach a document to an e-mail, he fails to respond to the required deviation of his routine: ‘ad recipients, type subject, type message, send’. This specific type of skill-based error is referred to as ‘stereotype fixation’(Kirwan,1992).

In contrast to slips, mistakes are characterized by misconceptions (rule-based) and ignorance (knowledge-based) on what actions are required for the intended outcome. For instance, a particular online game includes two options: 'return' and 'pause'. After clicking on the pause-button, a gamer mistakenly assumes that, *if* he clicks the return button, *then* this will bring him back in the game, right where he left it. Instead the 'return'-button will start a new game. This can be regarded as a rule-based error. A knowledge-based error for example occurs, when a person tries to find some information on the internet, using a search engine, while he doesn't know what keywords on his topic will deliver good search results. Part of the error classification by Reason (1990) as is introduced above, is depicted in figure 1.

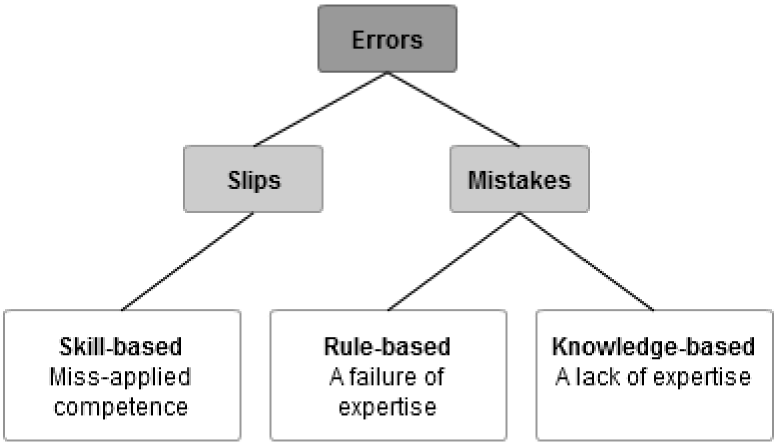


**Figure 1. SRK-classification of errors.**

Our goal was to investigate, whether recommendations following from usability test results account for all kinds of users, or whether usability examiners should consider structural differences between types of users in their analysis and advise.

*To what extend and in what way do differences between age, gender, education and experience between users relate to the kind of user problems (skill, rule, and knowledge based) experienced by these users?*

### 3.2 Research question 2: diversity in problems

As mentioned earlier in this paper, a frequently asked question in usability testing is:

*"How many users should one test to elicit all the different usability problems, that a specific product contains?"*

A debate on this topic started with Nielsen (2000), who argues that testing with only five users is enough. His assumption derives from applying a geometric series model, as introduced by Virzi (1992):

$$N(1-(1-L)^{n})$$

The model is meant to provide insight on what proportion of usability problems (N), is found with a given number of test participants (n), as a function of the average probability of uncovering a problem with a single participant (L) (Nielsen, 2000).

Nielsen and Landauer (1993) studied several separate usability tests. Averaging the varying L's over the different data led to the curve given in figure 2, which represents the proportion of usability problems that are found when testing a certain number of users. Based on this curve Nielsen (2000) argues that 15 users in a usability test are capable of bringing forward all existing usability problems. He adds that, in practice, it will be more efficient to test 5 users multiple times throughout the development and redevelopment of the product. Each time, these 5 users are claimed to discover around 85% of existing problems.

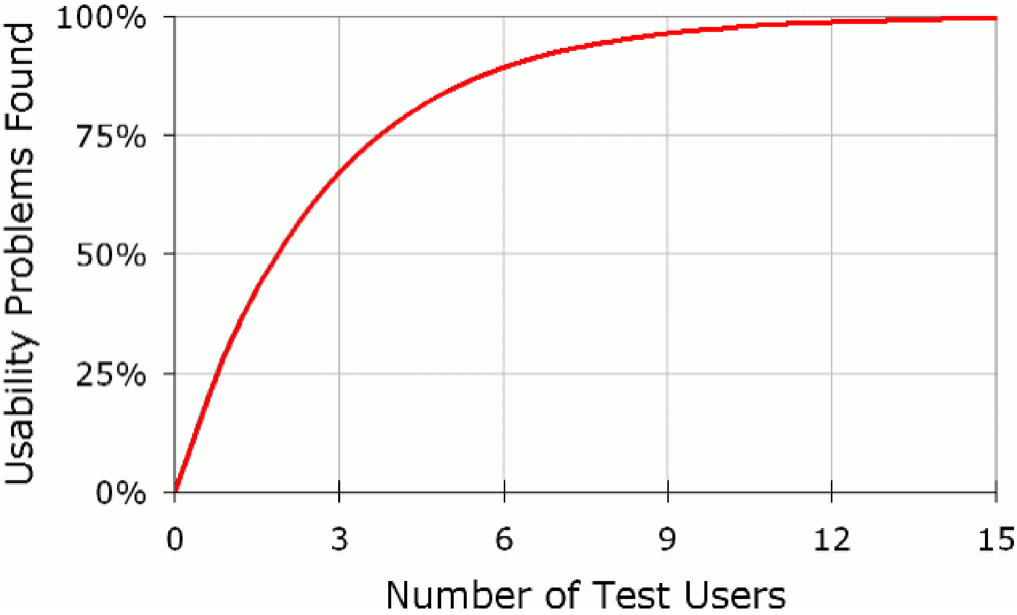


**Figure 2: The curve indicates what proportion of existing usability problems are found with a certain number of users that are included in a usability test. The curve derives from plotting the geometric series model with probability L=0,31, averaged over several studies. Adopted from Nielsen (2000).**

Nielsen's approach encounters various criticism. To start with, the mathematical model used may not be appropriate. In usability evaluation the fundamental underlying assumptions of the model are not fully met (Schmettow, 2011). One of the issues related to this argument is that not all usability problems are equally easy to discover; the data is not *homogeneous*. Furthermore, the observations are not *complete.* The number of unfound usability problems is typically unknown. The geometric series model does not take this into account. As a consequence the probability L is overestimated.

To overcome both the issue on homogeneity, and that on completeness, recently, Schmettow (2009) proposed an alternative for the geometric series model: the logit-normal binomial distribution ($LNB_{zt}$). In extension of the geometric series model the $LNB_{zt}$ introduces a prior distribution for the probability L. Comparing the two mathematical models Schmettow (2011) demonstrates that on several data sets the $LNB_{zt}$ model proves to be the better fit. This implies that discovering the majority of usability problems acquires much more test participants than assumed earlier, based on the geometric series model.

The number and kind of usability problems vary great depending on the nature of the evaluated application (Spool and Schroeder, 2001). Lewis (2011) therefore recommends to have no presumptions on how many users you need to test in advance of a specific product evaluation. An alternative is to first test a product with a couple of participants, and then, based on the data, estimate the number of users needed to elicit, say at least 80% of the existing usability problems. Afterwards an experimenter can evaluate whether this target was actually met.

Diverse users might also differ substantially in the kind and amount of problems they experience. In this light, testing users with distinct characteristics demands a larger sample size than proposed by Nielsen (2000). In this study we typically test a sample of users with distinct characteristics. Reasoned from the original standpoint of Nielsen we would hypothesize that:

*With 5-7 test participants we find 85% of problems*.

The second aim of this study is to investigate whether this applies to our data.

# 4. Method

## 4.1 The product under study

The object chosen for this study was one of the Dutch leading e-commerce websites: wehkamp.nl. Wehkamp.nl is an online department store selling a great variety of goods in fashion, living and hardware. The webshop is a pioneer in the field of Dutch e-commerce, has won several prices and has received certificates from ‘Stichting Certificering Thuiswinkel Waarborg’ and ‘Stichting Waarmerk drempelvrij.nl’. These foundations check compliance with general conditions as set by the Consumers Association and the accessibility of a web application, respectively.

To reduce the time needed for an in depth evaluation we chose not to test the whole website, but to focus on the product detail page as entry point of the checkout process.

## 4.2 Participants

Nineteen people took part in the usability test, of whom seventeen were existing customers of wehkamp.nl and two had never visited the website before. The customers of wehkamp.nl varied in the frequency they visited wehkamp.nl. Overall four participants visited wehkamp.nl less than once per month, eleven participants visited the webshop one to four times per month, and four participants visited wehkamp.nl more than four times per month.

Five subjects used the internet one hour or less a day. Eight subjects used the internet two to four hours a day. Six subjects used the internet more than four hours a day. Only one participant had never bought something online.

Fourteen of the participants were female, five were male. Age ranged from 24 to 61 years old, with an average of 42. Among the participants ten were educated on or above the Dutch level HBO (higher vocational education), while nine were educated below HBO.

## 4.3 Data gathering

A usability test often consists of three components (Rogers, Sharp and Preece, 2007) . (1) A (semi) structured interview, often held in advance of the usability test, can be useful to depict part of the context of use. Components (2) is the user test and component (3) a user satisfaction questionnaire.

### 4.3.1 Structured interview

Information about the participants was retrieved through a short structured interview. Rubin (1994) provides an overview for the selection and acquisition of participants. Among other things he describes how to set up a user profile. Rubin recognizes that specific characteristics which make up the user profile depend on the product. He anyhow provides the categories for a generic user characterization for a typical computer-based product. Using his ‘generic user characterization’ we formulated wehkamp.nl specific interview questions (appendix 1), divided in five topics:

1. Personal History
2. Educational History
3. Occupational History
4. Computer Experience
5. Product Experience

### 4.3.2 Usability test

A usability test measures the performance of users on specific tasks that should be representative of common user goals. With such a test, user problems concerning product effectiveness, efficiency and satisfaction can be brought into view. The usability test was conducted applying the Think Aloud Method (TAM).

*The Think Aloud Method*

The TAM is a method for uncovering cognitive processes. The basis for the TAM in usability testing lies within the classic writing about protocol analysis from Ericsson and Simon (1993). They discussed the use of introspective data in the study of task directed cognitive behaviour.

Especially the Concurrent Think Aloud Method (report of immediate thought) is widely used for depicting the behaviour and thought processes of users and analysis of occurring user problems within usability tests (Nielsen & Carsten, 2004).

Ericsson and Simon (1993) identified three levels of verbalization. These levels vary in the degree to which cognitive processes are needed to transform the thoughts into words before they are spoken. In level 1 verbalization one expresses his direct thoughts. In level 2 verbalization a single process takes place between short-term memory and verbalization; images or abstract concepts need to be transformed into words. In level 3 verbalizations there exist more cognitive demands than just those required for task performance and verbalization; a person must reflect on his own cognition, or retrieve information from long-term memory.

In the Concurrent Think Aloud Method, which was used in this study, participants are encouraged to express their direct thoughts going through the tasks. This way, level 1 and level 2 verbalizations are collected. Level 3 verbalizations are argued to be less reliable, since people are reported not to be able to correctly report on their own cognitive processes (Boren & Ramey, 2000).

*Task Selection*

A critical procedure in designing the Think Aloud usability test is the selection of tasks (Rubin, 1994). It is important to especially consider the representativeness of the task with respect to realistic user goals. Some points of interest were outlined, which helped to shape the tasks.

The tables in appendix 2 provide an overview per task, including the task description, the Uniform Resource Locator (URL) of the page on which the task starts, the criteria for successful completion, maximum duration of the task and the points of interest the task covers. As the test progress the tasks become more specific. For example, task descriptions one and three are as follows:

Task 1: “Find a new winter coat you like on this product overview page and order it.”

Task 3: “You want a new couch. Use the information and functionality on this product detail page to shape your opinion on this product. Order the couch if you have gained a positive impression of the product.”

*Setup*

We instructed the participants by reading a written directive together (see appendix 2 for the text of this directive). Herewith we guaranteed consistency within the subjects information at the start of the test. Contents of this directive were based on the procedure as summarized below.

There are differences in peoples capabilities to verbalise their thoughts. To overcome this problem we held a practice round before starting the actual tasks, in which we could stimulate the subjects to verbally express their thoughts more and better.

The test must take place in setting in which the subject feels at ease. The test sessions took place in a room in a behavioural science lab to be assured of a quiet environment. We explained participants that the object under study was the website and not them, and therefore there was no way for them to perform well or badly on the test.

Behaviour and prompting of an experimenter can influence test results (Boren & Ramey, 2000). We therefore only interrupted the participants during the tasks when they stopped talking or when the participants had questions or misconceptions regarding the tasks.

While the participants conducted the test, screen and audio were recorded. Tasks were performed on an Windows PC, using Internet Explorer as browser. Screen and audio were recorded with the freeware Camstudio.

### 4.3.3 Satisfaction questionnaire

The System Usability Scale (SUS) is a short and easy ten-item questionnaire (see appendix 3) with which participants can indicate how much they liked using the product (Brooke, 1996). The questionnaire uses a five-point Likert scale.

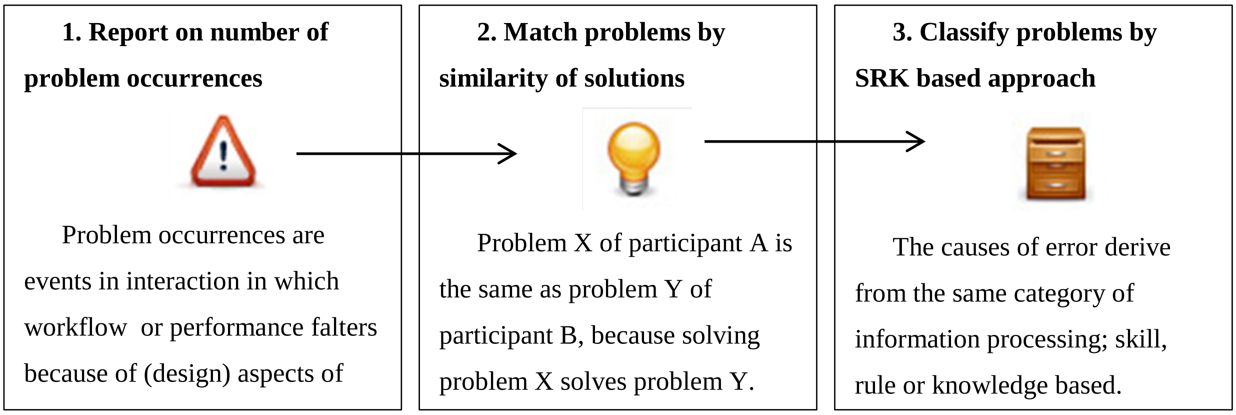
Because the participants were Dutch, we had to translate the English questionnaire. To validate the translation we had our Dutch version translated back to English by a resident of the United Kingdom. We then compared this translation with the original questionnaire. As this translation resulted very similar to the original, only a few adjustments were made regarding the tenses used.

## 4.4 Data analysis

### 4.4.1 Problem identification, matching and classification

For structuring the findings we used Techsmith Morae Manager, a software for analysing customer experience research data. Through this, we analysed the videos of the separate test-sessions.

The identification, matching and classification of user problems proceeded in three steps, as summarized in figure 4.1. First we reported on all the events in which the workflow of a participant was influenced by any kind of obstacle. Second, in order to know which identified problem occurrences were similar and which were not, a matching strategy needed to be applied. We matched problem descriptions by the similarity of solutions to the problem (Hornbæk & Frøkjær, 2008). We could then classify the problems according to the SRK based approach.



**1. Report on number of problem occurrences**

Problem occurrences are events in interaction in which workflow or performance falters because of (design) aspects of

**2. Match problems by similarity of solutions**

Problem X of participant A is the same as problem Y of participant B, because solving problem X solves problem Y.

**3. Classify problems by SRK based approach**

The causes of error derive from the same category of information processing; skill, rule or knowledge based.

**Figure 3: Problem identification, matching and classification**

The third, and last step in the process of clarifying the found problems was conducted by moving along two specific questions regarding the problem:

1. In what situation is the problem occurring?

   If the *situation is familiar* to the user he can respond to the demands for achieving a certain goal in an automatic, routine based way. If this is the case the user is in a state of skill-based information processing.

   If the *situation deviates* from the familiar circumstances and a behavioural adjustment of the user is required, then this normally indicates a state of rule-based processing.

   If the *situation is new* to the user this calls for gathering the right information and feedback to make sense of what should be the next step.

2. In what way does the behaviour of the user deviate from the desired response?

   This last question helps to make sense of how the user failed to rightly apply the required strategy in a specific situation. An error occurs when the user doesn’t respond appropriately to the situation at hand. In every level of processing there are a number of different errors that may occur that will lead users away from achieving their goals.

The underlying system we applied to support us in answering the questions above is given in figure 4.2. It comprises of a flowchart that helps to identify the flaws in the cognitive psychological process in interaction with a specific product (Rasmussen, Pedersen, Carnino, Griffon, Mancini, Gagnolet, 1981). In this way, we could assign one of three SRK problem categories to a problem. In the end, a subjects’ score on one of the problem categories was determined by dividing the number of occurrences of problems in that category by the total number of problems in that category.

### 4.4.2 Analysis of variance

A multivariate ANOVA was applied including four independent variables, one covariate and five dependent variables.

Independent variables:

1. gender
2. education,
3. internet experience
4. product experience

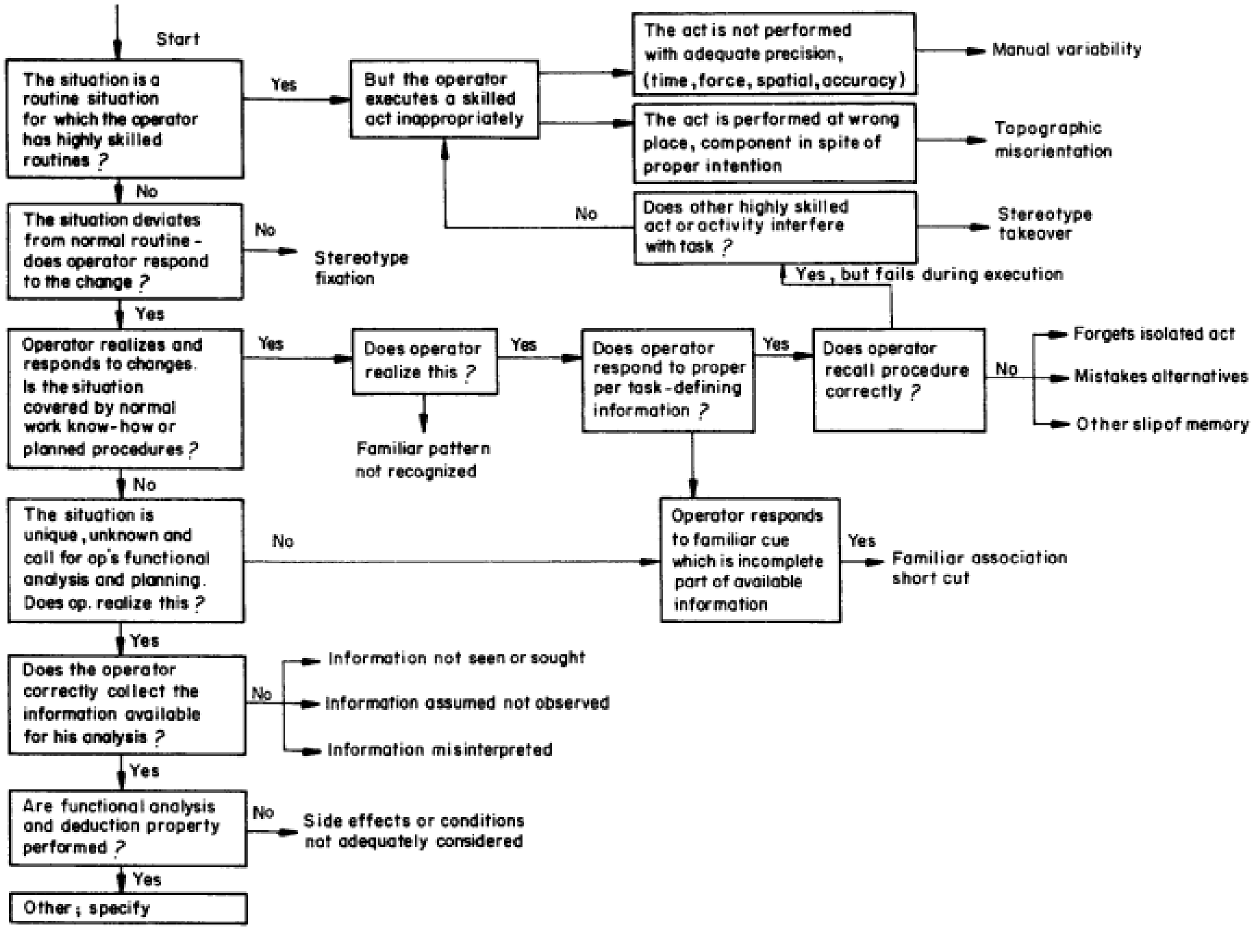

Start
The situation is a routine situation for which the operator has highly skilled routines ?
Yes
But the operator executes a skilled act inappropriately
The act is not performed with adequate precision, (time, force, spatial, accuracy)
Manual variability
The act is performed at wrong place, component in spite of proper intention
Topographic misorientation
No
Does other highly skilled act or activity interfere with task ?
Stereotype takeover
Yes, but fails during execution
No
The situation deviates from normal routine - does operator respond to the change ?
No
Stereotype fixation
Yes
Operator realizes and responds to changes. Is the situation covered by normal work know-how or planned procedures ?
Yes
Does operator realize this ?
Familiar pattern not recognized
Yes
Does operator respond to proper per task-defining information ?
Yes
Does operator recall procedure correctly ?
No
Forgets isolated act
Mistakes alternatives
Other slip of memory
No
The situation is unique, unknown and call for op's functional analysis and planning. Does op. realize this ?
No
Operator responds to familiar cue which is incomplete part of available information
Yes
Familiar association short cut
Yes
Does the operator correctly collect the information available for his analysis ?
No
Information not seen or sought
Information assumed not observed
Information misinterpreted
Yes
Are functional analysis and deduction property performed ?
No
Side effects or conditions not adequately considered
Yes
Other ; specify

**Figure 4: SRK error flowchart, adopted from Kirwan (1992).**

Covariate:

1. age

Following from earlier research we would expect that age (and life stage) is of influence on the test results. However, in this study we have to little subjects in the separate age groups to make noteworthy comparisons. To account for possible age related influences we included the continuous variable age as covariate.

Dependent variables:

1. the score on skill based problems
2. the score on rule based problems
3. the score on knowledge based problems
4. average time on tasks
5. the score on the SUS

### 4.4.3 SUS score

To calculate the SUS score, we summed the score contributions from each of the ten items. The score contribution of each item ranges from 0 to 4. Then we multiplied the sum of the scores by 2.5 to obtain the overall value of the SUS. The SUS scores have a range of 0 to 100. Note that, the lower the score, the better the usability of the product is judged.

### 4.4.4 LNBzt distribution

In order to know whether the usability test has revealed the majority of existing problems we estimated the found percentage of problems through the $\text{LNB}_{zt}$ distribution, as introduced by Schmettow (2009). Because the model has a complex mathematical basis it's components are only briefly summarized below.

The LNB probability distribution function  (pdf) reads as follows:

$$\text{pdf}_{LNB}(x; n, \mu, \sigma^2) = \binom{n}{x} \frac{1}{\sqrt{2\pi\sigma^2}} \int_0^1 (1-p)^{n-x-1} p^{x-1} \mathrm{e}^{-\frac{(logit(p)-\mu)^2}{2\sigma^2}}\, dp$$

Here, p stands for the probability of finding a problem with a single test participant. The unknown parameters $\mu$ and $\sigma^2$ determine the $N(\mu,\sigma^2)$ normal distribution of *p*. Parameter *x* stands for the number of observed usability problems. Where *n* indicates the number of users participating in the test.

In $\text{LNB}_{zt}$ the '$_{zt}$' stands for 'zero-truncation', which functions to estimate the number of still undiscovered usability problems:

$$\text{pdf}_{zt}(x; \pi\cdot) = \begin{cases} 0 & x = 0 \\ \frac{\text{pdf}(x;\pi\cdot)}{1-\text{pdf}(0;\pi\cdot)} & x > 0 \end{cases}$$

Central to the zero-truncation pdf function is the discrete random variable $X \in \{0,\ldots,n\}$, which consists of the number of times any problem is detected. It is distributed as $P(X = x|\pi) = \text{pdf}(x; \pi\cdot)$ (where $\pi\cdot$ are the model parameters). The $\text{pdf}_{zt}$ is obtained by setting the probability counts with X = 0 to zero and by readjusting the probability mass to one. The parameters $\pi\cdot$ of the $\text{pdf}_{zt}$ are estimated via the maximum likelihood method. Then, the number of unfound problems is estimated with the non-truncated pdf and the estimated parameter $\hat{\pi}$ :

$$P(X = 0|\pi\cdot) = \text{pdf}(0; \hat{\pi}\cdot)$$

# 5. Results

## 5.1 Problem identification, matching and classification

The 19 subjects revealed 147 problem occurrences with respect to the product detail related pages of wehkamp.nl. Matching those problem occurrences, in accordance to the 'similar solutions method', put forward 35 separate problems. Of those 35 problems 12 were skill based, 8 were rule based and 15 were knowledge based. See Table 5.1.1 for three examples of how the different problems were identified and classified.

**Table 1. Example of problem classification for a skill, rule and knowledge based problem.**

| Category | Problem description | Solution | Quote of participant |
|---|---|---|---|
| Skill | The participant automatically reads review grade as of 10-point, instead of 5-point scale. | Use other scale, show only in stars, or replace grade with a thumbs up or thumbs down icon. | "General review. Oh, that is not very high. […] Oh, wait! 'Four point six', that ís high, because there are only five stars." |
| Rule | The participant believes that answers to questions come from wehkamp.nl instead of other users and doesn't expect a (quick) answer to the question. | Provide more feedback on what is the intention of the possibility to place a question on the product detail page. | "At 'reviews', you can ask a question, but then you will actually not get a quick answer, will you?" |
| Knowledge | The page-item 'share with your friends' is never used before, hard to find and/or the participant would do it by copying the url into chat or e-mail. | More prominent position and clear indication of page item 'share with your friends'. | "To be honest, that was positioned very small. I only saw it at the last moment." |

## 5.2 Multivariate analysis of variance

Through a multivariate ANOVA F-test we could assess whether the test participants with different characteristics, also differ with regard to (1) their score on the different SRK-classes of error, (2) the average time they needed to fulfil the tasks and (3) to their score on the SUS questionnaire. In table 2. some statistics related to the test are summarized for the skill, rule and knowledge based problem experience, including the mean (M) scores of the separate groups, based on the modified population marginal mean, the standard errors (SE), the F-value (F) and the *p*-value (Sig.). Table 3. displays this data for the average time on task and the SUS score, whereas table 4. reports on the possible interaction effects for the skill, rule and knowledge based problems. Starting from an α-level of at least 0,1 we found some significant evidence to assume that the mean between the separate groups are not entirely the same with respect to 'Knowledge based problems' and the 'SUS score'. With regard to Knowledge based problem experiences, we found a significant effect among gender (p=0,082), internet experience (hours of use per day) (p=0,005), product experience (frequency of use per month) (p=0,015), and interaction effects of gender x product experience (p=,019), education x internet experience (p=,027), and internet experience x product experience (p=,010). The SUS-score varies significantly among gender (p=,054) and product experience (p=,043). In relation to the other dependents no significant variance was found between the separate groups of subject characteristics. We could also not make any claims on possible direct effects of age, because this factor was only significant as covariant.

**Table 2. Summary of statistics and significance of between-subject-effects tests for the dependent variables skill, rule and knowledge.**

| | Skill | | | | Rule | | | | Knowledge | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | M | SE | F | Sig. | M | SE | F | Sig. | M | SE | F | Sig. |
| Gender | | | 3,867 | ,188 | | | ,030 | ,878 | | | 10,730 | ,082+ |
| Male | ,280 | ,032 | | | ,202 | ,056 | | | ,286 | ,004 | | |
| Female | 0,136 | ,018 | | | ,173 | ,032 | | | ,285 | ,002 | | |
| Education | | | ,010 | ,928 | | | 1,297 | ,373 | | | ,007 | ,940 |
| High | ,200 | ,021 | | | ,220 | ,037 | | | ,337 | ,003 | | |
| Low | ,166 | ,020 | | | ,153 | ,035 | | | ,245 | ,002 | | |
| Internet Experience | | | ,633 | ,612 | | | 3,313 | ,232 | | | 208,697 | ,005** |
| <1-1 | ,198 | ,029 | | | ,238 | ,049 | | | ,324 | ,003 | | |
| 2-4 | ,179 | ,024 | | | ,090 | ,041 | | | ,199 | ,003 | | |
| >4 | ,167 | ,028 | | | ,236 | ,047 | | | ,351 | ,003 | | |
| Product Experience | | | 1,395 | ,418 | | | 8,066 | ,110 | | | 64,752 | ,015* |
| <1 | ,246 | ,031 | | | ,340 | ,053 | | | ,370 | ,004 | | |
| 1-4 | ,159 | ,023 | | | ,143 | ,040 | | | ,268 | ,003 | | |
| >4 | ,161 | ,034 | | | ,101 | ,059 | | | ,235 | ,004 | | |

+ Effect is significant at the 0,1 level

* Effect is significant at the 0,05 level

** Effect is significant at the 0,01 level

**Table 3. Summary of statistics and significance of between-subject-effects tests for the dependent variables average time on tasks, and the SUS-score. Possible interactions with no significant effects are excluded.**

| | Average Time on Task | | | | SUS-score | | | |
|---|---|---|---|---|---|---|---|---|
| | M | SE | F | Sig. | M | SE | F | Sig. |
| Gender | | | ,017 | ,907 | | | 17,205 | ,054+ |
| Male | 283,093 | 45,987 | | | 22,487 | 2,731 | | |
| Female | 214,304 | 26,133 | | | 10,527 | 1,552 | | |
| Education | | | ,033 | ,873 | | | ,079 | ,806 |
| High | 276,722 | 30,394 | | | 22,221 | 1,805 | | |
| Low | 203,973 | 28,987 | | | 8,076 | 1,721 | | |
| Internet Experience | | | ,599 | ,626 | | | 6,614 | ,131 |
| <1-1 | 292,715 | 40,741 | | | 17,504 | 2,420 | | |
| 2-4 | 204,144 | 34,022 | | | 10,055 | 2,021 | | |
| >4 | 216,874 | 39,158 | | | 16,076 | 2,326 | | |
| Product Experience | | | 2,467 | ,288 | | | 22,500 | ,043* |
| <1 | 302,778 | 43,647 | | | 30,939 | 2,592 | | |
| 1-4 | 222,116 | 32,977 | | | 7,16 | 1,958 | | |
| >4 | 196,193 | 48,598 | | | 11,780 | 2,886 | | |
| Gender x Product Experience | | | ,005 | ,951 | | | 13,199 | ,068+ |
| M <1 | 362,216 | 61,692 | | | 30,412 | 3,664 | | |
| 1-4 | 236,061 | 88,843 | | | 10,992 | 5,276 | | |
| >4 | 218,912 | 87,074 | | | 29,630 | 5,171 | | |
| F <1 | 243,340 | 61,544 | | | 31,466 | 3,655 | | |
| 1-4 | 217,467 | 31,876 | | | 5,895 | 1,893 | | |
| >4 | 188,620 | 58,719 | | | 5,830 | 3,487 | | |

+ Effect is significant at the 0,1 level

* Effect is significant at the 0,05 level

** Effect is significant at the 0,01 level

**Table 4. Summary of statistics and significance of interaction-effects for the dependent variables skill, rule and knowledge. Possible interactions with no significant effects are excluded.**

| | | Skill | | | | Rule | | | | Knowledge | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | M | SE | F | Sig. | M | SE | F | Sig. | M | SE | F | Sig. |
| Gender x Product Experience | | | | ,744 | ,479 | | | ,111 | ,771 | | | 50,761 | ,019* |
| M | <1 | ,327 | ,043 | | | ,369 | ,074 | | | ,404 | ,005 | | |
| | 1-4 | ,286 | ,062 | | | ,071 | ,107 | | | ,214 | ,007 | | |
| | >4 | ,172 | ,061 | | | ,130 | ,105 | | | ,196 | ,007 | | |
| F | <1 | ,164 | ,043 | | | ,310 | ,074 | | | ,336 | ,005 | | |
| | 1-4 | ,117 | ,022 | | | ,168 | ,038 | | | ,286 | ,003 | | |
| | >4 | ,157 | ,041 | | | ,091 | ,071 | | | ,247 | ,005 | | |
| Education x Internet Experience | | | | 1,648 | ,378 | | | 1,173 | ,460 | | | 35,937 | ,027* |
| H | <1-1 | ,275 | ,038 | | | ,231 | ,065 | | | ,361 | ,004 | | |
| | 2-4 | ,123 | ,049 | | | ,135 | ,084 | | | ,284 | ,006 | | |
| | >4 | ,167 | ,045 | | | ,288 | ,076 | | | ,355 | ,005 | | |
| L | <1-1 | ,083 | ,043 | | | ,250 | ,074 | | | ,267 | ,005 | | |
| | 2-4 | ,207 | ,038 | | | ,068 | ,066 | | | ,156 | ,005 | | |
| | >4 | ,168 | ,047 | | | ,201 | ,081 | | | ,348 | ,006 | | |
| Internet Experience x Product Experience | | | | 5,697 | ,153 | | | 1,072 | ,516 | | | 99,426 | ,010** |
| <1-1 | <1 | ,422 | ,061 | | | ,505 | ,105 | | | ,463 | ,007 | | |
| | 1-4 | ,136 | ,038 | | | ,148 | ,065 | | | ,295 | ,004 | | |
| | >4 | ,161 | ,061 | | | ,244 | ,105 | | | ,271 | ,007 | | |
| 2-4 | <1 | ,063 | ,086 | | | ,155 | ,148 | | | ,338 | ,010 | | |
| | 1-4 | ,252 | ,042 | | | ,085 | ,072 | | | ,153 | ,005 | | |
| | >4 | ,128 | ,043 | | | ,065 | ,074 | | | ,198 | ,005 | | |
| >4 | <1 | ,249 | ,049 | | | ,349 | ,084 | | | ,339 | ,006 | | |
| | 1-4 | ,056 | ,041 | | | ,225 | ,070 | | | ,402 | ,005 | | |
| | >4 | ,229 | ,086 | | | ,030 | ,148 | | | ,271 | ,010 | | |

* Effect is significant at the 0,05 level

** Effect is significant at the 0,01 level

Now that we have reason to assume that there exist some effects, we are interested in what these effects comprise of[1]. Based on comparing the separate mean scores of knowledge based problems and the SUS (see also the profile plots in figure 7), we suspect that the effects exist in the following directions:

[1] Notice that for internet-experience and product experience the F-test does not tell us what group differs significantly from the other.

1. Men experience slightly more knowledge based problems than women.
2. People that spend very little time on the internet (less than one hour a day) and people that spend a lot of time on the internet (more than four hours a day) experience more knowledge based problems than people that spend a medium amount of time a day on the internet (1-4 hours).
3. People that visit wehkamp.nl more often, experience less knowledge based problems.
4. Men generally score higher on the SUS than women do, which means that they are less satisfied with the product.
5. People that visit wehkamp.nl more frequently, score lower on the SUS, which means they are more satisfied with the product.

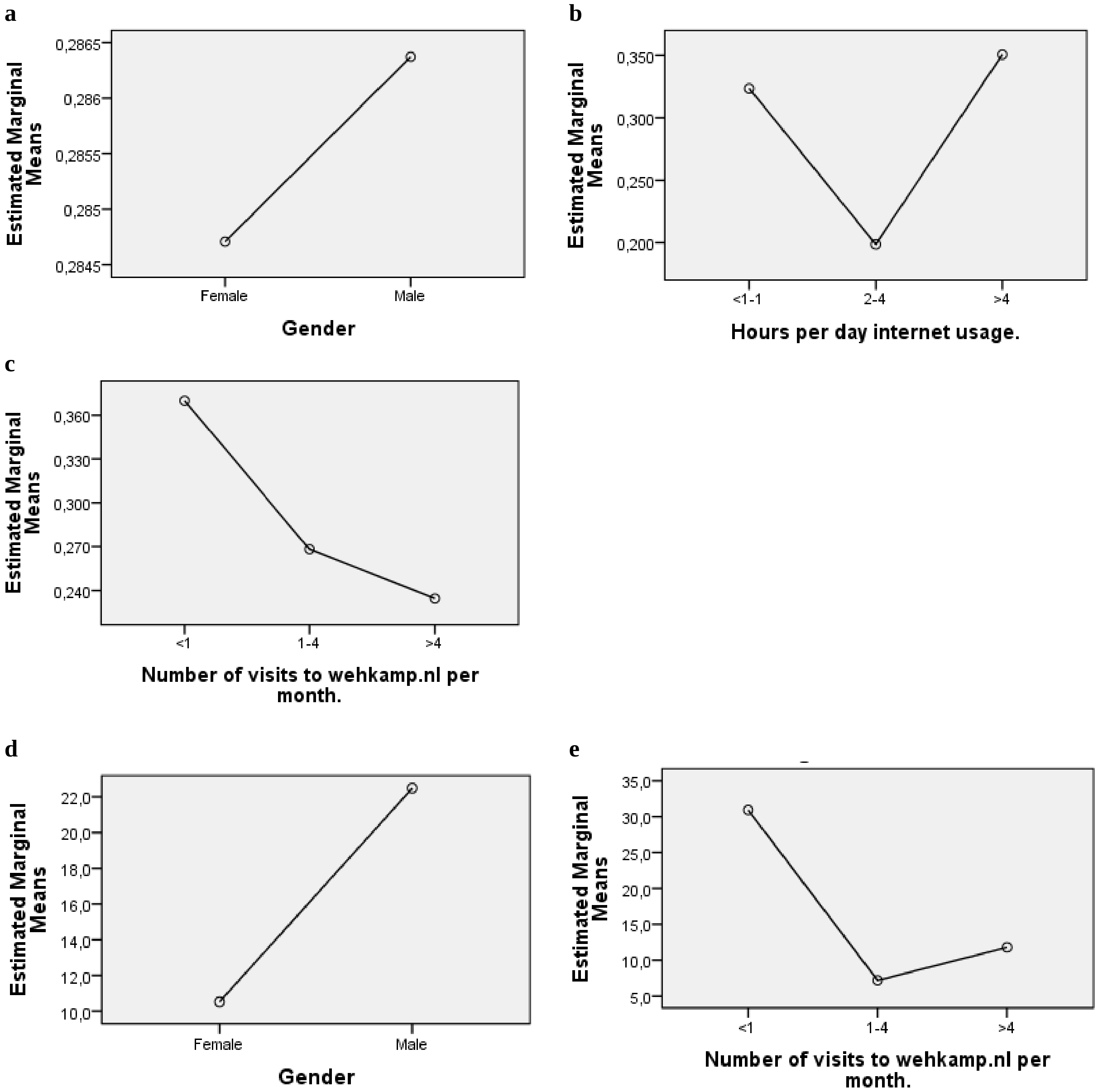


**Figure 8. Estimated marginal means of Knowledge based problems (a, b and c) and of SUS-score (d and e).**

On the basis of our findings we also hypothesize the following interaction effects (see figure 7. and 8. for the corresponding profile plots):

1. Female paying less frequent visits to wehkamp.nl experience less knowledge based problems than man paying less frequent visits to wehkamp.nl, whereas women paying more than four visits to wehkamp.nl per month experience more problems than men with the same amount of experience with the product.
2. There is a smaller difference with respect to knowledge based problems between the lower and higher educated, as the amount of internet experience they have increases.
3. People with a medium amount of product experience (visiting wehkamp.nl 1-4 times a month) vary the most in their performance over their level of experience with internet use.
4. Men that visit wehkamp.nl more often are considerably less satisfied with the product than women who have much product experience (the difference is less for the other levels of product experience).

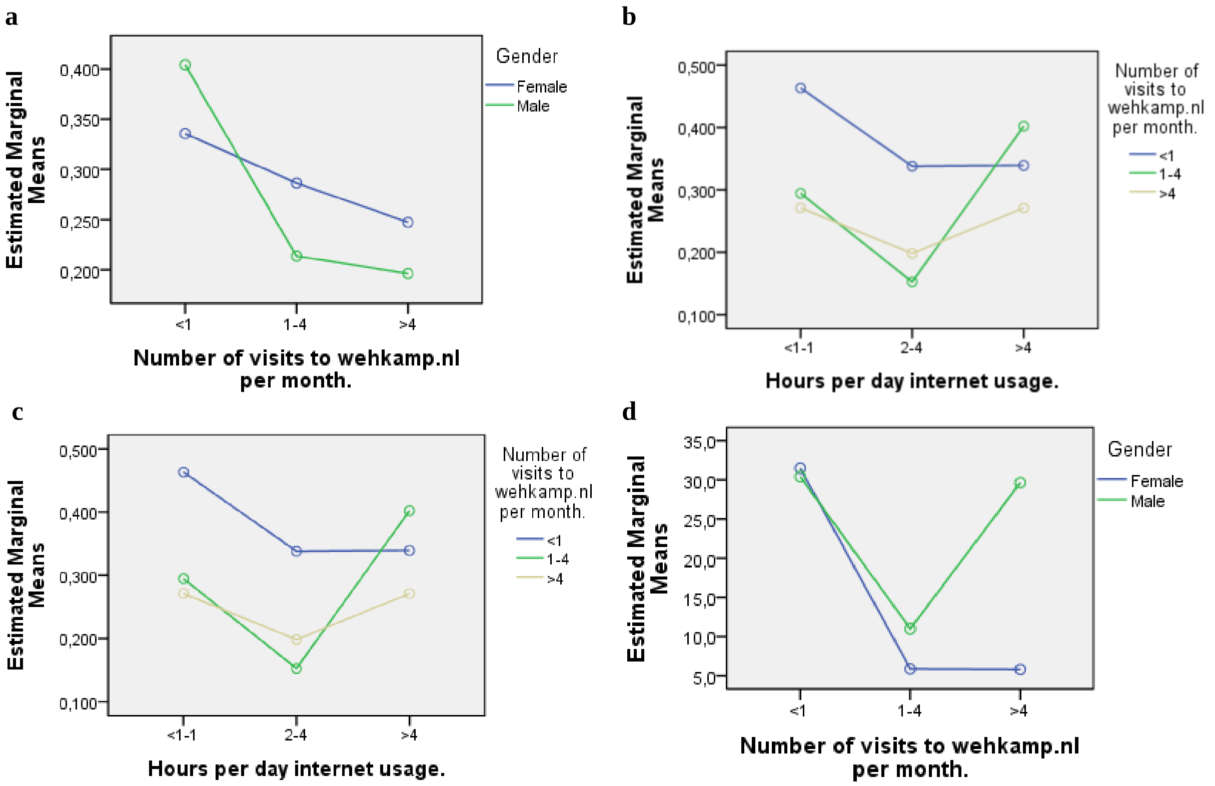


**Figure 9. Estimated marginal means of Knowledge based problems (a, b and c) and SUS-score (d) for interaction effects.**

*Normality of dependent variables*

For the results and hypothesis above to be of interest, it is important to consider whether our small sample is actually normally distributed with regard to our different independent variables. Applying the  Shapiro-Wilk test reveals that there is significant evidence that both ‘Skill based problems’ (*p*=0,155) and ‘Knowledge based’ (*p*=0,07) are normally distributed, whereas ‘Rule based problems’, the ‘Average time on tasks’ and the ‘SUS’-score are not (see Table 5). This implies that, for knowledge based problems, our analysis method was appropriate and therefore our findings may be relevant. Figure 10 reveals that the histogram and normal Q-Q plot for knowledge based problems still do not show an absolute perfect fit. However, the findings we included on the SUS-score are clearly much less accurate, as can be seen in figure 11. We should thus be careful with basing any general conclusions on our analysis of variance, since this test assumes that the dependent variables are normally distributed.

**Table 5: Tests of normality. The null hypothesis is tested that the sample is normally distributed. A *p*-level below the chosen α-level (0,05) rejects this claim.**

| | Shapiro-Wilk test | | |
|---|---|---|---|
| Dependent variable | Statistic | df | Significance |
| Skill based problems | ,927 | 19 | ,155 |
| Rule based problems | ,898 | 19 | ,044 |
| Knowledge based problems | ,909 | 19 | ,070 |
| Average time on tasks | ,795 | 19 | ,001 |
| SUS | ,745 | 19 | ,000 |

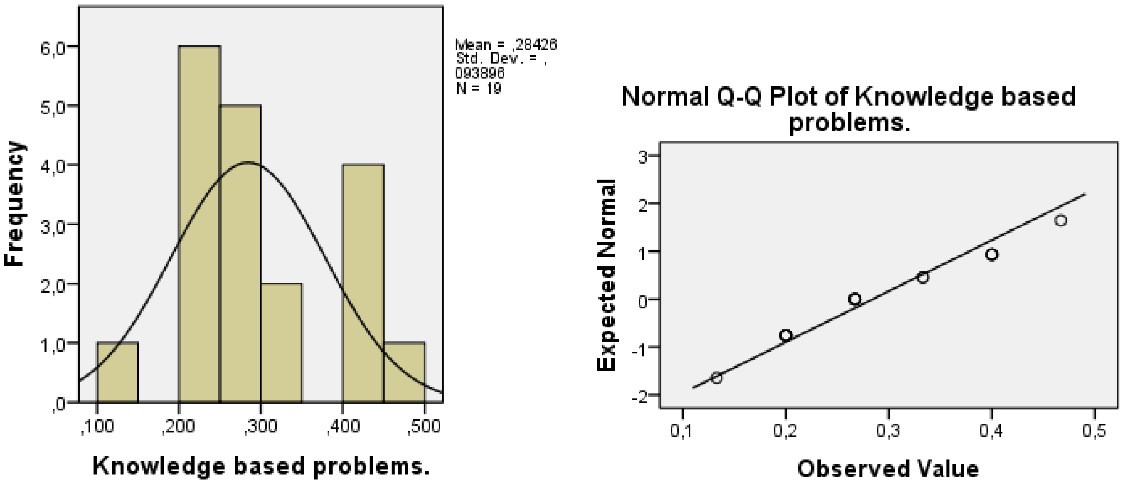


**Figure 10. Histogram with normal curve and normal Q-Q plots for Knowledge based problems.**

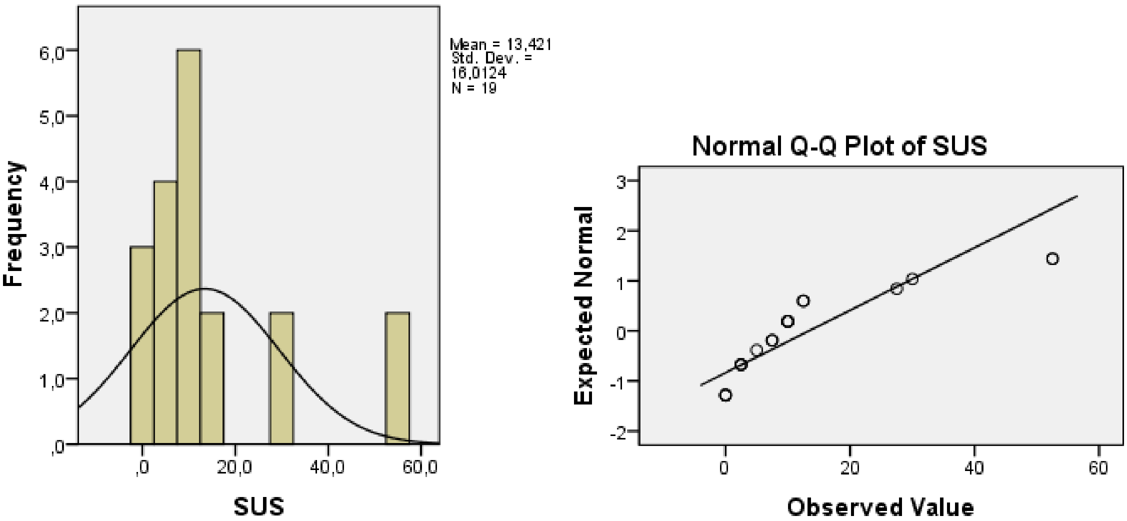


**Figure 11. Histogram with normal curve and normal Q-Q plots for SUS-score.**

### 5.3 Study progress

We estimated that the 35 problems we found, accounted for about 81% of total usability problems regarding the pages we have examined. The $LNB_{zt}$ distribution for our sample is displayed in figure 5, where the x-axis represents the number of times the same problem was discovered with separate users, and the y-axis encompasses the frequency with which problems were discovered a certain number of times. For example, you can read from the graph, that in our data there are five problems that were experienced by four separate users. According to the $LNB_{zt}$ model the estimated number of undiscovered problems (0 on the x-axis) is 8.

Corresponding to our data, the number of test participants, that we expect to detect a certain percentage of problems, is plotted in figure 6. In our specific study a confidence interval tells us that it is 90% likely that the probability of problems we discovered lies at least in between 62% and 94%. The curve and confidence range also tell us that testing with only five to seven users would only give us insight in 37%  to 72% of existing usability problems. The prediction of Nielsen that five user are enough to elicit 85% of usability problems does clearly not apply to our study, in which we considered that the data is not complete and not homogenous.

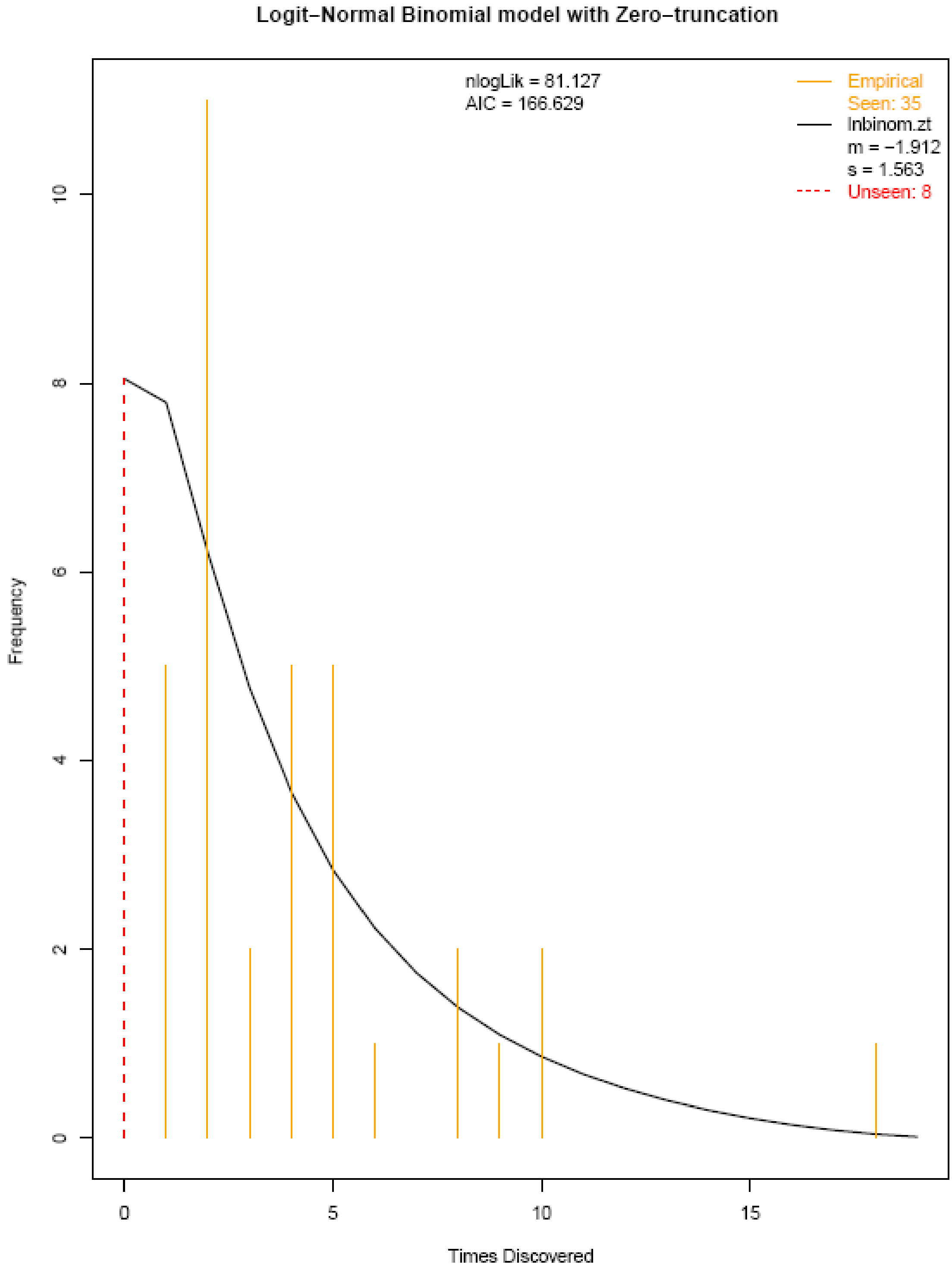


**Figure 5: Applying the LNB$_{zt}$ model brings forth the estimated number of 8 problems that were discovered 0 times.**

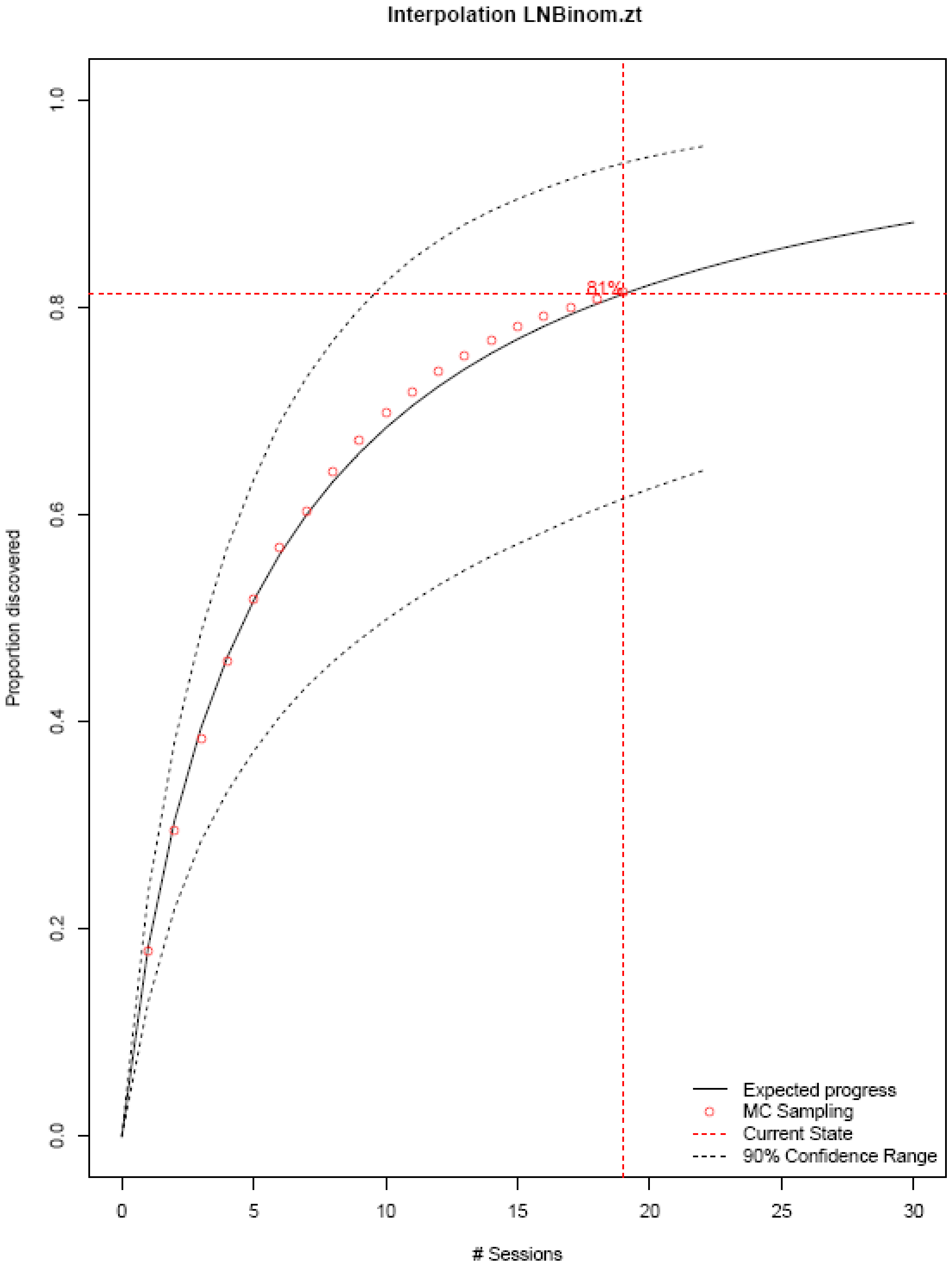


**Figure 6: The separate test session with 19 users, together elicited around 81% of existing usability problems. The 90% confidence interval ranges from 62% to 94%.**

# 6. Discussion

## 6.1 Conclusion

This study gives a view of the importance and the implications of taking into account differences between users in usability research. Results show that segmentation of participants on the basis of their characteristics may be of interest for the interpretation of test results, even with a relatively small sample.

### 6.1.1 Differences between users and their user problems

In this particular study we could reveal the following possible effects:

1. People who have more experience with the use of the product have less knowledge based problems and are more satisfied about the use of the product. Possibly, this finding is related to the nature of the classification we have used in this study. The more familiar one is with the use of  a specific product, the more this person shifts from knowledge based information processing to skill based information processing.
2. People that generally spend a medium amount of time (1-4 hours a day) on the Internet have less knowledge based problems than people that spend very little or, on the contrary, relatively much time on the internet.  Inexperienced users might have more knowledge based problems because  they have no rules or routines available to deal with the website. We can only explain the finding that also very experienced internet users have more knowledge based problems by considering that experienced internet users distinguish different websites better. They may recognise wehkamp.nl to be a uniquely functioning website compared to other websites visited. This reduces the tendency to use the website on a routine or rule basis, since experts often fail to transfer their skills to comparable, but nevertheless different domains (Anderson, 2005).
3. The website seems more suitable and attractive for female users. Men that have little experience with the product have slightly more knowledge based problems than women. Although men that visit wehkamp.nl frequently have less knowledge based problems than women, they judge the use of the product worse than women with the same amount of product experience. If wehkamp.nl aims to attract more men, according to this finding, the challenge may exist to design an appearance and functionality of their webshop that focuses more on men.

The observation that ‘knowledge based’ problems were found predominantly, and only this type of problem showed significant results, is in correspondence with the finding of Fu, Salvendy and Turley (2010). A usability test has shown to be the best suitable method for eliciting just these kind of problems. Heuristic evaluation is better at detecting the other types of problems: skill and rule based. An important question that arises from this fact, is whether the SRK based classification is the right one for the purpose of our study. People with more experience logically show a shift from knowledge based to more skill based information processing, which is also reflected in the types of problems they experience. Another type of classification, for example more on the side of required solutions, might provide more practical insights.

Although this study may have succeeded in eliciting 81% of usability problems of the product detail related pages of wehkamp.nl, the amount of 19 participants is minimal when the aim is to make comparisons over different users. A study with a comparable setup, but a larger number of respondents, might elicit more, stronger and even other effects between different user characteristics.

Nevertheless, we hope to have demonstrated through this article that practitioners in the field of the internet should not ignore the distinctions between users. We want to encourage usability specialists, web-designers and developers to structurally reflect individual differences in their conclusions and choices. If we better understand the versatility in internet use, we will be able to provide a better response to the needs of users, and exploit the potentials of humans interaction with the web.

### 6.1.2 Diversity of problems and the required number of test participants

This study clearly demonstrates that the hypothesis ‘*With 5-7 test participants we find 85% of problems*’ does not hold for every single study. Using a model that takes into account that the data is not homogenous and not complete, the estimated probability of unfound problems is much higher than it would be using the geometric series model, from which our hypothesis initially derives.

On the issue of required test participants we can still add the question how we can estimate the number of undiscovered problems of individual participants not represented in our test-sample. Generally, more participants are needed to find most character-specific problems. Accounting for individual differences requires that you either make sure all different characteristics are represented in the sample, or that you test only with (a) specific predefined target group(s).

## 6.2 Research challenges

Concerning the implications of user diversity, this study only lifts a corner of the veil. Much is still to be explored. In this section we give some incentives.

### 6.2.1 Practical implications

An on-going challenge lies in the translation of usability test results into concrete recommendations, and optimal design solutions. Where user diversity is concerned, this challenge becomes even more complicated. For example, if a young participant has trouble making the right choice between a long list of topics (Van Deursen & Van Dijk, 2009) suggesting to categorise the topics in different subsets in a hierarchical menu-structure would disadvantage elderly users (Freudenthal, 2004). On the basis of complete descriptions of different users and their problems, solutions must be sought that account for these different aspects.

### 6.2.2 Different methods for uncovering usability problems

Another factor that is of influence on our total estimation of required participants is the method we use to elicit the usability problems of a website. As emerged in this study, not every method in usability research brings about the same kind of problems (Fu et al.,2010). An estimation on the percentage of problems found will only say something about the problems that will be found through usability testing. For example, heuristic evaluation will elicit other usability problems and therefore increases the total number of actual number of problems that are not found with just a usability test. For the most valid results usability problems from different methods should be included in the discussion on overall found product-defects.

### 6.2.3 Generalizability over different nationalities

An interesting question that arises from cultural difference related research is if, besides differences in attitudes and usage between different cultures, there also exist differences in operational use of computers and therefore in experienced usability (Li & Kirkup, 2007). Can we learn from the usability findings from other countries? Or should we rely primarily on domestic specialists who test with domestic users? We acknowledge that further research on differences between usability problems among different cultures would be interesting for determining the generalizability of test results from other countries to local websites.

### 6.2.4 Usability problems over different kinds of internet applications

Apart from differences between users, there might also be differences ‘within users’ in the number and kind of problems they experience when using different applications (Spool & Schroeder, 2001). Does one user experience the same kind of user problems for governmental, e-commerce, educational and informational websites? And how about search engines, databases, e-mail and social media? A user might have different expectations regarding different digital services and adjust his behaviour. Future research may point out whether this is the case and what that might imply.

### 6.6.5 The absolute relevance of ‘usability’

Al-maskari and Sanderson (2011) raise another interesting issue with regard to differences in experienced usability. They showed that, although users with lower cognitive skills were less effective, they did not report to be less satisfied in their use of the product. If, in the end, our aim is to make internet applications equally usable for all different users, we should consider what aspect of ‘usability’ is decisive. Although users may not all be as fast and efficient in achieving their online goals as others, they might be equally effective. And even if they are not equally effective, they might be just as satisfied. We can imagine that for some applications it might not matter so much that users achieve the goals they initially visited a website for. For example, in e-commerce, a user might come to a website to quickly buy a t-shirt, but ends up browsing a webshop for hours to finally satisfactory order a pair of trousers, a couch and a juicer. No party would have been more happy if he had ‘effectively’ and ‘efficiently’ used the website for his original goal. Closer examination of the main factors for satisfactory internet use can shed light on whether it is worthwhile to make websites as effective and efficient as possible.

## 7. References

Al-Maskari, A., & Sanderson, M. (2011). The effect of user characteristics on search effectiveness in information retrieval. *Information Processing & Management*, *47*(5), 719-729. Elsevier Ltd. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S030645731100029X

Amiel, T., & Sargent, S.L. (2004). *Individual differences in internet usage motives.* University of Georgia, Department of Instructional Technology & Virgina Tech, Department of Communication.

Anderson, J.R. (2005). *Cognitive psychology and its implications*. New York: Worth Publishers.

Bevan, N. (2001). *International standards for HCI and Usability*. International Journal of Human-Computer Studies, 55, 533-552.

Bias, R.G., & Mayhew, D. J (2005). *Cost-justifying usability, an update for the internet age*. San Francisco: Morgan Kaufmann Publishers, Elsevier.

Boren, T., & Ramey, J. (2000). *Thinking aloud: reconciling theory and practice*. IEEE Transactions on Professional Communication, 43(3), 261-278.

Brooke, J. (1996). SUS – A Quick and Dirty Usability Scale. In Jordan, P.W., Thomas, B., Weerdmeester, B.A., & McClelland, I.L., *Usability evaluation in industry* (pp 189-194). London: Taylor and Francis Ltd.

Burnett, G.E., & Ditsikas, D. (2006). *Personality as a criterion for selecting usability testing participants.* University of Nottingham, School of Computer Science and Information Technology. Retrieved on the 16th of june, 2011, from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.98.4980.

Ericsson, K.A., & Simon, H.A (1993). *Protocol analysis, verbal reports as data*. Massachusetts: The MIT Press.

Eysenck, H. J., & Eysenck, M. W. (1985). *Personality and individual differences: A natural science approach.* New York: Plenum Press.

Faulkner, L., & Wick, D. (2005). Cross-user analysis: Benefits of skill level comparison in usability testing. *Interacting with Computers*, *17*(6), 773-786.

Freudenthal, D. (2001). *Age differences in the performance of information retrieval tasks.* Behaviour & Information Technology, 20, 9-22.

Helsper, E. J. (2010). Gendered Internet Use Across Generations and Life Stages. *Communication Research*, *37*(3), 352-374. Retrieved from http://crx.sagepub.com/cgi/doi/10.1177/0093650209356439

Hornbæk, K. (2005). Current practice in measuring usability: challenges to usability studies and research. International Journal of Human-Computer Studies, 64, 79-102.

Hornbæk, K., & Frøkjær, E. (2008). Comparison of techniques for matching of usability problem descriptions. *Interacting with Computers*, *20*(6), 505-514.

International Organization for Standardization (1998). *ISO9241-11: Guidance on Usability.* Genève: International Organization for Standardization.

Johnson, G. (2008). Cognitive processing differences between frequent and infrequent Internet users. *Computers in Human Behavior*, *24*(5), 2094-2106. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0747563207001550

Kim, K. S. (2001). Implications of user characteristics in information seeking on the World Wide Web. *International Journal of Human-Computer Interaction,13*(3), 323-340. Taylor & Francis.

Kirwan, B. (1992). Human error identification in human reliability assessment. Part 1: Overview of approaches. *Applied Ergonomics*, 23(5), 299-318.

Krug, S. (2006). *Don't make me think, a common sense approach to web usability.* California: New Riders Publishing.

Lia, N., & Kirkupb, G. (2007). Gender and cultural differences in Internet use: a study of China and the UK. Computers & Education, 48, 301-317.

Lewis, J.R. (2001). Evaluation of procedures for adjusting problem-discovery rates estimated from small samples. *International Journal of Human-Computer Interaction*, 13(4):445–479.

Nielsen, J. (2000). *Why You Only Need to Test with 5 Users.* Retrieved on the 3d of September, 2009, from http://www.useit.com/alertbox/20000319.html.

Nielsen, J. (2006). Variability in User Performance. Jakob Nielsen's Alertbox, retrieved from: http://www.useit.com/alertbox/performance_variability.html**.**

Nielsen, J., & Carsten Y. (2004). *What Kind of Information does an HCI Expert Want? - on Concurrent Usability Testing*. Copenhagen Business School, Department of Informatics.

Nielsen, J. & Landauer, T.K. (1993). A mathematical model of the finding of usability problems. In CHI '93: *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 206–213, New York, NY, USA. ACM Press.

Nørgaard, M., & Hornbæk, K. (2006). *What do usability evaluators do in practice? An explorative study of think-aloud testing.* University of Copenhagen, Department of Computer Science.

Papacharissi, Z., & Rubin, A.M. (2000). Predictors of internet use. Journal of Broadcasting & Electronic Media, 44, 175-196.

Rogers, Y., Sharp, H., & Preece, J. (2007). *Interaction design, beyond human-computer interaction*. Chichester: John Wiley & Sons Ltd.

Rubin, J. (1994). Hanbook of usability testing, how to plan, design and conduct effective tests. Indianapolis: Wiley Publishing.

Schmettow, M. (2009). Controlling the Usability Evaluation Process under Varying Defect Visibility. Passau University, Information Systems II.

Selwyn, N. (2004). Reconsidering Political and Popular Understandings of the Digital Divide. *New Media & Society*, *6*(3), 341-362.

Spool, J. & Schroeder, W. (2001). Testing web sites: five users is nowhere near enough. *CHI '01: CHI '01 extended abstracts on Human factors incomputing systems*, ACM, 285-286.

Thompson, S.H., & Vivien, K.G. (2000). *Gender differences in internet usage and task preferences*. Behaviour & Information Technology, 4, 283-295.

Van Deursen, J.A.G.M., & Van Dijk, A.J.A.M. (2009). *Using the Internet: Skill related problems in users' online behaviour* (Electronic version). Interacting with Computers. Retrieved from: http://www.utwente.nl/gw/mco/bestanden/Using%20the%20Internet-%20Skill%20related%20problems.pdf.

Virzi, R.A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors*, 34(4):457–468.

Weiser, E.B. (2000). Gender differences in internet use patterns and internet application preferences: a two-sample comparison. CyberPsychology & Behavior, 3, 167-178.

# 8. Appendix (Dutch)

## 1 Interview

Persoonskenmerken

Proefpersoonnummer: ___________

Geslacht: ___________

1. Wat is je leeftijd? ___________
2. Ben je links- of rechtshandig? ___________

Opleiding

3. Wat is je hoogst genoten opleiding?

_____________________________________________

4. Welke opleiding (richting / onderwerp)?

_____________________________________________

Werkervaring

5. Wat voor beroep beoefen je momenteel?

_____________________________________________

6. Hoe lang doe je dit werk al?

_____________________________________________

7. Heb je in het verleden nog andere functies bekleed? Zo ja, welke?

_____________________________________________

Computer Ervaring

8. Hoe vaak maak je gebruik van je computer en internet? (voor werk en privé)

___________________________________________________________________________

9. Hoeveel uur per dag? ___________
10. Wat voor besturingsysteem heeft je computer? ___________
11. Wat voor browser gebruikt je om te internetten? ___________

Product Ervaring

12. Hoe denk je over het algemeen over online winkelen?

___________________________________________________________________________

13. Hoe vaak bezoek je wehkamp.nl?

_______________________________________________

14. Wat doe je zoal op wehkamp.nl? (Rondkijken naar producten / bestellen / reviews schrijven / vragen stellen / informatie zoeken)

______________________________________________

15. Hoe vaak bestel je iets via wehkamp.nl en om wat voor producten gaat het?

**2 Taken:**
Instructie:
Voer de zes onderstaande taken uit. Elke taak duurt maximaal 10 min. De taak eindigt wanneer de onderzoeksleidster het sein geeft dat je naar een volgende taak kunt overgaan. Dit doet zij als de taak is volbracht, of als de 10 minuten zijn verstreken.
Elke taak begint op een specifieke pagina. Het internetadres van die pagina staat bij de taakomschrijvingen vermeld. Je kunt hierop klikken om de pagina te openen of het adres anders naar de adresbalk in de browser kopiëren.
Denk hardop terwijl je de taak uitvoert. De eerste taak zal een oefentaak zijn, waarbij de onderzoekleidster wat tips kan geven over hoe je hardop kunt denken. De onderzoeksleidster zal tijdens het uitvoeren van de overige taken alleen onderbreken als je gestopt bent met praten. Zij zal je dan aanmoedigen om hardop te blijven denken.

Denk eraan dat de website wordt getest, niet jij. Wat je doet om de taken uit te voeren is niet goed of fout. Probeer je wel zo veel mogelijk in de taken in te leven, alsof je je in een echte situatie bevindt, waarbij je een online aankoop wilt doen.
Tijdens het uitvoeren van de taken worden geluid en beeld opgenomen. Wij gaan vertrouwelijk met je gegevens om. Resultaten uit de test worden niet aan je naam verbonden.

Je ontvangt na afloop een waardebon ter waarde van 25 euro.

| **Oefentaak** | |
|---|---|
| Omschrijving | Sla een afbeelding op in 'mijn documenten', van een rode trui, die je mooi vindt. (Dit hoeft geen afbeelding van wehkamp.nl te zijn.) |
| Startstatus | www.google.nl |

| **Taak 1.** | |
|---|---|
| Omschrijving | Zoek op deze pagina naar een nieuwe winterjas naar jouw smaak en bestel de jas. |
| Startstatus | Voor dames:<br>http://www.wehkamp.nl/damesmode/jassen/jassen-jacks/C01_L06_L61/?PI=0<br>Voor heren:<br>http://www.wehkamp.nl/herenmode/jassen/jacks-jassen/C02_A01_A61/?PI=0 |

| **Taak 2.** | |
|---|---|
| Omschrijving | Je bent op zoek naar een ruime tweepersoons tent. Je belandt op deze pagina, maar je wilt weten wat wehkamp.nl nog meer voor tenten heeft.<br>a. Zoek verder naar andere tenten, maak je keuze en bestel.<br>b. Deel deze pagina met je kampeergenoot. |
| Startstatus | http://www.wehkamp.nl/spel-vrije-tijd/kamperen/tent/coleman-crestline-3-persoons-tent/C08_O05_O55_763239/?PI=0 |

| **Taak 3.** | |
|---|---|
| Omschrijving | Je wilt graag een nieuwe bank. Gebruik de informatie en functies op deze pagina om je mening over dit product te vormen. Bestel de bank als je een positief beeld van het product hebt gekregen. |
| Startstatus | http://www.wehkamp.nl/wonen/bank-fauteuil/hoekbank/hoekbank-elles/C10_R06_R64_508326/?PI=0 |

| **Taak 4.** | |
|---|---|
| Omschrijving | Bij het ene oor kan deze thermometer een andere temperatuur geven dan bij het andere oor, zo blijkt uit een review van janblauw op 10-11-2009. Je vraagt je af welke temperatuur nu klopt; de hoogste, de laagste of de gemiddelde. Stel je vraag via wehkamp.nl. |
| Startstatus | http://www.wehkamp.nl/fit-mooi/gezondheid-gewicht/meetapparatuur/medisana-fto-infrarood-oorthermometer/C04_Z0C_ZC7_670365/?PI=0 |

| **Taak 5.** | |
|---|---|
| Omschrijving | Je wilt deze laptop bestellen.<br>a. Kies een betalingswijze.<br>b. Bepaal de totaalprijs van uw bestelling.<br>c. Bepaal wanneer en hoe het product wordt geleverd. |
| Startstatus | http://www.wehkamp.nl/computers-telecom/laptop/mini-laptop/asus-eee-pc-1201ha-silver-mini-laptop/C09_I08_I44_734667/ |

| **Taak 6.** | |
|---|---|
| Omschrijving | Bekijk de pagina. Geef in de volgende situaties jouw vervolgstappen aan door met de muis naar de bijbehorende paginaonderdelen of hyperlinks te wijzen.<br>a. Wat is je vervolgstap als dit product net niet is wat je zoekt?<br>b. Wat is je vervolgstap als je naar inspiratie zoekt voor andere aankopen?<br>c. Wat is je vervolgstap als je meer informatie wilt over dit product?<br>d. Wat is je vervolgstap als je jouw mening over dit product wilt geven?<br>e. Wat is je vervolgstap als je meer wilt weten over kosten, betaling en levering van dit product? |
| Startstatus | Plaatje van product detail pagina. |

**3 Vragenlijst: System Usability Scale (SUS)**

Sterk
mee eens

Sterk
mee
oneens

1. Ik denk dat ik het leuk zou vinden om wehkamp.nl frequent te gebruiken.

| | | | | |
|---|---|---|---|---|
| □ | □ | □ | □ | □ |
| 1 | 2 | 3 | 4 | 5 |

2. Ik vond de website onnodig ingewikkeld.

| | | | | |
|---|---|---|---|---|
| □ | □ | □ | □ | □ |
| 1 | 2 | 3 | 4 | 5 |

3. Ik vond dat de website eenvoudig was om te gebruiken.

| | | | | |
|---|---|---|---|---|
| □ | □ | □ | □ | □ |
| 1 | 2 | 3 | 4 | 5 |

4. Ik denk dat ik ondersteuning nodig heb van een technisch persoon om wehkamp.nl te kunnen gebruiken.

| | | | | |
|---|---|---|---|---|
| □ | □ | □ | □ | □ |
| 1 | 2 | 3 | 4 | 5 |

5. Ik vond dat de verschillende onderdelen van deze website goed bij elkaar pasten.

| | | | | |
|---|---|---|---|---|
| □ | □ | □ | □ | □ |
| 1 | 2 | 3 | 4 | 5 |

6. Ik vond dat er te veel onsamenhangendheid was in de website.

| | | | | |
|---|---|---|---|---|
| □ | □ | □ | □ | □ |
| 1 | 2 | 3 | 4 | 5 |

7. Ik kan mij voorstellen dat de meeste mensen heel snel leren hoe ze wehkamp.nl kunnen gebruiken.

| | | | | |
|---|---|---|---|---|
| □ | □ | □ | □ | □ |
| 1 | 2 | 3 | 4 | 5 |

8. Ik vond de website erg lastig om te gebruiken.

| | | | | |
|---|---|---|---|---|
| □ | □ | □ | □ | □ |
| 1 | 2 | 3 | 4 | 5 |

9. Ik voelde me heel zelfverzekerd terwijl ik de website gebruikte.

| | | | | |
|---|---|---|---|---|
| □ | □ | □ | □ | □ |
| 1 | 2 | 3 | 4 | 5 |

10. Ik moest veel dingen leren voordat ik met de website aan de gang kon.

| | | | | |
|---|---|---|---|---|
| □ | □ | □ | □ | □ |
| 1 | 2 | 3 | 4 | 5 |

### 4. SPSS Syntax

```
GET
  FILE='C:\Users\Inge\Documents\Studie\Afstuderen
MPS\resultaten\results1.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.
GLM SkillPrblmCat RulePrblmCat KnowProblmCat TimeonTasks SUS BY Gender
Education InetExperience PrdctExperience WITH Age
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /PLOT=PROFILE(Gender Education InetExperience PrdctExperience
Gender*PrdctExperience Education*InetExperience
InetExperience*PrdctExperience)
  /EMMEANS=TABLES(Gender) WITH(Age=MEAN) COMPARE ADJ(BONFERRONI)
  /EMMEANS=TABLES(Education) WITH(Age=MEAN) COMPARE ADJ(BONFERRONI)
  /EMMEANS=TABLES(InetExperience) WITH(Age=MEAN) COMPARE ADJ(BONFERRONI)
  /EMMEANS=TABLES(PrdctExperience) WITH(Age=MEAN) COMPARE ADJ(BONFERRONI)
  /EMMEANS=TABLES(Gender*PrdctExperience) WITH(Age=MEAN)
  /EMMEANS=TABLES(Education*InetExperience) WITH(Age=MEAN)
  /EMMEANS=TABLES(InetExperience*PrdctExperience) WITH(Age=MEAN)
  /CRITERIA=ALPHA(.05)
  /DESIGN=Age Gender Education InetExperience PrdctExperience
Gender*Education Gender*InetExperience Gender*PrdctExperience
Education*InetExperience Education*PrdctExperience
InetExperience*PrdctExperience Gender*Education*InetExperience
Gender*Education*PrdctExperience Gender*InetExperience*PrdctExperience
Education*InetExperience*PrdctExperience
Gender*Education*InetExperience*PrdctExperience.


EXAMINE VARIABLES=SkillPrblmCat RulePrblmCat KnowProblmCat TimeonTasks SUS
  /PLOT BOXPLOT STEMLEAF NPPLOT
  /COMPARE GROUPS
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.

* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=KnowProblmCat
MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: KnowProblmCat=col(source(s), name("KnowProblmCat"))
  GUIDE: axis(dim(1), label("Knowledge based problems."))
  GUIDE: axis(dim(2), label("Frequency"))
  ELEMENT: interval(position(summary.count(bin.rect(KnowProblmCat))),
shape.interior(shape.square))
  ELEMENT: line(position(density.normal(KnowProblmCat)))
END GPL.


* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=SUS MISSING=LISTWISE
REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: SUS=col(source(s), name("SUS"))
  GUIDE: axis(dim(1), label("SUS"))
  GUIDE: axis(dim(2), label("Frequency"))
  ELEMENT: interval(position(summary.count(bin.rect(SUS))),
shape.interior(shape.square))
  ELEMENT: line(position(density.normal(SUS)))
END GPL.
```