Indicator-Based Evaluation

A Standard to Evaluate the Usability of Dutch E-Commerce Websites

Thesis

For the degree of Master of Information Science
at the Radboud University of Nijmegen

Paul Peeters

August 2006

Acknowledgements

First of all, I would like to thank my supervisor, Peter Achten, for his very enthusiastic support. Without his guidance this thesis would not be what it is.

I am also very grateful to Jan Tretmans for his very useful suggestions during the final stages of this work.

Many thanks also to Marja Peeters, Bo Peeters, Erik van Beijnen, Bob Gits and Paul Rozestraten for taking part in the tests. Their contributions have been invaluable to me.

Finally, I would like to thank my family and friends for their constant support and encouragement.

Abstract

A Standard to Evaluate the Usability of Dutch E-Commerce Websites

On the internet, E-Commerce websites play an important role, and in the future their role will only become bigger. People use E-Commerce websites for purchasing goods, or for personal or company use. If E-Commerce websites lack good usability, no one will visit them because they are not visitor-friendly.

In this thesis, a new method to evaluate the usability of E-Commerce websites is presented and evaluated. As a guidance aid, the ISO 9126(+) standard will be used. In this ISO standard, all of the characteristics of usability are presented. With the help of indicators, which are part of the usability characteristics, a new method of validating Dutch E-Commerce websites will be presented and compared to existing usability evaluations.

That is why each of the usability characteristics and their indicators will be thoroughly discussed, and their relevance in relation to E-Commerce and other websites will also be presented. After this discussion, the Indicator-Based Evaluation method will be presented. This approach offers advantages over the existing evaluation methods. For instance, the Indicator-Based Evaluation method restricts itself to E-Commerce websites only, while many existing methods have a broader scope.

The main perspective of this thesis is: E-Commerce websites should be evaluated on their usability with the help of the Indicator-Based Evaluation method, which is based on the ISO 1926+ standard.

Finally, this new method will also be evaluated by a test group. The results of this evaluation are as well part of this thesis. A comparison of the scores of the Indicator-Based Evaluation method with the expected results (based on statistics or other evaluation methods) shows that the Indicator-Based Evaluation method is a valid evaluation method.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 The origin of Indicator-Based Evaluation

On the internet, E-Commerce websites play an important role, and in the future their role will only become bigger. People are broadly using E-Commerce websites to purchase goods, for personal use or for their company. When E-Commerce websites would lack good usability, no one would visit these websites because they don't work easily. A good usability score is therefore recommended for an E-Commerce website to be successful. It is stated that an estimated 90% of existing sites does not provide adequate usability [Van02], and a number of these websites are E-Commerce websites. Unfortunately, the exact percentage of E-Commerce websites was not mentioned in Vanderdonckt's article.

The problem to be solved by this thesis is: "In the field of E-Commerce websites, usability plays an important role. There are ways to evaluate the usability of these websites, but how good are they? And can a better way of evaluating E-Commerce websites be created (and validated)?"

Secondly, there should be examined if a usability test of E-Commerce websites, based on the ISO 9126+ standard for usability [Zei96], gives better results than existing usability tests, which are mostly based on best practices.

The domain of the thesis is large, namely all the E-Commerce websites that are available nowadays. It is not possible to look at all the E-Commerce websites, so making a selection of representative E-Commerce websites to evaluate is needed. To come to a representative group, knowledge about the objectives of (Dutch) E-Commerce websites (why is an E-Commerce website used by a company) and about the target groups of these websites (for who is an E-Commerce website made) is required.

This is quite a broad problem statement, so this problem statement is divided into sub-problem [*]statements. In these sub-problem statements, the terms of the main problem statement are described, such as: what is an E-Commerce website, and how can usability be measured. In this thesis, the chapters cover each of these sub-problem statements.

[*] Anywhere in this thesis where "He" is written, "She" can also be read

To clarify some of the technical terms used in this thesis, the definitions of the terms that are used are summed up below:

Quality: Totality of characteristics of an entity that bear on its ability to satisfy stated and implied needs [ISO00]

Software: Computer programs, procedures, and possibly associated documentation and data pertaining to the operation of a computer system [IEEE90]

Software Quality: Conformance to explicitly stated functional and performance requirements, explicitly documented standards, and implicit characteristics that are expected of all professionally developed software [Gal04]

E-Commerce: Electronic exchange of information between a company and external stakeholders, with the purpose of supporting the business processes, […] for instance the business process of selling goods [Cha02]

1.2 How to perform the Indicator-Based Evaluation

The main part of this thesis deals with the evaluation of E-Commerce websites by means of their usability. To evaluate website usability, the ISO 9126+ standard [Zei96] can be used as a guidance aid, which thoroughly describes the quality characteristic of usability. The ISO 9126+ standard is an extension of the ISO 9126 standard [ISO04]. This ISO 9126+ standard uses indicators that make it possible to evaluate the usability sub-characteristics of software products in general. Because E-Commerce websites are a kind of software products, this evaluation may need to be altered to become more suitable for E-Commerce websites.

In this thesis there is a restriction applied to the field of E-Commerce websites, namely there will only be looked at Dutch E-Commerce websites. It can be said that the restriction of looking only at Dutch E-Commerce websites doesn't mean that this method can also be used for most West-European websites and websites of culturally related societies. To rule out cultural biases, this thesis restricts itself to Dutch E-Commerce websites is the following: this specific selection excludes (possible) cultural biases (Dutch culture has many similarities with most West-European cultures, but there are some differences. Looking at the cultural differences between countries is beyond the scope of this thesis). For this thesis, a representative selection of E-Commerce websites has been made. This selection covers the

majority of the E-Commerce websites domain, containing general E-Commerce websites (websites where one can buy various kinds of goods) and specific E-Commerce websites (specialized in one single product group). Visitor numbers have also been an important factor in selecting the Dutch E-Commerce websites. When choosing a relevant selection, the time of research to examine a broader group of websites can be limited.

Purpose of this thesis is to find out if, with the use of the ISO 9126+ usability sub-characteristics and indicators, a usability test that is suitable to evaluate the usability of Dutch E-Commerce websites can be made. It could be possible that this usability test can be used to evaluate the usability of most of the West-European E-Commerce websites as well, because most West-European E-Commerce websites are quite similar, apart from the language and other possible cultural biases.

Finally, an evaluated website will receive an Indicator-Based Evaluation score. This score tells us what the usability of the website is compared to other E-Commerce websites. The final score is not the most meaningful part of the evaluation; with the Indicator-Based Evaluation method it is also possible to compare the usability of E-Commerce websites for each indicator separately. A scale is used to determine the Indicator-Based Evaluation score for an E-Commerce website. The Indicator-Based Evaluation method makes it possible to present a first impression for an E-Commerce website, for a more precise meaning of the score one can look at the scores of the individual indicators.

1.3 Objectives of the Indicator-Based Evaluation

In this thesis the methods of validating a Dutch E-Commerce website, by using existing usability evaluations and by developing and validating a new usability evaluation, are examined. This thesis examines whether it is possible to evaluate all different kinds of E-Commerce websites with one single evaluation method.

There are, in the field of software quality, some general software usability test methods and best practices (e.g. Nielsen [Nie99], Shneiderman [Shn04], Shackel [Sha91], Holmes [Hol02], QUINT research group [Zei96]). This thesis examines if it is possible to use some of the more general test methods, and if it is possible to modify these methods for the purpose of evaluating Dutch E-Commerce websites. Other studies on the same topic - usability and E-Commerce websites - offer several different approaches to evaluate the usability of E-

Commerce and other websites. In Chapter 2 of this thesis, the most relevant of these approaches, considering this thesis, are thoroughly described. In Section 4.6 the related approaches are compared to the Indicator-Based Evaluation method.

1.4 Summary of chapters

Chapter 2 deals with the related research. First, the ISO 9126+ standard is discussed. Next, a closer look is taken at existing methods for testing and evaluating the usability of E-Commerce websites. For each of the researched methods, the advantages and disadvantages are summed up. The (possible) relations of the related research with the ISO 9126+ standard are also described.

Chapter 3 deals with quality characteristics and quality attributes. In this chapter all of the quality characteristics, usability sub-characteristics and quality attributes are described. The usability sub-characteristics that are used are derived from the Extended ISO 9126 model. All of these sub-characteristics are described with the help of indicators. These indicators provide a method to evaluate the usability of software products (e.g. an E-Commerce website). Because of the generality of the indicators (they apply to all software products), these indicators sometimes need to be altered to apply correctly to E-Commerce websites. All of the relevant indicators to measure the quality characteristics, quality sub-characteristics and quality attributes will be classified according to the importance of the indicator to the field of Dutch E-Commerce websites.

Chapter 4 introduces the Indicator-Based Evaluation method. It covers the topic of how to make and verify a test (based on a preliminary research to possible existing tests more or less related to the domain of E-Commerce websites and usability) in which adults of a Dutch target group assess E-Commerce websites with the help of the most important quality characteristics, quality sub-characteristics and quality attributes.

Chapter 5 covers the domain of Dutch E-Commerce websites and the target groups of these websites. Questions which are answered are: what kinds of E-Commerce websites are there and what are the differences and similarities between these E-Commerce websites? This chapter also covers the selection of E-Commerce websites that are evaluated for this thesis.

Chapter 6 deals with the results of the evaluation of the selected websites from Chapter 5. The evaluation results are discussed and the conclusions of the Indicator-Based Evaluation are given.

In Chapter 7, the main conclusion of this thesis is presented. In the conclusion it is stated that the Indicator-Based Evaluation method is valid and fast, and that this method has advantages over existing evaluation methods.

Chapter 2: Related Research

This chapter deals with the related research done to provide background information for this thesis. In the following sections, several existing (E-Commerce) website usability evaluation methods are discussed. A closer look will be taken at the fact whether these methods use the guidance of the ISO 9126(+) standard or whether (and why) another method is used instead of the ISO 9126(+) standard.

Section 2.1 deals with the ISO 9126+ standard. First the relation with ISO 9126 is described, following by the extensions made on this standard. These extensions make it possible to measure the quality (and therefore also the usability) of software products.

Section 2.2 deals with the current usability evaluation methods. The question answered in this section is: which evaluation methods for usability are there, and what are the advantages and disadvantages of these methods?

Section 2.3 covers the topic of existing methods for testing the usability of websites. Six testing methods are discussed. Their advantages and disadvantages are also described.

Section 2.4 relates the usability methods stated in Section 2.3 with the ISO 9126+ standard. A closer look will be taken on the possible relations of each of the usability methods with ISO 9126+.

Section 2.5 summarizes the advantages and disadvantages of the current usability evaluation methods of Section 2.3

2.1 ISO 9126+

In this thesis, the ISO 9126+ standard [Zei96] plays an important role. The ISO 9126 standard is universal, which makes comparison of one software product (i.e. an E-Commerce website) with another easier, by looking at the quality characteristics. The advantage of the ISO 9126+ standard over the ISO 9126 standard is that the ISO 9126+ standard adds indicators to the standard. This makes the evaluation of E-Commerce websites possible, because of the indicators provided. These indicators offer the ability to evaluate the usability sub-characteristics.

The ISO 9126 standard [ISO04-2] is the first effort that the International Standardization Organization has made to fulfil the need for a quality standard for software products. In this standard, six quality characteristics are given (Usability, Functionality, Reliability, Efficiency, Maintainability and Portability). For each of these characteristics, sub-characteristics are given. The sub-characteristics for Usability are explained in Section 3.1 of this thesis. The ISO 9126 standard is a derivation of the McCall model [Gal04], which has been proposed as the basis for an international standard for software quality measurement.

The ISO 9126 standard, however, has some shortcomings. The most important shortcoming is that the standard does not provide metrics (indicators) or methods for measurement, rating and assessment [ISO04 and ISO04-2].

The ISO 9126+ standard, also known as the Extended ISO 9126 model, does provide these metrics (also known as indicators), to measure, rate and assess the quality of software products.

That is why in this thesis, the ISO 9126+ standard is used as a basis for evaluating E-Commerce websites, by means of the usability sub-characteristics, using the indicators from this ISO 9126+ standard.

2.2 Current major usability evaluation methods

To evaluate the usability of a software product, there are several methods to work with. These methods are very general and are (with little modification) applicable to any kind of software product, for instance a computer desktop application or an E-Commerce website. Ivory and Hearst [Ivo00] have developed a set of usability evaluation methods, especially written for website evaluation. This set can be categorized into the following major method classes:

- Testing (involving an evaluator observing users interacting with an interface)
- Inspection (involving an evaluator using a set of heuristics to identify potential usability problems in an interface)
- Inquiry (involving users providing feedback on an interface via interviews, surveys, etc.)
- Analytical modelling (involving an evaluator employing user and interface models to generate usability predictions)
- Simulation (involving an evaluator employing user and interface models to mimic a user interacting with an interface and report the results of this interaction)

A disadvantage of using either the testing or the inquiry method for evaluation of the usability of E-Commerce websites, is the fact that E-Commerce websites are frequently updated and redesigned, which makes the recurring costs of hiring test users for the evaluation quite high. It is also important to emphasize that users of E-Commerce websites are most often customers, which is different compared to traditional computer applications.

These evaluation methods are considered to be a framework and do not exactly describe what a test (in depth) looks like. Nevertheless, it is useful to categorize *how* a website evaluation can be done. According to the available resources (time, money, capacity, people, etc.) one of the methods can be chosen. The majority of usability evaluation techniques can be categorized into one of the evaluation methods proposed by Ivory and Hearst.

In the next section, existing methods of usability evaluation testing are discussed.

2.3 Existing methods for testing the usability of websites

In the following subsections six usability evaluation and testing methods are discussed.

2.3.1 Nielsen's "Think aloud" method

Tests of E-Commerce websites which used the "think aloud" method [Nie93], described by Jakob Nielsen, cover a larger scope of the research area compared to this thesis (the "think aloud" method covers all sorts of websites, instead of just E-Commerce websites). The "think aloud" method is a form of testing, regarding the major evaluation methods provided by Ivory and Hearst. In this form of user testing, a single user's interaction with a product is videotaped and analyzed to improve the product. Advantages of this method are a very close approximation to actual individual usage and it requires few test persons (it is not very expensive). "Think aloud" has one great disadvantage: the act of the user describing a task may possibly alter the task performance. "Think aloud" is also a subjective evaluation method, which can be objectified by having many users evaluate the system [Fin05].

Although this method is quite old, it is still used. In 2005, the Danish DialogDesign group [DD05] carried out a usability test using Nielsen's "think aloud" method for the website of Tower Records.com [TR05].

2.3.2 Protocol analysis

In a study carried out by Benbunan-Fich [Ben01], evaluating E-Commerce websites is done by protocol analysis. The protocol analysis method is a form of testing, according to the major evaluation methods provided by Ivory and Hearst. Ivory and Hearst state that despite the popularity of E-Commerce, little evidence of methodical evaluation of the usability of commercial websites is found. Protocol analysis is similar to Nielsen's "think aloud" method and looks at 15 usability principles and 3 evaluation parameters (content, navigation and interactivity). Some of these principles (about 5 of them) are derived from the ISO 9126 usability standard, but she did not directly look at the quality characteristics of this standard. For instance, the principle "Obviousness of order buttons/links" is derived from the usability sub characteristic *Clarity*. Or "Support for understanding the requirements of the ordering process" is derived from *Understandability*. Some best-practices (such as the question "what do users want while visiting an E-Commerce website?") are also considered to be a principle. An advantage of Benbunan-Fich's method is that it especially covers the field of E-Commerce

websites. The main disadvantage is the same as for Nielsen's "think aloud" method: it is unnatural for the tester to be videotaped while performing the evaluation and it therefore can result in inaccurate results [Ben01].

2.3.3 Factors affecting usability of E-Commerce websites (Tilson method)

Another method for evaluating usability was given by Tilson et al [Til98]. Looking back to the methods for usability testing proposed by Ivory and Hearst this form of testing is considered to be an inquiry. Tilson and his group evaluated four E-Commerce websites by looking at the following UI principles: simplicity, support, obviousness, encouragement, satisfaction, accessibility, versatility and personalization. A test group of 18 people evaluated four E-Commerce websites, regarding the principles mentioned above. In this evaluation, the evaluators systematically varied the presentation order of the selected websites so that they could check for order effects. This means that person A sees website 1 first, followed by website 2, while person B sees website 2 first, followed by website 1. After evaluating the four sites, the testers rated the significance of factors that would influence their decision of whether to purchase an item online. While using each site, the participants stated what they liked and disliked. Drawbacks of using Tilson's method for this thesis are: the "age" of the study (8 years old) and, more importantly, the fact that no formal usability standards such as ISO 9126 were used

2.3.4 Mathematical approach by Kauffman and Hahn

A mathematical approach of evaluating E-Commerce websites is given by Kauffman and Hahn [Kau05]. This mathematical approach is a form of analytical modelling, according to the major evaluation methods provided by Ivory and Hearst. In their research study they conceptualize human-computer interaction during online shopping as an economic production process. This view makes it possible to formulate a new perspective on website usability: the ability to transform inputs into outputs. That is why concepts of Cybernetics are used. In addition, the proposed method makes extensive use of actual customer-website interaction data using web server logs. Data Envelopment Analysis (DEA) is used as a method for evaluating usability of the E-Commerce websites. Although this approach is very interesting, it is not really useful for this thesis because the proposed method is grounded in the *economic* theory of production and thus explicitly puts business/commercial interests to the foreground. It gives thorough information about the mathematical aspects of DEA and less information about usability standardization as in ISO 9126 [ISO04 and ISO04-2].

2.3.5 Scenario-based walkthrough

To overcome the problem of the frequent updates on E-Commerce websites, as mentioned in paragraph 2.2, Scenario-Based Walkthrough can be used [Seg05]. This method is a form of inspection, regarding the major evaluation methods proposed by Ivory and Hearst. The Scenario-Based Walkthrough method, developed by Segawa, Sugimura and Ishigaki, consists of two phases. In the first phase, the evaluator systematically extracts case episodes of actual site use by users (described as scenarios) to define the objectives of the target website. In many cases these objectives are business objectives. By making a so-called "issue tree", the usability characteristics are derived and can be evaluated. The issue tree method sequentially breaks down solutions for a key task to the level of detailed measures to be taken and then arranges and expresses the solutions in a tree structure. In the second phase, the evaluator answers questions about the steps of operation needed to perform each scenario to evaluate problems that hamper attainment of a website's objectives.

This is an interesting approach to evaluation, but it has the disadvantage that, for evaluation, all the scenarios need to be available. Otherwise, not all of the E-Commerce website's (usability) aspects are evaluated. The best way of using the Scenario-Based Walkthrough is when, during the development of a website, the scenarios are stored for later evaluation.

2.3.6 Automated web-evaluation

Automated web evaluation is a form of analytical modelling, according to the major evaluation methods, proposed by Ivory and Hearst. Vanderdonckt and Beirekdar [Van05] propose an automated web evaluation by reviewing usability and accessibility guidelines. This is done by performing a static analysis of a website's HTML code against these guidelines. The evaluation process is divided into two autonomous phases: specifying formal guidelines (which is achieved only once before any evaluation) and evaluating a website (which is conducted every time an evaluation is performed). A Guideline Definition Language (GDL) is used to specify the evaluation guidelines. An example of a GDL is: select colours that will make your page easy to read by people with colour blindness. This is a usability guideline (readability). In a study done by Murch [Vdh97] several good (and bad) website colour combinations are given.

By using these GDLs an automated evaluation method can be made. These guidelines give Vanderdonckt and Beirekdar's automated evaluation some advantages over other automated evaluation methods. The most important advantages are targeted guidelines (traditional evaluation tools cannot evaluate any guideline outside precompiled set of guidelines hard coded in the evaluation engine of the tool) and the flexibility of the evaluation process (separating evaluation logic from the evaluation engine in independent phases gives many new evaluation possibilities). The advantage is that, when the tool is set up and running, the evaluation is fast. A disadvantage of this evaluation approach is that it is not easy to set up, because it requires substantial website design knowledge (DHTML, HTML, XML).

2.4 Testing and evaluation methods related to ISO 9126+

During the related research part for this thesis, several usability testing and evaluation methods were found, most of them already specified for E-Commerce and other websites.

None of the methods was directly using the ISO 9126+ usability sub-characteristics [Zei96] for testing and evaluation purposes. In most cases a combination of these sub-characteristics and best practices was used. Most of the existing usability evaluations use either the evaluation method of testing or the evaluation method of inquiry, neither of which is very reliable (see section 2.2) for the purpose of E-Commerce website evaluation because these methods are quite unnatural for a test person. This thesis will evaluate the usability of E-Commerce websites by explicitly using the usability sub-characteristics stated in the extended ISO 9126 standard, and by using the evaluation method of inspection.

Because there is no testing scheme available which evaluates specifically the field of E-Commerce websites, a new evaluation scheme (the Indicator-Based Evaluation method) needs to be created. This scheme will be derived from the usability sub-characteristics and their indicators. These indicators can be found in the ISO 9126+ standard. In the next chapter these usability sub-characteristics are explained.

By means of classification, the relevant usability sub-characteristics are put in order of importance.

2.5 Advantages and disadvantages of the evaluation methods

Each of the evaluation methods mentioned in this chapter have advantages and disadvantages. In this section, a summary is given of the most important advantages and disadvantages of evaluating E-Commerce websites with one of the evaluation methods. This summary is shown in table 2.1.

Section	Evaluation Method	Advantage(s):	Disadvantage(s):
2.3.1	Think Aloud (Nielsen)	- Uses few users; - Close approximation to actual individual usage.	- The act of the user describing a task may possibly alter the task performance; - Unnatural for the user to be videotaped during evaluation.
2.3.2	Protocol Analysis (Benbunan-Fich)	- Covers the field of E-Commerce websites; - The use of evaluation parameters.	- Unnatural for the user to be videotaped during evaluation.
2.3.3	Factors affecting usability of E-Commerce websites (Tilson)	- Lets users rate significance of factors by themselves. What do they (not) like on a website?	- The age of the study. The websites examined in the study were quite old; - User decides positive and negative aspects of a website. Evaluator needs to judge what aspects are usability-related.
2.3.4	Mathematical approach (Kauffman)	- Uses cybernetics to identify the importance of economics in E-Commerce activities.	- Deals thoroughly with economics and mathematics and is not about usability characteristics.

Section	Evaluation Method	Advantage(s):	Disadvantage(s):
2.3.5	Scenario-Based walkthrough (Segawa)	- Define the E-Commerce website objectives by using scenarios.	- For comprehensive evaluation, all the scenarios need to be available; - During website development, scenarios must be stored.
2.3.6	Automated web evaluation (Vanderdonckt)	- Targeted guidelines; - Flexibility of evaluation process; - Fast evaluation method.	- Website design language knowledge required.

Table 2.1: Advantages and disadvantages of usability evaluation methods

The advantages of the evaluation methods of the related work show that evaluation parameters (Benbunan-Fich), in combination with the ISO 9126 usability quality characteristics, can be useful for this thesis. Each of the UI principles (see paragraph 2.3 of this thesis) can be linked with one of the evaluation parameters (Content, Navigation, Interactivity). These UI principles are derived from the research done by Tilson. The evaluation parameters mentioned by Benbunan-Fich are three different but complementary aspects of web-based systems, namely: presentation of the information, navigational cues and nature of the interaction between the user and the site. The latter two of these parameters are the most related with usability.

For this thesis, the other evaluation methods have more disadvantages than advantages.. Some methods are out of the scope, while others are too time consuming.

Chapter 3: Quality Characteristics and Quality Attributes

This chapter covers all of the relevant quality characteristics, quality sub-characteristics and quality attributes related to the topic of usability. To be more specific, a closer look is taken onto the quality characteristics and quality attributes that are related to the quality of software. These quality characteristics and quality attributes related to software quality are also related to E-Commerce websites, because the (E-Commerce) website is a piece of software.

First, the most common quality characteristics and quality attributes that deal with the topic of usability will be discussed. For each quality characteristic and quality attribute the relevance with the specific topic of usability of E-Commerce websites is evaluated. These usability characteristics are derived from the ISO/IEC 9126 standard [ISO04].

Secondly, a look will be taken at how the usability quality characteristics and quality attributes in relation to E-Commerce websites can be classified. Some of the quality characteristics and attributes are more important for E-Commerce websites than others. An explanation is given why a particular kind of classification is chosen.

3.1 Usability Quality Characteristics dealing with software quality

Considering the Extended ISO model (see figure 1, from [Hen00]), six quality characteristics can be found. Usability is one of them, and can be split up in four sub-characteristics (understandability, learnability, operability and explicitness).

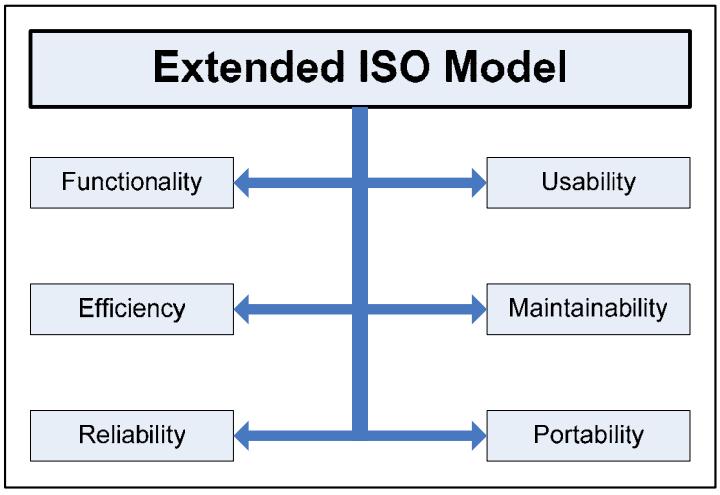

figure 3.1: The Extended ISO Model

The list of four usability sub-characteristics can be enlarged to a list of nine sub-characteristics if we add a few extra useful usability sub-characteristics derived from software projects from the QUINT (Quality in Information Technology) group [Zei96]. In this group several parties co-operated to work on a general framework for the quality of software products. These five additional sub-characteristics, added to the four from the ISO 9126 standard, are more than relevant when looking at E-Commerce websites. Figure 2 shows this complete list of nine usability sub-characteristics. This list of nine usability sub-characteristics appears also in the ISO 9126+ standard, which contains additions to the ISO 9126 standard. This ISO 9126+ standard is introduced and thoroughly described by QUINT, in cooperation with the Software Engineering Research Centre (SERC) [Zei96].

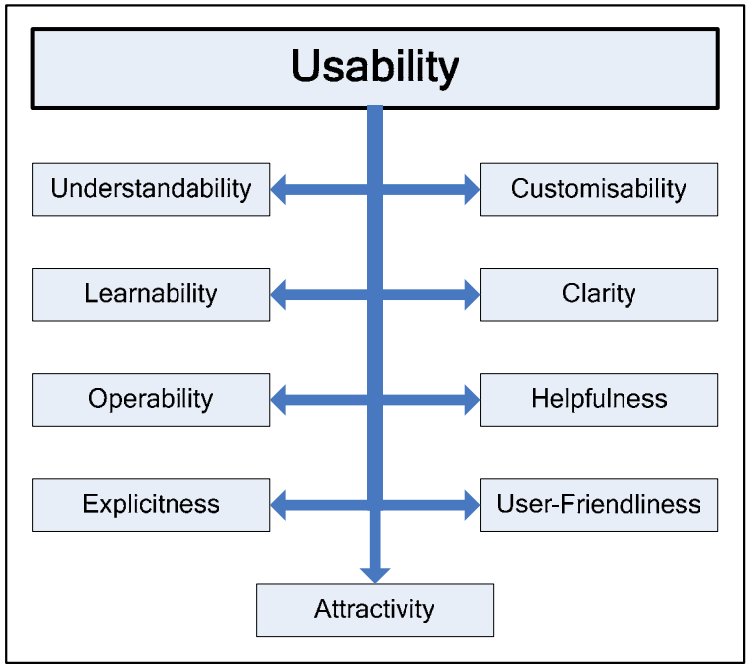

figure 3.2: Usability sub-characteristics

As stated before, an E-Commerce website is also a software product. All of these quality sub-characteristics are therefore relevant for this thesis. Each of these (usability) quality characteristics will be discussed and explained below. The extension of the ISO 9126 standard with extra sub-characteristics, as it is done in the ISO 9126+ standard, make it possible to evaluate a software product more thoroughly, and are therefore a good addition to the ISO 9126 standard.

Understandability

Definition: the capability of a software product to enable the user to understand whether the software is suitable, and how it can be used for particular tasks and conditions of use [ISO04-2].

Regarding the ISO 9126 standard [ISO04], five indicators for understandability can be described. These indicators are:

- Concept clearness
- Availability of demonstration software
- Usage clearness
- Availability of input/output data items list
- Recognisability of modifiable parameters

Each of these indicators will be explained and discussed in the context of E-Commerce websites:

Concept clearness is quite an important indicator, because it represents the degree to which functions and conventions of a software product (an E-Commerce website) are explained through models using familiar concepts from the everyday world [ISO04]. Especially the last part of the previous sentence is interesting, where it says that a certain degree of familiarity is needed to make an E-Commerce website understandable. This could possibly be further related to a form of uniformity in E-Commerce websites (in a way that E-Commerce websites tend to have the same features on their site than the E-Commerce website of a competitor). For this reason this indicator is considered to be relevant for this thesis.

The second indicator mentions *demonstration software*. For E-Commerce websites, these demonstration software systems can be prototypes of not entirely complete E-Commerce websites, in which some of the functionalities of the website to-be are already developed for testing purposes [ISO04]. In this thesis, prototypes are not considered to be in the research, therefore this indicator will be ignored.

Usage clearness, the third indicator, deals with the ratio of functions explained or by using clear models or presented to the user through demonstration software or described anyway [ISO04]. One of the goals for websites in general is that all the functionalities within a

website are clear for the user. So the requirements determining functionality that are found before building the website must be thoroughly examined. In most cases, the usage clearness indicator deals with these requirements of a website and is mostly applicable for website developers and not for website visitors. Because website development is beyond the scope of this thesis, this indicator will be ignored.

The *availability of an input/output data items list* is the fourth indicator. This kind of list describes the number of input/output items for functions with input or output operations [ISO04]. For E-Commerce websites an example of such a function is a search function, where a visitor can fill in a query (input) that generates a list of results (output). Another function is a calculation tool to calculate the total costs of a purchase. Visitors enter their requirements (e.g. gift wrap, discount vouchers, fast delivery fee), which can be interpreted as input, and the E-Commerce website returns the total costs to the visitor (output). The two examples mentioned can be placed on a input/output data list for an E-Commerce website. This indicator is relevant because the input/output routine occurs often in E-Commerce websites.

The fifth ISO understandability indicator is *the recognisability of modifiable parameters*. This means: the proportion of parameters that are identified as being either modified or fixed [ISO04]. This applies when a software product suffers a lot of modifications or "improvements". The user can, with the help of this indicator, identify these modifications. This is not relevant for this research, because the scope of this thesis does not deal with the aspect of (possible) modifications or fixes within E-Commerce websites.

QUINT [Zei96] mentions two additional indicators which are derived from their research, based on best-practices. These indicators are relevant to the quality of E-Commerce websites and they mainly deal with the user aspect. These indicators are:

- Rated understandability
- Readability score

Each of these indicators will now be explained and discussed in relation to E-Commerce websites:

Rated understandability is defined as follows: the understandability of the instructions, menus, commands, pictograms, icons, help information, instructions, manuals, etc., of the software product as rated by the user [Zei96]. E-Commerce websites contain a large number

of menus, commands, pictograms, icons, help information and instructions, so this is a relevant indicator. With the rated understandability indicator the understandability of these items can be properly measured.

The *readability score* can be interpreted as follows: the rating of the readability of the software product (on-screen messages, documents, pictograms, etc.) [Zei96]. Most E-Commerce websites contain pictograms (such as shopping carts or product pictograms), documents (such as company policies or company news) and on-screen messages (such as the latest products or discounted articles). That is why this indicator is relevant for this research. By measuring the readability of these texts with the use of the readability score indicator the difficulty of the text on the website and therefore the understandability of these texts can also be measured.

To summarize the usability sub characteristic of understandability: seven indicators are given, of which four are relevant for this thesis. These indicators are:

- Concept clearness
- Availability of input/output data items list
- Rated understandability
- Readability score

Learnability

Definition: attributes of software that bear on the users' effort for learning its application (for example, control, input, output) [ISO04-2].

The ISO standard mentions nine indicators dealing with learnability. These indicators are listed below:

- Rate of user errors
- Time required to learn operation
- Effort required to learn one operation
- Availability of help functions
- Average learning time
- Usability of manual
- Availability of manual
- Availability of pre-learning functions

- Availability of learning functions

Each of these indicators will now be explained and discussed in relation to E-Commerce websites:

The *rate of user errors*, the first indicator, is defined as the proportion of user claims resulting from incorrect operations caused by misunderstanding of the programs or manuals [ISO04]. For E-Commerce websites this indicator applies when a webmaster of an E-Commerce website receives a lot of complaints from visitors about, for instance, errors occurring with the payment of products (e.g. the visitor selects to pay when the package arrives at his address but after selecting this he is asked for his credit card number), and this error occurs because of the ambiguity of the page on which the visitor chooses a payment method. (It must be stated that it is quite well possible that some visitors won't report these errors, and go to a competitor's website instead.) The webmaster (or the person(s) who have built and maintain the website) must take care that the rate of user errors is reduced to a minimum by designing the website in such a way that it is always clear for the user and that no ambiguity can occur. For this reason this indicator is relevant for this research.

The second indicator, the *time required to learn an operation* is defined as follows: the usage time from the first usage until the usage at which the operating time is shortened to x/y of the first use to perform the same specific task [ISO04] (x means: the time a person needs to perform an action for the first time; y means: the time a person needs to perform an action this person is already acquainted with). This is a relevant indicator, because an E-Commerce website must be made in a way that a user can find his way easily through the website. That is why there must be a kind of familiarity with other E-Commerce websites, so that the time required to "learn" to use the site is very short.

The *effort to learn one operation*, the third indicator, is defined as follows: the ratio of time required to learn one operation for a specific task and operation time [ISO04]. This indicator is relevant for this research because the introduction of a new feature in an E-Commerce website is something that must be done carefully. One must think of the effort to learn this new feature. If a company doesn't want a shopping cart (metaphor) for its website, and "invents" something new that does the same, but in another way, users easily become confused [Ayk05]. They are used to the shopping cart, and now they must learn something new that in fact does the same.

The *availability of help functions*, the fourth indicator, is defined as follows: for a given range of software functions, the ratio of the number of available help functions to the number of necessary help functions [ISO04]. Necessary help functions are help functions that must be on an E-Commerce website, for instance a help function for a feature that is (quite) unique for a specific E-Commerce website (that feature is not available on many other websites). This is also a relevant indicator within the scope of E-Commerce websites, because if there, for instance, is a new feature introduced on the website, there must be good assistance in the form of a help function which covers all the things this feature does. This help function must be available and therefore be placed at a logical place on the website.

The *average learning time* is the fifth indicator. The definition of this indicator is: the time an average end-user needs to learn to work with the software product, plus the amount of guidance time needed [ISO04]. This indicator is not considered to be relevant for this research, while the two previously mentioned indicators are more specific (and can be more specified to deal with the topic of E-Commerce websites). It is taken for granted that most visitors of E-Commerce websites are more or less familiar with the basics of web-browsing.

The sixth and seventh indicator are both about manuals (usability of manual and availability of manual). The *usability of the manual* is defined as the proportion of examples, index entries, illustrations and tables per command, and/or the proportion of references, chapters, sections and subheadings per page [ISO04]. The *availability of a manual* is defined as the degree of availability of reference manuals, on-line user manuals and self-tuition documents such as operation manuals, grammar reference manuals, installation manuals, etc. for software functions [ISO04]. Although on-line user manuals are mentioned here, it is quite well possible that an appropriate help function (fourth indicator) covers this aspect of the sub characteristic learnability of E-Commerce websites. And because a visitor of an E-Commerce website doesn't (often) see or use a paper manual of the E-Commerce website both of these indicators are not considered to be relevant for this thesis.

The *availability of pre-learning functions*, the eighth indicator, can be defined as follows: for a given range of software functions, the degree of availability of materials for pre-learning (before use of the system), such as self-tuition manuals, automated tutorials or instructions [ISO04]. Because E-Commerce websites are software products which are already running on

the internet, this indicator is not relevant for this research. However, for testing the usability of a prototype of an E-Commerce website this indicator could be relevant.

The ninth indicator deals with the *availability of learning functions*. The definition of this indicator is: for a given range of software functions, the ratio of the number of available learning functions to the number of necessary learning functions [ISO04]. This indicator is not considered to be relevant for this research, because the help function (fourth indicator) should be able to cover this aspect of the sub characteristic learnability of E-Commerce websites. In the help function of the E-Commerce website, a tutorial for first-time users can be included.

To summarize the usability sub characteristic of learnability: nine indicators are given, of which four are relevant for this thesis. These indicators are:

- Rate of user errors
- Time required to learn operation
- Effort required to learn one operation
- Availability of help functions

Operability

Definition: attributes of software that bear on the users' effort for operation and operation control [ISO04-2].

In the ISO standard, no less than 24 indicators for operability are defined. QUINT adds three indicators to this large number. These combined 27 indicators are:

- Command uniformity
- Consistency of terms in message
- Message clearness
- Skill level adaptability
- Mean time between human error operations
- Expert judgement on operability
- Operability compared with sample
- Operability in practice
- Setup installation time
- Setup installation procedures with human interaction

- Availability of setup performance
- Ease of setup
- Availability of setup installation restart
- Availability of setup installation preparation
- Availability of setup installation confirmation
- Default value availability ratio
- Uniformity of screen manipulating operations
- Stability of input/output areas on display
- Number of keystrokes
- Availability of reduced effort for repeated operations
- Ultimate operation time
- Time for shutdown operation
- Guide function availability ratio
- Human error operations cancellability ratio
- Ability to emphasize expressions
- Response time for user
- Display time

Each of these indicators will now be explained and discussed in relation to E-Commerce websites:

Command uniformity is defined as follows: the proportion of operating commands having uniform formats, which are based on “common sense” and comprehensible rules [ISO04]. This indicator is considered to be relevant because regarding, for instance, an order placed on an E-Commerce website, the buyer (user of the site) must be guided through the payment process. For this reason the steps he must take must be based on “common sense” and comprehensible rules, for instance in the same way as on other E-Commerce websites.

The *consistency of terms in message*, the second indicator, is defined as follows: the proportion of system message terms that are standardized [ISO04]. This means that the system messages that appear on (for instance) an E-Commerce website are standardized, which means that these messages (if necessary) also appear on other E-Commerce websites. This indicator is related to the next one, *Message clearness*. Message clearness is defined as

follows: the proportion of system messages from software or system in which causes and corresponding action are clearly identified by the user who received those messages [ISO04]. The consistency of terms in message and the message clearness are both relevant indicators for evaluating the usability of E-Commerce websites. If, for instance, a user makes a mistake, the E-Commerce website (the system) must respond in a clear way, so that the user knows what went wrong. And the user must, after reading the response, know what to do next.

Skill level adaptability is defined as: the proportion of functions for which operating methods can be selected to correspond to the user's level of skill [ISO04]. This is a relevant indicator for E-Commerce websites. Take for instance the E-Commerce website of bol.com in The Netherlands [BL05]. People can usually navigate through the site in different ways, for instance by using the menu structure or the search function. Users who are more familiar with the site have a higher skill level. Bol.com has taken these user differences into acount by making the payment of an ordered product accessible for different skill levels, so that people who are already familiar with the website can pay their product in a faster way (fast ordering, "bestel snel" in Dutch), while first time visitors usually take the more time-consuming way of paying their product on the website, maybe with some guidance.

The *mean time between human error operations* indicator is defined as: the average time interval between one human error (human error operation, a mistake by the user while performing an operation) and the next [ISO04]. This is also a relevant indicator for E-Commerce website usability because the website developers must accomplish building a website where there are very few human errors. Some errors are not so important, for instance, typing a wrong search query. But if a user enters a wrong address or bank account number, the impact is larger. When measuring the mean time between these larger errors, E-Commerce website builders can analyse and hopefully improve the website to increase this mean time between human error operations.

Expert judgement on operability is an indicator, introduced by QUINT in the ISO 9126+ standard, and is defined as follows: the extent to which the software product presents functionality to the user without hindrance, as judged by a team of experts in the field [Zei96]. Since it is not possible for this research, in the time given for working on this thesis, to ask for expert judgement on E-Commerce websites, this indicator will not be used.

Operability compared with sample is also introduced by QUINT and can be defined as: the operability of the software product is compared to the operability of a predetermined sample product [Zei96]. To measure this indicator, QUINT suggests that a group of experts is used. As said before, the use of experts is beyond the scope of this research. Nevertheless, the subject of comparing an E-Commerce website with other E-Commerce websites is covered (in a way) in some other indicators, such as command uniformity (operability) and concept clearness (understandability).

Operability in practice is an indicator that is also introduced by the QUINT research group and can be defined as: the extent to which the software product presents functionality to the user without hindrance, as judged by users after a period of use [Zei96]. This indicator is relevant for E-Commerce websites, because with this indicator one can determine the skill level of an E-Commerce website user. This indicator has therefore a link with the indicator skill level adaptability mentioned earlier.

The next seven indicators deal with the setup process of a software product. Since an E-Commerce website is already working, these indicators are not within the scope of this thesis.
The *default value availability ratio* is defined as follows: the ratio of operating commands having default values to the total number of operating commands [ISO04]. This is not a relevant indicator for E-Commerce websites, because there aren't any default values to be found on an E-Commerce website. The website visitor can navigate freely through the website and isn't helped (or hampered) by default values, given by the system (the website).

The *uniformity of screen manipulating operations* is described as: the proportion of types of screen-manipulating operations using common basic conventions or patterns [ISO04]. This indicator is not considered to be relevant for this research, as it has much more to do with the field of software application use, such as word-processing software or accountancy software.

The *stability of input/output areas* on display is defined as follows: the proportion of input/output screen formats designed with standardized formats in which the position and form of input/output fields are commonly laid out [ISO04]. This indicator is not considered to be relevant for E-Commerce websites, because the input/output fields can be customized for each E-Commerce website and therefore these input/output fields don't need to be standardized, and within one single E-Commerce website the input/output areas are used for

very different purposes (such as searching, filling in personal details) and are shown on different places on the site (different pages), so the standardization is not that necessary.

The *number of keystrokes* is defined as: the number of keystrokes for an operation a user has to make to carry out the work. Those keystrokes include strokes of key click, (mouse) button click, screen touch, pen move, mouse move, etc [ISO04]. This indicator is not considered to be relevant, because the number of keystrokes used on an E-Commerce website can vary very much from user to user (per operation). Some users are fast and familiar with the E-Commerce website. They make fewer keystrokes than people who are just looking for a wide variety of products (for comparison, for instance).

The *availability of reduced effort for repeated operations* is defined as: the ratio of strokes required to repeat operation to the strokes required for the first operation to perform a specific task [ISO04]. This indicator is not considered to be relevant for E-Commerce websites, because this subject is already (and better) covered by other indicators (skill level adaptability under operability, concept clearness under understandability). More experienced users can go through the website faster compared to less experienced users, because they know their way on the E-Commerce website.

The *ultimate operation time*, that is the time required for the operation (of a software product, for instance an E-Commerce website), that is ultimately reduced and cannot be reduced anymore by further improvements, is the next indicator [ISO04]. This indicator is not considered to be relevant, because the operation time for E-Commerce websites depends on the uptime of the server(s) hosting the website. And the uptime of the website server is not relevant within the scope of this research.

The *time for shutdown operation* is defined as: the average time between the input of the system shutdown command at the beginning of shutdown operation and the time when shutdown is completed [ISO04]. This is not relevant for this thesis, because users are not allowed to shut down the E-Commerce website.

The *guide function availability ratio* is defined as follows: the ratio of available guide functions to the required ones for a given set of functions. This indicator is not considered to be that relevant, because the help function on E-Commerce websites should cover all the

(guidance) problems users experience. The help function is mentioned earlier under the sub characteristic of learnability (availability of help functions).

The *human error operation cancellability ratio* is defined as: the ratio of command/data entries that can be cancelled to the total command/data entries [ISO04]. This is a relevant indicator for evaluating the usability of E-Commerce websites. Good E-Commerce websites are able to cope with users who, after making a mistake, cancel an operation. Amazon [AZ05] has a function called "1-click ordering" where users can cancel their purchases within one hour after order confirmation. This is a very good example of dealing with operation cancellation. In a broader context, a good return policy (something which most E-Commerce websites have) is also a way of cancelling an operation.

The *ability to emphasize expressions* is described as: the ratio of number of actually implemented means to required ones which are provided to emphasize expressions to the user for a given set of functions [ISO04]. It is considered that colour, sound, brightness and animation are means to emphasize expressions. this is not a relevant indicator for E-Commerce websites, because the expressions mentioned are not useful in the context of E-Commerce websites. For example, the colour chosen on an E-Commerce website is mostly the company's colour and the user is not allowed to change that.

The *response time for user* is defined as follows: the elapsed time from the user input request or command to the nearest response that is requested by the software or system [ISO04]. The first response to the request may be a state or progress report expressing that the software or the system is processing the request. At least mean, minimum and maximum time should be measured. This is not a relevant indicator, because this indicator deals more with… performance instead of E-Commerce website usability. This indicator is considered to be more relevant for large software programs for large companies or businesses.

Finally, *the display time.* This indicator is defined as: the elapsed time from the current display on screen after request to change, to the next complete display on screen [ISO04]. At least mean, minimum and maximum time should be measured. Just like the previous indicator, this indicator is not considered to be relevant because of the reliability and high speed of Dutch E-Commerce website servers.

To summarize the usability sub characteristic of operability: twenty-seven indicators are given, of which seven are relevant for this thesis. These indicators are:

- Command uniformity
- Consistency of terms in message
- Message clearness
- Skill level adaptability
- Mean time between human error operations
- Operability in practice
- Human error operation cancellability ratio

Explicitness

Definition: attributes that bear on the clarity of the software product with regard to its status (progression bars, etc.) [ISO04-2].

The ISO 9126 standard mentions only one explicitness indicator, which is not really relevant for this thesis:

- Status or progress report availability ratio.

This means the following: the ratio or status or progress reports that are available to the ones required for a given set of functions [ISO04]. For E-Commerce websites, the number of required status or progress reports is low. The only occurrence of a (desired) status report is during the ordering process. Other software products, such as desktop software, can desire more status or progress reports.

QUINT mentions in their ISO 9126+ standard four more indicators:

- Expert judgement on explicitness
- Insecure time
- Status report ratio
- Insecure time in practice

Expert judgement on explicitness means: explicitness of the software product as judged by a team of experts [Zei96]. As mentioned earlier, expert judgement for this thesis cannot be used, as time (and money) do not allow it. For this reason, this indicator is ignored.

The second indicator is the *insecure time*, and is defined as follows: the period of time the user is left insecure about the status of the software product [Zei96]. This is a relevant indicator for evaluating the usability of E-Commerce websites. When a user of an E-Commerce website performs an action, he must not have to wait too long to receive a response. For instance, when he places an order by clicking on the "order" button, the order must be placed instantly or the user must receive a status message which could say something like "your order is being processed". If it is common for the E-Commerce website that it takes a long time to process the order (for instance because of a slow server or much network traffic) and there is no progress indicator, the customer may think that the order has not been processed and clicks again on the "order" button, so that the order is placed twice.

The *status report ratio* (meaning: the ratio of all possible actions, that is immediately followed by a status report of the software product after activation [Zei96]) is not so relevant for E-Commerce websites, because it is not always necessary and desirable to show a status screen. A user does not like to see a status screen after he has clicked on a menu button to visit another section of the website. Status screens can be a nuisance to users, in particular if they appear each time they want to visit another section of the website by using the menu.

Finally, the *insecure time in practice* indicator. This means the following: the period of time the user is left insecure about the status of the software product, reported by the users after a period of use [Zei96]. This indicator is linked strongly with the insecure time indicator. With the insecure time indicator one can measure the insecure time of an action, and with the insecure time in practice one can measure the insecure time of a software product as a whole, by measuring how often and how long insecurities about the status of the software product (for instance an E-Commerce website) occur. So this is a relevant indicator.

To summarize the usability sub characteristic of explicitness: five indicators are given, of which two are relevant for this thesis. These indicators are:

- Insecure time
- Insecure time in practice

Customisability

Definition: attributes of software that enable the software to be customized by the user to reduce the effort required for use and increase satisfaction with the software [ISO04-2].

The sub characteristic customisability appears in the ISO 9126+ standard, which is an addition to the ISO 9126 standard. QUINT derived two indicators from this 9126+ standard:

- Configurability ratio
- Configurability effort

The *configurability ratio* is defined as: the ratio of parts of functionality that can be changed without changing the code to the total number of parts of functionality [Zei96].
The *configurability effort* is defined as: the effort required to change the functionality of the software product [Zei96].

Both indicators are relevant for evaluating the usability of E-Commerce websites. There are examples of E-Commerce websites (e.g. http://www.altavista.co.uk or http://www.bol.com) which deal with the usability sub characteristic of customisability. On these websites you can register as a customer and when you visit the website you can see a personalized page containing offers that may be of interest to you. The customized website is configured in a way that the appearance is different for every user of the E-Commerce website.

Attractivity
Definition: attributes of software that bear on the satisfaction of latent user desires and preferences, through services, behaviour and presentation beyond actual demand [ISO04-2].
The only indicator, mentioned in the ISO 9126+ standard, for this usability sub characteristic is:

- User judgement on attractivity

User judgement on attractivity is defined as: the list of functions that are present in the software product, especially the availability of additional services, behaviour and presentation, as judged by the user [Zei96]. For E-Commerce websites this is a very important indicator. If a visitor finds an E-Commerce website not attractive, he will easily go to the E-Commerce website of a rival company which sells the same products. So, to keep customers to a E-Commerce website, or to attract new customers, a E-Commerce website should be considered attractive by the majority of its users.

Clarity

Definition: attributes of software that bear on the clarity of making the user aware of the functions it can perform [ISO04-2]. Clarity is related to explicitness in a way that explicitness can be interpreted as: clarity as a consequence of being explicit.

Clarity is mentioned as a usability sub characteristic in the ISO 9126+ standard. It has two indicators:

- Function recognition ratio
- Function use ratio

Function recognition ratio means: the ratio of functions that an average unpracticed user distinguishes in the software product without help, within a period of time [Zei96]. This is a very important usability indicator for evaluating the usability of E-Commerce websites. It is absolutely necessary that the functions of the websites, such as the order process, the navigation structure and the payment process, are clear to the E-Commerce website user. The use of metaphors, such as shopping carts, is a good method to improve clarity. This is because these metaphors can be found on many other E-Commerce websites. An E-Commerce website visitor must also be able to recognize most of the functions the E-Commerce website contains. So this visitor can judge if the E-Commerce website can fulfil his needs or not.

Function use ratio is defined as: the ratio of functions of the software product that an average user actually employs after a certain period of time [Zei96]. This is also an important indicator. A webmaster of an E-Commerce website can, by measuring the use ratio of all the pages of the website, determine which pages (functions) are scarcely visited. Possibly there are functions on the website that no user is interested in (or in fact is interested in, but it is not clear to the user what the function means). The webmaster can evaluate whether the function can be left out of the website or he can rename the function so that it is clear what the purpose of the function is.

Helpfulness

Definition: attributes of software that bear on the availability of instructions for the user on how to interact with it [ISO04-2].

For software products, and for E-Commerce websites too, it is necessary that the user gets assistance if he needs it, for instance when he makes a mistake or when he needs to perform a difficult process. For this reason, the ISO 9126+ standard introduced the usability sub characteristic of helpfulness. Helpfulness contains two indicators:

- Ratio of expounding (i.e. elucidating) text
- Normalized ratio of expounding text

The *ratio of expounding text* is defined as: the ratio of the amount of expounding text (including error messages) available in the software product to the total amount of text which can be presented on screen [Zei96]. For E-Commerce websites, this indicator means the following. The ratio of helpful text, including error messages, must be of a certain amount (lines of text on the screen). This means that there must not be too few of these explanatory text, but neither too much. So, a balance must be found. This indicator deals solely with the amount of expounding text and not with the content of the text, which is covered by the usability sub characteristic of understandability.

The *normalized ratio of expounding text* is defined as: the ratio of the amount of expounding text (including error messages) available in the software product to the size of the software product [Zei96]. This indicator is not relevant for this thesis, because it is not easy to determine (for an E-Commerce website visitor) what the normalized ratio of expounding text is. One can measure this for instance by looking at the lines of code for building the website.

User-Friendliness

Definition: attributes of software that bear on the users' satisfaction [ISO04-2]. This also means that a software system (i.e. a website) must be easy to learn and easy to use to be considered user-friendly.

The ISO 9126+ standard gives three indicators concerning the usability sub characteristic of user-friendliness:

- Rated user-friendliness
- Expert judgement on user-friendliness
- User-friendliness compared to sample

Rated user-friendliness is defined as: the ratio of users that prefer the new software product over the previous, after a certain period of practical use [Zei96]. For E-Commerce websites, this indicator is an important one. The rated user-friendliness deals with the aspect of preferring one E-Commerce website above another from the viewpoint of a random user. The new software product is the E-Commerce website the visitor prefers, and the previous software product is another (less preferred) E-Commerce website.

Expert judgement on user-friendliness is defined as: user-friendliness, as judged by a team of experts on topics such as screen composition, vocabulary, application of colour and sound [Zei96].

User friendliness compared to sample means: user-friendliness compared to a sample product, as judged by a team of experts [Zei96]. The experts decide to what extent user-friendliness of the software product matches the sample product. These two indicators are not used for this thesis, because the time (and money) to hire a group of experts to judge the user-friendliness of E-Commerce websites are not available.

3.2 Classification of Usability Quality Characteristics and Quality Attributes

In the previous section, all of the usability sub-characteristics and their indicators have been introduced. In this section, all of the relevant sub-characteristics and indicators will be classified. In the previous section, it was explained why these sub-characteristics and indicators are chosen to be relevant. An overview of these relevant sub-characteristics can be found in Table 4.1, in the next chapter. Now the indicators of each usability sub characteristic by their importance for E-Commerce websites will be classified. That is why the following symbols, as shown in table 3.1, will be used:

Symbol:	**Meaning:**
+	Less important
++	Important
+++	Very important

Table 3.1: Importance of usability indicators

This measurement is based upon judging the importance of the indicators for E-Commerce websites. Not each indicator is as important as another. Following each usability sub

characteristic and its indicators, there will be explained why there is chosen for that specific importance classification.

Understandability:

- Rated understandability (+++)
- Readability score (+++)
- Concept clearness (++)
- Availability of input/output data items list (+)

The classification above is chosen because the rated understandability is an indicator that is very important for E-Commerce websites, because E-Commerce websites contain a wide range of menus, icons, etc. The readability score is a good indicator for E-Commerce websites, because the readability of the website must be appropriate and clear; if not, people have difficulties understanding the website. The concept clearness indicator is useful for E-Commerce websites because of the familiarity website visitors can have while comparing different E-Commerce and other websites. In other words, they know what the concept of an E-Commerce website is and what its purpose is. Finally, the availability of an input/output data items list is a relevant indicator, because there are examples of data lists to be found in E-Commerce websites (see previous section). The importance of this indicator is not so high, because the availability of an input/output data items list is not necessary for an E-Commerce website to be understandable for the user.

Learnability:

- Rate of user errors (+++)
- Time required to learn operation (++)
- Effort required to learn one operation (++)
- Availability of help functions (++)

The classification above is chosen because the rate of user errors is a very important indicator for E-Commerce websites. An E-Commerce website must be made in a way that there is (ideally) no ambiguity within their functionalities, so that a user won't make an error easily. The time required to learn the operation is an important indicator which deals with the familiarity of using an E-Commerce websites. This is related to the concept clearness indicator of the sub characteristic understandability. It is not as important as the rate of user errors: (handling with) errors on a website is more important than learning time. The effort

required to learn one operation is a useful indicator for E-Commerce websites, because users like to see the same metaphors on various E-Commerce websites (for instance, the shopping cart mentioned in the previous paragraph). Because this is related to the time required to learn an operation, this indicator is equally important. Finally, the availability of help functions is relevant, because an E-Commerce website needs a good help function for users who cannot find their way intuitively through the E-Commerce website. Help functions are important, but are more considered as "nice-to-haves", because if a website is well built the availability of help functions is not necessary. For this reason this indicator is the least important one.

Operability:

- Command uniformity (+++)
- Operability in practice (++)
- Consistency of terms in message (++)
- Human error operation cancellability ratio (++)
- Message clearness (++)
- Skill level adaptability (+)
- Mean time between human error operations (+)

The classification, as stated above, is chosen because command uniformity is a very important indicator. Some commands, such as the shopping cart, or the "order" button (and the following ordering process) are preferably uniform on all E-Commerce websites. If one website should have a different command that does the same as the generally used command, website visitors can become confused. Operability in practice is an important indicator because it deals with the functionality of the E-Commerce website, even for visitors without technical knowledge. Consistency of terms in message and message clearness are also important, because the user must be able to react on messages given by the system. The human error operation cancellability ratio is a very useful indicator that makes it possible for users to correct their mistakes. These four indicators are not as important as command uniformity, because the familiarity of operations (commands) on various websites makes an E-Commerce website much more operable for a user. If there were operations that are not common for the user on an E-Commerce website, the user will sooner go to another website. Skill level adaptability is a useful function that E-Commerce websites can offer their visitors [BL05], and the mean time between human error operations is an interesting indicator for website developers and webmasters. These two indicators are more "nice-to-haves" for an E-

Commerce website than “must-haves”. That is why these indicators score lower on importance than the others.

Explicitness:

- Insecure time (++)
- Insecure time in practice (++)

The classification above is chosen, because the insecure time is as important as the insecure time in practice. This is because the insecure time describes the time one is left insecure about a function on the website, which is as important as the time one is left insecure about all of the functions of the website. The difference is to be found in the way it is measured, insecure time on a ratio scale (measure the length of the interval), and the insecure time in practice on an ordinal scale (high, average, low) [Zei96]. These two indicators are closely related to each other, and therefore they score both the same.

Customisability:

- Configurability ratio (+++)
- Configurability effort (++)

The configurability ratio is a very neat feature that E-Commerce websites can offer to their first-time and returning visitors. The use of personalized pages is already discussed in the previous paragraph and is a good example of configurability. It is certainly an important feature for evaluating the usability of E-Commerce websites. The effort to configure certain parts of an E-Commerce website is not so important, because customization of E-Commerce website is mostly considered as an extra function for E-Commerce websites, and not a necessity.

Attractivity:

- User judgement on attractivity (++)

What users find attractive is important for evaluating the usability of E-Commerce websites. If users find an E-Commerce website attractive, it is very likely they use this website to purchase their goods, and they are also more likely to return to this website.

Clarity:

- Function recognition ratio (+++)
- Function use ratio (++)

Users must recognize the functions the E-Commerce website offers. If users cannot find a function on the website, they are likely to go to another E-Commerce website. So it is very important that the users discover all or most of the functions of the website. If, for instance the search function is not easy to find, users can become frustrated and leave the website.

It is also important that visitors of the E-Commerce website use all of the functions the website offers. If some functions are never or hardly used, they may be unclear to (the majority) of website visitors and need to be improved or left out. But the recognition of functions is more important than the use of functions, because different users can use different functions (if they recognize the meaning of the function).

Helpfulness:

- Ratio of expounding (i.e. elucidating) text (+)

The quantity of expounding text is useful for evaluating the usability of E-Commerce websites, because it helps the user to find their way through the website, or helps him to explain what a function or command means. Because many functions and commands are used on different websites in the same way, expounding text is not always necessary and could be placed in the help section of the E-Commerce website.

User-Friendliness:

- Rated user-friendliness (+++)

User-friendliness is a very important indicator to evaluate the usability of E-Commerce websites. If users find an E-Commerce website easy to use, it is very likely they will use this website to purchase their goods, and they are also more likely to return to this website.

With this classification of usability sub-characteristics and their indicators, the E-Commerce websites for this research will be evaluated (with the indicators mentioned above).

Chapter 4: Indicator-Based Evaluation

In the previous chapter, a classification of usability quality characteristics was discussed. In this chapter, an evaluation method is given to test the usability of E-Commerce websites, according to the classification of the usability indicators.

Section 4.1 explains how the classification of the usability quality characteristics can be used for a test.

Section 4.2 provides information about the test group. It gives information about the size of the group, the backgrounds of the test persons and what the tasks are for the group to carry out.

Section 4.3 introduces the Indicator-Based Evaluation method. Here will be explained how E-Commerce websites are evaluated by this Indicator-Based Evaluation method.

Sections 4.4 and 4.5 explain how the evaluation is carried out. This is done with the use of evaluation forms.

Section 4.6 compares the Indicator-Based Evaluation method with the evaluation methods from the related research.

4.1 How to use the classification of usability quality characteristics for a test?

In the classification of the usability quality characteristics, three degrees of importance for the usability indicators are given: less important, important and very important. This classification helps to understand the difference in importance between the indicators.

When a test has been made upon these indicators, the importance needs to be dealt with in an appropriate way. This can be done by giving the “very important” indicators more weight over the “important” and “less important” indicators. To measure the relative importance of each indicator, a certain weight for each indicator group (“very important”, “important” and “less important”) must be assessed. The weight of these indicator groups will be 1/3 (for the less important indicators); 2/3 (important indicators); 1 (very important indicators).

4.2 The test group

The test is designed to be carried out by a group of Dutch adults. The group consists of at least five people with with different backgrounds (e.g. education level) and various degrees of internet knowledge. This group will be extended with an evaluator, who co-ordinates the test group and carries out some of the evaluations by himself. The size of the group is not that large, this is because the time to carry out the evaluation was limited.

The members of the test group will evaluate a selection of Dutch E-Commerce websites (in random order), by means of looking at the usability indicators, without knowing the classification of the usability indicators. So, they do not know which indicators are more important. This is done because the test persons must not think that one indicator is more important than another. If they know this, they can probably put more effort in judging the more important indicators, and less effort in judging the less important ones.

The evaluation will go as follows: the evaluator gives an evaluation form, which contains all of the indicators that are to be evaluated by the test persons, to a test person. The indicators are made easy to understand and are put in a mostly closed multiple choice question format. For instance, the indicator “Concept Clearness” can be put in the evaluation form as follows: “Do you think this E-Commerce website has familiarity with other E-Commerce websites that you have (recently) visited?”. The test person answers Yes or No to this question. Many of these questions are derived from the QUINT research group, who developed the ISO 9126+ standard [Zei96]. The indicators that are evaluated by the evaluator are omitted in this evaluation form and are placed in a separate evaluation form.

The indicators that are assessed in the evaluation forms for the test persons and the evaluation form for the evaluator are presented in Section 4.4. These evaluation forms are written in Dutch, because the test group consists of Dutch native speakers and the E-Commerce websites that are evaluated are also in the Dutch language. Appendices A and B contain the evaluation forms for the evaluator (Appendix A) and the test persons (Appendix B). These forms are translated into English.

4.3 The method of evaluating each of the indicators

Table 4.1 shows all of the 24 relevant indicators with their importance, as they were explained in Section 3.2 of this thesis.

No	Indicator	Usability characteristic	Importance
1	Rated understandability	Understandability	+++ (Very important)
2	Readability score	Understandability	+++
3	Concept clearness	Understandability	++ (Important)
4	Availability of input/output data items list	Understandability	+ (Less important)
5	Rate of user errors	Learnability	+++
6	Time required to learn operation	Learnability	++
7	Effort required to learn one operation	Learnability	++
8	Availability of help functions	Learnability	++
9	Command uniformity	Operability	+++
10	Operability in practice	Operability	++
11	Consistency of terms in message	Operability	++
12	Human error operation cancellability ratio	Operability	++
13	Message clearness	Operability	++
14	Skill level adaptability	Operability	+
15	Mean time between human error operations	Operability	+
16	Insecure time	Explicitness	++
17	Insecure time in practice	Explicitness	++
18	Configurability ratio	Customisability	+++
19	Configurability effort	Customisability	++
20	User judgment on attractivity	Attractivity	++
21	Function recognition ratio	Clarity	+++
22	Function use ratio	Clarity	++
23	Ratio of expounding text	Helpfulness	+
24	Rated user-friendliness	User-Friendliness	+++

Table 4.1: Usability indicators and their importance

With the help of these indicators and their importance, an evaluation method for the indicators can be developed. These evaluation methods are discussed below. These methods and the proposed answering methods are derived from the ISO 9126+ standard [Zei96].

1. Rated understandability (+++):

1) Show the menus, icons, etc. to a test person
2) Let the test person decide on understandability, just by answering yes or no

3) Appreciate understandability as:

High: 80% or more of the test persons found the menus, icons, etc. understandable

Average: 60% to 80% of the test persons found the menus, icons, etc. understandable

Low: less than 60% of the test persons found the menus, icons, etc. understandable

This indicator can be easily included in an evaluation form. A test person can carry out the evaluation of this indicator. The proposed evaluation method is derived from the ISO 9126+ standard [Zei96].

2. Readability score (+++):

1) Determine the average number of words (on the homepage) per sentence
2) Determine the percentage of familiar words according to a standard dictionary
3) Calculate the readability score according to the Dale-Chall Readability formula [Höc84]:

   Readability score = 0.1579 nfw + 0.0496 sl + 3.6305

   With *nfw* = percentage unfamiliar words and *sl* = average sentence length

This indicator is quite difficult for the test person to evaluate, so it needs to be evaluated by the evaluator himself. The outcome of the readability score formula must be interpreted as follows: a higher score denotes a text which is less easy to read. A score between 1 and 8 can be explained as a elementary school reading level, a score between 9 and 12 can be explained as a high school reading level, and a score between 13 and 16 can be explained as a college reading level. These reading levels are derived from the American school and college system. To come to a good evaluation of readability score, the readability score should be below 8. To be an average evaluation, the value of the readability score should be between 9 and 12, to receive a a low evaluation of the readability score, the score should be 13 or higher. The proposed evaluation method is derived from the ISO 9126+ standard [Zei96].

3. Concept Clearness (++):

1) Ask the test person if the E-Commerce website is similar to other E-Commerce websites he has recently visited

2) Appreciate concept clearness as:

High: 80% or more of the test persons found the website familiar with other E-Commerce websites

Average: 60% to 80% of the test persons found the website familiar with other E-Commerce websites

Low: less than 60% of the test persons found the website familiar with other E-Commerce websites

This indicator can be included in an evaluation form. A test person can carry out the evaluation of this indicator by himself. The proposed evaluation method is derived from the rated understandability metric (see above).

4. Availability of input/output data list (+):

1) Determine if there are input/output operations to be found on the E-Commerce website (such as “search” operation or “price calculator”). Just answer yes or no

This indicator can be evaluated by the evaluator himself. There is no need for the test person to do this as well. The proposed evaluation method is derived from the ISO 9126-4 [ISO04] standard.

5. Rate of user errors (+++):

1) Determine how many times the test person encounters incorrect behaviour by misunderstanding an operation on the E-Commerce website

2) Appreciate rate of user errors as:

High: 50% or more of the test persons encountered incorrect behaviour by misunderstanding two or more operations on the E-Commerce website

Average: 50% or more of the test persons encountered incorrect behaviour by misunderstanding one operation on the E-Commerce website

Low: at least 50% of the test persons did not encounter incorrect behaviour by misunderstanding any of the operations on the E-Commerce website

The proposed evaluation method is derived from the “rate of user errors” definition from the ISO 9126-4 [ISO04] standard.

6. Time required to learn operation (++):

1) The evaluator prepares a random action that can be taken on an E-Commerce website, for instance the action to purchase an article
2) The test person carries out this action on the E-Commerce website and measures the time required to perform this action for the first time (x)
3) An hour later, the test person carries out the same action on the same E-Commerce website and again measures the time required to perform this action (y)
4) Appreciate time to learn operation as:

Very short:	80% or more of the test persons have a (x/y) > 2 result*
Short:	between 50% and 80% of the test persons have a (x/y) > 2 result
Average:	50% or more of the test persons have a (x/y) > 1.2 result
Long:	50% or more of the test persons have a (x/y) < 1.2 result
Very long:	80% or more of the test persons have a (x/y) < 1.2 result

* = "(x/y) > 2" means that the user carries out the same action twice as fast the second time, compared to the first time.

The proposed evaluation method is derived from the "Time required to learn operation" definition from the ISO 9126-4 [ISO04] standard. The QUINT research group has developed the weights of the results by doing their own research.

7. Effort required to learn one operation (++):

1) The evaluator selects a not very common operation on the E-Commerce website, for instance the "add items to wish list" operation on the Amazon website [AZ05]
2) The test person carries out this operation on the E-Commerce website and measures the time required to perform this action for the first time (x)
3) An hour later, the test person carries out the same operation on the same E-Commerce website and again measures the time required to perform this action (y)
4) Appreciate effort to learn operation as:

Very little effort: 80% or more of the test persons have a (x/y) > 1.5 result*

Little effort: between 50% and 80% of the test persons have a (x/y) > 1.5 result

Average effort: 50% or more of the test persons have a (x/y) > 1 result

Much effort: 50% or more of the test persons have a (x/y) < 1 result

Very much effort: 80% or more of the test persons have a (x/y) < 1 result

* = "(x/y) > 1.5" means that the user carries out the same action the second time 150% faster, compared to the first time.

The proposed evaluation method is derived from the "Effort required to learn operation" definition from the ISO 9126-4 [ISO04] standard.

8. Availability of help functions (++):

1) The test person counts the times he does not know how to perform an action on the E-Commerce website and requires some assistance
2) The test person counts the times that a help function was available to assist him
3) Appreciate availability of help functions as:

High: 80% or more of the test persons found a help function when they needed one

Average: 60% to 80% of the test persons found a help function when the needed one

Low: less than 60% of the test persons found a help function when the needed one

The proposed evaluation method is derived from the "Availability of help functions" definition from the ISO 9126-4 [ISO04] standard.

9. Command uniformity (+++):

1) The test person counts the number of commands (actions, buttons) where he thinks that the commands don't have a uniform format, but which perform a common task. (For instance, the uniform shopping cart icon is replaced by less common icon.)

2) Appreciate the command uniformity as:

High: 80% or more of the test persons found no non-uniform formatted commands

Average: 60% to 80% of the test persons found no non-uniform formatted commands

Low: less than 60% of the test persons found no non-uniform formatted commands

The proposed evaluation method is derived from the "Command uniformity" definition from the ISO 9126-4 [ISO04] standard.

10. Operability in practice (++):

1) Let a test person express their judgement in one of the following options:

High: Little or no appeal is made to the tester's technical knowledge

Average: The tester has to perform non-functional (i.e.) technical operations

Low: High appeal is made to the tester's technical knowledge

This indicator can be included in an evaluation form. A test person can carry out the evaluation of this indicator by himself. The proposed evaluation method is derived from the ISO 9126+ standard [Zei96].

11. Consistency of terms in message (++):

1) The test person determines whether the system messages that he encounters are common (standard) messages. This means that these messages can be found on other E-Commerce and other websites as well

2) Let the test person express consistency of terms in message as:

High: No system messages were found by the test person, or all of the system messages (i.e. pop-up messages) were consistent, compared to system messages found on other E-Commerce and other websites

Average: Few inconsistent system messages compared to system messages found on other E-Commerce and other websites

Low: Many inconsistent system messages compared to error messages found on other E-Commerce and other websites

3) The evaluator calculates the average consistency of terms in message after all of the test persons have performed the evaluation.

The proposed evaluation method is derived from the "Consistency of terms in message" definition from the ISO 9126-4 [ISO04] standard.

12. Human error operation cancellability ratio (++):

1) The test person tries different operations on the E-Commerce website and looks if his actions can be cancelled. An example of an operation is to buy an article, but to cancel the order in a very late stage of the ordering process
2) Appreciate human error operation cancellability as:
   Very High: All operations can be cancelled; even some operations which were finalized can be cancelled (i.e. return policy after delivery)
   High: All operations can be cancelled, but finalized operations can't be cancelled or were not evaluated
   Average: More than 50% of the operations can be cancelled
   Low: Few or none of the operations can be cancelled
3) The evaluator calculates the average human error operation cancellability ratio after all of the test persons have performed the evaluation

The proposed evaluation method is derived from the "Human error operation cancellability ratio" definition from the ISO 9126-4 [ISO04] standard.

13. Message clearness (++):

1) The test person indicates whether messages that he receives are clear to him just by navigating on the E-Commerce website. The cause of the message and the action that must be taken after the message must be clear to the test person
2) Appreciate message clearness as:
   High: All messages were clear, the cause of the message and the actions to be taken were all clear
   Average: Most of the messages were clear, the cause of the message and the actions to be taken were most of the time clear

Low: Few or none of the messages were clear, the cause of the message and the actions to be taken were most of the time not clear

3) The evaluator calculates the average message clearness after all of the test persons have performed the evaluation

The proposed evaluation method is derived from the "Message clearness" definition from the ISO 9126-4 [ISO04] standard.

14. Skill level adaptability (+):

1) The evaluator looks on the E-Commerce website if there are different "levels" of navigation accessible on the E-Commerce website. For instance: a more skilled person (someone who has already ordered many products from E-Commerce websites) can order a product in a faster way than a person who has never bought something online
2) Appreciate skill level adaptability as:

High: The E-Commerce website has taken many measures for more skilled persons in a way that they can use the E-Commerce website in an easier way

Average: The E-Commerce website has taken some measures for more skilled persons in a way that they can use the E-Commerce website in an easier way

Low: The E-Commerce website has taken few or no measures for more skilled persons in a way that they can use the E-Commerce website in an easier way

This indicator can be evaluated by the evaluator himself. There is no need for the test person to do this as well. The proposed evaluation method is derived from the ISO 9126-4 [ISO04] standard.

15. Mean time between human error operations (+):

1) The evaluator looks for major error possibilities on the E-Commerce website. An example of a major error is typing in a wrong bank account or credit card number

2) The evaluator looks whether the E-Commerce website is able to detect this error and whether the E-Commerce website provides feedback to the user
3) Appreciate mean time between human error operations as:
   High: The major errors were detected and were properly dealt with by giving feedback
   Average: Most of the major errors were detected and were properly dealt with by giving feedback
   Low: Few of the major errors were detected and were properly dealt with by giving feedback

Note: it is difficult to measure the exact time between the errors, because each E-Commerce website is different and so these times can vary largely. It is more important that the E-Commerce website can detect human errors, such as typing an incorrect bank account number and to give feedback to the user before the order of a product is finalized.

This indicator can be evaluated by the evaluator himself. There is no need for the test person to do this as well. The proposed evaluation method is derived from the ISO 9126-4 [ISO04] standard.

16. Insecure time (++):

1) The evaluator selects a representative set of actions on the website
2) The tester performs these actions
3) The tester measures the length of the interval before status information is displayed
4) The evaluator determines the average interval for each set, tested by each tester

A possible action set is the purchase process (select item → place in basket → order → confirm/cancel order). During these actions, the customer interacts with the server. If the server is slow, the customer will have to wait a long time. The proposed evaluation method is derived from the ISO 9126+ standard [Zei96].

17. Insecure time in practice (++):

1) Ask a tester to write down how often and how long he has been kept insecure about the status of the E-Commerce website
2) The evaluator collects these notes
3) The explicitness, based on these notes, is determined in terms of:
   High: Users are hardly kept insecure after the start of an action, and if so, only for a short time
   Average: Same, but more often and for a longer period
   Low: Same, but most of the time and for a longer period

The difference with Insecure time is that Insecure time in practice is applicable to the entire website (and the entire evaluating process), while insecure time only covers a set of actions.
The proposed evaluation method is derived from the ISO 9126+ standard [Zei96].

18. Configurability ratio (+++):

1) The evaluator counts the number of functions (actions, operations) which can be performed to configure the E-Commerce website in his own way. An example of a configurable function is the personalized profile function
2) Appreciate configurability ratio as:
   High: Many functions can be customized (compared to other E-Commerce websites. More than on the average E-Commerce website)
   Average: Some functions can be customized (compared to other E-Commerce websites. About as much as on the average E-Commerce website)
   Low: Few functions can be customized (compared to other E-Commerce websites. Less than on the average E-Commerce website)

The proposed evaluation method is derived from the "Configurability ratio" definition from the ISO 9126-4 [ISO04] standard.

19. Configurability effort (++):

1) The evaluator selects a configurable function (action, operation) on the E-Commerce website
2) The test person measures the configurability effort to perform the customization of the website by using the function presented by the evaluator, just by saying *easy*, *average* or *difficult*

The proposed evaluation method is derived from the “Configurability effort” definition from the ISO 9126-4 [ISO04] standard.

20. User judgement on attractivity (++):

1) Show the tester the extra options, extra functionalities, available on the website
2) Let the tester express these extras as:
   - *High:* The tester believes all desirable features are applied
   - *Average:* The tester believes only a limited number of features are applied
   - *Low:* The tester believes very few features are applied
3) The evaluator determines the average appreciation (based on all test reports)

The proposed evaluation method is derived from the ISO 9126+ standard [Zei96].

21. Function recognition ratio (+++):

1) Present a test person a list of available functions on the E-Commerce website
2) Ask the test person to mark the functions that he found during a certain period of use
3) The evaluator calculates the average ratio of functions found to the total number of functions

The proposed evaluation method is derived from the ISO 9126+ standard [Zei96].

22. Function use ratio (++):

1) Provide a tester a list of available functions on the E-Commerce website
2) Ask the tester to mark the functions that he actually used on the website
3) The evaluator calculates the average ratio of functions found to the total number of functions

To receive better evaluation results, the function use ratio indicator can be done prior to the function recognition ratio. The proposed evaluation method is derived from the ISO 9126+ standard [Zei96].

23. Ratio of expounding text (+):

1) The evaluator determines the total amount of text to be displayed on screen in number of words
2) The evaluator determines the total amount of expounding text in number of words
3) The evaluator calculates the ratio of expounding text

This indicator can be evaluated by the evaluator. There is no need for the tester(s) to do this too. The proposed evaluation method is derived from the ISO 9126+ standard [Zei96].

24. Rated user-friendliness (+++):

1) The evaluator asks the test group to fill in a questionnaire about the E-Commerce website's user-friendliness
2) The evaluator calculates the ratio

The proposed evaluation method is derived from the ISO 9126+ standard [Zei96].

4.4 The evaluation forms based on ISO 9126+ indicators

In section 4.2, two evaluation forms were proposed. One evaluation form is used by the evaluator to assess the indicators which are for him to evaluate. The other evaluation form is used by the test person(s) and includes all of the indicators that need to be assessed by the test group. Tables 4.2 and 4.3 show the indicators that need to be assessed by the evaluator (4.2) and the indicators that need to be assessed by the test group (4.3).

Indicator		
Readability score	Skill level adaptability	Mean time between human error operations
Configurability ratio	Ratio of expounding text	Availability of input/output data items list

Table 4.2: Indicators to evaluate by the evaluator

Indicator		
Rated understandability	Concept clearness	Rate of user errors
Time required to learn operation	Effort req. to learn one operation	Availability of help functions
Command uniformity	Operability in practice	Consistency of terms in message
Message clearness	Insecure time	Insecure time in practice
Rated user-friendliness	Configurability effort	User judgment on attractivity
Function recognition ratio	Function use ratio	
Human error operation cancellability ratio		

Table 4.3: Indicators to evaluate by the test group

The indicators that are on the evaluation form for the evaluator have no specific order in which they need to be evaluated. The evaluator must also prepare six of the indicators for the test group, before they can evaluate an E-Commerce website. These six indicators are:

- Time required to learn operation;
- Effort required to learn one operation;
- Insecure time;
- Configurability effort;
- Function recognition ratio;
- Function use ratio.

In the previous section was explained what the evaluator has to do to prepare these indicators.

The indicators that are on the evaluation form for the test group have a specific ordering. This is, because some indicators must be evaluated first. For instance, the function use ratio needs to be evaluated before the function recognition ratio.
Additionally, some indicators, such as “time required to learn operation”, need to be evaluated twice (in the beginning of the evaluation, and at the end). For this reason an ordering is made. Table 4.4 shows this ordering.

No (see table 4.1)	**Indicator**
6	Time required to learn operation (first evaluation)
7	Effort required to learn one operation (first evaluation)
3	Concept clearness
5	Rate of user errors
17	Insecure time in practice
8	Availability of help functions
16	Insecure time
19	Configurability effort
22	Function use ratio
21	Function recognition ratio
12	Human error operation cancellability ratio
9	Command uniformity
10	Operability in practice

No (see table 4.1)	Indicator
13	Message clearness
11	Consistency of terms in message
1	Rated understandability
24	Rated user-friendliness
20	User judgment on attractivity
6	Time required to learn operation (second evaluation)
7	Effort required to learn one operation (second evaluation)

Table 4.4: Ordering of indicators for evaluation by the test group

Some remarks need to be made on some of the indicators. “Message clearness” (13) and “Consistency of terms in message” (11) can only be evaluated if there are messages appearing on the E-Commerce website. If there are no messages shown on the E-Commerce website, these indicators can be omitted (left blank).

For “The rate of user errors” (5), the “Insecure time in practice” (17) and the “Availability of help functions” (8) must be said that these indicators need to be evaluated during the entire evaluation process. The test person counts occurrences of specific terms or items to assess these indicators.

4.5 The scenario of the evaluation

In this paragraph it is explained how the Indicator-Based Evaluation takes place. In short, there are three phases: preparation, testing and evaluation.

4.5.1 Preparation phase

In the preparation phase, the evaluator selects the E-Commerce websites which will be evaluated. For this thesis, five Dutch E-Commerce websites have been selected (see Section 5.3). For each of the selected websites the evaluator fills in the evaluator evaluation form, containing the six indicators of Table 4.2. After this, the evaluator prepares the four indicators which are specific for each website. These indicators are:

- Time required to learn operation;
- Effort required to learn operation;
- Insecure time;
- Configurability effort.

For the Time required to learn operation and the effort required to learn operation, the evaluator needs to choose an operation that can be found on the specific E-Commerce website. This operation may be found on any of the E-Commerce websites that will be evaluated, but that is not necessary. In other words, it may be possible that for each website another operation is chosen.

For the Insecure time, the evaluator selects a representative set of actions that can be found on the specific E-Commerce website. It is also possible that for each website another set of actions is chosen. The evaluator also measures the insecure time of the set of actions himself. This value will be the average value, what will be compared with the times the test persons have measured.

For the Configurability effort, the evaluator selects a configurable function (if possible) on the specific E-Commerce website. It is also possible that for each website another configurable function is chosen.

The evaluator will select a group of test persons. For each of the test persons, the evaluator notes age, gender, computer experience, online shopping experience, the highest level of education completed and the internet connection speed at the location where the test person did the evaluation. This parameter is important because some of the indicators deal with time, and when a test person has a slow internet connection (for instance a dial-up modem), the test results can be influenced by this.

The evaluator tries to create a group of test persons that have different profiles, so that not each person for instance has the same computer experience and online shopping experience.

4.5.2 Testing phase

During this phase, the test persons will evaluate the selected E-Commerce websites by using their evaluation form. In the preparation phase, the evaluator has selected the E-Commerce websites and has prepared some of the indicators.

The procedure for each website is presented below:
The test person picks one of the E-Commerce websites and evaluates this website by filling in the evaluation form that goes with this website. These evaluation forms are not generic,

because the evaluator has placed some website-specific questions on the forms. There is a specific ordering in filling in the questions on the evaluation form. This ordering is presented in Table 4.4. Each question on the evaluation form corresponds with one of the indicators. After evaluating the first E-Commerce website, the test person picks another one and fills in the corresponding evaluation form. The test person has finished his evaluation when all of the E-Commerce websites are evaluated and therefore all of the evaluation forms are completely filled in.

There is no specific time limit for the test person to complete the entire evaluation. The evaluator can suggest a date by which the evaluation needs to be complete. It is not to be preferred that the test person evaluates all of the E-Commerce websites directly after each other, because the attention of the test person can drop and the (later) evaluations are not that accurate anymore.

4.5.3 Evaluation phase

After all the test persons have completed their evaluations, the evaluator collects the evaluation forms and orders them by website. Now, for each website, the evaluator looks at the answers the test persons have given for each question. The evaluator calculates the average value for each indicator. Here follows the conversion scheme for each of the values the indicators can have (the number between brackets corresponds with the numbers of Table 4.1):

Rated understandability (1), Concept clearness (3), Availability of help functions (8), Command uniformity (9), Operability in practice (10), Consistency of terms in message (11), Message clearness (13), Skill level adaptability (14), Mean time between human error operations (15), Insecure time in practice (17), Configurability ratio (18), User judgment on attractivity (20):

- High = 1
- Average = 2/3
- Low = 1/3

Readability score (2):

- Score lower than 9 = 1

- Score between 9 and 12 = 3/5
- Score higher than 12 = 1/5

Availability of input/output data list (4):

- Yes = 1
- No = 0

Rate of user errors (5):

- High = 1/3
- Average = 2/3
- Low = 1

Time required to learn operation (6):

- Very Short = 1
- Short = 4/5
- Average = 3/5
- Long = 2/5
- Very Long = 1/5

Effort required to learn operation (7):

- Very little effort = 1
- Little effort = 4/5
- Average effort = 3/5
- Much effort = 2/5
- Very much effort = 1/5

Human error operation cancellability ratio (12):

- Very High = 1
- High = 8/10
- Average = 6/10
- Low = 3/10

Insecure time (16):

- Faster than evaluator time = 1
- The same as evaluator time = 8/10
- Slower than evaluator time = 1/2
- Very much slower = 2/10

Configurability effort (19):

- Easy = 1
- Average = 2/3
- Difficult = 1/3

Function recognition ratio (21), Function use ratio (22):

- The values are computed as follows: the average ratio of functions found to the total number of functions

Ratio of expounding text (23):

- If much expounding text is found = 1
- If little expounding text is found = 1/2
- If no expounding text is found = 0

Rated user-friendliness (24):

- Tester finds website friendly to use = 1
- Tester finds website not friendly to use = 0

The proposed values are based upon the proposed evaluation methods from the QUINT research group [Zei96] and are normalized to values between 0 and 1.

When the evaluator has got the average values for each of the indicators the test persons have evaluated, he receives the values of the indicators the evaluator himself has evaluated for the website.

Now the evaluator looks at the importance of each indicator (Table 4.1). For each indicator the evaluator multiplies the average evaluation value with the importance of the indicator (see

Section 4.1). The evaluator adds all of these values, and the number that the evaluator finally gets is the Indicator-Based Evaluation score for the E-Commerce website. The higher this score is, the better the E-Commerce website's usability is, according to this evaluation method.

It may be possible that for one or more indicators, an E-Commerce website provides no evaluation material. For instance, if there is no user customization possible on the E-Commerce website, the Configurability effort and Configurability ratio indicators cannot be evaluated. The questions regarding these indicators will then be omitted from the evaluation form(s). The scores for these omitted questions and also for their indicators will be ½. The value of ½ is the most fair weight of an omitted question, because a value of 0 (zero) implies that when one of the indicators is missing, the website suffers a lot of "damage" because of the relatively low Indicator-Based Evaluation score. Giving the missing indicator a score of 1 is also unfair, because this implies that a missing indicator is good (you receive the maximum score for something that is not on the website). For this reason a score of ½ has been given.

4.6 Indicator-Based Evaluation compared to the related research approaches

Looking at the evaluation methods from the related research, the following can be said: the proposed evaluation method which is based on the ISO 9126+ standard has the same advantages as Nielsen's method (Section 2.3.1): it can be done with 5 (or more) test persons and gives a close approximation to actual individual usage. Furthermore, it doesn't have the disadvantages of Nielsen's "think aloud method". The evaluation method is not unnatural and the test persons do not have to describe the tasks they do.

Just like Benbunan-Fich's Protocol analysis method (Section 2.3.2) the evaluation method covers the field of E-Commerce websites. Protocol analysis also makes use of evaluation parameters. The Indicator-Based Evaluation method doesn't use these evaluation parameters.

Tilson (Section 2.3.3) gives users the opportunity to rate the significance of factors by themselves, so that they can say what they like and don't like on the website to measure the user-friendliness and the attractivity of the website. The Indicator-Based Evaluation method does this too, but in a more strict way. The indicators rated user-friendliness and user judgment on attractivity deal with this topic. Indicator-Based Evaluation makes use of the ISO 9126+ indicators to determine what the testers find about the E-Commerce website, while

Tilson uses the direct input from the test persons (not put in a certain framework, such as the ISO 9126+ standard).

The Indicator-Based Evaluation method doesn't have any clear relation with Kauffman's Mathematical approach evaluation method (Section 2.3.4). No cybernetics is used to identify the importance of economics in E-Commerce activities.

Segawa's Scenario-based walkthrough method (Section 2.3.5) makes use of scenarios to define the E-Commerce website objectives. The Indicator-Based Evaluation method doesn't make use of scenario's; it is more natural, because the testers are more free in their choices (they are not bound to the scenarios, but can navigate freely through the website). Every person navigates through an E-Commerce website in their own way.

Automated web evaluation (Section 2.3.6) is a fast evaluation method and makes use of targeted guidelines. By using indicators, the Indicator-Based Evaluation method uses some kind of guidelines. The test person is given direction by means of these indicators to evaluate (a specific part of) an E-Commerce website.

Chapter 5: E-Commerce websites, Objectives and the selection of these websites

In this chapter it is explained what the objectives are of different Dutch E-Commerce websites. For whom is a certain E-Commerce website designed and made, and who buys products from this site? Are there many E-Commerce websites with the same objective, or are there many differences? (Section 5.1)

It will also be explained what the specific properties of Dutch E-Commerce websites are, with the emphasis on the usability of these websites (Section 5.2).

Finally it will be discussed which E-Commerce websites are used for the Indicator-Based Evaluation (Section 5.3) by the test group.

5.1 Dutch E-Commerce websites, the Universe of Discourse

There are many different kinds of E-Commerce websites to be found in the Netherlands. There are very specialized websites where a customer can buy very specific goods, for instance ink for an inkjet printer. But there are also more general websites where one can buy several different kinds of goods, such as clothing and sports gear. In this paragraph a relevant cross-section will be made of the field of Dutch E-Commerce websites.

Two aspects are important to determine the relevance of the E-Commerce websites: popularity of the E-Commerce website (in terms of number of visits) and the kind of E-Commerce website (is it a specialized website or a general one).

Looking at the Multiscope list of most frequently visited Dutch websites of 2005 [Mul05], five E-Commerce websites are placed in the top 20. These websites are: Marktplaats.nl (position 4), Speurders.nl (position 8), Ebay.nl (position 10), nl.Bol.com (position 15) and Wehkamp.nl (position 20). The list of these most frequently visited websites is presented in Figure 5.1.

It attracts immediate attention that the three highest placed websites are all online trading places, where people can place and advert on the website or respond to an already placed advert. These adverts are of all kinds, from clothing to bicycles and from pets to art objects.

The other two E-Commerce websites in the top 20 are also quite general online shops. Bol.com sells books, CDs, DVDs and computer games, while Wehkamp.nl sells clothes, sports gear, furniture and electronics.

2005		2004	Site
1	↑	3	Google.nl
2	↓	1	Startpagina.nl
3	↓	2	Msn.nl '
4	↑	6	Marktplaats.nl
5	↓	4	Detelefoongids.nl
6	↑	10	Msn.com
7	→	7	Postbank.nl
8	●	99	Speurders.nl
9	↓	8	Anwb.nl
10	↑	20	Ebay.nl
11	↑	15	Rtl.nl
12	→	12	Rabobank.nl
13	↓	9	Startkabel.nl
14	↓	5	Ilse.nl
15	↓	11	Nl.bol.com
16	●	23	Staatsloterij.nl
17	●	21	Nu.nl
18	↓	14	Planet.nl
19	→	19	Ns.nl
20	↓	18	Wehkamp.nl

Legend
↑ = Site went up in ranking
↓ = Site went down in ranking
→ = Site is stable
● = Site is new in top 20

Figure 5.1: Top 20 of best visited websites (source: Multiscope)

Not in the top 20 are the more specialized E-Commerce websites as Kruidvat.nl (photography), InkClub.nl (printer ink), Kijkshop.nl (electronics, watches), Dell.nl (computers), Bruna.nl (books and computer games) and Conrad.nl (electronics). In a research of 2003 by Nielsen Netratings [Nie03] Kijkshop and Bruna are also mentioned as growing E-Commerce websites (in terms of visitors).

The websites of Conrad, Bol.com and Wehkamp were given the “Thuiswinkel Publieksprijzen 2006” (awards from the people visiting Dutch E-Commerce websites) in their categories. [Thu06] Conrad was the best specialized shop, Bol.com was the best multimedia shop and Wehkamp was the best department store.

It can be said that the more general E-Commerce websites are visited the most, while more specialized websites are staying behind. This is because the target group of the more general websites is larger. A specialized E-Commerce website as InkClub will attract only customers who are interested in buying fairly priced inkjet printer ink, while a general website as Marktplaats will attract all kinds of people. This is no surprise, but interesting to keep in mind while selecting a range of E-Commerce websites (not choose only too general websites, but also choose some specialized websites).

5.2 Properties of E-Commerce websites

Looking at the properties of the two major classes of E-Commerce websites, the general E-Commerce websites and the specific E-Commerce websites, there are many similarities to be found. Obviously, all of these websites offer the possibility to buy goods from their website. The way in which this can be done is mostly the same: the customer selects an article he wants to buy and selects how he wants to pay for this article and then the order is finalized.

Other properties that can be found are:

- Large banner or pop-up containing the newest product or a special offer;
- A (interactive) FAQ section where customers can look for an answer or ask a question themselves;
- Shopping basket;
- Prize contest;
- Personalized pages;
- Search function;
- Banners / Links to other websites;
- “Thuiswinkel waarborg” (home shopping hallmark);
- Order status;
- … and many more.

Some of these properties are directly linked with usability, such as the personalized pages (configurability), the search function (clarity) and the FAQ section (learnability). Others are more indirectly linked with usability (for instance a shopping cart is a metaphor. This metaphor is related to the operability sub characteristic).

5.3 The selection of Dutch E-Commerce websites

To make a good selection of the E-Commerce websites for this thesis, it is important to cover a large field of websites. In this field, general and specialized websites must be selected to come to a more diverse intersection of the domain. A balance between general and specific websites must be found, and the number of visitors (popularity of the website) also plays an important role, because a well-visited website should be popular, and therefore the usability of this website should be high, according to the Indicator-Based Evaluation method.

For the Indicator-Based Evaluation a selection of five websites largely covers the field of Dutch E-Commerce websites. This selection contains websites which are popular and/or general and/or specialized. The five websites are:

- www.nl.bol.com;
- www.conrad.nl;
- www.kruidvat.nl;
- www.marktplaats.nl;
- www.wehkamp.nl.

Here follows an explanation why each of these websites is included to this selection, by giving the most important properties of the websites:

www.nl.bol.com:

- General website;
- Won the “Thuiswinkel Award 2006” in the Multimedia category;
- 15th position of best visited websites 2005 (4th E-Commerce website).

www.conrad.nl

- Specialized website (electronics);

- Won the “Thuiswinkel Award 2006” in the Specialized shop category.

www.kruidvat.nl:

- Specialized website (mostly photography);
- Growing in popularity [Mul05].

www.marktplaats.nl:

- General website;
- 4th position of best visited websites 2005 (1st E-Commerce website);

www.wehkamp.nl:

- General website;
- 20th position of best visited websites 2005 (5th E-Commerce website).

From the top-20 of best visited websites, a top-5 of best visited E-Commerce websites can be made. From this top-5, the places 1, 4 and 5 are included to this selection. The reason why the 2nd and 3rd placed websites are not included is the following: Speurders.nl and Ebay.nl are having the same purpose as Marktplaats.nl, but are less popular. For this research is it not relevant to include more than one online-trading website.

With this selection of five Dutch E-Commerce websites, a good representation of the field of Dutch E-Commerce websites is given. It includes most of the Universe of Discourse.

Because of the limited time in which this thesis was written, it was not possible to examine the Universe of Discourse more thoroughly.

Chapter 6: The results of the Indicator-Based Evaluation

For this thesis the five E-Commerce websites presented in the previous chapter are evaluated by a test group of five persons. This group was extended with one evaluator.
In this chapter, the results of the evaluation are presented. The structure of this chapter is described below.

In the first section, the test group is presented. What are the backgrounds of the test persons regarding computer experience, online shopping experience and education?

In section 6.2 the evaluation results of one of the websites is thoroughly described. Each of the questions (on the evaluator form and on the test person form) will be discussed.

In section 6.3 the (summarized) results of the four other websites are presented. Because the method of evaluation for these websites is the same as for the website chosen for section 6.2, there is no need to discuss each of the questions again.

In section 6.4 the interpretation of the website scores is explained. In this chapter is discussed what the Indicator-Based Evaluation score means, and what a good (or bad) score is.

In the final section a conclusion regarding the evaluation (results) is given.

6.1 The test group

The test group consisted of five persons and one evaluator. The evaluator role was carried out by the author. The reason to do so was the following: The questions that the evaluator has to answer are mostly very trivial. There was no need for another person to act as evaluator as well. It is assumed that the effect of appointing someone else for the evaluator role doesn't have much effect on the evaluation results (the answers on the questions would be the same).

The details of the evaluator are shown in Table 6.1:

Age	26
Gender	Male
Computer experience	Professional
Online shopping experience	Very high
highest level of education	HBO (university of professional education)
Internet speed	Cable modem

Table 6.1: Details of the evaluator

The details of the test group is presented in Table 6.2:

	Test Person 1	**Test Person 2**	**Test Person 3**	**Test Person 4**	**Test Person 5**
Age	53	50	42	22	24
Gender	Male	Female	Male	Male	Male
Computer experience	Average / Little	None	Professional	Professional	Professional
Online shopping experience	Average / Little	None	Average	High	Average
highest level of education	HBO (university of professional education)	VWO (pre-university education)	MBO (senior secondary vocational education)	HBO (university of professional education)	VWO (pre-university education)
Internet speed	Cable modem	Cable modem	Cable modem	ADSL modem	Cable modem

Table 6.2: Details of the test group

The test group consists of people with different experience regarding online shopping and internet use. The age of the test group varies from 22 to 53 years. Their highest level of education completed is quite high and varies from senior secondary vocational education (MBO) to university of professional education (HBO). This means that the test group consists of people with different backgrounds, which makes the evaluation more fair.

6.2 The extended evaluation results of one website

In this section the evaluation results of the website www.conrad.nl are presented. First the answers on the questions of the evaluation form are discussed. Thereafter the results of the test persons are given.

6.2.1 The evaluator form results

Here the results of the evaluator are presented. For the complete questions corresponding with these answers, please see Appendix A of this thesis. In Section 4.5.3, the interpretation of the results is given (the answers are converted to (numeric) values, to make it possible to come to a general evaluation score).

Question 1a: Average number of words = 4

Question 1b: Percentage familiar words = 50

Question 1c: Readability score = 11.72

This leads to a (converted to numeric) readability score of **0.6** (see Section 4.5.3).

Question 2: Skill level is High.

This leads to a (converted to numeric) skill level score of **1**.

Question 3: Error handling is High.

This leads to a (converted to numeric) error handling score of **1**.

Question 4: Yes, there were input/output operations to be found on the Conrad website.

This leads to a (converted to numeric) input/output operations score of **1**.

Question 5: Personalization is Average.

This leads to a (converted to numeric) personalization score of **0.67**.

Question 6a: Number of words = 200

Question 6b: Number of words = 20

This leads to a (converted to numeric) explanatory text score of **0.5**.

Question 7: Action chosen = “Search, in the Model Train product section (which can be found on the main menu), for the Digital Start Set N 89395”

Question 8: Operation chosen = “Search for and download the calibration list”

Question 9: Action chosen = “Choose three products, each of an other product group, and place these in the shopping cart. After this, look in the shopping cart”
Time needed to perform this action = 1 minute, 40 seconds

Question 10: Function chosen = “Place three randomly chosen products in the shopping cart. Look in the shopping cart and look on the same page if the offers shown down on the page are (quite) relevant to the products placed in the shopping cart”

Question 11:

- Links “Contact” up to “Nieuwsbrief” on top of the page
- Links “Snel zoeken”, “Bestelnr zoeken”, “Inloggen” and “Winkelwagen” on top of the page
- Navigation by using the horizontal menu bar (tab structure) on top of the page
- (Red) link with Home on the top left of te page (under the Conrad logo)
- Menu on the left side of the page
- Links entirely below on the page
- The “Bladercatalogus”
- The “Conrad Infocenter”
- Special offers on the homepage
- The “Catalogusaanvraag”
- The “Special Shops”
- The “Catalogus Download Infocenter”
- See the General Conditions
- Sign up as “Affiliate”
- See Press Releases about Conrad
- View Shopping Cart
- Order a product
- “Direct bestellen van een product” (fast ordering)

This completes the results of the evaluator form. For the first six questions, which are the questions regarding the indicators that are ONLY evaluated by the evaluator, points are given. The total score for the evaluator part is: **0.6 + 1 + 1 + 1 + 0.67 + 0.5 = 4.77**

Now, the importance of each indicator (Table 4.1) is used to compute the Indicator-Based Evaluation score for the evaluator part: (0.6 x 1) + (1 x 0.33) + (1 x 0.33) + (1 x 0.33) + (0.67 x 1) + (0.5 x 0.33) **= 2.43**

The questions 7 up to 11 are questions the evaluator makes for the test group. The scores that go with these questions are included in the test person evaluation, and not in the evaluator evaluation.

6.2.2 The test person form results

In this section the results of the evaluation, done by the five test persons, of the Conrad website is presented. For the complete questions corresponding with these answers, please see Appendix B of this thesis. In Section 4.5.3, the interpretation of the results is given (the answers are converted to numeric values, to make it possible to come to a general evaluation score).

Before presenting the results, the time needed to evaluate the Conrad website is presented in Table 6.3.

Test person	**Time needed**
Test person 1	29 minutes
Test person 2	41 minutes
Test person 3	30 minutes
Test person 4	25 minutes
Test person 5	30 minutes

Table 6.3: Time needed to evaluate the Conrad website

The average evaluation time is approximately **30 minutes**. The variance of this evaluation time is **32 minutes and 36 seconds**. The standard deviation of the evaluation time for the Conrad website is **5.71**.

Question 1: Average time needed to perform action = 150 seconds and the mean deviation is 48 seconds.

(Test person 1: 90 seconds; Test person 2: 180 seconds; Test person 3: 240 seconds; Test person 4: 120 seconds; Test person 5: 120 seconds)

Question 19 (corresponds with question 1):

Average time needed to perform action = 195 seconds and the mean deviation is 173 seconds.

(Test person 1: 30 seconds; Test person 2: 600 seconds; Test person 3: 90 seconds; Test person 4: 60 seconds; Test person 5: 60 seconds)

This leads to a Time required to learn operation score of *Very short*, and can be converted to the numeric value of **1**. (see Section 4.3).

Question 2: Average time needed to perform operation = 300 seconds and the mean deviation is 344 seconds.

(Test person 1: 70 seconds; Test person 2: operation not found after 900 seconds; Test person 3: 180 seconds; Test person 4: operation not found; Test person 5: operation not found)

Question 20 (corresponds with question 2):

Average time needed to perform action = 15 seconds* and the mean deviation is 4 seconds.

(Test person 1: 12 seconds; Test person 2: operation not found; Test person 3: 20 seconds; Test person 4: operation not found; Test person 5: operation not found)

This leads to a Time required to learn operation score of *Average*, and can be converted to the numeric value of **0.6**. (see Section 4.3).

* = This score is not *High* because three of the five persons weren't able to find the operation.

Question 3: Similarity with other E-Commerce websites?

Yes: 80 % (test persons 1, 2, 4 and 5);

No: 20 % (test person 3).

This leads to a Concept clearness score of *High*, and can be converted to the numeric value of **1**. (see Section 4.3).

Question 4: Number of times of unusual behaviour of the E-Commerce website:

0 : Test persons 2, 4 and 5;

2 : Test persons 1 and 3;

This leads to a Rate of user errors score of *low*, and can be converted to the numeric value of **1**. (see Section 4.3).

Question 5: Average number of times the user was insecure = 1.4 and average waiting time after being insecure = approx. 1 minute:

(Test person 1: 3x insecure with average waiting time of 45 seconds; Test person 2: 0x insecure; Test person 3: 4x insecure with average waiting time of 10 seconds AND one time endless waiting time; Test person 4: 0x insecure; Test person 5: 0x insecure)

This leads to a Insecure time score of *average*, and can be converted to the numeric value of **0.67**. (see Section 4.3).

Question 6: Average number of times the user didn't know how to perform an action = 0.4, average number of times that there was support = 0.

(Test persons 1, 3 and 5 did always know how to perform an action and didn't need any support; Test persons 2 and 4 both didn't know how to perform one (1) action, and there wasn't any support)

This leads to an Availability of help functions score of *low*, and can be converted to the numeric value of **0.33**. (see Section 4.3).

Question 7: Average time before status information is shown = 120 seconds and the mean deviation is 42 seconds.

(Test person 1: 105 seconds; Test person 2: 120 seconds; Test person 3: 180 seconds; Test person 4: 20 seconds; Test person 5: 150 seconds)

This leads to a Insecure time score of *Slower than evaluator time* (100 seconds; see Section 6.2.1), and can be converted to the numeric value of **0.5** (see Section 4.3).

Question 8: Customizability of the website. For this question it was required that the test person has a Conrad profile. None of the test persons has a Conrad profile, so this question is skipped.

The configurability effort score for this question will therefore be *Average*, which leads to a converted numeric value of **0.67** (see Section 4.3). The value of *Average* is the most fair weight for this question, because a value of 0 (zero) implies that when this indicator is missing, the website suffers a lot of "damage" because of the relatively low Indicator-Based Evaluation score for this indicator. Giving the missing indicator a score of 1 is also unfair, because this implies that a missing indicator is good (you receive the maximum score for something that is not on the website). That is why a score of 0.67 is been given.

Question 9: Average number of functions found = 9.4 of a total of 18 functions, with a mean deviation of 4.2.

(Test person 1: 12 functions found; Test person 2: 15 found; Test person 3: 7 found; Test person 4: 11 found; Test person 5: 2 found)

This leads to a Function use ratio score of 9.4 / 18 = **0.52** (see Section 4.3).

Question 10:

Average number of functions found in 5 minutes = 16.4 of a total of 18 functions, with a mean deviation of 1.3.

(Test person 1: 18 functions found; Test person 2: 15 found; Test person 3: 15 found; Test person 4: 18 found; Test person 5: 16 found)

This leads to a Function recognition ratio score of 16.4 / 18 = **0.91** (see Section 4.3).

Question 11: Average possibility to cancel an operation on the website is to be found high.
(Test persons 1 and 2: Very high; Test persons 3, 4 and 5: High)

This leads to a Human error operation cancellability ratio of *High*, and can be converted to the numeric value of **0.8** (see Section 4.3).

Question 12: Average number of unknown commands = 2.2
(Test persons 1, 2, 4 and 5: none; Test person 3: 11 unknown commands)

This leads to a Command uniformity score of *High*, and can be converted to the numeric value of **1** (see Section 4.3).

Question 13: Average functionality without technical appeal = Average
(Test persons 1 and 5: High; Test person 2: Average; Test persons 3 and 4: Low)

This leads to a Command uniformity score of *Average*, and can be converted to the numeric value of **0.67** (see Section 4.3).

Question 14: Average message clearness = Average
(Test persons 2 and 5: High; Test persons 1 and 4: Average; Test person 3: Low)

This leads to a Message clearness score of *Average*, and can be converted to the numeric value of **0.67** (see Section 4.3).

Question 15: Average consistency of terms in message = High
(Test persons 1, 2, 4 and 5: High; Test person 3: Average)

This leads to a Consistency of terms in message score of *High*, and can be converted to the numeric value of **1** (see Section 4.3).

Question 16a: Are the menus on the website understandable?

Yes : Test persons 1, 2, 3 ,4 and 5

No : -

Question 16b: Are the icons/images on the website understandable?

Yes : Test persons 1, 2, 4 and 5

No : Test person 3

Question 16c: Is the navigational structure on the website understandable?

Yes : Test persons 1, 2, 4 and 5

No : Test person 3

This leads to a Rated understandability score of *High*, and can be converted to the numeric value of **1** (see Section 4.3).

Question 17a: Testers' positive comments on the user-friendliness of the (Conrad) website:

Test person 1: - well-organized
- good classification of product groups
- a wide range of products

Test person 2: - good tab structure
- clear catalogue
- clear ordering process

Test person 3: - Nothing found ("due to irritation about the Conrad website", the test person said)

Test person 4: - Clear menu structure
- Always clear where you are on the website (via "You are here" on every page)

- The use of just one base colour (Blue) and not an entire rainbow of colours. This makes the website easier to use

Test person 5: - well-organized
- calm (not busy)
- fast

Question 17b: Testers' negative comments on the user-friendliness of the (Conrad) website:

Test person 1: - Nothing found
("Just a good website if you are looking for these kinds of products")

Test person 2: - Not everything can be found on the website ("Calibration list" of question 2)

Test person 3: - Make the icons/pictures more consistent
- Integrate the infocenter, or at least put it in the same window and in the same colour set. I thought I came on a website of another company
- Reduce the number of adverts and/or product choices

Test person 4: - Improve message feedback. In the first question I've put some products in the shopping cart, but there wasn't a message that the products were really put in the shopping cart. So, you are waiting for a message which won't appear
- On the home page there are two menus which are the same (above and left on the page). The space on the home page is already limited, so this is not necessary
- Vivid home page with images and adverts which can be a lot smaller.

Test person 5: - Double information or lists on two places
- There is no general search function for the entire website

The test group (in general) thinks that the navigation of the Conrad website contributes to the user-friendliness of the website. The fact that the website is not so clear, because of the large images and adverts, is not good for the user-friendliness of the website, according to the test group.

In general it can be said that there are more (better) positive remarks on the Conrad website, so the Rated user-friendliness score for the website will be **1** (see Section 4.3).

Question 18a: Testers' positive comments on the attractivity of the (Conrad) website:

Test person 1: - Couldn't find anything in particular

Test person 2: - Nothing found

Test person 3: - Nothing was positive

Test person 4: - Nothing found

Test person 5: - Nothing found. Most of the things on the website are so common and can be found on any E-Commerce website

Question 18b: Testers' negative comments on the attractivity of the (Conrad) website:

Test person 1: - Didn't miss anything in particular

Test person 2: - Nothing found

Test person 3: - I missed clearness, consistency and a clear navigation structure

Test person 4: - Due to double menus and many pictures the overview on the home page is not clear

Test person 5: - No great failures concerning attractivity found

In general it can be said the website was not very attractive. The test group couldn't find anything positive about the Conrad website. This means that the Conrad website is not extra attractive. On the other hand, test persons 3 and 4 have mentioned some negative parts concerning attractivity of the website.

For this reason ir can be said that there are more negative remarks on the Conrad website, so the Rated attractivity score for the website will be *Low*, which can be converted to the numeric value of **0.33** (see Section 4.3).

On the evaluation form, the test person was able to fill in some comments about the questions or the evaluation itself. For the Conrad website, these are the comments of the test persons.

Test person 1: - A good evaluation form. It is also a way to deal with this kind of websites in a easier and faster way

Test person 2: - Calibration list not found (questions 2 and 20)

- I'd prefer to go to a real shop instead of shopping online

Test person 3: - No comments

Test person 4: - Calibration list not found (questions 2 and 20)

Test person 5: - Question 7 was not very clear, because it wasn't clear whether I was supposed to write down the total time or the average waiting time

These personal comments don't affect the score for the Conrad website, but might be useful to improve the Indicator-Based Evaluation method, for instance by modifying the questions to prevent ambiguity.

This completes the results of the test group forms. For each of the questions, points varying from 0 to 1 are given.

The total (average) score for the test group part of this evaluation is: **1 + 0.6 + 1 + 1 + 0.67 + 0.33 + 0.5 + 0.67 + 0.52 + 0.91 + 0.8 + 1 + 0.67 + 0.67 + 1 + 1 + 1 + 0.33 = 13.67**

Now, the importance of each indicator (Table 4.1) is used to compute the Indicator-Based Evaluation score for the test group part:

(1 x 0.67) + (0.6 x 0.67) + (1 x 0.67) + (1 x 1) + (0.67 x 0.67) + (0.33 x 0.67) + (0.5 x 0.67) + (0.67 x 0.67) + (0.52 x 0.67) + (0.91 x 1) + (0.8 x 0.67) + (1 x 1) + (0.67 x 0.67) + (0.67 x 0.67) + (1 x 0.67) + (1 x 1) + (1 x 1) + (0.33 x 1) **= 10.89**

The total score for the evaluator part was: **2.43**

This leads to a total Indicator-Based Evaluation score for the Conrad website of:

2.43 + 10.89 = 13.32

6.3 The summarized evaluation results of the other four websites

In this section the results of the evaluation of the other four E-Commerce websites are presented. The evaluation of these websites took place in the same way as for the Conrad website (see Section 6.2). That is why, only the scores of these websites, and some striking comments of the test group are presented.

At the end of this section, a table presents the overview of the evaluation results (Table 6.3).

6.3.1 Wehkamp

The score for the evaluator part of this website is:

1 + 0.67 + 1 + 1 + 0.67 + 0.5 = 4.84

(the ordering of these scores corresponds with the questions on the evaluation form)
Now, the importance of each indicator (Table 4.1) is used to compute the Indicator-Based Evaluation score for the evaluator part: (1 x 1) + (0.67 x 0.33) + (1 x 0.33) + (1 x 0.33) + (0.67 x 1) + (0.5 x 0.33) **= 2.72**

The average time needed for the test group to carry out the evaluation is **30 minutes**.
The (average) score for the test group of this evaluation is:
0.8 + 1 + 0.33 + 1 + 1 + 1 + 0.8 + 0.67 + 0.64 + 0.92 + 0.8 + 0.33 + 1 + 1 + 1 + 1 + 1 + 1 = 15.29
(the ordering of these scores corresponds with the questions on the evaluation form)
Now, the importance of each indicator (Table 4.1) is used to compute the Indicator-Based Evaluation score for the test group part:
(0.8 x 0.67) + (1 x 0.67) + (0.33 x 0.67) + (1 x 1) + (1 x 0.67) + (1 x 0.67) + (0.8 x 0.67) + (0.67 x 0.67) + (0.64 x 0.67) + (0.92 x 1) + (0.8 x 0.67) + (0.33 x 1) + (1 x 0.67) + (1 x 0.67) + (1 x 0.67) + (1 x 1) + (1 x 1) + (1 x 1) **= 11.98**

This leads to a total Indicator-Based Evaluation score for the Wehkamp website of:
2.72 + 11.98 = 14.70

Remarkable comments from the test group are listed below:

- “logical, recognisable classification” (Test person 1, Question 17a)
- “this was my first visit of an E-Commerce website, and therefore I couldn’t compare this website with other E-Commerce websites” (Test person 2, General comment)
- “I clicked on Customer Support, but I was lead to the My Profile page. This is not good” (Test person 2, Question 17b)
- “Slightly “boxed” structure due to the strict use of frames” (Test person 3, Question 17b)
- “Clear navigational structure. You can navigate to a product group in different ways” (Test person 3, Question 17a)
- “Bright photography, except on the “Claire” clothing for firm girls (I cannot see these firm girls).” (Test person 3, Question 17b)
- “You can search for a comparable (similar) article. This is very useful” (Test person 3, Question 18a)

6.3.2 Kruidvat

The score for the evaluator part of this website is:

1 + 0.33 + 0.33 + 1 + 0.33 + 0 = 2.99

(the ordering of these scores corresponds with the questions on the evaluation form)

Now, the importance of each indicator (Table 4.1) is used to compute the Indicator-Based Evaluation score for the evaluator part: (1 x 1) + (0.33 x 0.33) + (0.33 x 0.33) + (1 x 0.33) + (0.33 x 1) + (0 x 0.33) **= 1.87**

The average time needed for the test group to carry out the evaluation is **1 hour, 5 minutes**.

The (average) score for the test group of this evaluation is:

0.4 + 0.8 + 0.67 + 0.67 + 0.67 + 0.67 + 0.5 + 0.5 + 0.27 + 0.85 + 0.6 + 0.67 + 0.67 + 0.67 + 0.67 + 1 + 0 + 0.33 = 9.61

(the ordering of these scores corresponds with the questions on the evaluation form)

Now, the importance of each indicator (Table 4.1) is used to compute the Indicator-Based Evaluation score for the test group part:

(0.4 x 0.67) + (0.8 x 0.67) + (0.67 x 0.67) + (0.67 x 1) + (0.67 x 0.67) + (0.67 x 0.67) + (0.5 x 0.67) + (0.5 x 0.67) + (0.27 x 0.67) + (0.85 x 1) + (0.6 x 0.67) + (0.67 x 1) + (0.67 x 0.67) + (0.67 x 0.67) + (0.67 x 0.67) + (1 x 1) + (0 x 1) + (0.33 x 1) **= 8.28**

This leads to a total Indicator-Based Evaluation score for the Kruidvat website of:

1.87 + 8.28 = 10.15

Remarkable comments from the test group are listed below:

- "I'd prefer another icon for the shopping cart" (Test person 1, Question 17b)
- "I could find the puzzle tool (Question 2), but I wasn't able to move a picture in this tool" (Test person 2, General comment)
- "It wasn't clear to me how I should order pictures" (Test person 2, Question 18b);
- "A bad, unclear website, no form of assistance available" (Test person 2, General comment)
- "The "Ik wil snel naar" (go to) function on the website sometimes disappears" (Test person 3, Question 17b)
- "Website is sometimes slow, because of the downloading of plugins" (Test person 3, Question 17b)

- “Picture rotation function within the puzzle function worked surprisingly well!” (Test person 3, Question 18a)
- “There was no possibility to view my ordered products directly from the home page” (Test person 4, Question 18b)
- “No pop-ups” (Test person 5, Question 17a)
- “Error messages in the German language (although it is a Dutch E-Commerce website) on the photo department of the website” (Test person 5, Question 18b)

6.3.3 Marktplaats

The score for the evaluator part of this website is:

0.6 + 0.67 + 0.67 + 1 + 0.67 + 0.5 = 4.11

(the ordering of these scores corresponds with the questions on the evaluation form)

Now, the importance of each indicator (Table 4.1) is used to compute the Indicator-Based Evaluation score for the evaluator part: (0.6 x 1) + (0.67 x 0.33) + (0.67 x 0.33) + (1 x 0.33) + (0.67 x 1) + (0.5 x 0.33) **= 2.21**

The average time needed for the test group to carry out the evaluation is **30 minutes**.

The (average) score for the test group of this evaluation is:

1 + 1 + 1 + 1 + 1 + 1 + 0.8 + 0.67 + 0.36 + 0.82 + 0.8 + 0.67 + 1 + 1 + 1 + 1 + 1 + 1 = 16.12

(the ordering of these scores corresponds with the questions on the evaluation form)

Now, the importance of each indicator (Table 4.1) is used to compute the Indicator-Based Evaluation score for the test group part:

(1 x 0.67) + (1 x 0.67) + (1 x 0.67) + (1 x 1) + (1 x 0.67) + (1 x 0.67) + (0.8 x 0.67) + (0.67 x 0.67) + (0.36 x 0.67) + (0.82 x 1) + (0.8 x 0.67) + (0.67 x 1) + (1 x 0.67) + (1 x 0.67) + (1 x 0.67) + (1 x 1) + (1 x 1) + (1 x 1) **= 12.61**

This leads to a total Indicator-Based Evaluation score for the Marktplaats website of:

2.21 + 12.61 = 14.82

Remarkable comments from the test group are listed below:

- “It was possible to enter an invalid postal code” (Evaluator, Question 3 on the Evaluator’s form)
- “Once you’re used to the website it is very easy to find your way. During the last few years, nothing has changed on the website” (Test person 1, Question 17a)

- "The product groups are very clear" (Test person 2, Question 17a)
- "Very much a text-based website. This is good for consistency" (Test person 3, Question 17a)
- "Use of colours might be a bit more vivid" (Test person 3, Question 17b)
- "There was (just) one moving image on the website. It would be better if this image didn't move" (Test person 3, Question 18b)
- "Clear menu structure", Test person 4, Question 17a)
- "The use of the available space is very good. The menu is not too large, so that there is more than sufficient room for the items next to the menu" (Test person 4, Question 17a).

6.3.4 Bol

The score for the evaluator part of this website is:

1 + 1 + 0.67 + 1 + 0.67 + 1 = 5.34

(the ordering of these scores corresponds with the questions on the evaluation form)

Now, the importance of each indicator (Table 4.1) is used to compute the Indicator-Based Evaluation score for the evaluator part: (1 x 1) + (1 x 0.33) + (0.67 x 0.33) + (1 x 0.33) + (0.67 x 1) + (1 x 0.33) **= 2.88**

The average time needed for the test group to carry out the evaluation is **45 minutes**.

The (average) score for the test group of this evaluation is:

1 + 1 + 1 + 0.67 + 1 + 0.33 + 1 + 0.67 + 0.4 + 0.83 + 0.8 + 1 + 1 + 1 + 1 + 1 + 1 + 0.67 = 15.37

(the ordering of these scores corresponds with the questions on the evaluation form)

Now, the importance of each indicator (Table 4.1) is used to compute the Indicator-Based Evaluation score for the test group part:

(1 x 0.67) + (1 x 0.67) + (1 x 0.67) + (0.67 x 1) + (1 x 0.67) + (0.33 x 0.67) + (1 x 0.67) + (0.67 x 0.67) + (0.4 x 0.67) + (0.83 x 1) + (0.8 x 0.67) + (1 x 1) + (1 x 0.67) + (1 x 0.67) + (1 x 0.67) + (1 x 1) + (1 x 1) + (0.67 x 1) **= 12.00**

This leads to a total Indicator-Based Evaluation score for the Bol website of:

2.88 + 12.00 = 14.88

Remarkable comments from the test group are listed below:

- “Good background information on the articles” (Test person 1, Question 17a)
- “The slogan “lang leve de luiheid” (long live laziness) doesn’t make me very enthusiastic” (Test person 1, Question 18b)
- “The helpdesk really works!” (Test person 2, Question 17a)
- “The ordering process and the way to cancel your purchase are clear and easy to use” (Test person 2, Question 17a)
- “Not all of the terminology used on the website is understandable for me” (Test person 2, Question 17b)
- “Clear and comprehensive search engine/function” (Test person 3, Question 17a)
- “There are several ways to find your product”, Test person 3, Question 17a)
- “The font use can be more consistent” (Test person 3, Question 17b)
- “The website is neither surprisingly good nor bad” (Test person 3, Question 18a)
- “Good use of (one base) colour” (Test person 4, Question 17a)
- “Too much information on the home page. This makes this page a bit unclear” (Test person 4, Question 17b)
- “Quite fast website” (Test person 5, Question 17a)

In the previous section (Section 6.2) and this section, the results of the Indicator-Based Evaluation (IBE) were presented. As an overview, Table 6.4, shows these results.

Website:	**Evaluation time*:**	**Evaluator score:**	**Test group score*:**	**Total IBE score:**
Conrad	30 minutes	2.43	10.89	13.32
Wehkamp	30 minutes	2.72	11.89	14.70
Kruidvat	1 hour, 5 minutes	1.87	8.28	10.15
Marktplaats	30 minutes	2.21	12.61	14.82
Bol	45 minutes	2.88	12.00	14.88

Table 6.4: Results of Indicator-Based Evaluation

(* = Average time/score)

The fact that three of the four IBE scores are quite close to each other, doesn’t mean the Indicator-Based Evaluation method doesn’t have much discriminating power. When looking at the separate indicator scores, these sites score each on different indicators.

6.4 The interpretation of the Indicator-Based Evaluation scores

The minimal score an E-Commerce website can receive by using the Indicator-Based Evaluation method, is **4.13 points** (0.75 points for the evaluator part and 3.38 points for the test group part). The highest possible Indicator-Based Evaluation score for an E-Commerce website is **17.03 points** (3.32 points for the evaluator part and 13.71 points for the test group part).

Knowing these scores, the website Indicator-Based Evaluation scores can be put in perspective.

The Kruidvat website had the lowest score (10.15) of the five evaluated websites. This is a significant lower score than the other four E-Commerce websites. Their scores are quite close to each other, especially the scores of the Wehkamp, Marktplaats and Bol websites. That this is purely coincidental can be verified by looking at the individual scores per question (see Sections 6.2 and 6.3).

Looking at a scale, based on the lowest possible and highest possible scores for Indicator-Based Evaluation, a judgement can be made concerning the usability of the E-Commerce websites that were evaluated. This scale is listed below in table 6.5. The scale can be interpreted as a nominal scale, which implies a certain kind of classification. This judgement is based on the test done by the test group and the Indicator-Based Evaluation score. The purpose of this scale is that an exclamation of for instance “This E-Commerce website has a very good usability, according to the Indicator-Based Evaluation” makes more sense than “This E-Commerce website scores 15.10 on the Indicator-Based Evaluation scale”. It is easier to understand when the numbers are “labelled” with a text. This score gives a first impression (the site is good, okay, poor), for a more precise meaning of the score one can look at the scores of the individual indicators.

Indicator-Based Evaluation score:	**Usability of the website is:**
Lower than 7.50	Very poor
Between 7.50 and 9.99	Poor
Between 10.00 and 12.49	Okay, Reasonable
Between 12.50 and 14.99	Good
Higher than 15.00	Very good

Table 6.5: Interpretation of Indicator-Based Evaluation scores

According to table 6.5, the following can be said about the usability of the five E-Commerce websites which were evaluated: The Kruidvat website, with an Indicator-Based Evaluation score of 10.15 has a Reasonable usability. The other four websites scored all between 12.50 and 14.99 and have therefore a Good usability, according to the Indicator-Based Evaluation method.

6.5 Conclusion of the evaluation

Looking at the selection of these five websites (Section 5.3), the Kruidvat website was the only E-Commerce website which didn't win an award or didn't appear on a list of best visited websites. This is a possible explanation why the Kruidvat E-Commerce website scored rather low. The fact that the other four websites either have won an award or have been present on a list of best visited websites suggests that the owners of these websites have spent time (and money) on (improving the) usability of their websites.

Looking at the test group profiles, it can be said that the person with the least computer experience (and, as well, the least E-Commerce experience) needed the most time for evaluating the E-Commerce websites. It also seems that age and education are less important than computer experience. Test person 1, aged 53, has average/little computer experience and their highest level of education completed. He is never the fastest person during the evaluation, while the three computer professionals are always the fastest persons. A test person's gender appears not to play a role, looking at the evaluation results.

Chapter 7: Conclusion

In this thesis a new usability evaluation method for E-Commerce websites, called Indicator-Based Evaluation, is presented.

First, a closer look was taken at the existing evaluation methods. Some advantages of the Indicator-Based Evaluation method over these existing methods are speed (fast evaluation method), availability of targeted guidelines and the use of evaluation parameters (by means of indicators). The Think Aloud method has the advantage that it can be carried out by a small number of people, which makes the evaluation inexpensive. The Indicator-Based Evaluation method can be carried out by 5 or more people, and is also a inexpensive evaluation method. Some undesirable features of the existing methods are the unnatural way the evaluation takes place and the scope of the methods (broader scope than just E-Commerce websites). These disadvantages do not occur in the Indicator-Based Evaluation method.

For the Indicator-Based Evaluation method, the ISO 9126(+) standard is used as a guidance aid. This standard consists of usability characteristics and indicators. These indicators are the most important part of the evaluation method. Each of these indicators covers a specific part of the (E-Commerce website) usability and have a certain weight according to their importance.

By evaluating all of these indicators, an Indicator-Based Evaluation score can be calculated. This score shows how good or bad the usability of the evaluated E-Commerce website is. The ISO 9126(+) standard, with its indicators, covers the field of usability of E-Commerce websites. The Indicator-Based Evaluation method is therefore an effort to come to a formalized way to evaluate the usability of E-Commerce websites.

After evaluating five E-Commerce websites with the Indicator-Based Evaluation method, the evaluation results are the results that were more or less expected. Because four of the five websites chosen received either an award or were present on the list of best-visited websites, it would be reasonable to assume that these websites have an adequate (or good) usability. The Indicator-Based Evaluation method proves this assumption: these four websites got a good score. The fifth website (Kruidvat) didn’t appear on a list of best-visited websites nor

has it won an award. So it can be assumed that this website should receive a lower usability score. The Indicator-Based Evaluation method proves this assumption: the usability of the Kruidvat website is not so good.

This means that the Indicator-Based Evaluation method proves to be a valid evaluation method.

The Indicator-Based Evaluation score tells us what the usability of a website is, compared to other E-Commerce websites. The final score is not the most meaningful part of the evaluation; with the Indicator-Based Evaluation method it is also possible to compare the usability of E-Commerce websites for each indicator separately.

The evaluator gave the test group two weeks time to evaluate the E-Commerce websites, so that the test group could spread the evaluation. This meant that the test persons were not obliged to evaluate all five websites without a pause. This made evaluating websites by using the Indicator-Based Evaluation method not so stressful.

The Indicator-Based Evaluation method is also proven to be a fast evaluation method: it takes about 40 minutes for a test person to evaluate one E-Commerce website. To fill in all five evaluation forms, the test person needed approximately 3 hours and 20 minutes. The Indicator-Based Evaluation method is especially created to evaluate E-Commerce websites (and not a broader scope). These are great advantages over the existing (E-Commerce) website evaluation methods.

Another advantage of the Indicator-Based Evaluation method is the way the evaluation takes place. The evaluation process is more natural than the unnatural videotaped sessions (when the test person is monitored by a video camera and must say with a loud voice what he is doing) of other methods (Nielsen, Benbunan-Fich).

References

[Ayk05] Aykin, N. *Usability and Internalization of Information Technology*, Laurence Erlbaum Associates London, 2005

[AZ05] www.amazon.nl / www.amazon.co.uk, website, amazon.com, Seattle, USA, 2005

[Ben01] Benbunan-Fich, R., Using protocol analysis to evaluate the usability of a commercial website, *Information and Management*, Volume 39, Number 2, December 2001, pp. 151-163

[BL05] www.nl.bol.com, website, bol.com bv, Vianen, The Netherlands, 2005

[Cha02] Chaffey, D., *E-Business and E-Commerce management: Strategy, Implementation and Practice*, Edinburgh Gate, Harlow: Pearson Education Limited, 2002

[DD05] R. Molich (DialogDesign), *TowerTest Report, Usability Test of www.TowerRecords.com*, Denmark, 2005

[Fin05] Finlay, A., Dix, A., Evaluation techniques (presentation), University of Western Australia, 2005

[Gal04] Galin, D., *Software Quality Assurance, From theory to implementation*, Pearson Education Limited, 2004

[Hen00] Hendriks, P., *Kwaliteitszorg van software-ontwikkeling*, Software Engineering Research Centre (SERC), 2000 (Dutch)

[Höc84] Höcker, H. et al., Comparative descriptions of Software Quality Measures, *GMD-Studien nr. 81*, Gesellschaft für Mathematik und Datenverarbeitung GmbH, Bonn, 1984

[Hol02] Holmes, M., *Web Usability and Navigation (a beginner's guide)*, McGraw-Hill Companies, 2002

[IEEE90] Institute of Electrical and Electronics Engineers (IEEE), *IEEE standard 610.1-12, Standard Glossary of Software Engineering*, IEEE, 1990

[ISO00] International Organization for Standardization (ISO), *Quality management systems, fundamentals and vocabulary -- ISO 9000:2000*, ISO, 2000

[ISO04] International Organization for Standardization (ISO), International Electrotechnical Commission (IEC), *Software engineering -- Product quality -- Part 4: Quality in use metrics – ISO/IEC 9126-4*, ISO, 2004

[ISO04-2] International Organization for Standardization (ISO), International Electrotechnical Commission (IEC), *Software engineering – Product quality – Part 1: Quality Model – ISO/IEC 9126-1*, ISO, 2004

[Ivo00] Ivory, M., Hearst, M., *The state of the art in automated usability evaluation of user interfaces (Report)*, Computer Science Division, University of California, 2000

[Kau05] Kauffman, R., Hahn, J., A Methodology for Business Value-Driven Website Evaluation, working paper, Purdue University, 2005

[Mul05] Multiscope, press release, http://www.multiscope.nl/persbericht-volledig.phtml?uniek=20051201162408, Multiscope, 2005

[Nie93] Nielsen, J., *Usability Engineering*, Academic Press, London, 1993

[Nie99] Nielsen, J., *Designing Web Usability, the practice of simplicity*, New Riders Press, 1999

[Nie03] Nielsen Netratings, website, http://www.nielsen-netratings.com/pr/pr_030121_netherlands.pdf, Nielsen Netratings, 2003

[Seg05] Segawa, S., Sugimura, M., Ishigaki, K., New Web-Usability Evaluation Method: Scenario-Based Walkthrough, Oxford Journals, 2005

[Sha91] Shackel, B., *Usability – context, framework, definition, design and evaluation*, Cambridge University Press, 1991

[Shn04] Shneiderman, B., *Designing the User Interface: Strategies for Effective Human-Computer Interaction, Fourth Edition*, Addison-Wesley, 2004

[Thu06] http://www.thuiswinkel.org/onderdeel/awards/, website, Thuiswinkel.org, 2006

[Til98] Tilson, R., Dong, J., Martin, S., Kieke, E., Factors Affecting the Usability of E-Commerce Sites, *4th Conference on Human Factors and the Web*, IBM, 1998

[TR05] www.towerrecords.com, website, MTS incorporated, 2005

[Van02] Vanderdonckt, J., Beirekdar, A., Noirhomme-Fraiture, M., A Framework and a Language for Usability Automatic Evaluation of Web Sites by Static Analysis of HTML Source Code, *Kluwer Academics Pub*, 2002, pp 337-348

[Van05] Vanderdonckt, J., Beirekdar, A., Automated Web Evaluation by Guideline Review, Université catholique de Louvain (UCL), School of Management (IAG), 2005

[Vdh97] Vanderheiden, G. C., E.a., *Unified Website Accessibility guidelines*, Version 7.2, Trace Center, University of Wisconsin-Madison, 1997

[Zei96] van Zeist, B., Hendriks, P., Paulussen, R., Trienekens, J., *Kwaliteit van softwareprodukten*, Deventer, Kluwer Bedrijfswetenschappen, 1996 (Dutch)

Appendix A: Evaluator evaluation form

On the following five pages the evaluator evaluation form is shown.

Indicator-Based Evaluation

Evaluator form

This form is meant for the evaluator to evaluate a Dutch E-Commerce website. With this form, the evaluator evaluates the indicators that are for him to evaluate. The evaluator also prepares some of the indicators for the test group with this form.

The indicators that are evaluated or prepared with this form are (in brackets the number of the corresponding question on this form is shown):

Evaluated by the evaluator:

- Readability score (1);
- Skill level adaptability (2);
- Mean time between human error operations (3);
- Availability of input/output data items list (4);
- Configurability ratio (5);
- Ratio of expounding text (6).

Prepared by the evaluator:

- Time required to learn operation (7);
- Effort required to learn one operation (8);
- Insecure time (9);
- Configurability effort (10);
- Function use ratio/function recognition ratio (11).

Please fill in these details before starting the evaluation on the following pages:

Date:	
Evaluator Name:	
Website:	

If there were any questions that you couldn't answer, please write down which question(s) you couldn't answer and explain why answering was not possible.

Question no(s):	
Reason(s):	

I Indicators to evaluate ONLY by the evaluator

Readability of the website

1a) Determine the average number of words on the homepage per sentence;

Average number of words = ……..

1b) Determine the percentage of familiar words, according to a standard dictionary;

Percentage familiar words = ……..

1c) Calculate the readability score as follows:
Readability Score = 0.1579 *nfw* + 0.0496 *sl* + 3.6305
With *nfw* = percentage unfamiliar words and *sl* average sentence length

Readability score = ……..

Skills for website visitors

2a) Look on the website if there are different "levels" of navigation accessible on the E-Commerce website. For instance: a more skilled person (who has ordered already many products from E-Commerce websites) can order a product in a faster way than a person who has never bought something online.

2b) Appreciate the skill level as:

High: The E-Commerce website has taken many measures for more skilled persons in a way that they can use the E-Commerce website in an easier way;

Average: The E-Commerce website has taken some measures for more skilled persons in a way that they can use the E-Commerce website in an easier way;

Low: The E-Commerce website has taken few or no measures for more skilled persons in a way that they can use the E-Commerce website in an easier way.

Skill level is: High / Average / Low *

(* = circle the term that is applicable)

Erroneous website?

3a) Look for major error possibilities on the website. An example of a major error is typing in a wrong bank account or credit card number;

3b) Look whether the website is able to detect these errors and whether the website provides feedback to the user;

3c) Appreciate the error handling as:

High: The major errors were detected and were properly dealt with by giving feedback;

Average: Most of the major errors were detected and were properly dealt with by giving feedback;

Low: Few of the major errors were detected and were properly dealt with by giving feedback.

Error handling is: High / Average / Low *

(* = circle the term that is applicable)

Input vs Output

4) Determine if there are input/output operations to be found on the E-Commerce website (such as "search" operation or "price calculator"). Just answer Yes or No.

Yes / No *

(* = circle the term that is applicable)

Personalization of the website (1)

5a) Count the number of functions (actions, operations) which can be performed to configure the E-Commerce website in your own way. An example of a configurable function is the personalized profile function.

Number of functions: ……..

5b) Appreciate the personalization possibilities as:

High: Many functions can be customized (compared to other E-Commerce websites. More than on the average E-Commerce website);

Average: Some functions can be customized (compared to other E-Commerce websites. About as much as on the average E-Commerce website);

Low: Few functions can be customized (compared to other E-Commerce websites. Less than on the average E-Commerce website).

Personalization is: High / Average / Low *

(* = circle the term that is applicable)

Explanatory text

6a) Determine the total amount of text to be displayed on the homepage in number of words.

Number of words = ……..

6b) Determine the total amount of explanatory text displayed on the homepage in number of words.

Number of words = ……..

II Indicators the evaluator has to PREPARE for the test group

Learnability of the website

7) Look on the website for a random action that can be taken on the E-Commerce website. An example of an action is the process of purchasing an article.

Action chosen = ……..……..……..

8) Select a not very common operation that can be found on the E-Commerce website, for instance the "add items to wish list" operation.

Operation chosen = ……..……..……..

Insecurity while navigating

9) Select a set of actions on the E-Commerce website (for instance the purchasing process already mentioned in question 6. It is to be preferred that another action is chosen)

Action chosen = ……..……..……..
Time needed to perform this action = ……..……..……..

Personalization of the website (2)

10) Select a configurable function (action, operation) that can be found on the website.

Function chosen = ……..……..……..

Functions on the website

11) Make a list containing the functions (actions, operations) that are available on the E-Commerce website.

III Questions, remarks

Are there any questions or remarks you want to make on this evaluation form? Or do you have difficulties with answering one of the questions?
Please write down your comments, so that this evaluation form can be improved when necessary.

……..……..……..……..……..……..
……..……..……..……..……..……..
……..……..……..……..……..……..
……..……..……..……..……..……..
……..……..……..……..……..……..
……..……..……..……..……..……..
……..……..……..……..……..……..
……..……..……..……..……..……..
……..……..……..……..……..……..

Thank you for filling in this evaluation form!

Appendix B: Test person evaluation form

On the following seven pages the evaluation form for the test person is shown.

Indicator-Based Evaluation

Test person form

This form is meant for a test person to evaluate a Dutch E-Commerce website. With this form, the test person evaluates the indicators that are dedicated to him.

Please fill in these details before starting the evaluation on the following pages:

Date:	
Test person's age:	
Test person's gender:	Male / Female
Computer experience:	Professional / High / Average / Low / None
Online shopping experience:	Very high / High / Average / Low / Very low / None
Highest finished form of education:	
Internet connection speed:	Analogue Modem / ADSL or Cable modem / LAN / Unknown

The evaluator that co-ordinates this evaluation has prepared some of the questions you need to answer. The evaluator has also selected the website that you need to evaluate with this form. This website is:

---Here follows the name of the website---

Please answer the questions in the order that the questions are presented. So, start with question 1, then answer question 2, etc.
Mind you, some questions are applicable to the entire time you are performing this evaluation.

If there were some questions that you couldn't answer, please write down which question(s) you couldn't answer and explain why answering was not possible.

Question no(s):	
Reason(s):	

I The evaluation of the website

Learnability of the website

1) The evaluator has prepared an action that occurs on this E-Commerce website. This action is:

   ---Here follows the action chosen by Evaluator---

   Carry out this action on the website and measure the time you needed to perform this action, from the beginning until the end of the action. Fill in the time you needed below:

   Time required = ……..

2) The evaluator has selected a not so common operation that occurs on this E-Commerce website. This operation is:

   ---Here follows the operation chosen by Evaluator---

   Carry out this operation on the website and measure the time you needed to perform this operation, from the beginning until the end of the operation. Fill in the time you needed below:

   Time required = ……..

Clearness, obviousness of the website

3) Do you think this E-Commerce website is similar to other E-Commerce websites that you have (recently) visited?

   Yes / No *

   (* = circle the term that is applicable)

4) Determine how many times you have encountered incorrect behaviour by misunderstanding an operation on the E-Commerce website while performing this evaluation. (So, also count the times you encounter incorrect behaviour when answering the next questions!)

   Number of times you've encountered incorrect behaviour = ……..

5) Determine how long and how often you have been kept insecure about the status of the E-Commerce website. For instance: you have pressed a button, but it takes very long until something happens.

   Number of occurrences of insecurity = ……..
   Average waiting time = ……..

6a) Count the times that you don't know how to perform an action on the E-Commerce website and you would like to have some assistance.

Times you don't know how to perform an action = ……..

6b) Count the times that a help function was available to assist you when you wanted some assistance.

Times a help function was available = ……..

7) The evaluator has prepared a representative set of actions that can be found on this website.

---Here follows the set of actions selected by the evaluator---

Perform this action and measure the length of interval before status information is displayed.

Length of interval = ……..

8) The evaluator has prepared a configurable function on this website.

---Here follows the configurable function selected by the evaluator---

Determine the effort you experienced by performing this customization on the website, just by choosing one of the three terms below:

Easy to customize / Average / Hard to customize *

(* = circle the term that is applicable)

9) Below a list of available functions on this website is shown:

---Here follows the list of available functions---

Mark the functions that you have used on the website, or that you think you will use when you visit this website with the intention of buying a product on this website.

10) Look again at the list presented in the previous question.
Give yourself 5 minutes to find as much of these functions as you can.

Number of functions found in 5 minutes time = ……..

Operation of the website

11) Just browse the website and try to cancel some of the actions you have taken. For instance: placing a product in a shopping cart and try to remove this product, or buy a product but try to cancel the order in a very late stage of the ordering process.

Appreciate the cancel possibility as:

Very high: All operations can be cancelled, even some operations which were finalized can be cancelled (i.e. return policy after getting purchased product home);

High: All operations can be cancelled, but finalized operations can't be cancelled or were not evaluated;

Average: Most of the operations can be cancelled;

Low: Few or none of the operations can be cancelled

Cancel possibility is: Very high / High / Average / Low *

(* = circle the term that is applicable)

12) Count the number of commands (actions, buttons) on the website that are unfamiliar to you, but which perform a common (familiar) task. For instance, the shopping cart icon is replaced by another not so common icon.

Number of unfamiliar commands with a common task = ……..

13) Do you have to rely on technical (internet) skills to use this website in a easy way, for instance by performing an action on the website?

Appreciate functionality without technical appeal as:

High: Little or no appeal is made to your technical knowledge;

Average: You have to perform non-functional (i.e. technical) operations;

Low: High appeal is made to your technical knowledge.

Functionality without technical appeal is: High / Average / Low *

(* = circle the term that is applicable)

14) While navigating on the E-Commerce website, you can receive some messages from the website. Is the cause of you receiving these messages and the action you must take after these messages appeared clear to you?

Appreciate the message clearness as:

High: Yes, all messages were clear, the cause of the message and the actions to be taken were all clear to me;

Average: Most of the messages were clear, the cause of the message and the actions to be taken were most of the time clear to me;

Low: Few or none of the messages were clear, the cause of the message and the actions to be taken were most of the time not clear to me.

Message clearness is: High / Average / Low *

(* = circle the term that is applicable)

15) While navigating on the E-Commerce website, you can receive some *system* messages from the website. Are these messages common to you, have you also found these messages for instance on other E-Commerce and other websites? In other words, are these system messages consistent, compared to other E-Commerce and other websites.

Appreciate the consistency of the messages as:

High: No system messages were found by you, or all of the system messages were consistent, compared to system messages found on other E-Commerce and other websites;

Average: Few system messages were not consistent compared to system messages found on other E-Commerce and other websites;

Low: Many system messages were not consistent compared to error messages found on other E-Commerce and other websites.

Consistency of messages is: High / Average / Low *

(* = circle the term that is applicable)

General questions about the website

16a) Do you find the menus on the website understandable?
Yes / No *

16b) Do you find the icons on the website understandable?
Yes / No *

16c) Do you find the navigation on the website understandable?
Yes / No *

(* = circle the term that is applicable)

17a) What is your opinion about the user-friendliness of the website?
Please try to name three aspects you think that are user-friendly on this website:

1. ……..……..……..……..……..
……..……..……..……..……..
……..……..……..……..……..
2. ……..……..……..……..……..
……..……..……..……..……..
……..……..……..……..……..
3. ……..……..……..……..……..
……..……..……..……..……..
……..……..……..……..……..

17b) Please try to name three aspects you think that can be improved to make this website more user-friendly:

1. ……..……..……..……..……..
……..……..……..……..……..
……..……..……..……..……..
2. ……..……..……..……..……..
……..……..……..……..……..
……..……..……..……..……..
3. ……..……..……..……..……..
……..……..……..……..……..
……..……..……..……..……..

18a) Are there features that you specifically like on the website? In other words, is this website attractive to you because of these features? If so, name some of these features.

……..……..……..……..……..
……..……..……..……..……..
……..……..……..……..……..
……..……..……..……..……..
……..……..……..……..……..
……..……..……..……..……..

18b) Are there features that you miss on the website, or that are not desirable on this website? In other words, is this website unattractive to you because of these missing or undesirable functions?

……..……..……..……..……..
……..……..……..……..……..
……..……..……..……..……..
……..……..……..……..……..
……..……..……..……..……..
……..……..……..……..……..

Back to the first two questions

The following two questions are the same as the first two questions of this evaluation form. The reason you receive these question again is the following: Do you “learn” from performing an action twice during one evaluation?

19) The evaluator has prepared an action that occurs on this E-Commerce website. This action is:

---Here follows the action chosen by Evaluator---

Carry out this action on the website again and measure the time you needed to perform this action, from the beginning until the end of the action. Fill in the time you needed this second time below:

Time required = ……..

20) The evaluator has selected a not so common operation that occurs on this E-Commerce website. This operation is:

---Here follows the operation chosen by Evaluator---

Carry out this operation on the website again and measure the time you needed to perform this operation, from the beginning until the end of the operation. Fill in the time you needed this second time below:

Time required = ……..

II Questions, remarks

Are there any questions or remarks you want to make on this evaluation form? Or do you have difficulties with answering one of the questions?
Please write down your comments, so that this evaluation form can be improved when necessary.

……..……..……..……..……..……..
……..……..……..……..……..……..
……..……..……..……..……..……..
……..……..……..……..……..……..
……..……..……..……..……..……..
……..……..……..……..……..……..
……..……..……..……..……..……..
……..……..……..……..……..……..
……..……..……..……..……..……..
……..……..……..……..……..……..
……..……..……..……..……..……..
……..……..……..……..……..……..
……..……..……..……..……..……..
……..……..……..……..……..……..
……..……..……..……..……..……..
……..……..……..……..……..……..
……..……..……..……..……..……..

Thank you for filling in this evaluation form!