National Chiao Tung University

Institute of Management of Technology

Dissertation

文化差異對於台灣製造業與西方消費者在電子產品使用上所造成的影響

The Influence of the Cultural Difference between Taiwanese Manufacturers and Western Consumers on Usability

Student: Steve Wallace
Advisor: Hsiao-Cheng Yu

January, 2011

文化差異對於台灣製造業與西方消費者在電子產品使用上所造成的影響

The Influence of the Cultural Difference between Taiwanese Manufacturers and Western Consumers on Usability

研 究 生：史蒂芬　　Student: Steve Wallace

指導教授：虞孝成　　Advisor: Hsiao-Cheng Yu

國 立 交 通 大 學

科 技 管 理 研 究 所

博 士 論 文

A Thesis

Submitted to Institute of Management of Technology

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in

Management of Technology

January 2011

Hsinchu, Taiwan, Republic of China

中華民國一百年一月

The Influence of the Cultural Gap between Taiwanese Manufacturers and Western Consumers on Usability

Student: Steve Wallace Advisors: Dr. Hsiao-Cheng Yu

Institute of Management of Technology
National Chiao Tung University

ABSTRACT

Designing products for usability is crucial for the success of Taiwanese manufacturers competing in overseas markets such as the U.S. To compete in overseas markets, Taiwanese companies need to know how overseas users understand usability and the usability problems overseas users experience when using their products. To support these goals, Taiwanese usability evaluators need usability evaluation tools that are effective in testing their products for overseas markets.

To answer these questions, a model of culture, usability and usability measurement is proposed based on ISO9124-defined usability attributes; effectiveness, efficiency and user satisfaction; and cultural dimensions identified by Hofstede and the World Values Survey. Studies are conducted which support the model and also suggest usability measurement tools which may be effective for Taiwanese producers interested in improving the usability of their products, particularly through heuristic evaluations and severity rating methods.

Keywords: usability, culture, culturability, usability attributes, usability evaluation

Table of Contents

List of Tables

List of Figures

I. Introduction

Designing products for usability is crucial for the success of Taiwanese manufacturers competing in overseas markets such as the U.S. The advantages are many, with improvements to product branding, sales, and ultimately profits, and for the development of the Taiwanese export sector as whole. On the other hand, ignoring usability could mean lost competiveness, with increasing losses to the Taiwanese economy as a result.

To compete in overseas markets, Taiwanese companies need to understand their customers, and they need to know how to make their products usable for these customers. Companies therefore need the answers to the following questions.

- Do overseas users understand usability the same way as Taiwanese customers?
- Do overseas users experience the same usability problems to the same degree as Taiwanese users?
- Are common usability measurement tools effective in supporting Taiwanese usability evaluators who test products designed for overseas markets?
- If not, what solutions may support Taiwanese designers of products for overseas markets?

To answer these questions, we need to examine the interaction of culture and usability. This field of study is also called "culturability", a term first coined by Barber and Badre (1998). Since its inception, the study of culturability has studied the use of a variety of products such as web sites, software, and consumer electronics, by users from a number of countries, particularly the U.S. and the Asian region, to identify the effect on usability as indicated by attributes such as product performance and user attitudes.

In the study of culturability, definitions of usability vary, but the ISO9241 definition of usability is perhaps one of the most commonly used in the HCI field (Clemmensen and Roese, 2010). ISO9241 defines usability in terms of usability attributes: effectiveness, efficiency, and user satisfaction (ISO, 1998). This definition was applied in this study to help understand culture's effect on usability in more detail. Each ISO9241-defined usability attribute was examined to identify the effect culture has on usability attributes and usability as a whole.

In addition, understanding culture's effect on usability is of little use without a means of solving the usability problems behind any differences that may exist. So, to solve usability problems for overseas users, this dissertation also examined the usefulness of usability

measurement tools in cross-cultural situations, focusing especially on heuristic evaluations and severity rating methods. These usability evaluation methods were selected for study as they are both based largely on evaluator opinion, and hence may be likely than other methods to be influenced by culture.

Based on this research, a model of culturability was developed, based on previous research and the results of this study. This model shows the relationship between culture, usability factors, usability attributes, and the measurement of usability. In so doing, this dissertation hopes to both add to the field of culturability, and provide practical support to Taiwanese manufacturers of products aimed at overseas markets such as the U.S.

1.1 Research Background

Taiwan is a major producer of consumer electronic products, most of which are sold overseas. However, to succeed in international markets, Taiwanese companies must realize the importance of good design. Designing for usability is clearly one way to sharpen Taiwan's competitive edge. Mayhew and Mantei (1994) describe improvements in usability resulting in increased sales, improved branding, decreased training costs, and less need for after-sales support. However, if the same product in different countries differs in its effectiveness, efficiency, and user satisfaction, then product designers need to be aware of these differing customer requirements and focus more on improving usability in a way that more effectively focuses on the usability requirements of the target market.

1.2 Research Motivation

The study of culture and usability has generally taken three approaches (Fitzgerald, 2004). The first is the cultural dimension approach, where quantifiable differences in attitudes due to culture are examined for their effect on usability. The second approach studies the effect on usability of features of product design which represent cultural values, such as a marker of social status on a web site. The final approach studies how users with different cultural values use the same product for different tasks. In these approaches, culture is typically represented by cultural values or dimensions. On the other hand, none of these approaches have a standard approach to the representation of usability.

The accurate representation of usability is a problem shared by usability literature as a whole, not just culturability. Hornbaek (2006), in his literature survey of 180

studies into usability, suggests that usability may not always have been accurately measured. Similarly, Frandsen-Thorlacius et al. (Frandsen-Thorlacius, et al., 2009) suggest that many studies only focus on single indicators such as user preference or efficiency as an indicator of usability as a whole. However, as shown by Frøkjær et al. (Frøkjær, et al., 2000), it cannot be certain that any single aspect of usability is correlated with usability as a whole.

It is important, therefore, when studying the effect of culture on usability, to consider the effect of culture on all aspects of usability – effectiveness, efficiency, and user satisfaction. As well as the impact of culture on absolute values of usability attributes, the impact of culture on the relative values of usability attributes needs to be identified. Studies such as that of Frandsen-Thorlacius et al. (2009) suggest that culture has an impact on the relative importance of usability attributes. By combining an understanding of how cultures weight usability differently with knowledge of differences in measures of usability attributes between cultures, we can more accurately focus product design on areas valued by overseas customers and avoid usability problems shown to be serious for overseas customers.

If culture influences usability and usability attributes, it is likely that it also influences usability evaluation. There is evidence that supports the idea that culture affects not only test subjects' perception of usability, but also that of usability evaluators (Clemmensen, et al., 2009), as well as the usability evaluation process (Shi and Clemmensen, 2008, Vatrapu and Pérez-Quiñones, 2006). It is likely that other aspects of the usability evaluation process which involve a high degree of subjectivity on the part of evaluators, such as heuristic evaluations and severity ratings, may also be influenced by culture. Inaccuracies in usability measurement caused by culture therefore, must be identified, as well as possible solutions to this problem.

1.3 Research Objective

The purpose of this research was, therefore, to identify how the cultural background of a user affects the usability of a product, and the evaluation of its usability. The study aimed to examine the effect of a user's culture on the relative importance of these factors when determining overall usability, and as well as on their absolute values, and to more effectively measure and understand the usability problems of Taiwanese-made products in a cross-cultural environment.

1.4 Research Method

To attain these objectives, four studies are carried out.

1. Users from different cultures were surveyed to determine the effect of culture on the relative importance of effectiveness, efficiency, and user satisfaction.
2. The effect of culture on these aspects of usability was investigated by usability testing users from different cultures as they used a Taiwanese-made consumer electronic device.
3. The effect of culture on usability evaluation was indicated by comparing severity ratings made by Taiwanese usability evaluators with those of U.S. users.
4. The usefulness of heuristic evaluations as a means of discovering usability problems in a cross-cultural context was assessed by using two heuristic evaluations to identify the usability problems of Taiwanese-made software documentation.

1.5 Organization of Dissertation

The dissertation is organized into nine main parts.

1. Literature relating to the topic of culture and usability is reviewed in order to identify relevant definitions and models.
2. A model of culturability is proposed.
3. An appropriate methodology is discussed.
4. The influence culture has on the user's relative valuation of efficiency, effectiveness, and user satisfaction is examined.
5. The influence culture has on the effectiveness, efficiency, and user satisfaction of the product is examined.
6. The severity of usability problems of Taiwanese made software documentation is measured using Taiwanese evaluators in order to identify a severity rating method that is both more effective and consistent with definitions of usability.
7. The effectiveness of using heuristic evaluations to identify usability problems of Taiwanese-made software documentation is measured based on a framework of usability attributes.
8. Results from research into the effect of culture on usability are summarized and discussed with regard to their implications for the model of usability proposed in this dissertation.
9. A conclusion to the dissertation is given.

II. Literature review

In this section usability and culture are discussed, and relevant definitions and models are presented.

2.1 Usability

As a social construct, usability is a notion that is difficult to agree on. Definitions of usability typically attempt to identify the aspects of usability that encapsulate its meaning. Models of usability are based on these definitions but attempt to incorporate factors that may influence usability such as context variables or variables affecting the measurement of usability.

2.1.1 Definitions

One approach to defining usability is to examine how users themselves understood the concept of usability. McGee et al. (2004) looked at how 42 users rated 64 usability characteristics. They found users defined usability as a combination of consistency, efficiency, ease of use, effectiveness, controllability, usefulness, expectability, and naturalness. Satisfaction and style were excluded. Similarly, Hertzum et al. (2007) discovered that a user's concept of usability also varies between cultures. Danish and Indian users described usability as including concepts such as 'easy-to-use', 'intuitive', and 'liked'. On the other hand, Chinese users offered a range of issues related to security, task types, training, and system issues. Such studies have not resulted in accepted usability definitions, no doubt because by nature, such usability definitions differ across cultures and between individuals, and cannot become a standard.

An additional attribute, aesthetics has been suggested. Tractinsky (1997) and Kuroso, M. and Kashimura, K. (1995) found significant correlations between user's perceptions of beauty and perceptions of usability for Israeli (r=0.921) and Japanese (r = 0.59) users respectively. Tractinksy therefore called for a more holistic understanding of usability which includes aesthetics. This measure, however, is only intended as an addition to existing definitions of usability. In addition, it is possible that aesthetics may be similar to satisfaction, a commonly used attribute.

Shackel attempted to provide a standardized definition of usability. Shackel defined usability within a framework of user, task, tool and environment. (Shackel,

1991). Usability for individual users was determined by subjective ease of use and objective performance measures such as effectiveness, learnability, flexibility, ease of use, and user attitude. Shackel included a variety of possible measurement methods for these usability metrics such as number of errors, time taken to learn a task, and attitude measurement questionnaires. However, problems exist with implementation of some attributes. Learnability is not applicable to products that are used a single time only. For example, many websites are encountered only once. In addition, the flexibility of a product may be difficult to measure.

Nielsen suggested similar attributes to describe usability. Nielsen's definition of usability includes learnability, efficiency, memorability, errors, and satisfaction (Nielsen) However, Nielsen's definition has been criticized for its blending of usability measurement methods, such as number of errors, and usability attributes, such as efficiency, which could be measured a number of ways (Winter, et al., 2008).

The Consolidated Model of Usability (Witold, et al., 2003) was based on a combination of usability attributes found in the ergonomics standard set by the International Standards Organization - ISO9241 - and additional attributes considered important in other definitions. Witold et al. incorporate learnability from Shackel and Nielsens' definitions, despite issues with single use products. In addition, the model incorporates security from the ISO9126 software quality standard. However, while security may be an important feature in software design, it is not essential in all products, while usability is.

ISO/IEC 9126-1 is a software quality standard which incorporates a definition of usability. The definition describes usability as, "the capability of the software product to be understood, learned, used and attractive to the user, when used under specified conditions" (ISO, 2001). This definition aims at defining user interface qualities to ensure overall software quality, and so is not intended to provide a standalone definition of usability. As an example, effectiveness is not included as an attribute of usability in ISO9126. This definition appears to be non-intuitive, as, to be usable, a product must at least do the job for which it was intended. However, effectiveness appears in ISO9126 as an attribute of software quality, supporting the aim of this standard, which is to define aspects of software quality, rather than usability.

On the other hand, ISO9241 is an ergonomics standard, widely accepted in the HCI (Human-Computer Interaction) community (Frandsen-Thorlacius, et al., 2009), which provides a standalone definition of usability aimed at user (and business) needs.

The standard defines usability as "the extent to which a product can be used by specified users to achieve specified goals with efficiency, effectiveness and user satisfaction in a specified context of use" (Bevan, 2001). ISO9241 further defines usability measures as follows.

- Effectiveness: the accuracy and completeness with which users achieve specified goals.
- Efficiency: the resources expended in relation to the accuracy and completeness with which users achieve goals.
- Satisfaction: the comfort and acceptability of use.

Of all usability definitions, ISO9241 most succinctly encapsulates actual product performance in terms of process and results, and perceptions of performance in terms of user satisfaction. For that reason and because of its widespread use in HCI community it is adopted for use in this dissertation.

One problem with the ISO9241 definition is that there may be little correlation between usability attributes. Hornbæk and Law (2007) examined 73 published usability studies and concluded that attributes showed little or no correlation with each other. This was supported by Frøkjær et al.'s work, which investigated the correlation between usability measures of a document retrieval system and found little evidence of a correlation between usability attributes (Frøkjær, et al., 2000). Walker et al. (Walker, et al., 1998) report a similar lack of correlation between usability measures in their study of a voice controlled email interface.

However, these findings have been challenged by the results of research by (Sauro and Lewis, 2009) who identified correlations between measures of usability in their analysis of the results of 90 unrelated usability tests. By examining actual usability tests rather than academic studies, and by examining results at the task-level, rather than for a usability test as a whole, they were able to identify correlations ranging from -0.44 (errors vs. task satisfaction) to 0.6 (errors vs. time), all much higher than those identified by Hornbaek and Law or Frøkjær et al. They therefore argued that a usability construct exists, and that a single measure of usability is possible, based on usability metrics. However, their results also show such a measure would only account for 54% of variance in raw scores, which they acknowledge would result in the loss of a great deal of valuable information. Consequently, it is argued in this paper, that even if a single measure of usability were possible as Sauro and Lewis argue, the loss of information is too high to warrant combining measures, or to argue

that one usability measure sufficiently represents usability as a whole. Measurement of all usability attributes is still required to understand the effect of culture on usability.

2.1.2 Models

Another criticism of the ISO9241 usability definition is its high level of abstraction (Winter, et al., 2008). Refinements therefore have been attempted in order to enhance definitions of usability to support a clearer distinction of usability's attributes, usability problem classification, and usability measurement methods. For example, the User Action Framework (Andre, et al., 2001) provides a taxonomy of usability problem types intended to support the identification of usability problems, and so of usability overall. (Folmer and Bosch, 2004) are interested in usability insofar as it promotes good design. Consequently they suggest a four level model of usability, with high-level usability definitions described by usability indicators in the following level. The last two levels describe usability properties such as consistency or user control, with accepted usability solutions and heuristics or principles underlying these at the lowest level of the model. While Folmer and Bosch's model succeeds in its goal of incorporating usability indicators and solutions, it still contains multiple definitions of usability, avoids attempts to identify which definition of usability is best, and fails to identify factors affecting usability.

Table II-1 Folmer and Bosch's Four Level Model of Usability

	Usability			
Usability Definitions	ISO9241 Effectiveness Efficiency Satisfaction	Shackel Effectiveness Learnability Flexibility Attitude	Nielsen Learnability Efficiency Memorability Errors Satisfaction	Others Throughput Learnability Evolvability Attitude
Usability Indicators	e.g. completeness, time to learn, errors, speed, satisfaction			
Usability Properties	Consistency, error management, feedback, guidance, user control			
	Problem-solution boundary			
Design Knowledge	Usability patterns		Heuristics/ guidelines design techniques	

On the other hand, the European Community ESPRIT project's MUSIC model (Measuring the Usability of Systems In Context) identifies variables that affect usability (Bevan, 1992), such as task, user, product, and the environment.

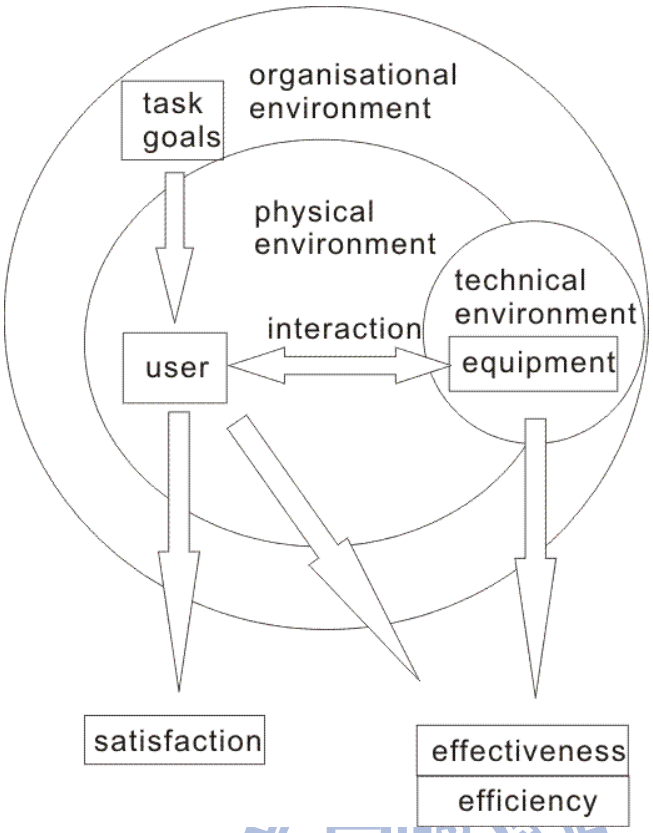


Figure II-1 The MUSIC Usability Model

- The user
  - Personal details
  - Skills and knowledge
  - Physical and mental attributes
- Equipment
  - Basic description
  - Specification
- Tasks
  - Job characteristics
  - Task characteristics
- Organizational environment
  - Structure
  - Attitudes and culture
  - Worker control
- Technical Environment
  - Hardware, software

- Reference materials
- Physical environment
  - Workplace conditions
  - Workplace design, safety

The MUSIC model bases itself on the ISO9241 definition of usability by incorporating usability attributes effectiveness, efficiency and user satisfaction, but expands on this definition by describing the interaction of task, user, product, and environment. However, it is still limited by a lack of usability metrics, such as errors or time taken, and usability solutions.

Hornbaek's Model of Usability Aspects (2006), is similar to the MUSIC model in that is based on ISO9241, also includes usability metrics, a feature missing in the MUSIC model. The model is based on a literature survey of the more common approaches to usability in HCI literature and is perhaps the most comprehensive attempt so far to identify what researchers commonly understand is usability and how it is measured. Based on a survey of 180 articles on usability, Hornbaek distinguishes between measures of effectiveness, efficiency, and user satisfaction. Furthermore, he finds that effectiveness measures are related to outcomes of using a product, while efficiency measures are related to the process of interacting with the product. In addition, he finds differences in measuring subjective and objective usability. For example, time taken to perform a task is often measured as an indicator of efficiency, but a user's perception of the time taken may differ from this measure. Measures of effectiveness may be similarly divided into actual effectiveness and the user's perception of his/her effectiveness.

Table II-2 Hornbaek's Model of Usability Aspects

Usability Aspects	Objective Measures	Subjective Measures
Outcomes (effectiveness)	Expert assessment, comprehension	User's perception of outcome
Interaction process (efficiency)	Time, usage patterns, learnability	Subjectively experienced duration, mental workload, perception of task difficulty
Users attitudes and experiences (satisfaction)	Physiological usability, reflex responses	Validated questionnaires

Based on these findings Hornbaek therefore defines usability as a construct comprising a user's perceptions of product effectiveness, efficiency and user

satisfaction, combined with the actual effectiveness and efficiency of the product. As it is based on a literature survey, Hornbaek's model has the advantage of consistency with other HCI research. Furthermore, it is a clarification of the ISO9241 definition of usability, and comes a long way in helping resolve the issue of connecting usability's definition with its measurement. However, it fails to incorporate the effect of usability factors such as user, task, and product as described in the MUSIC model.

2.2 Culture

2.2.1 Definitions

The concept of culture is difficult to describe precisely, and there is little agreement on culture's definition. By 1951, 164 different definitions for culture had been identified (Olie, 1995), suggesting that currently the number of definitions for culture must be significantly higher. Of these definitions, Straub, et al. (Straub, et al., 2002) have identified three groups of definitions of culture based on either values or cognitive style, or all-encompassing definitions of culture.

Studies of culturability however, base themselves on slightly different definitions of culture. These include differences either in nationality, communication style, meanings, or a comprehensive set of differences including cognition and values. Badre (2000) loosely describes culture as no more than "a means of distinguishing between different countries", indicating that culture is simply based on the nationality of the user. Cultural differences have also been based on differences in communication (e.g. (Hall, 1976), or meaning (e.g. (Bourges-Waldegg and Scrivener, 1998, Marcus, 2006). For Marcus, culture is based on differences in the meaning of group "behaviors, leaders/followers, values, artifacts, and signs". Another type of definition of culture is more comprehensive and attempts to combine cognitive, emotional, communicational, and behavioral differences that help to distinguish groups of people (e.g. (Borgman, 1986, Hofstede, 1984). For Hofstede, culture is "the collective programming of the mind that distinguishes the members of one group or category of people from another, where the mind stands for thinking, feeling and acting, with consequences for beliefs, attitudes and skills". In studies of culturability, the concept of culture is commonly separated into objective and subjective culture (Ford and Kotzé, 2005, Hoft, 1996). Objective culture includes expressions of culture such as institutions, language, and lifestyle etc, while subjective culture includes internal, psychological manifestations of culture, such as attitudes held in common with others.

2.2.2 Models

Models of culture used in studies on culturability usually describe dimensions along which national cultures vary (Badre, 2000, Choi, et al., 2005, Ford and Kotzé, 2005, Marcus and Gould, 2000). Typically, such studies are based on Hofstede's cultural dimensions (Clemmensen and Roese, 2010). However, given the large number of cultural dimensions identified (Lee et al. (2008) reported 36 such dimensions), discussion of the relative merits of the more common cultural dimensions is required. These include Hall (1976), Triandis (1972), Trompenaars (1994), Schwartz (2004), Hofstede, and the World Values Survey.

A number of these models suggest descriptions of culture that are difficult to quantify. For example, Triandis suggests cultures vary according to their complexity, for example, a traditional agrarian society may have a simple structure compared to that of a society based on a developed service-oriented economy. Criteria such as these would be extremely difficult to assess numerically. Similarly Hall describes a number of characteristics by which a culture can be classified, such as cultural complexity or tightness, but no means to gauge the degree a culture can be associated with such a characteristic.

Table II-3 Models of Culture (1)

Hall	Triandis	Trompenaars
• Context • Space • Time	• Cultural complexity • Cultural tightness, • Individualism • Collectivism	• Universalism vs. particularism (rules or relationships) • Neutral or emotional • Individualism vs. collectivism • Specific vs. diffuse (assignment of responsibility) • Achievement vs. ascription (how is status obtained) • Sequential vs. synchronic (how tasks are performed) • Internal vs. external (control over environment)

On the other hand, a number of models provide quantifiable descriptions of culture that can be used to investigate culture's effect on usability. For example, Trompenaars provides seven dimensions of culture, five relating to interpersonal relationships, the other two relating to time and space. However, apart from individual-collective dimensions, Trompenaar's dimensions have little correlation with those of Schwartz, Hofstede, and the World Values Survey. The dimensions provided by the latter three researchers overlaps in the area of individual-collective values, hierarchical values, and control-acceptance values. Schwartz's model of cultural dimensions concludes that global culture can be described by three bipolar axes along

which culture varied. These axes consisted of embeddedness versus autonomy, hierarchy versus egalitarianism, and mastery versus harmony. Embeddness vs. autonomy refers to the extent an individual is embedded in a group. Hierarchy vs. egalitarianism naturally deals with acceptance of differences based on status. Mastery vs. harmony concerns attitudes towards controlling one's social and physical environment. While not exactly matching Hofstede's results, Schwartz's showed some consistency with Hofstede's values of individualism-collectivism, power-distance, and masculinity-femininity, respectively. Similarly, Hofstede found that the WVS dimensions well-being versus survival correlated strongly with his own dimensions individualism and masculinity, and that the WVS dimension secular-rational versus traditional authority correlated negatively with power distance. This overlap between these three models of culture suggests confirmation of the accuracy of these dimensions.

Table II-4 Models of Culture (2)

Schwartz	Hofstede	World Values Survey
• Embeddedness vs. autonomy • Hierarchy vs. egalitarianism • Mastery vs. harmony	• Power Distance • Masculinity vs. Femininity • Individualism vs. Collectivism • Uncertainty Avoidance • Time Orientation	• Need for survival vs. the need for self-expression • Traditional vs. secular-rational values

Given the similarity of these three cultural models, it is worth questioning why Hostede's work is so prevalent in academic literature compared to others. Hofstede's study was conducted among IBM employees in 49 countries in 1967 and 1973. Hofstede describes five main values by which cultures can be distinguished. These are power-distance, masculinity-femininity, individual-collective, uncertainty avoidance, and long-term orientation. Power-distance refers to the degree to which people with less power in a group accept the authority of more powerful members. Masculinity-femininity describes the degree to which the masculine values in a culture, such as assertiveness, independence, and ambitiousness, differ from feminine values, such as caring and cooperation. Individual-collective values describe the extent to which individuals identify themselves as a group member or as an individual. Uncertainty avoidance refers to the need for an individual to have certainty, for example, provided by rules or tradition. Additionally, Hofstede describes long term orientation, identified

in cooperation with Bond (Hofstede and Bond, 1998), which describes certain Confucian-oriented values related to long and short term views of the world such as persistence and the need to save face.

A number of criticisms of Hofstede's methodology have been made, which perhaps have implications for other cultural dimension models as well. One point raised by McSweeney (2002) was that the idea of national culture being shared by all of a country's citizens ignores the fact that there exist a number of distinctive cultures within one nation. The existence of national subcultures, McSweeney argues, makes it hard to say that there exists for each country a single shared culture that makes that country's culture unique. In fact, cultures exist on every level of social interaction, below and above the level of a nation, and distinctive cultures within a nation help to make up the culture on a national level. For example, the cowboy culture of the American West, while clearly distinct from suburban culture in America's cities, is still an important part of American national culture. In addition, it could be said that there exists a North American regional culture, comprising the cultures of the USA and Canada. To study the impact of culture on usability, therefore, researchers need to consider the possibility that nationality may not be the most reliable indicator of cultural background.

In addition, many of McSweeney's criticisms concern Hofstede's methodology. Hofstede's conclusions were based on a survey of IBM employees, who do not necessarily represent the values of the country they represent. In fact, in a developing country, the values of educated, well-off IBM employees may even be that of a minority for that country. McSweeney also notes that the sample size of the survey varies between countries. For example, a country such as Singapore with a smaller local office provided much fewer results per head of population, compared to the USA with its large number of IBM offices.

Lastly, McSweeney argues that Hofstede's survey of company employees may be overly influenced by the working context it was conducted in. Alternative values, such as family or societal values may have been ignored. In fact, Lee et al. suggest that instead of considering differences in work-related attitudes, or other attitudes such as those concerned with family or society, usability research should focus on product related attitudes. They raise the point that cultural dimensions which measure values related to human-to-human interaction may not be relevant to interaction that is, in fact,

human-to-interface. To support their argument they identified ten cultural dimensions which vary between machine types as well as between cultures.

Unfortunately, later, similar research by Schwartz on cultural dimensions may have repeated the methodological errors made by Hofstede. Although Schwartz's model of culture was based on more recent research over a larger sample than Hofstede's, with research conducted students in 67 countries in the period 1988-2000, Schwartz restricted his survey to school teachers and college, thus making it hard to say his surveys represented a cross-section of society and contexts.

On the other hand, research conducted by the World Values Society (WVS) (http://www.worldvaluessurvey.org/) shows greater methodological rigor than both Schwartz's and Hofstede's studies. The WVS conducted research into global values in five waves of surveys between 1981 and 2008, with the latest survey covering 64 countries. Participants were chosen randomly, once quotas were decided, based on the demographic features of that country. National sub-cultures were also investigated. The greater spread of WVS survey suggests that the cultural values identified by the WVS more accurately represent the culture of the country surveyed than those identified by Hofstede. The World Values Survey identified two axes of variance in global values – the need for survival vs. the need for self-expression, and traditional vs. secular-rational values. Societies with low incomes cannot take survival for granted and values in such societies are based on the need for economic and physical security. On the other hand, countries with sufficient income can afford to focus on other needs such as personal well-being, an improved quality of life, and the need for self-expression. Traditional values are strongly influenced by religion and include such values as the importance of traditional family values and national pride. Secular-rational values are based less on traditional religions and attitudes, and are more liberal on social issues such as divorce or euthanasia.

Given the differences in methodology between Hofstede, Schwartz, and the WVS, it is interesting to note that their results support each other. Schwartz's (2004) research into regional groupings of cultural values shows that regions identified by the WVS, Hofstede and Schwartz's own research are similar, and that the values identified by each study, while differing in nature, show some correlation with each other. This point supports the reliability of cultural dimension models such as those developed by Hofstede, Schwartz, and the WVS. Research conducted by the WVS showed the greatest methodological accuracy, yet arrived at similar conclusions to Hofstede's.

Therefore, it appears reasonable to select cultural dimensions developed by WVS for study into the effect of culture on usability. In addition, the widespread use of Hofstede's model of culture allows comparison between studies, thus providing two models of culture for use in this study.

2.3 Culturability

Given the evidence indicating the effect culture has on usability it is surprising that there are so few models which incorporate culture's role, such as the MUSIC model (Bevan, 1992). Two recent models make an attempt to do this. The first, Ford and Kotzé's Conceptual Model of Usability (Ford and Kotzé, 2005), is an attempt to comprehensively model usability, incorporating many variables which influence usability, as well as the culture variable. The second, Clemmensen's Cultural Model Theory of Usability (CM-U) focuses purely on culturability, focusing only on the interrelationships between culture, a user or evaluator's understanding of usability, the product itself and its use, and perceptions of usability (Clemmensen, 2009).

Ford and Kotzé (2005) describe a conceptual model of usability made up of user, task, and environmental factors. In this model culture plays two roles. It is a factor influencing the characteristics of users, which in turn affect product usability. In addition, culture is a factor affecting product design, which in turn affects usability and the technical environment.

Table II-5 Overview of Ford and Kotzé's Conceptual Model of Usability

Conceptual model of usability	User context	**User characteristics**
		User knowledge
	Task context	Task characteristics
		Task execution
	Environment context	Organization environment
		Technical environment
		Physical environment

User characteristics can be further divided into cultural, physical, and psychological factors.

Table II-6 Cultural Variables Affecting User Characteristics in Ford and Kotzé's Model

User characteristics	Culture	Subjective	Strengths	
			Interplays	
			Relative impact	
		Objective	Nationality and ethnicity	Economic system
				Social customs
				Political structures
				Arts, crafts and literature
			Religion	
			Social class	
	Physical	*	*	*
	Psychological	Attitude	Relative importance of usability measures	
		*	*	

* Additional items omitted for clarity

The technical environment can be further divided into hardware and software, the usability of which is affected by a number of cultural factors.

Table II-7 Cultural Variables Affecting the Technical Environment in Ford and Kotzé's model

Technical environment	*	*	*	Hardware	Hardware platform	
				Software	Usability	Partial representation of cultural dimensions
						Nature of cultural dimensions
						Principles and heuristics
						Relative impact of components

* Additional items omitted for clarity

While Ford and Kotzé's Conceptual Model of Usability extensively models culture and usability, it omits usability attributes such as efficiency, effectiveness, and user satisfaction, which are incorporated in models such as the MUSIC model. If these factors are mentioned at all by Ford and Kotzé's model, it is by implication, when the model indicates that the relative importance of usability measures is an attitudinal variable, influenced by a user's individual, psychological user characteristics.

More recently, Clemmensen (2009) developed a theory which connects culture with usability, called the Cultural Model Theory of Usability (CM-U). This model shows how perceptions of usability are derived from the interaction between culture and the user, artifact (product), and understanding of usability. For Clemmensen the product has fixed functionality based on its design, which may be influenced by cultural context (1 – see diagram). The user's interaction with the product is based on the user's set of goals, actions, and emotions, which are also influenced by culture.

The user's goals, actions, and emotions correspond to the effectiveness, efficiency, and satisfaction achieved when using the product (2). Clemmensen's model also shows that culture also influences how usability is understood, not just by the user, but also the HCI practitioner (3). Based on an understanding of usability and the product's actual usability, perceptions of usability are derived (4).

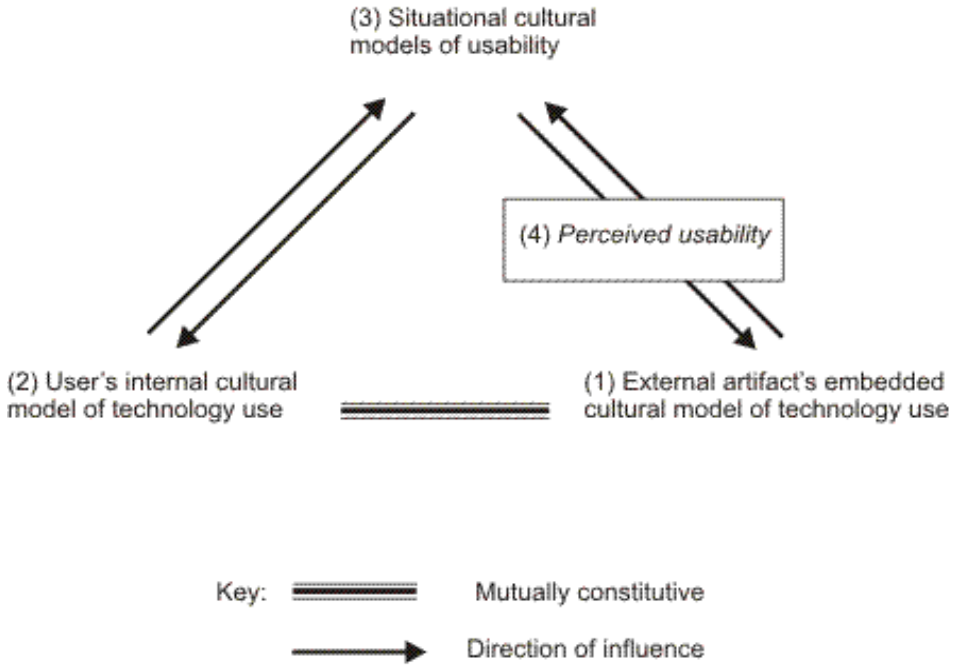


Figure II-2 Clemmensen's Cultural Model Theory of Usability (CM-U)

The strength of Clemmensen's model lies in its succinct summation of the interaction of culture, user (or usability practitioner), product, and perceived usability. The question it raises is exactly how one factor affects another. Consequently, this study follows on from Clemmensen's model and attempts to identify more exactly the nature of the influence of culture on usability, in terms of both the relative values of usability attributes that determine our understanding of usability, and also absolute values of objective and subjective measures of usability attributes.

2.4 Approaches

There are multiple approaches to understanding the effect of culture on usability. As mentioned, Fitzgerald (2004), in an overview of research on culture and website usability, identified approaches which may be equally suitable for studying culturability generally. Among these approaches, Fitzgerald described cultural dimension models, cultural marker models, and cultural behavior models. The cultural

behavior approach looks at how users from different cultures may have different uses for a product, for example, Chau et al. (Chau, et al., 2002). Cultural dimension studies examine the usability of a product based on the one of the many available models of culture described earlier, for example, Choi et al. (2005), and Marcus & Gould, (2000) of which Hofstede's culture model predominates (Clemmensen and Roese, 2010). Studies on the interaction between culture and product design look at the cultural markers, sometimes called attractors, of a product. These are product features that signify or represent a culture, such as flags, or colors (Barber and Badre, 1998). In fact, it should be noted that the cultural dimension and cultural marker approaches can easily be combined. For example, Ford and Gelderblom (2003) include both the cultural dimensions of the user and culture-specific design features as part of their research. Models of culturability such as that of Ford and Kotzé (2005) or Clemmensen (2009) attempt to the incorporate the results of these four approaches, as must this dissertation.

III. Proposed Model of Culturability

In this dissertation several definitions and models of usability, culture, and culturability are adopted and used in a proposed model of culturability, as shown in the table below.

Table III-1 Definitions and Models Summary

Usability	ISO9241	Effectiveness, efficiency, and user satisfaction.
	MUSIC model	Interaction of user, task, and product in a context of use influences usability and usability attributes.
	Hornbaek's Usability Aspects Model	Subjective and objective effectiveness, efficiency, and user satisfaction
Culture	Hofstede	Differences in "thinking, feeling and acting, with consequences for beliefs, attitudes and skills"
	Hofstede's cultural dimensions	Individualism vs. collectivism, power-distance, masculinity vs. femininity, long-term vs. short-term
	World Values Survey	Traditional vs. modern, survival vs. self-expression
Culturability	Ford & Kotzé Conceptual Model of Usability	Effect of culture on user characteristics and technical environment.
	Clemmensen's Cultural Model Theory of Usability	The effect of culture on the user and evaluator's understanding of usability.

The proposed model of culturability is based on a synthesis of usability models and definitions. First, a model of usability is derived from ISO9241, the MUSIC model and Hornbaek's Usability Aspects Model. Usability variables described in the MUSIC model – context, user, product, and task (A) interact to provide a level of usability. A product's usability is expressed in terms of usability attributes identified in ISO9241 (C). Subjective and objective aspects of these attributes are described in Hornbaek's Usability Aspects Model (B), as are possible ways to measure these attributes.

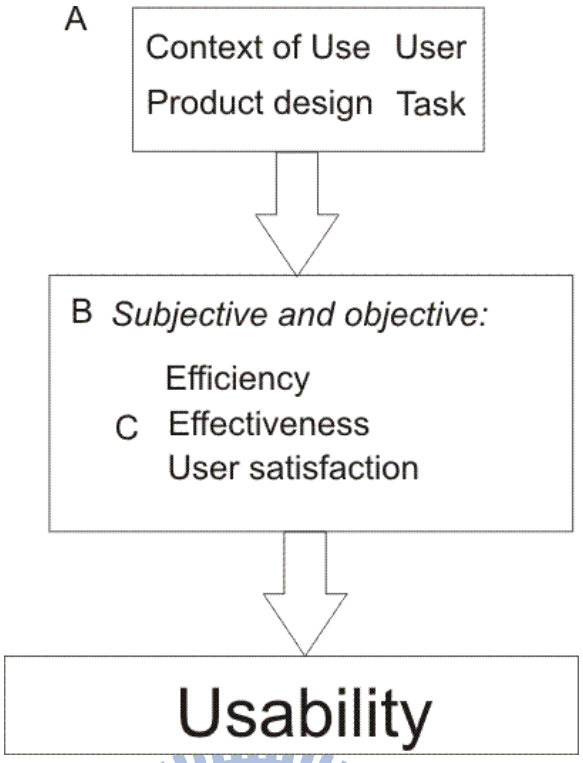


Figure III-1 Usability Factors

Second, the effect of culture on usability factors is derived from models of culturability. The effect of culture on MUSIC usability variables is described in both Ford and Kotzé and Clemmensen's culturability models. Ford and Kotzé's model shows culture affects user and product design and Clemmensen describes how culture affects task. Culture clearly must affect context of use. But it is still not known how culture affects usability attributes (B) and (C). This is the focus of the first two studies of this dissertation.

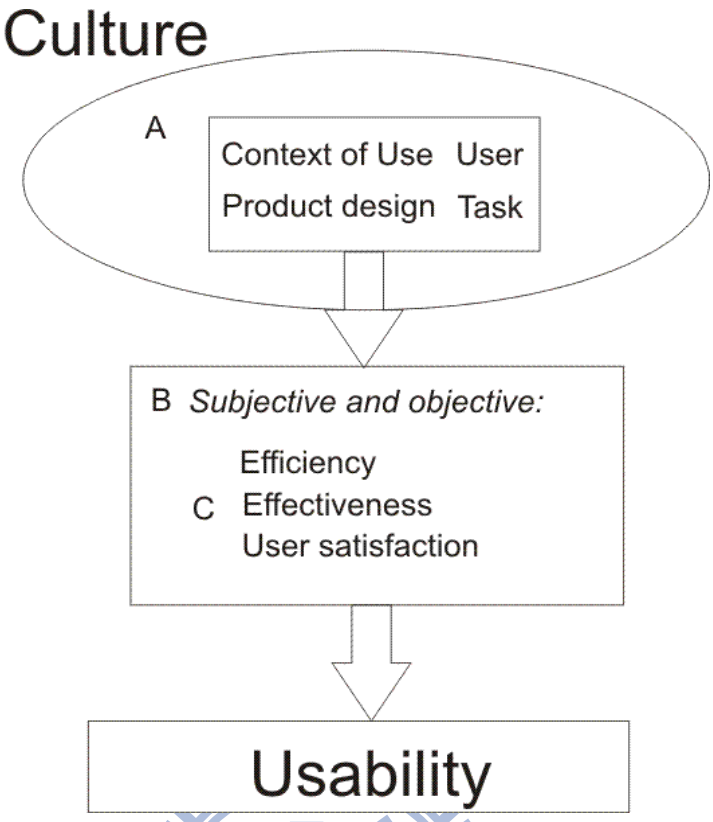


Figure III-2 Culture and Usability Factors (1)

Third, the effect of culture on usability evaluation is derived from culturability models. Clemmensen describes how culture influences a user or usability evaluator's understanding of usability (A). An evaluator's understanding of usability must clearly affect its assessment (B). But it is still not known in what aspects culture may affect the assessment of usability. This is the focus of the last two studies of this dissertation.

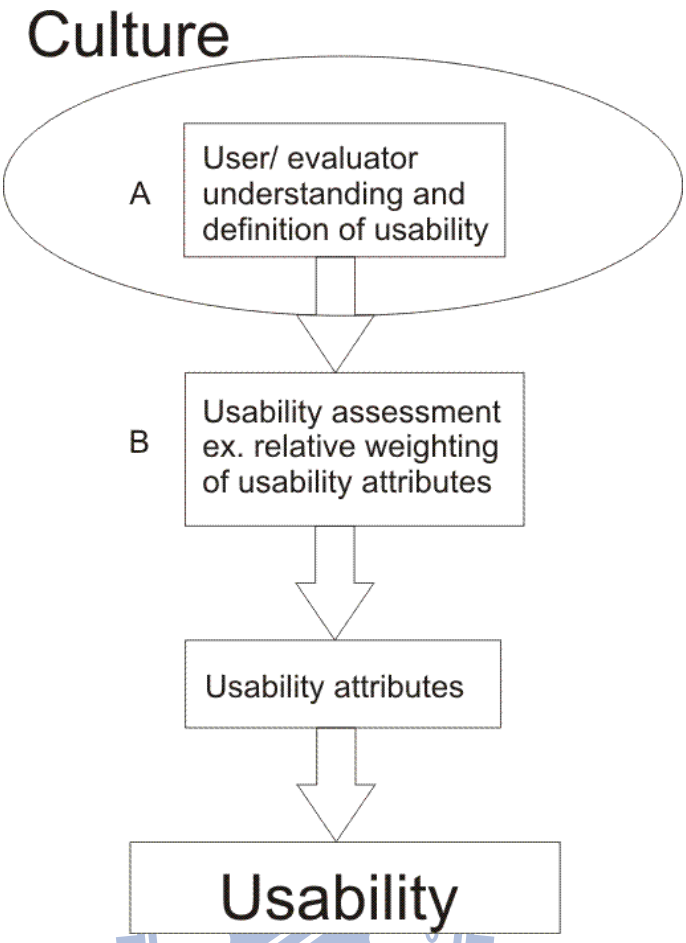


Figure III-3 Culture and Usability Factors (2)

Finally, by adding aspects of Clemmensen's model together with Ford and Kotzé's, we can see how both understandings of usability (A) and contextual variables (B) are affected by culture. To confirm this proposed model of culturability, the effect of culture on usability assessment (C) and usability attributes (D) needs to be examined.

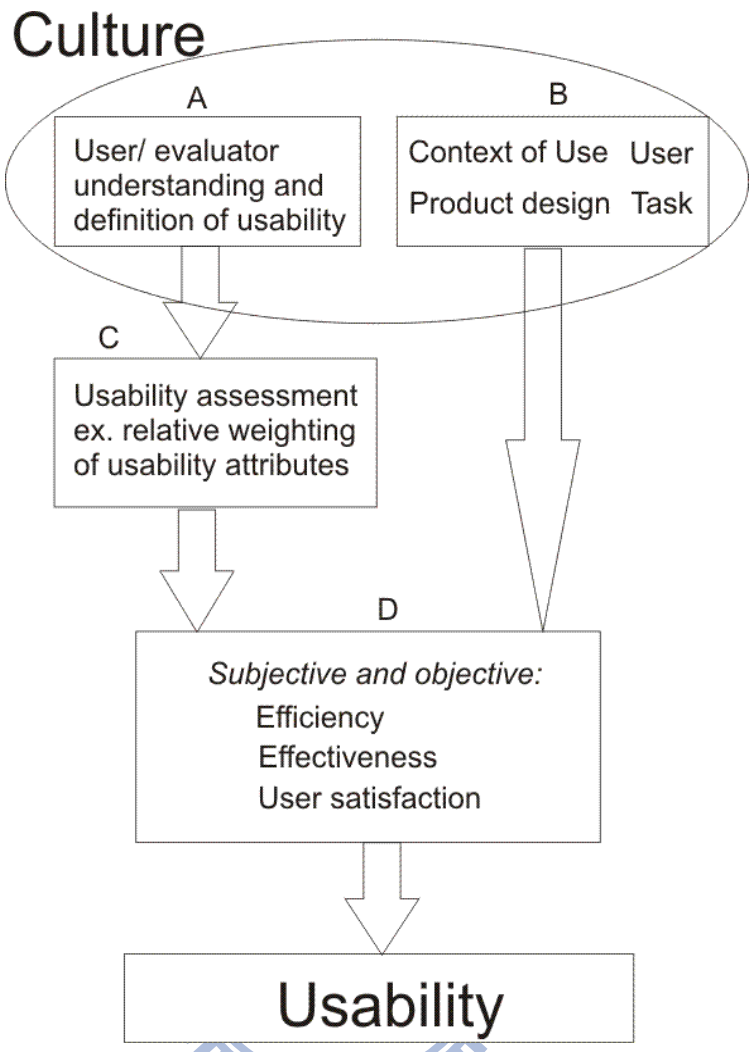


Figure III-4 Culture and Usability Factors (3)

This dissertation proposes a model of culturability which describes the influence of culture on understanding of usability (A), context, user, product design, and task (B), usability assessment (C), and usability attributes (D).

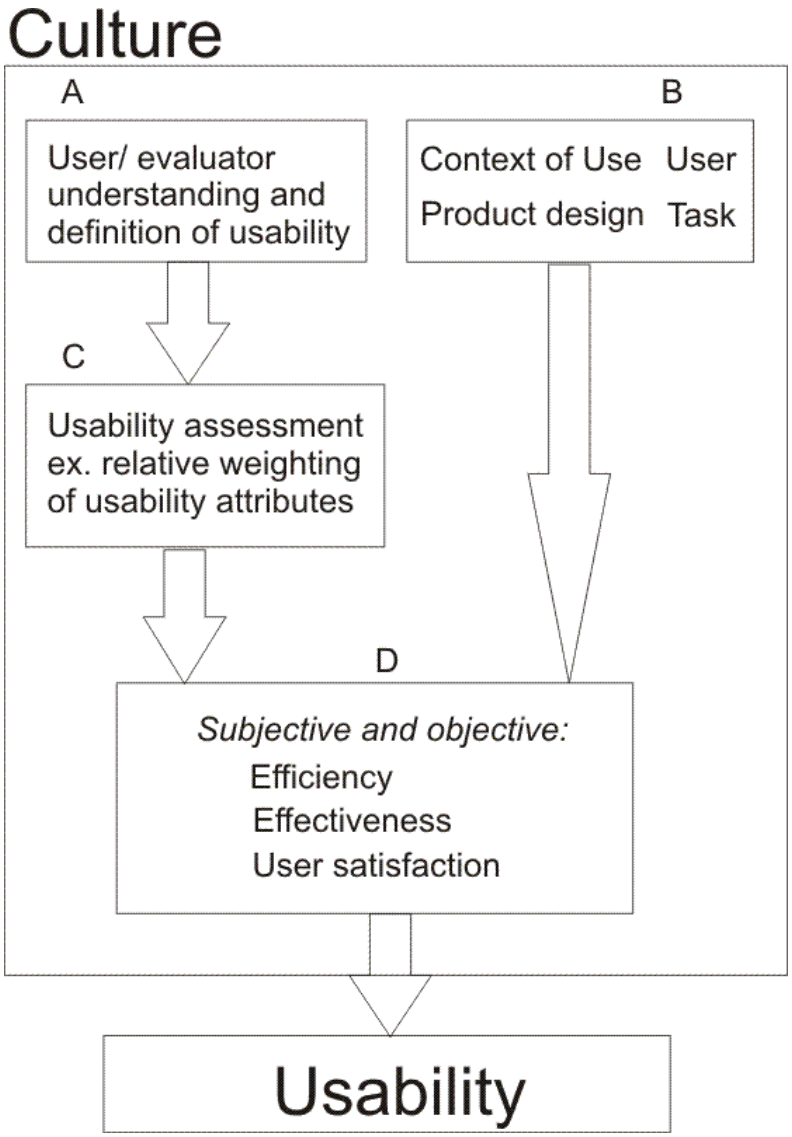


Figure III-5 Culture and Usability Factors (4)

IV.Methodology

To confirm the proposed model and identify the effect of culture on usability attributes and assessment, methodological issues must be discussed.

4.1 Cultures

In determining the effect of culture and cultural dimensions on usability and usability measurement the choice of cultures to be studied is an important one. Clemmensen and Roese, in a literature survey of culturability in the last decade (2010),

identified a number of studies which focus on English speakers, especially the U.S., and with Asia becoming a greater focus, especially the Chinese speaking region. The focus on the China region and the U.S. has the advantage of allowing for a comparison of two very different cultures. On the other hand, limiting the number of cultures investigated to two main cultural groups does not allow for the effect of variation on a cultural dimension to be examined with any statistical reliability. In addition, it omits investigation into culturability in alternative cultural regions.

4.2 Products

The choice of product to be studied is also likely to have some impact on results. Culturability studies generally examine three main kinds of products – websites, software, and consumer items, particularly consumer electronics. Research by Lee et al. (2008) indicates the selection of product for study has implications for the results of any study into the effect of cultural dimensions on usability. Their study into the influence of culture on attitudes to mobile phones, MP3 players, refrigerators, and LCD TVs, indicates that culturally specific attitudes vary between products. For example, MP3 players and mobile phones both elicited an individualistic response from users, while refrigerators and TVs elicited collective values, as they were generally shared by family members. The level of difference in values varied between countries, so an MP3 player elicited individualistic responses from all countries surveyed, but refrigerators and TVs elicited highly collective responses particularly from Russians, compared to other countries. Based on their research, care is therefore required before concluding a particular cultural dimension is responsible for a result, without considering the values elicited by the product examined in the study.

4.3 Measurement Methods

As mentioned, approaches to culturability use a variety of methods to measure usability and its attributes, making results inconsistent and the effect of culture on usability unclear. The cultural behavior approach provides information on an important factor affecting usability, product use, but not usability itself. In the cultural marker approach Shen et al. (Shen, et al., 2006) and Choi et al. (2005) examined user preference to identify usability, and ignored other usability measurements. Alternatively, Ford and Kotzé (2005) measured speed, accuracy, and user satisfaction to determine the effect of cultural markers on usability, and left user perceptions of

efficiency and effectiveness unmeasured. On the other hand, Evers and Day (1997), when examining the effect of culture on software usability, looked at subjective measures of usability such as perceptions of usefulness and ease of use, and user satisfaction, but did not record objective measures of usability, such as speed or accuracy. However, without information on how cultural dimensions affect all usability attributes, objective and subjective, we cannot understand how usability as a whole is affected, given the lack of correlation between usability measures.

To compare usability across cultures evaluation methods typically include questionnaires and usability tests (Clemmensen and Roese, 2010). However, both methods vary in their application. Hornbaek identified a number of commonly used measures of effectiveness in usability testing such as success counts, accuracy, completeness, or quality. Efficiency is commonly measured by time required, and also measures of effort, such as clicks required, or effort spent learning the system (Hornbæk, 2006). Similarly, subjective aspects of usability are measured in a variety of ways. Many researchers develop their own survey instrument, focusing on wide range of aspects of user satisfaction, such as fun, attractiveness, etc. The lack of consistency may partly be explained by the wide range of usability questionnaires developed for specific use. Questionnaires may be product specific, focusing on websites (WAMMI: Website Analysis and Measurement Inventory), software (SUMI: Software Usability Measurement Inventory). Alternatively they may focus on usability attributes such as user satisfaction (QUIS: Questionnaire for User Interface Satisfaction) or perceived effectiveness and efficiency (PUEU: Perceived Usefulness and Ease of Use). Only one standardized survey appears to be relevant to all product types, to cover all usability attributes, and this is the USE survey (Usefulness, Satisfaction, Ease of Use). More widespread use of this survey will allow the comparison of results between studies, and support future research into culturability.

Cross-cultural usability evaluation presents specific problems (Clemmensen, et al., 2009). There are a number of differences to be aware of when usability testing or interviewing Western or Eastern subjects, such as differences in tendency to talk aloud, and express surprise. To avoid such problems this study does not use the talking aloud approach to identify usability problems, and instead aims to record objective measures of efficiency and effectiveness, such as speed and number of errors. To identify subjective attitudes, the USE survey instrument will be used.

V. Culture and the Relative Importance of Usability Attributes, Effectiveness, Efficiency, and User Satisfaction for New Zealand, Philippine, Taiwanese, and U.S. Users of Cell Phones.

5.1 Introduction

Studies of usability across cultures (also termed cultural usability or culturability) in recent years have focused on a number of issues such as cultural differences, globalization, localization, product design, product testing, and user-product interaction (Clemmensen and Roese, 2010). In addition, the concept of usability itself may be worthy of consideration. Usability is socially constructed concept made up of attributes which may not be equally valued across cultures. It is possible that culture affects how usability is understood, thus adding an additional factor to our understanding of culture's effect on usability. To investigate this possibility, this study surveyed the attitudes of cell phone users from a variety of cultural groups to identify the differences in importance they placed on a number of usability attributes, and whether cultural values were a factor in these differences. Findings indicate significant in-country and between-country differences, and relationships between the importance of usability attributes and cultural dimensions.

Previous studies suggest that users' valuation of usability attributes varies between cultures. A study into the usability of software across cultures (Evers and Day, 1997) found Chinese users valued the effectiveness of the software more highly than its ease of use, while Indonesian users valued its ease of use more highly. Variations in the importance placed on effectiveness, efficiency and satisfaction are also reported by Vöhringer-Kuhnt. His study reports differences in how HCI practitioners rate the importance of concepts such as "minimal effort" or "reaching goals", which represent usability attributes, effectiveness, efficiency and satisfaction. In his study he reports that Australians and South Africans place more importance on efficiency and satisfaction than do Americans and Europeans who value effectiveness more (Vöhringer-Kuhnt 2002). Most recently, Frandsen-Thorlacius et al. investigated between-country and in-country differences in the relative importance Chinese and

Danish users of email and word-processing software placed on usability attributes. For example, they found that Danes preferred efficiency over satisfaction and effectiveness over ease of use. On the other hand, Chinese users preferred ease of use over effectiveness. They also found between-country differences with Danish respondents showing a higher preference for effectiveness and efficiency than Chinese did, and Chinese respondents preferring satisfaction more than Danes did (Frandsen-Thorlacius, et al., 2009).

Knowledge of how culture affects users' preference for usability attributes is of both practical and theoretical importance. By identifying attributes that are valued in a target market, resources can be more optimally utilized to focus on product qualities that customers value. In addition, usability evaluators can make more accurate measurements of usability, thus improving a producer's competitive position. Understanding the effect of culture on the relative importance of usability attributes may also help to support and inform the few models of culturability that exist. Ford and Kotzé's Conceptual Model of Usability is perhaps the only usability model that incorporates the weighting of usability attributes, while not ascribing this to culture (Ford and Kotzé, 2005). Clemmensen's Cultural Model of Usability, while not specifically mentioning the importance different cultures place on usability attributes, describes how a culturally specific understanding of usability influences perceptions of usability (Clemmensen, 2009).

5.2 Definitions and Models

Usability has been defined in a number of ways, mostly as a construct based on a variety of attributes. Usability attributes suggested by Nielsen include, among others, learnability, efficiency, memorability, errors, and satisfaction (Nielsen), while Shackel proposed effectiveness, learnability, flexibility, ease of use, and user attitude (Shackel, 1991). Attractiveness has also been identified as an additional attribute (Kurosu and Kashimura, 1995, Tractinsky, 1997). By looking at how users themselves defined usability, McGee et al. identified a number of attributes including consistency, efficiency, ease of use, effectiveness, controllability, usefulness, expectability, and naturalness (McGee, et al., 2004).

It is also possible that different cultures see usability as comprising different sets of usability attributes. In a comparison of user attitudes towards computer applications regularly used by survey participants from China, Denmark, and India, Hertzum et al.

discovered a variance between cultures, which, the authors argue, is linked to a difference in understanding of usability (Hertzum, et al., 2007). Danish and Indian users made distinctions between applications based on concepts such as easy-to-use, intuitive, and liked. Study of Chinese users elicited a range of concepts related to security, task types, training, and system issues.

However, to compare attitudes to specific usability attributes across cultures, a standard set of attributes is required. The International Organization for Standardization (ISO) has provided a number of definitions of usability depending on the purpose of the industry standard. For example, ISO9126-1 describes usability as, "the capability of the software product to be understood, learned, used and attractive to the user, when used under specified conditions" (ISO, 2001). ISO9126 is a software quality standard which incorporates a definition of usability. It defines usability in order to define software quality, and, so, is not intended to stand alone, or to describe products other than software. On the other hand, ISO9241 defined usability as "the extent to which a product can be used by specified users to achieve specified goals with efficiency, effectiveness and user satisfaction in a specified context of use" (Bevan, 1992, ISO, 1998). Attributes of usability implied within this definition included effectiveness, efficiency, and satisfaction, all of which were further defined in ISO9241 as follows:

- Effectiveness: the accuracy and completeness with which users achieve specified goals.
- Efficiency: the resources expended in relation to the accuracy and completeness with which users achieve goals.
- Satisfaction: the comfort and acceptability of use.

ISO9241 is an ergonomics standard, widely accepted in the HCI (Human-Computer Interaction) community (Frandsen-Thorlacius, et al., 2009), which provides a definition of usability aimed at user (and business) needs. The definition is both simple and comprehensive in its incorporation of subjective and objective usability by considering both user satisfaction and product performance, and its description of performance in terms of process (efficiency) and outcomes (effectiveness). For these reasons this study looks at the interaction between usability and culture in terms of ISO9241 usability attributes effectiveness, efficiency, and user satisfaction.

In studies of culturability, culture is commonly separated into subjective and objective culture (Ford and Kotzé, 2005, Hoft, 1996). The latter includes expressions of culture such as institutions, language, and lifestyle, while subjective culture includes internal, psychological manifestations of culture, such as attitudes held in common with others. It is this part of culture that Hofstede is describing when he describes culture as the "collective programming of the mind which distinguishes members of one group from people from another" (Hofstede, 1984). Although objective culture may influence usability as part of the context of use of a product, subjective culture may be more of a factor in attitudes towards usability, which is the focus of this study.

Cultural research so far has identified a large number of dimensions intended to describe subjective culture (Lee, et al., 2008). However, in research on cultural usability, Hofstede's model is dominant (Clemmensen and Roese, 2010). Hofstede describes five main values by which cultures can be distinguished. These are power distance, masculinity/femininity, individual/collective, uncertainty avoidance, and long term orientation. Power distance refers to the degree to which people with less power in a group accept the authority of more powerful members. Masculinity/femininity describes the degree to which the masculine values in a culture, such as assertiveness, independence, and ambitiousness, differ from feminine values, such as caring and cooperation. Individual/collective values describe the extent to which individuals identify themselves as a group member or as an individual. Uncertainty avoidance refers to the need for an individual to have certainty, for example, provided by rules or tradition. Additionally, Hofstede describes long term orientation (identified in cooperation with Bond (Hofstede and Bond, 1998)), which describes certain Confucian-oriented values related to long and short term views of the world such as persistence and the need to save face.

However, a number of criticisms of Hofstede's methodology have been made. McSweeney cited weaknesses in Hofstede's methodology in the following areas.

1. The research ignores subcultures such as different ethnic groups, or regional cultures.
2. It focuses on employees of a single company, who do not necessarily represent the values of the country they represent.
3. The sample size relative to population varies between countries.
4. The research ignores familial or societal values by focusing on business contexts.

Research carried out by the World Values Survey (WVS) (http://www.worldvaluessurvey.org/) attempts to avoid the methodological problems encountered by Hofstede. On the one hand, Hofstede's study was conducted among IBM employees in 49 countries in 1967 and 1973. On the other hand, the WVS conducted research into global values in five waves of surveys between 1981 and 2008, with the latest survey covering 64 countries. WVS survey participants were chosen randomly, once quotas were decided based on the demographic features of that country. The greater spread of the WVS survey suggests that the cultural values identified by the WVS more accurately represent the culture of the country surveyed than those identified by Hofstede.

The WVS identified two axes of variance in global values – the need for survival vs. the need for self-expression, and traditional vs. secular-rational values. Societies with low incomes cannot take survival for granted and values in such societies are based on the need for economic and physical security. On the other hand, countries with sufficient income can afford to focus on other needs such as personal well-being, an improved quality of life, and the need for self-expression. Traditional values are strongly influenced by religion and include such values as the importance of traditional family values and national pride. Secular-rational values are based less on traditional religions and attitudes, and are more liberal on social issues such as divorce or euthanasia.

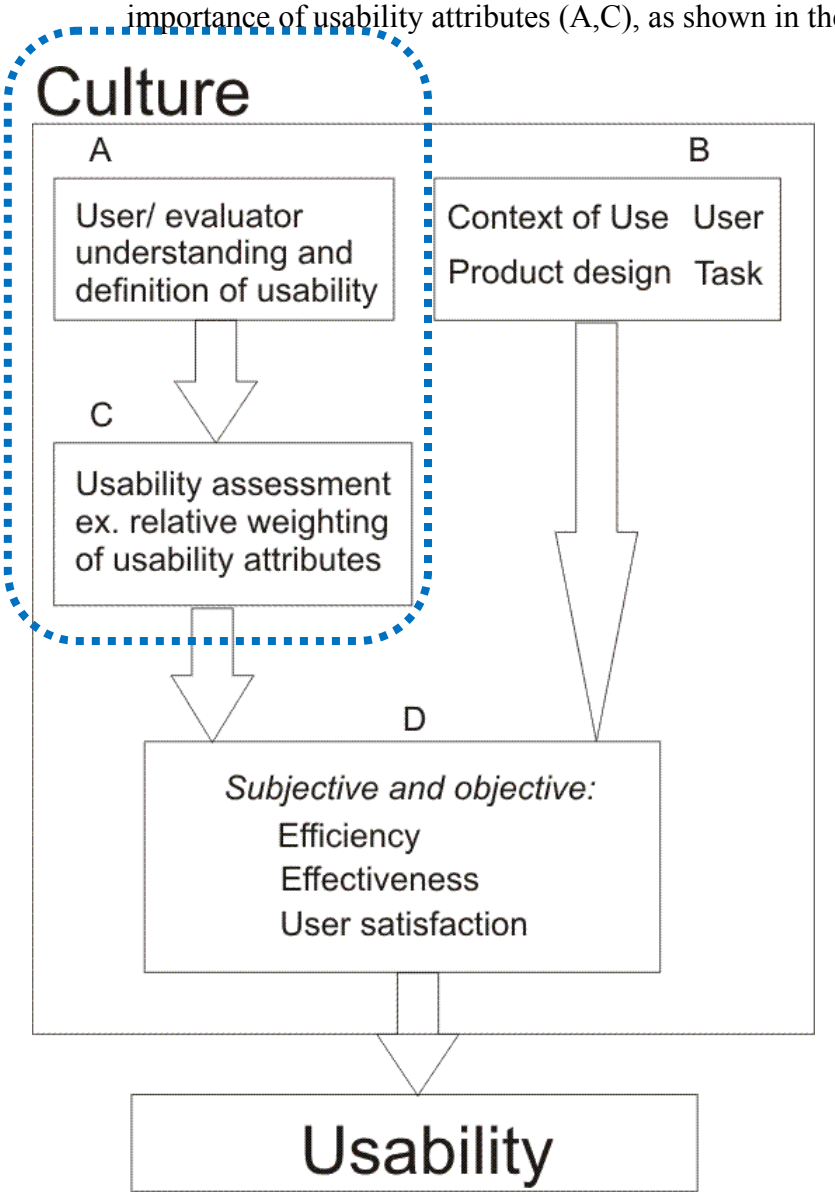
Because of the methodological advantages of the WVS, and the dominance of Hofstede's results in the culturability field, cultural dimensions identified by both Hofstede and the WVS are employed in this study.

5.3 Research Propositions

This study therefore considers the effect of culture on user weighting of usability attributes in terms of the ISO9241 definition of usability, and the cultural dimension models of Hofstede and the WVS. The following possibilities are examined.

1. There is a significant difference in how users from a single country value usability attributes, efficiency, effectiveness, and satisfaction.
2. There is a significant difference in how users from different countries value a single usability attribute.
3. There are one or more significant relationships between cultural dimensions identified by Hofstede or the WVS and users' valuation of usability attributes.

By better understanding these relationships we can ascertain the accuracy of an important part of our model, the effect of culture on user attitudes towards the importance of usability attributes (A,C), as shown in the diagram below.



5.4 Methodology

To identify differences in preferences for usability attributes between cultures, and their connection to cultural dimensions, cell phone users from New Zealand, the Philippines, Taiwan, and the U.S. were surveyed using a variant of the USE (Usefulness, Satisfaction, Ease of Use) survey developed by Lund (Lund, 2001). These nationalities were chosen to represent a range of cultural groups, including English-speaking countries, South-East Asia, and East Asia.

5.4.1 Survey Participants

Participants were found by contacting educational institutions and social and sports groups in New Zealand, Taiwan, and the U.S. Participants comprised local and overseas university students in Taiwan, the U.S. or New Zealand, or recent graduates working in Taiwan.

Table V-1 Sample Characteristics

Cultural background	New Zealand	Philippines	Taiwan	USA
Median age group	45-54	25-34	25-34	25-34
Gender ratio (M=male, F=female, NG=not given)	M:68% F:32%	M:51% F:39% NG:10%	M:66% F:30% NG:4%	M:60% F:40%
Median education level achieved	Under-graduate	Under-graduate	Under-graduate	Under-graduate
Median reported level of skill with cell phones (on a scale rising from 1-5)	3	4	4	4
Sample size (total 144)	28	41	50	25

5.4.2 Cultural Dimensions

The range of nationalities represented in this survey allows a range of cultural dimensions to be tested. The following tables show values for the cultural dimensions identified by Hofstede (http://www.geert-hofstede.com/) and the WVS (http://www.worldvaluessurvey.org/) for each country with upper and lower values bolded.

Table V-2 Cultural Dimensions Described by the World Values Survey†

Country	Traditional/secular- rational*	Survival/self-expression**
New Zealand	0.00	**1.86**
Philippines	**-1.21**	-0.11
Taiwan	0.66	**-0.81**
United States	**0.81**	1.76

†Values are indicated on a scale from -2 to 2.

* Negative values measure traditional attitudes, while positive values represent secular-rational attitudes.

** Negative values measure survival attitudes, while positive values represent self-expression values.

Table V-3 Cultural Dimensions Described by Hofstede†

Country	Power Distance	Individualism/ Collectivism*	Masculinity/ Femininity**	Uncertainty Avoidance	Long Term Orientation***
New Zealand	**22**	79	58	49	30

Philippines	**94**	32	**64**	**44**	**19**
Taiwan	58	**17**	**45**	**69**	**87**
United States	40	**91**	62	46	29

†Values are measured on a scale from 0 to approximately 100.

* High values indicate greater individualism, while low values indicate greater collectivism.

** High values indicate more masculine values, while low values indicate greater feminine values.

*** High values indicate more long-term values, while low values indicate more short-term values.

Cultural dimensions are assigned to users based on their nationality. This is in line with views expressed by Hofstede that cultural values represent a society rather than individuals. Hofstede argues that, while there may be individual differences, cultural traits should not be calculated on an individual level (Hofstede, et al., 2008).

5.4.3 Product

The topic of the survey concerned respondents' preferences for aspects of usability in cell phones. This product was chosen for the reason that it is widely used around the world, in both personal and work-related contexts. The range of cell phones on offer also allows users to express their preference for usability attributes. No specific model was indicated in the survey on the grounds that users are being surveyed to identify their attitudes on the importance of features which would provide greater or less effectiveness, efficiency, or satisfaction. Restricting users to a single model with a fixed level of effectiveness, efficiency and satisfaction-provoking qualities would preclude such an investigation.

The choice of other products is entirely possible since this research is not focusing on any product in particular, but on how attitudes towards usability differ across cultures. However, the results of research into one product may not be generalizable to other products. An investigation by Lee et al. into the cultural dimensions elicited by refrigerators, MP3 players, TVs, and cell phones showed how user values varied depending on the product in question. Similarly, Frandsen-Thorlacius et al. observed Danish preference for effectiveness was higher for a word processing program than for an email application. Without knowing more about the interaction between product type, cultural dimensions, and usability attribute preference, the results of this study cannot be applied to alternative product types.

5.4.4 Survey and Procedure

The survey used in this research was based on items initially developed by Lund (2001) in the USE (Usefulness, Satisfaction, Ease of Use) survey. The USE survey aimed to measure overall perceptions of usability and was based on research by Lund which identified 30 items which he categorized into three aspects of usability: usefulness, ease of use, and user satisfaction. A pilot survey conducted by the authors involving 111 respondents suggested that the number of items could be reduced without seriously affecting internal consistency. Table 4 shows the items selected and their internal consistency as indicated by Cronbach's alpha (and eigenvalues where Cronbach's alpha falls below the standard 0.70). Item categories which Lund labeled usefulness and ease of use were relabeled effectiveness and efficiency in line with ISO9241 terminology. The authors believe the terms ease of use and usefulness refer to the process and outcome of using a product and so are consistent with ISO9241 definitions of effectiveness and efficiency.

Table V-4 Internal Consistency of Survey Items

Item categories	Items	Cronbach's alpha
Effectiveness	meets my needs saves me time useful makes the things I want to do easier to get done	0.6618 (but with a single eigenvalue over 1)
Efficiency	easy to use I can recover from mistakes quickly and easily user friendly easy to learn how to use it	0.8057
Satisfaction	fun to use satisfying wonderful I feel I need to have it	0.7311

The USE survey had an additional advantage in that it is applicable to a wide range of products of services, thus facilitating the replication of research on other products. However, the survey was long, with 30 questions, and used Likert questions, which resulted in a small spread of values with participants rating all attributes as important instead of choosing among them. The survey was therefore redesigned, and the items were arranged so participants would select three among twelve that were most important, and then rate these items out of one hundred, with the total for all

three items making up one hundred. This constant sum question format forced participants to choose which item was most important, thus increasing the spread of results.

Administration of the survey took place in a number of locations, wherever it could be arranged with respondents. The survey required less than two minutes to complete, and respondents received no compensation.

5.4.5 Analysis

There were two sets of data. One was a list of attributes selected by users, as represented by the items they identified as important, the other was the rating of importance of these attributes. Data on the attributes selected by users was analyzed using Chi-square analysis to identify whether differences in the attributes selected by users from each country was significant. Data on the ratings users gave for each attribute was compared using MANOVA to identify significant overall differences in ratings for all countries, and then using one-way ANOVA to identify whether in-country and between-country differences in mean ratings for each attribute existed, with Tukey's HSD used for post-hoc analysis to identify specific differences. Finally, the data was tested to identify correlations between cultural dimensions and attribute ratings, and regression analysis was used to identify relationships between attribute ratings and sets of cultural dimensions.

5.5 Results

Data from the attributes selected showed no significant association between nationality and attributes selected, while data from the rating of attributes showed clear differences in attribute ratings in and between countries.

5.5.1 Selected Attributes

Overall, the association between country and selected attributes is not statistically significant ($c^2 = 11.68$, $p = 0.07$, alpha = 0.05), although it is close to being significant. For New Zealand, Taiwan, and the U.S. there was a significant association between attribute category and the selection of attributes by users ($p = 0.001$**, $p = 0.001$**, and $p = 0.000$** for each country respectively). For each of these three countries, results for satisfaction and to a lesser degree, effectiveness, differed most from expected results. When compared across countries, counts of effectiveness, efficiency,

and satisfaction each showed significant differences (p = 0.016*, p = 0.02*, and p = 0.000** for each attribute respectively).

5.5.2 Attribute Ratings

MANOVA analysis of user ratings of usability attributes showed a significant differences in attribute ratings based on nationality (Wilk's criterion = 0.93, p = 0.001**). Variance in attribute rating based on nationality is 11.20%. Figure 1 shows how the ratings of an attribute varied between users from different countries, and how users from a country showed different preferences for usability attributes.

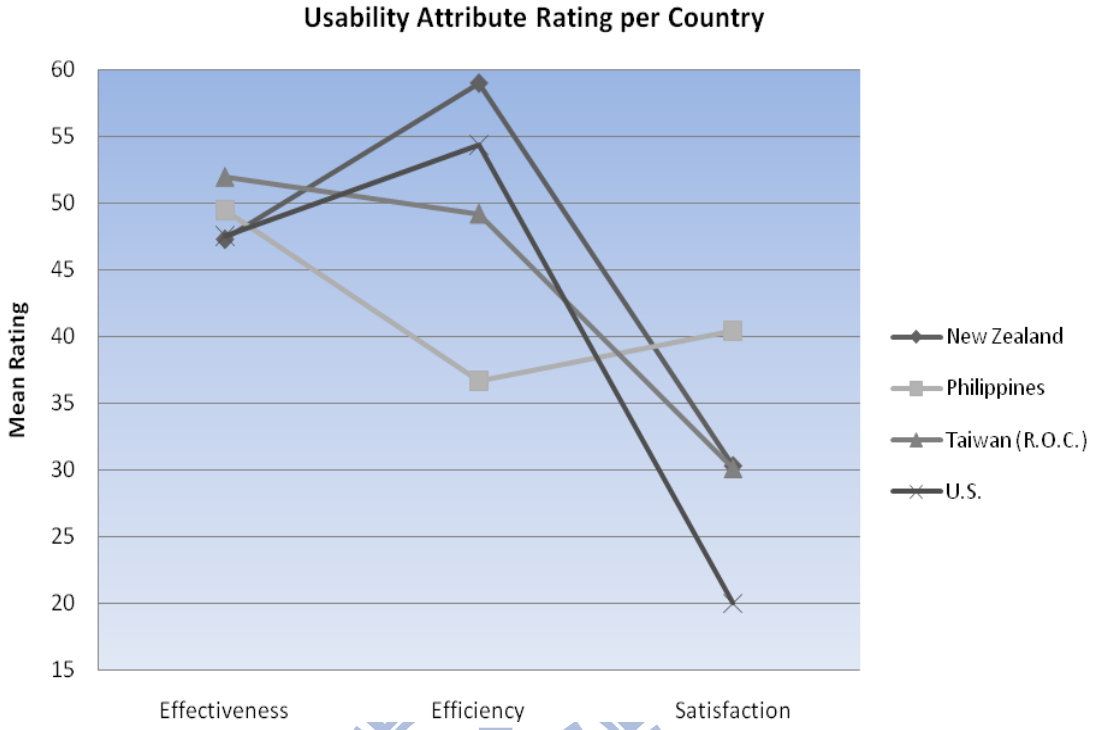


Figure V-1 Usability Attribute Ratings by Country

Table V-5 Usability Attribute Ratings by Country

	New Zealand		Philippines		Taiwan (R.O.C.)		United States	
	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Effectiveness	47.34	27.88	49.50	22.33	51.96	22.28	47.55	25.00
Efficiency	59.04	26.23	36.67	15.37	49.21	22.77	54.37	24.75
Satisfaction	30.36	21.04	40.46	19.28	30.13	14.68	20.00	6.45

There were significant differences within each nationality concerning their relative preference for each usability attribute. One-way ANOVA indicates that for both U.S. and Taiwan respondents a significant difference in attribute ratings exists (p = 0.007** and p = 0.000** respectively), and Tukey's HSD shows that for both countries effectiveness and efficiency are valued more highly than satisfaction. New Zealand respondents showed a significant difference (p = 0.041*), with efficiency valued more satisfaction. Respondents from the Philippines also valued usability

attributes significantly differently (p=0.026*), with effectiveness valued more highly than efficiency.

One-way ANOVA indicates significant difference between nationality-based user groups in their rating of efficiency (p = 0.003**) and satisfaction (p = 0.025**). Post-Hoc analysis using Tukey's HSD indicated test participants from New Zealand had a significantly greater preference for efficiency than did their Filipino counterparts. Filipino respondents also had a significantly greater preference for satisfaction than did test participants from the U.S.

5.5.3 Relationships between Usability Attributes and Cultural Dimensions

Significant and weak or moderate correlations between usability attributes and cultural dimensions were identified in the research.

Table V-6 Correlations between Usability Attributes and Cultural Dimensions

	Effectiveness		Efficiency		Satisfaction	
	correlation	p-value	correlation	p-value	correlation	p-value
Power distance	0.035	0.698	**-0.350**	**0.000****	**0.323**	**0.009****
Individual/ Collective	-0.076	0.395	**0.219**	**0.021***	-0.181	0.153
Masculine/ Feminine	-0.064	0.476	-0.094	0.327	0.175	0.166
Uncertainty Avoidance	0.068	0.448	0.065	0.499	-0.164	0.195
Long-term Orientation	0.068	0.449	0.064	0.506	-0.175	0.165
Traditional/ Secular-rational	0.054	0.542	0.146	0.127	-0.197	0.119
Survival/ Self-expression	-0.005	0.953	-0.311	0.001	0.335	0.007

Cultural dimensions that worked together to influence efficiency and satisfaction were able to account for a significant yet low amount of variation, approximately 10%.

Table V-7 Relationships between Usability Attributes and Cultural Dimensions Sets

	Cultural dimension set	Coefficients	Coefficient p-values	ANOVA p-value	R-squared (adj)
Efficiency	Individual/Collective Masculine/Feminine	0.311 - 0.929	0.000** 0.004**	0.001**	10.3%
	Individual/Collective Uncertainty Avoidance	0.330 0.706	0.000** 0.005**	0.001**	10.0%
	Individual/Collective Long-term Orientation	0.319 0.249	0.000** 0.006**	0.002**	9.5%
	Individual/Collective Traditional/ Sec-Rational	0.273 7.19	0.001** 0.003**	0.001**	10.6%
	Masculine/Feminine Uncertainty Avoidance Long-term Orientation	- 11.9 - 15.1 2.41	0.000** 0.002** 0.042*	0.003**	9.9%
Satisfaction	Individual/Collective Masculine/Feminine	- 0.240 0.708	0.013* 0.014*	0.017*	9.6%
	Individual/Collective Uncertainty Avoidance	- 0.260 - 0.564	0.009** 0.012*	0.015*	10.1%
	Individual/Collective Long-term Orientation	- 0.257 - 0.210	0.009** 0.010*	0.013*	10.5%
	Individual/Collective Traditional/ Sec-Rational	- 0.208 - 5.21	0.023* 0.019*	0.022*	8.8%

5.6 Discussion

The hypotheses in this study can be confirmed. Significant differences in preference for usability attributes exist in some countries and between some countries. Levels of preference for usability attributes are related to one or more cultural dimensions identified by both Hofstede and the WVS.

5.6.1 In-country and Between-country Differences

By grouping countries into cultural regions such as Asia, Europe and Anglophone countries, consistencies in past and current research on attribute preferences become clearer. As shown by the following table, research by Evers and Day, Frandsen-Thorlacius et al, Vöhringer-Kuhnt, together with the results of our study, suggests that,

in most countries studied, effectiveness is most important and efficiency is more important than satisfaction.

Table V-8 Reported In-Country Differences in Usability Attribute Valuation

	Evers and Day (software interfaces)	Vöhringer-Kuhnt (company website)	Frandsen-Thorlacius et al. (email, word-processing software)	Study findings (cell phones)
Asia	Usefulness influences satisfaction (China) Ease of use influences satisfaction (Indonesia)		Efficiency > satisfaction, visual appearance Ease of use > effectiveness (China)	Effectiveness and efficiency > satisfaction (Taiwan) Effectiveness > efficiency (Philippines)
Europe		Navigation > reaching goals, completing tasks > hedonistic quality (Germany) Reaching goals > navigation > hedonistic quality > quality, design (Holland)	Efficiency > satisfaction, visual appearance Effectiveness > ease of use (Denmark)	
Anglophone countries	Satisfaction not influenced by usefulness or ease of use (Australia)	Minimal effort/resources > reaching goals, learnability > error resistance, feedback (U.S.A) Reaching goals, navigation most important (U.K) Navigation, minimal resources, and hedonistic quality > reaching goals (Australians and South Africans)		Effectiveness and efficiency > satisfaction (U.S.A.) Efficiency > satisfaction (New Zealand)

By comparing attribute rankings between countries, we can see efficiency is more important in the Western countries studied than for the Asian countries and satisfaction is more important in the Asian countries studied than in Western countries. Australia and South Africa are exceptional for Western countries for their lower preference for effectiveness and the greater importance they place on satisfaction.

Table V-9 Reported Between-Country Differences in Usability Attribute Valuation

	Effectiveness	Efficiency	Satisfaction
Vöhringer-Kuhnt (company website)	Reaching goals: U.S.A. and Europeans > Australians and South Africans	Minimal resources: Americans, Australians, South Africans > U.K. > Germans, Dutch, Danish Navigation: Germans, Australians, S. Africans, Danish, Dutch and British > Americans	Hedonistic quality: Aust. > Germany Least for Denmark and Holland
Frandsen-Thorlacius et al. (email, word-processing software)		Efficiency: Denmark > China	Satisfaction, fun, visual appearance: China > Denmark
Study findings (cell phones)		Efficiency: New Zealand > Philippines	Satisfaction: Philippines > U.S.

5.6.2 Cultural Dimensions and Usability Attributes

Our final hypothesis, that user preference for usability attributes is influenced by cultural dimensions, is confirmed in our research by the identification of significant relationships. As the importance of effectiveness does not appear to vary significantly across the cultures surveyed, no relationship with cultural dimensions can be identified. However, efficiency and satisfaction appear to be influenced by a number of cultural dimensions. Without further (perhaps qualitative) research, the causal basis for these relationships is pure conjecture. However, some patterns can be observed.

The importance of efficiency and satisfaction are influenced by different values and value sets. Correlation analysis indicated that preferences for efficiency were negatively influenced by values of power distance and positively influenced by survival values. Individualism in relation to a number of values is also associated with higher preferences for efficiency. These values include feminine values, uncertainty avoidance, long-term orientation, and secular-rational values. On the whole, such values relate to the achievement of individual goals, with feminine values making the exception. In contrast, preference for satisfaction is moderately correlated with self-expression values. Satisfaction is also related to sets of values including collective, traditional, masculine values, short-term orientation, and reduced levels of uncertainty

avoidance. These values are related to non-task oriented goals such as status and face-saving needs, and maintaining group relationships.

5.6.3 Limitations of this Research

The use of self-reported rather than observational data may have been a cause of error. Observing test participants as they selected cell phones which represented various usability attributes may more effectively indicate users' true attitudes, although this may require a great deal more time to observe the same number of users.

A number of possible factors in attribute preference have not been taken into account. These include the different values elicited by different products, as identified by Lee et al. (2008). Another possible variable may be differences in age, gender, or skill level of users. Finally, the task intended for the product may also affect which attribute users consider important. Clearly, work-related tasks will require greater effectiveness than tasks related to a user's own entertainment.

A more serious limitation in this study is the small number of cultures studied. While results are clearly significant for survey participants, it cannot be said that the countries surveyed are a representative sample of the cultures existent in the world. To improve the accuracy of this research, cultures representing the major cultural regions should all be surveyed.

5.6.4 Research Implications

As mentioned by Frandsen-Thorlacius et al., the results of studies of this type have implications for usability evaluation and design. In international HCI practice, users, evaluators, and designers may have different understandings of usability. Designers need to keep in mind those usability attributes most values by users. Evaluators need to employ usability measurement methods that support the measurement of usability attributes, and to incorporate the preferences of users into their findings and recommendations.

There is also great scope for future research in this area. By researching a greater number of cultures a clearer picture of the effect of cultural dimensions may be drawn. Study of countries which express a greater range of cultural values such Arabic speaking, European, ex-Communist, Latin, and sub-Saharan African countries may further our understanding of this question. The cultural variables chosen for study need not be restricted to cultural dimensions such as those identified by Hofstede or

the WVS, as objective cultural variables such as income or population need also to be examined. Finally, the effect of product type needs to be examined. As shown by Lee et al. (2008), attitudes towards different products vary across cultures. It may be interesting to ask how much variation in user attitudes is due to product type and how much is due to cultural variables.

Finally this confirms an important part of the model proposed in this dissertation, the relationship between culture and the importance of usability attributes. The likelihood of this possibility has implications for usability evaluators, who necessarily are influenced by the same cultural forces as users. This connection will be investigated later in this dissertation. The relationship between culture and the relative importance users place on usability attributes also raises the question whether there is also a difference in actual levels of effectiveness, efficiency, and user satisfaction - a question that will be investigated in the next chapter.

5.7 Conclusion

Identifying what different cultures mean by a word or a symbol is an essential part of cross-cultural understanding, and the term "usability" is no exception. This study shows that users from different cultures place different weight on the elements of usability – effectiveness, efficiency and user satisfaction, thus indicating a difference in how usability is perceived across cultures. Differences in cultural dimensions are a likely reason for variation in importance of usability attributes, although the influence of other factors may also be important. The results however, need to be combined with research into the effect of culture on actual levels of effectiveness, efficiency and user satisfaction in order to more clearly understand the effect of culture on overall usability.

VI.Culture and Usability Attributes- Effectiveness, Efficiency, and User Satisfaction - for North American and Taiwanese Users of a Taiwanese-made MP3 Player.

6.1 Introduction

Taiwan is a major producer of consumer electronic products, most of which are sold overseas. However, to succeed in the international market Taiwanese companies must realize the importance of good design. Designing for usability is clearly one way to sharpen Taiwan's competitive edge. Mayhew and Mantei (1994) describe improvements in usability resulting in increased sales, decreased training costs and less need for after-sales support. But is making a product usable in a local market the same as designing for usability in a global market? Can it be presumed that an electronic product that is usable in the Taiwanese market is equally usable in other major markets such as the North American market? To answer this question, this article attempts to analyze how the culture of users from North America and Taiwan affects the usability of a new touch-screen MP3 player produced by Taiwanese company ErgoTech.

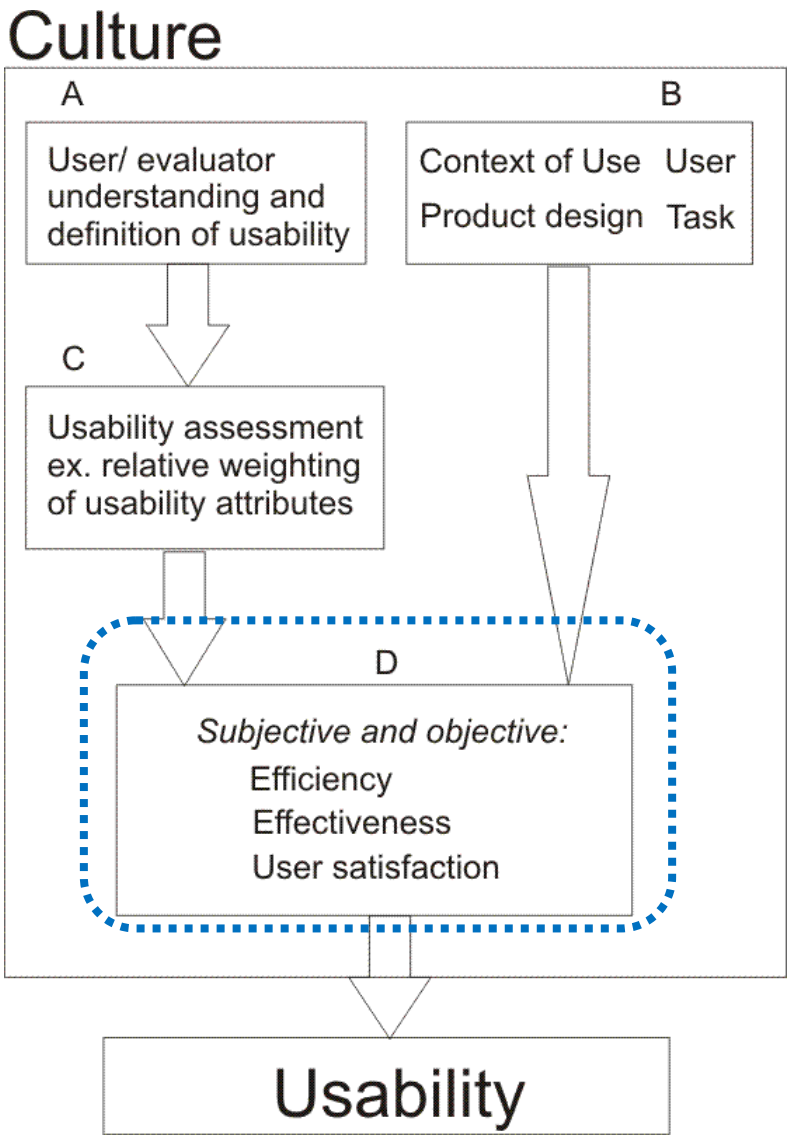
The definition of usability has evolved over the last two decades. Earlier definitions of usability have usually focused on performance-related criteria such as ease of use and effectiveness (Shackel, 1991), but more recently the concept of usability has been expanded to include a subjective aspect, user satisfaction. It is now an ISO standard defined in terms of product effectiveness, efficiency and user satisfaction (ISO, 1998). A literature survey by Hornbaek (2006) of 180 usability articles suggests that the ISO definition is commonly accepted, and that there is also a distinction made in the literature between perceptions of usability and actual usability. Based on this distinction we can therefore define usability as a construct comprising a user's perceptions of product effectiveness, efficiency and user satisfaction, combined with the actual effectiveness and efficiency of the product. To measure and compare the usability of a product across two cultures, it is these factors which must be measured and compared.

Research on web site and software usability has helped describe much about the relationship between culture and usability, or "culturability" as it is described by Barber and Badre (1998). It has been shown that web sites vary across cultures based on culturally-specific characteristics, or cultural markers (Singh, 2003) initially described by Hofstede (1984), and that users' preference for a web site is affected by the cultural features of a web site (Badre, 2000). Users from different cultural backgrounds are likely to use a web site for different purposes (O'Keefe, et al., 2000). In addition, the use of translation in a multilingual web site (even when done expertly) affects user satisfaction (Nantel and Glaser, 2008), as does the use of culturally familiar/unfamiliar icons (Shen, et al., 2006) in software applications. In a more comprehensive study of usability and culture Evers and Day (1997) found culture to be an important factor regarding the interrelationships of perceptions of efficiency, effectiveness, satisfaction, and user behavior when using a software application. In short, culture is likely to influence many elements affecting the usability of a product.

However, while many studies identify the effect culture has on many important aspects of usability, it cannot be certain that any single aspect of usability is correlated with the overall usability of a product (Frøkjær, et al., 2000). It is important therefore, when studying the effect of culture on usability, to consider the usability of a product as a whole. In this study this is done by identifying the effect of culture on the elements Hornbaek (2006) found as commonly considered factors in usability: perceptions of efficiency, effectiveness and satisfaction, as well as performance-based measures of efficiency and effectiveness.

6.2 Hypothesis

This study aims to identify whether levels of subjective and objective effectiveness, efficiency, and user satisfaction differ significantly due to differences in the cultural background of the user. The existence of such a difference would support the second part of the proposed model of culturability proposed by this dissertation - the connection between culture and nominal levels of usability attributes.



6.3 Methods

Hornbaek (2006) describes the main components of usability as being both subjective and objective. So to measure subjective impressions of product efficiency, effectiveness and user satisfaction, this study surveys users to evaluate their perceptions of usability. To measure the actual efficiency and effectiveness of the product, users are observed as they use the product.

6.3.1 The Product

The product is an MP3 player from a Taiwanese electronic company called ErgoTech. The player includes a touch screen with a graphical interface. While the graphical interface reduces the use of text, text is still required for many functions. The interface can be configured in a variety of languages, including Traditional Chinese (for Taiwanese users) and English (for North American users).

An MP3 player was chosen for a number of reasons. This kind of product is popular among both young Taiwanese and North Americans, thus reducing the chance one group would be less familiar with the product than the other. There are also a small number of functions, reducing the chance that different cultures would use the product for different purposes – a cultural difference identified by O'Keefe et al. (2000). A touch screen interface which relied mainly on icons was chosen in order to reduce the influence of translated text on usability, as translated text may result in a reduced level of usability for users (Nantel and Glaser, 2008). Finally, it is hoped that results of such research will have implications for manufacturers and exporters of consumer electronic products in Taiwan and elsewhere.

6.3.2 Survey Sample Characteristics

Participants in the study were chosen for being representative of the target market. Analysis of marketing materials and the product's design and features indicated the MP3 player was aimed at a young and international audience, interested in the latest hi-tech gadgets, many of whom would be students. In total 23 people were selected, based on their match with the target market. Of those tested, 13 had a Taiwanese cultural background and 10 had a North American (U.S.A. and Canada) cultural background. Subjects were compensated for their time in the form of food and refreshments. Apart from cultural background other factors such as age, gender ratio, education levels and experience with the product or other similar products were at similar levels between cultural groups.

The following table indicates the characteristics of the sample.

Table VI-1 Sample Characteristics

Cultural Background	Taiwan	North America
Average age	26.08	27.9
Gender ratio	4 females/9 males	4 females/6 males
Average education level	Bachelors degree	Bachelors degree
Average reported experience with similar products	Somewhat – very experienced.	Somewhat experienced
Average reported level of confidence with consumer electronics (on a scale 1-5, 1 is low, 5 is high)	3.38	3.35
Sample size	13	10

6.3.3 Measuring User Performance

Subjects were then given a series of tasks to complete using the MP3 player. Using scenario-based analysis a number of tasks were identified as being likely tasks that the target user would perform. These tasks included listening to the radio, playing a song, recording a voice, playing a game and adjusting settings. The subjects were observed using the product as they carried out these tasks. As the subject was using the product s/he was observed to see how effective and efficient the product is to use.

Effectiveness was measured by recording whether or not a user could complete a task. Since time was a variable in this study no time limits were given. The user either completed the task or announced s/he was unable to complete the given task. The binomial (yes/no) result was then recorded and summated for all tasks attempted by the user.

Efficiency was measured using the time taken and the number of errors made. Errors were defined as an attempt to click on the screen or other hardware attachment that would not result in completing the task assigned. The resulting scores on the time taken and number of errors made however, were not combined into a total score for efficiency. As noted by Hornbaek (2006), this would result in two errors. Firstly, in order to combine the scores a weighting for each score would have to be identified, which is beyond the scope of this article. Secondly, combining scores may lead to the overlooking of important patterns in data representing each score. Finally, by separating the time taken and number of errors made we may better identify the usefulness of these variables as a measurement of efficiency.

User satisfaction was not measured by empirical means as the methods available (facial or verbal expressions) were considered too variable across cultures to be a culture-neutral method of measurement.

These measurements were then analyzed using a t-test (based on a small sample size and assumptions of a normal distribution) to identify whether the average result differed based on cultural background.

6.3.4 Measuring User Perceptions

Subjective measurements of effectiveness, efficiency and user satisfaction were carried out using a Likert survey. The survey items were initially developed by Lund

(2001) in the USE survey on usefulness, ease of use and user satisfaction, but are adapted in this survey to indicate efficiency, effectiveness and user satisfaction. This survey was selected among others for a number of reasons.

- It can be used when testing a variety of products, compared to other usability measurement methods such as the QUIS or SUS tests which focus on web site or software usability.
- The variables covered in the survey closely follow those identified by Hornbaek (2006) as being a component of usability, as well as being the industry standard used by ISO.
- The survey aims to become a commonly accepted usability measurement tool. Using the survey helps to address the problem identified by Hornbaek (2006) of the lack of replication and comparability of studies of usability measurement.

The USE survey was found to be a highly reliable indicator of user perceptions as indicated by Cronbach's alpha. Similarly, Lund (2001) reports high levels of Cronbach's alpha when designing the survey. The following table shows the high level of internal consistency of this survey.

Table VI-2 Survey Reliability

Perceived Usability Factor	Cronbach's Alpha
Efficiency	0.86
Effectiveness	0.92
User satisfaction	0.88
Total usability	0.96

The survey responses for each variable (efficiency, effectiveness and user satisfaction) were summated and then analyzed for differences based on cultural background. Both the Mann-Whitney U test and the t-test were applied. The t-test is commonly used in studies such as this where the sample size is small and a normal distribution is assumed. However, one set of results, the number of tasks completed, was not found to be normally distributed, so the Mann-Whitney U test was also applied to the data. These tests gave similar results, so for convenience only the t-test results are given in this article.

In this section of the study one set of hypotheses is examined. The null hypothesis is that the mean or median values of all usability factors are the same for Taiwanese and North American users. The alternative hypothesis is that they differ.

6.3.5 Measuring Correlations

The correlation between variables was also estimated using Pearson's correlation co-efficient and the Spearman Rank Correlation. The latter measure was used for the reason mentioned earlier, that the number of completed tasks is not normally distributed. However, results for both measures are similar, so to allow comparison only measures of Pearson's correlation coefficient are used in this article.

Correlations are shown in two aspects. Firstly, the correlation between subjective and objective measures of usability and users' cultural background is examined. While focusing on the connection between culture and usability, other correlations were also estimated for the purpose of comparison. Secondly, a comparison is made of correlations between usability factors within each culture group – Taiwanese and North American users.

In this section of the study two sets of hypotheses are made. Firstly, correlations between culture and other usability factors are examined across both culture groups. The null hypothesis is there is no significant correlation between culture and any usability factor. The alternative hypothesis is that there is a statistically significant correlation between culture and one or more usability factor.

Secondly, correlations between usability factors in one culture group are examined to see whether they are significantly different from the other. The null hypothesis is that there is no significant difference between one correlation and it's equivalent in either group. The alternative hypothesis is that one or more correlations in a culture group differ significantly from the equivalent correlation(s) in the other culture group.

6.4 Results

The results from the analyses conducted support the first two of the hypotheses made but found no statistically reliable evidence to support the third.

1. The first test showed that there are many statistically significant differences in the mean values of the usability factors obtained from each culture group.
2. The second test confirmed our hypothesis that correlations exist between cultural background and usability factors.
3. The third test found no statistically reliable evidence to suggest that correlations between usability factors differ significantly between culture groups.

6.4.1 Comparing Averages for the Two Groups

Results from the t-test show that the average perceptions of usefulness, ease of use and user satisfaction differ significantly between Taiwanese and North American users. Culture clearly is associated with perceptions of usability. However, the average levels of performance do not differ so clearly between Taiwanese and North American users. There is no clear difference in average number of tasks completed between groups, indicating that the culture of the user may not have an impact on the actual effectiveness of a product. The link between culture and product efficiency is also not clear. The average number of errors made differed significantly according to culture, while the other measure of product efficiency, the time required to complete a task showed no clear difference in averages.

The table below summarizes the difference in distributions between the samples.

Table VI-3 Comparison of Mean Measures of Usability between Cultures

	Taiwan	North America	p-value (t-test)	Significance
User Perceptions (mean score as rounded percent of total possible)				
Effectiveness	70%	41%	< 0.001	Highly significant
Efficiency	64%	45%	< 0.01	Very significant
User satisfaction	63%	42%	< 0.01	Very significant
Total perceived usability	62%	44%	< 0.05	Significant
User Performance (mean score)				
Tasks completed (max 16)	13.46	13.60	> 0.05	Not significant
Time required (seconds)	667.08	788.80	> 0.05	Not significant
Errors made	93.92	155.70	< 0.01	Very significant

6.4.2 Correlations between Usability and Culture

It is also interesting to see how significantly culture is correlated with the variables associated with usability. The following table indicates the correlations (*r*) between the variables measured in this study. The degree of correlation is indicated by the closeness of the correlation measure to 1 or -1. A result of 0 indicates an absence of correlation, while a result of over 0.5 or -0.5 shows a moderate to strong association.

There is disagreement about where exactly the cutoff for a moderate or a strong correlation lies. In this study *r* values from 0.5 to 0.7 or from -0.5 to -0.7 indicates a moderate correlation, while *r* values above 0.7 or below -0.7 (shown in bold) indicate a strong correlation. Weak correlations, an absence of correlation, or statistically insignificant correlations (p >= 0.05) are not shown in this table.

Table VI-4 Correlations between Factors in Usability for Both Culture Groups

	Culture	No. of Errors	Effectiveness	Efficiency	Satisfaction
User Perceptions					
Total perceived usability	-0.501*	-0.590**	0.634**	**0.755*****	**0.744*****
User satisfaction	-0.633**	-0.687***	**0.787*****	0.694***	
Efficiency	-0.579**	-0.593**			
Effectiveness	**-0.774*****				
User Performance					
No. of Errors	0.588**				

* = p < 0.05

** = p < 0.01

*** = p <0.001

The results in table 4 confirm the alternative hypothesis. Culture clearly is correlated with many of the factors which make up usability. Firstly, culture is moderately to strongly associated with perceptions of usability. In particular, culture has a strong association with perceptions of effectiveness, but was only moderately associated with other perceptions of usability. The negative correlation shown in some cells reflects the numbers used to represent the different cultures for statistical purposes. Taiwanese and North American cultures were represented by the numbers "1" and "2" respectively. So the negative correlation between culture and effectiveness reflects the impression North American subjects had of the lack of effectiveness of the product.

Secondly, culture is also directly linked to a user's actual efficiency when using the device, as shown by the number of errors. It is also possible that culture indirectly affected perceptions of usability. The number of errors had a negative correlation with

variables representing perceptions of usability. As the number of errors increased, perceptions of effectiveness, satisfaction and overall perceptions of usability fell.

While not strictly related to this study, it is also clear that perceptions of usability shared some correlation. Perceptions of effectiveness and efficiency were moderately positively correlated with user satisfaction. Understandably, total perceived usability was correlated with the variables that it comprises – effectiveness, efficiency and satisfaction. It also shows an extremely moderate relationship with cultural background and the number of errors, possibly as a result of its indirect relationship with these variables.

Interestingly, one measure of efficiency, task completion time, showed no significant correlation with any other variables measured in this study, while the other indicator of efficiency - the number of errors - showed a strong connection to many perceptions of usability. The lack of correlation between two measures of efficiency – time taken and number of errors, combined with the lack of correlation between time taken and any other variable raises questions as to the usefulness of this variable as a measure of efficiency.

6.4.3 Comparing correlations between usability factors across cultures

This study also examines the correlations between usability factors within each culture. The tables below show moderate to strong, statistically significant ($p < 0.05$) correlations between usability factors for North American and Taiwanese users. Because the sample size for each culture group is roughly half that of the total sample size, there are fewer statistically significant observations available.

There are marked differences in the correlations observed in each culture group. For example, the number of errors is clearly correlated with user perceptions in both groups. For the Taiwanese the number of errors is correlated with perceptions of efficiency. This is in contrast to North American users for whom the number of errors was strongly correlated with lower levels of satisfaction. Unfortunately, because these correlations are for different sets of variables, these correlations cannot be compared.

In fact, the correlation between overall usability and efficiency is the only correlation found in both culture groups. For Taiwanese users there are moderate to strong correlations between perceptions of effectiveness, efficiency, and user satisfaction. These values were also strongly correlated with overall usability. For North American users however, only a moderate correlation between perceptions of

efficiency and overall usability was observed. However, we cannot say with a 95% confidence level that the correlation between perceived efficiency and total perceived usability is statistically different between cultures.

In short, no statistically significant comparisons of correlations can be made between these two user groups. Because of this the hypothesis that there is no difference between usability correlations for Taiwanese and North American users still stands.

Table VI-5 Correlations between Usability Factors for North American Users

	No. Errors	Efficiency
Total Perceived Usability		0.578*
Satisfaction	**-0.730****	

* = p < 0.05

** = p < 0.01

Table VI-6 Correlations between Usability Factors for Taiwanese Users

	No. Errors	Efficiency	Efficiency	Satisfaction
Total Perceived Usability		**0.814****	**0.877****	**0.906*****
Satisfaction		**0.824****	0.641*	
Efficiency	-0.672*			

* = p < 0.05

** = p < 0.01

*** = p <0.001

6.5 Discussion

This study shows that the cultural background of the user is a likely factor in determining the usability of a consumer electronic product, such as ErgoTech's MP3 player. Most aspects of usability identified by Hornbaek (2006) in his survey of usability literature are affected by the cultural background of participants, in particular, users' perception of effectiveness, efficiency and levels of satisfaction. In addition, efficiency as measured by the number of errors is also clearly connected to culture.

However, in this study efficiency was measured by both the time taken and the number of mistakes made when performing a task. It is to be expected that, as users make more mistakes they also require more time to complete tasks. However, this was

not the case. North American users required similar amounts of time to their Taiwanese counterparts, but made more errors within that time. One possible explanation for this could be a different problem-solving style. It was often observed during tests that, when faced with a problem using the MP3 player, North American users sometimes became more active or even clearly frustrated, which may have been the reason for the higher number of errors. The number of errors then correlated with low levels of user satisfaction and perceptions of efficiency among North American users. On the other hand, the lack of correlation between completion time and any other variable raises questions as to the usefulness of this variable as a measure of efficiency.

While efficiency may have been affected by culture, effectiveness as measured by the number of tasks completed is similar for both cultural groups. That there is little connection between efficiency and effectiveness is supported by Frokjaer et al. (2000) who found little correlation between efficiency and effectiveness. As the study by Frokjaer et al. note, the lack of correlation between efficiency and effectiveness means any attempt to measure usability must take into account the range of factors which make up usability.

So, on the whole, culture affects usability, but how much does usability vary according to culture? So far this study has only considered whether there existed a relationship between culture and the perceived and actual components of usability – efficiency, effectiveness and user satisfaction. Given that there is a relationship, is it possible to also say how much culture affects usability as a whole? While it may be possible to give a total value for usability by combining the results for each component, the question is how much weight should be given to each component. This question is made more difficult since culture may also affect the weighting given to the components of usability. Research on Indonesian and Chinese users show higher importance placed on efficiency by Indonesians and more emphasis placed on effectiveness by the Chinese (Evers and Day, 1997). To arrive at a total measure of usability for each culture more research is needed to compare the weightings different cultures have for each component of usability.

Why is culture such an important variable in usability? This study does not attempt to answer the "why" in culturability. For suggestions as to possible reasons, studies on software and web site usability have implications. Nantel and Glaser (2008) argue that translated text reduces web site usability, Shen et al. (2006) finds evidence

that software interfaces using icons with a relevance to their own culture are more satisfying for users, and Badre (2000) finds the cultural content of web sites affects preference for a web site. How do these elements of a product affect overall usability for each culture? Much research has also been done on the overall effect of individual elements of consumer electronic product design, such as size, look and feel, etc on usability (e.g., Han et al., 2000). However, this has been done in a single culture. It would be useful to compare the effects such design elements have on usability across cultures.

6.6 Implications

The study therefore supports our hypothesis - that culture does have an effect on absolute levels of subjective and objective effectiveness, efficiency and user satisfaction. This result has implications for the culturability model proposed in this dissertation, and combined with previous findings on the effect of culture on relative importance of usability attributes, suggests that overall culture cannot be ignored when investigating the usability of products intended for overseas markets.

6.6.1 Recommendations

This study provides strong evidence that culture affects usability as a whole. In so doing, it is hoped this study provides a methodology to be used in further research into the effect of culture on usability. It is recommended that future studies along similar lines consider the following.

- All aspects of usability must be assessed including subjective and objective measures of efficiency, effectiveness and user satisfaction. It is not certain that a measurement of a single aspect of usability will accurately represent levels of usability as a whole.
- A standard measuring instrument to measure attitudes should be used when measuring usability. A tool such as the USE instrument allows results to be compared across different products including software, web sites and consumer electronics.

This study also raises the possibility of further research in the following areas.

- To evaluate the total level of usability of a product it is necessary to understand the weighting a culture gives to the components of usability – efficiency, effectiveness and satisfaction. For example, Evers and Day (1997) show Indonesians place more importance on efficiency while Chinese users place more importance on effectiveness. By rating the usability of a product based on the

weighting given for each aspect of usability a culturally specific value of total usability could be derived, which could then be compared across cultures.

- Single culture usability research has isolated a range of product design factors affecting usability such as size and look and feel (Han, et al., 2001). A similarly detailed comparison of the effect of product characteristics on usability across cultures may help to identify reasons why culture affects usability so strongly.
- Many companies design for overseas markets. What solutions have they found already to the problem of culturability?

6.7 Conclusion

This study shows good evidence of a link between culture and usability. To understand usability as a whole, both subjective and performance-based measures of usability were considered. To gather data, commonly accepted usability measurement tools, such as the USE survey, and standard measures of performance, such as tasks complete, time taken and errors made, were used. By doing so, it is hoped similar studies can be made in future, to further develop understanding of the influence of culture on usability.

VII. Culture and the Accuracy of Judgment-Driven and Data-driven Severity Rating Methods

7.1 Introduction

Rating the severity of a usability problem (UP) is a problematic task for usability evaluators, particularly when the cultural background of users and evaluators differ. Yet this is an important and common task for evaluators. When conducting usability evaluation methods such as heuristic evaluations or empirical usability testing, evaluators are called upon to rate the severity of the usability problems they have identified in their tests. Severity ratings are then incorporated into usability reports upon which downstream users, such as software developers, make decisions on the allocation of resources to solve usability problems. However, given the impact cultural differences can have on perceptions of usability (Clemmensen, 2009), the process of rating the severity of UPs that are encountered by users from another culture may be prone to error. In this study, severity ratings prepared by Taiwanese evaluators of a video editing application are compared against ratings given by users from the target market, the U.S. By comparing the different ratings, we were able to identify more accurate and more reliable severity rating methods for use in a cross-cultural context.

Severity rating is an important tool for describing usability problems. Capra and Smith-Jackson (2005), in a study involving 74 experienced usability practitioners, suggest that usability practitioners on the whole feel that a description of the severity or impact of a problem is helpful, relevant to describing usability problems, and should be required in a usability report. However, along with providing solutions to UPs, usability practitioners judge rating the severity of a problem as the most difficult task in producing usability reports.

The difficulty of severity rating is shown in the results. Jacobsen et al. (1998) observed in one usability study that not one UP known to be severe was agreed to be severe by all evaluators. Nielsen (1992) does show that the inter-rater reliability of severity ratings increases as the number of evaluators goes up. However, the reliability of a single evaluator is so low, that the likelihood of the rating being that agreed on by most evaluators is only 55%.

As well as being inconsistent between evaluators, severity ratings are also often inconsistent with objective measures of problem severity. Hertzum (2006) surveyed six studies to identify correlations between frequency of problem occurrence and severity ratings and found correlations ranging from -0.29 (Woolrych and Cockton, 2001) to 0.46 (Virzi, 1992). Hassenzahl (2000) shows severity ratings based on problem-handling time do not correlate with severity ratings. John and Marks (1997) found only a low correlation between the number of users who experience a problem and the severity rating assigned to the UP by the evaluator.

An important reason for the lack of consistency in severity rating results is the subjectivity of usability evaluation methods, resulting in what is known as the evaluator effect. For example, Law and Hvannberg(2004) found that evaluators used their own experiences and preferences to derive results. Similarly, Hassenzahl found subjective factors such as the inability of evaluators to imagine the consequences of a UP influenced severity ratings assigned by evaluators, while an objective factor such as the frequency of a problem does not. Overall, Hertzum and Jacobsen (2001) argue the main reason for the evaluator effect is that "usability evaluation is a cognitive activity which requires that the evaluators exercise judgment".

It is important therefore to find severity rating methods that reduce the impact of the evaluator effect, especially in cross-cultural contexts. Culture also has been shown to influence usability measurement in terms of the test process, evaluator-subject interaction, and UPs identified (Shi and Clemmensen, 2008, Vatrapu and Pérez-Quiñones, 2006). Rating the severity of usability problems encountered by users from a culture different from that of the evaluator may therefore present an even greater risk of error from the evaluator effect.

Several severity rating methods are available for use by evaluators. Severity rating methods are typically judgment-driven or data-driven, depending respectively on whether the evaluator bases a rating on primarily subjective impressions of the UP's severity or uses empirical data to derive a severity rating. All methods employ criteria for rating a UP, however, Nielsen's criteria appear easiest to measure and quantify (Nielsen, 1992), which may offer a way to reduce bias caused by subjectivity. For this reason Nielsen's severity rating method is one method tested in this study.

Table VII-1 Severity Rating Methods

	Nielsen (1992)	Rubin (1994)	Dumas and Redish (1999)
Number of levels	0-4	1-4	1-4
Level descriptors	Based on priority of fix	Based on descriptions of user behavior and effort required to use product (combined with frequency to give a "criticality" rating).	Based on task completion and user emotional response
Severity sub-categories	Frequency – problem frequency Impact – effort required to overcome a problem Persistence – ability of user to learn how to overcome a problem		Scope – global and local

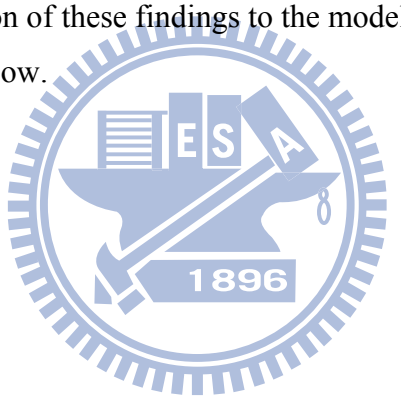
Usability engineering also suggests a severity rating method. According to Nielsen, the identification of UPs should be based on specified usability measures such as time taken, accuracy, etc (Nielsen, 1993). If usability measures can be used to identify UPs, then they can also be used to measure their severity. Although, as mentioned, the use of single objective usability measure has proven to be inaccurate, a better selection of usability measures may offer another solution to this problem. Guidance as to which usability measures should be employed may be offered by the ISO9241 definition of usability. ISO9241 defines usability as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and user satisfaction in a specified context of use" (ISO, 1998). As suggested by Hornbaek (2006), many possible measures of effectiveness, efficiency and user satisfaction exist, such as binomial success measures, number of errors made on task, and user satisfaction surveys, all of which may be useful indicators of problem severity. The second severity rating method tested in this study is therefore based on selected measures of effectiveness, efficiency, and user satisfaction.

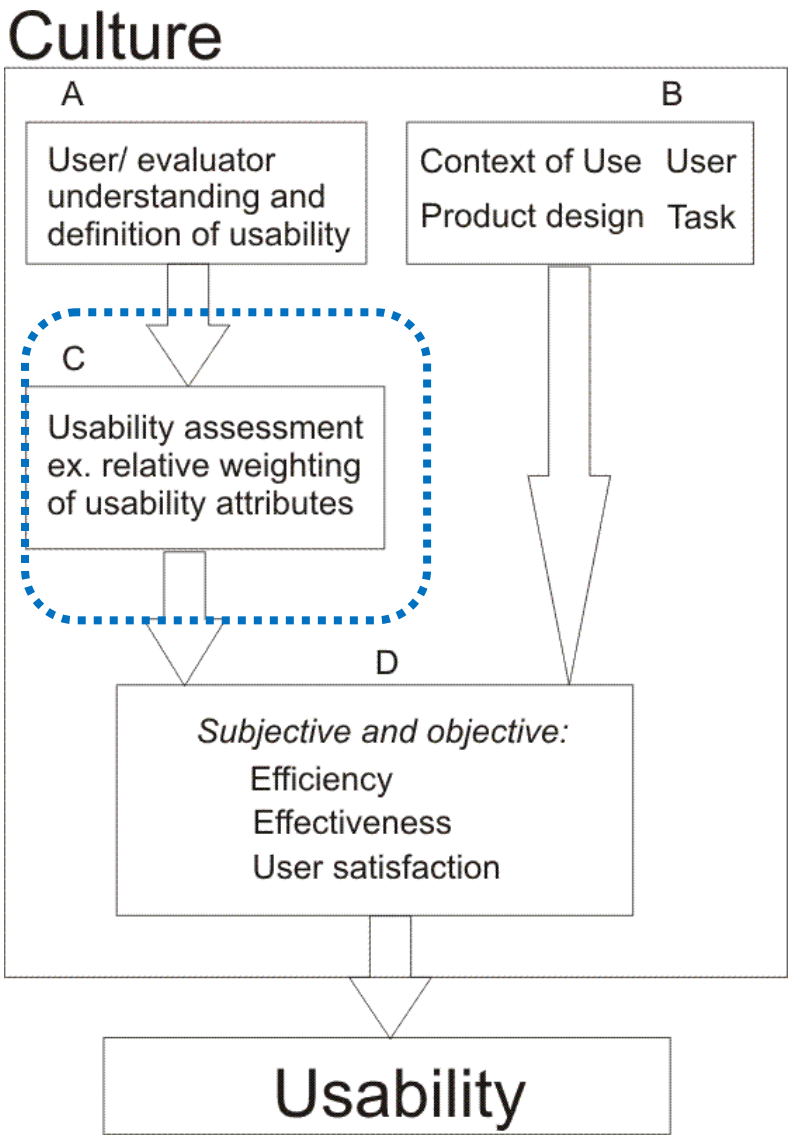
7.2 Research *Problem and Hypothesis*

The goals of this study are therefore to identify severity rating methods that may lead to greater accuracy, particularly in a cross-cultural context where subjectivity can easily lead to bias. This leads us to the following hypotheses.

1. Severity ratings made by Taiwanese usability evaluators differ significantly from those made by U.S. users.
2. Data-driven severity ratings made by Taiwanese usability evaluators more accurately predict U.S. user ratings than judgment-driven severity ratings.
3. Severity ratings made by Taiwanese usability evaluators based on usability attributes effectiveness, efficiency, and user satisfaction more accurately predict U.S. user ratings than those based on UP frequency, impact, and persistence.

This section of the dissertation seeks to ascertain the effect of culture on usability evaluation. If it is found that culture indeed has an effect on the results of usability evaluators, then possible solutions to minimize the effect of such bias may be presented. The relation of these findings to the model proposed by this dissertation is shown in diagram below.





7.3 Methodology

To confirm or deny these hypotheses, severity ratings provided by Taiwanese usability evaluators are compared to a gold standard based on severity ratings provided by U.S. users of the product. It is argued in this study that, while there is no objectively correct set of severity ratings available for comparison, severity ratings provided by users is closest to being accurate as possible. The users have experienced the UPs, and in the business context, it is their view of the usability of the product that evaluators need to understand to support the business needs of usability measurement.

7.3.1 Product

The product chosen for this study is a video-editing application called MagicSports, produced by a Taiwanese company, Cyberlink (www.cyberlink.com). The software automatically detects highlights in a sports video, and lets users create

their own video comprising the sports highlights. Initial inspection of the product indicated a number of evident usability problems on which to base this research.

7.3.2 Subjects

The experiment required participants from the software's target market, the U.S., and usability evaluators from a different culture. As the test was carried out in Taiwan, participants were contacted locally. Graduate HCI students from the National Taiwan University of Science and Technology acted as evaluators. Locally based U.S. users who matched the description of a target user were contacted through local social groups and paid a small amount for their participation in user testing.

The typical user of the sports editing software was judged to be a sports fan, confident with computers and video equipment, with previous video editing experience. It was likely the user would be male (although likely female users were found), tertiary educated, and working, with sufficient income to support luxuries such as hobby related computer hardware and software. The anticipated age group was therefore between 25 and 55. The user's nationality was most likely to be the U.S., given the size and income of its market and as the software was only available in English.

7.3.3 Procedure

Initial inspection of the software indicated twelve usability problems of interest. Fifteen US users then completed typical tasks using the software during which users encountered all of the identified UPs. Based on user performance, data was derived on frequency, impact, and persistence of twelve UPs. Also the UPs' effect on effectiveness, efficiency and user satisfaction was derived from measures of task completion, time taken on task, and a satisfaction survey.

During completion of a task, after experiencing a UP, users were asked to rate its severity. The ratings were based on a five point scale as follows.

1. An extremely minor problem
2. A minor problem
3. A somewhat important problem
4. A major problem
5. An extremely major problem

Afterwards, twenty one Taiwanese usability evaluators rated the same UPs, also with the scale shown above, using both judgment-driven and data-driven methods,

with data drawn from the user tests described earlier provided to evaluators when using data-driven rating methods.

Evaluators were walked through each UP, with each problem described in terms of when and where users encountered the problem, its cause, and resulting user behavior. Explanations were given first in English, then again in Chinese, although the evaluators' English was of a high level. Of the twelve UPs, users assessed the first six with no empirical data, i.e., the assessments were judgment-driven, while the second six were data-driven with empirical data provided. Of each six judgment and data-driven severity ratings, three were based on UP frequency, impact, and persistence criteria, while the others were based on measures of effectiveness, efficiency, and satisfaction.

The results were processed in a number of ways. To compare whether the mean severity rating for each UP differed significantly between evaluators and users, the MANOVA test was applied to results after removing outliers in user ratings. Results were also compared for internal consistency, as to be effective, ratings must also agree with each other, as well as being accurate in their ratings. By setting user median user ratings (after removing outliers) as a gold standard to which evaluator ratings can be compared, I was able to use attribute agreement analysis to assess the accuracy of the different severity rating methods. Finally, I compared the observed number of accurate ratings based on the gold standard with expected number using chi-squared analysis.

Table VII-2 Severity Rating Methods

Question Type	Method description
Method 1	Judgment-driven, rating based on UP frequency, impact, and user persistence
Method 2	Judgment-driven, rating based on the effect of the UP on product effectiveness, efficiency and user satisfaction.
Method 3	Data-driven, rating based on UP frequency, impact, and user persistence
Method 4	Data-driven, rating based on the effect of the UP on product effectiveness, efficiency and user satisfaction.

7.4 Results

Overall, results show greater accuracy, if not consistency, for data-driven methods and for methods based on measures of effectiveness, efficiency, and user satisfaction.

MANOVA analysis shows an overall significant difference between ratings by evaluators and users, as well as significant differences for Methods 1-3. The smaller Wilks' lambda for Method 1 indicates a large difference between mean user and evaluator ratings, while Methods 2 and 3 give evaluator ratings that are statistically significant and closer to user ratings.

Table VII-3 Association between Method and Rating Accuracy

Question Type	Wilks' lambda	p-value
Method 1	0.18	0.000**
Method 2	0.69	0.016*
Method 3	0.80	0.049*
Method 4	0.77	0.112
Overall	0.22	0.004**

*significant (alpha = 0.05)

** highly significant (alpha = 0.01)

The inter-rater consistency of each severity rating method, as indicated by Kendall's coefficient of concordance, shows raters using Method 3 agree with each other much more, while raters using Method 4 are least consistent.

Table VII-4 Rating Method and Inter-Rater Agreement

Question Type	Kendall's coefficient of concordance	p-value
Method 1	0.41	0.0002**
Method 2	0.49	0.0000**
Method 3	0.79	0.0000**
Method 4	0.40	0.0002**
Overall	0.53	0.0000**

The results of attribute agreement analysis, as indicated by Kendall's coefficient of concordance, show significant and moderate to high levels of accuracy for most methods, apart from Method 1. Overall, data-driven methods were more accurate than judgment-driven methods, while effectiveness, efficiency and user satisfaction proved a more reliable guide to rating UP severity than did UP frequency, impact, and persistence.

Table VII-5 Rating Method Accuracy

Rating Method	Kendall's coefficient of concordance	p-value
Method 1	0.07	0.6858
Method 2	0.47	0.0045*
Method 3	0.77	0.0000**
Method 4	0.56	0.0001**
Overall	0.44	0.0000**

Table VII-6 Rating Method Accuracy (judgment vs. data-driven)

Rating Method Type	Kendall's coefficient of concordance	p-value
Judgment-driven	0.41	0.0000**
Data-driven	0.50	0.0000**

Table VII-7 Rating Method Accuracy based on UP Severity Criteria

Rating Method Type	Kendall's coefficient of concordance	p-value
Based on frequency, impact, persistence	0.33	0.0000**
Based on measures of effectiveness, efficiency, satisfaction	0.47	0.0000**

By comparing the number of ratings by evaluators that matched or did not match median ratings by users, significant differences in the accuracy of the rating methods were identified, X^2 (3, N = 252) = 33.341, p = 0.000. The association between rating method and accuracy is weak as indicated by the lambda value (λ=0.22), perhaps because Method 2 and 3 gave results similar to what would be expected. However, Method 1 resulted in a much lower number of accurate ratings than expected, while Method 4 was more accurate than expected.

Table VII-8 Rating Method Accuracy

Question Type	Accurate Evaluator Ratings	Expected Count	Contribution to C^2	Inaccurate Evaluator Ratings	Expected Count	Contribution to C^2
Method 1	12	29.50	10.381	51	33.50	9.142
Method 2	32	29.50	0.212	31	33.50	0.187
Method 3	30	29.50	0.009	33	33.50	0.008
Method 4	44	29.50	7.127	19	33.50	6.276

7.5 Discussion

My hypotheses are suggested, if not confirmed, by the results. As a type, data-driven ratings are more accurate than judgment-driven ratings, and methods based on measures of effectiveness, efficiency, and user satisfaction are more accurate than those based on UP frequency, impact, and persistence.

For each method, however, discrepancies are found. Method 4, which is both data-driven and based on ISO9241 usability measures, is much less accurate than Method 3, which is data-driven and based on frequency, impact, and persistence. The reason for this discrepancy may lie in the observation that Method 4 also shows lower levels of inter-rater agreement, suggesting the greater spread of results for Method 4. This suggests that Method 4 may produce more accurate ratings, but inaccurate assessments are spread much further away from correct ratings than they are for Method 3. The greater familiarity of HCI students with methods based on UP frequency, impact, and persistence may have a part to play in the greater consistency of Method 3. If this is so, with training and increasing familiarity, improvements in inter-rater consistency and also accuracy could be achieved with Method 4.

7.5.1 Limitations of This Research

As mentioned, culture has been shown to influence the process and results of usability measurement. It seems likely therefore that culture may also impact on severity ratings. However, while this study shows a gap between U.S. users and Taiwanese evaluators, it does not conclude that this is the result of culture. The same gap may exist between users and evaluators from the same culture. Until ratings between evaluators from different cultures are compared, it cannot be presumed culture is a factor in severity ratings.

7.5.2 Research Implications

Nonetheless, the study has implications for severity ratings in cross-cultural contexts. A significant difference between how Taiwanese evaluators and U.S. users rate the severity of UPs has been indicated. Evaluators rating the severity of products intended for users from different cultural backgrounds from their own should consider severity rating methods which reduce the chance of bias due to culture. The use of empirical data in data-driven severity rating methods is recommended in cross-cultural

contexts. In addition, the selection of criteria by which the severity of UPs is rated is also important. It is argued that ISO9241-based methods are more accurate, and that with greater exposure to this method on the part of the evaluators, greater inter-rater agreement could be achieved.

In fact, the more widespread use of ISO9241-based criteria in severity rating in general ought to be considered. ISO9241 is an ergonomics standard, widely accepted in the HCI (Human-Computer Interaction) community (Clemmensen and Roese, 2010), which provides a definition of usability aimed at user (and business) needs. The definition is both simple and comprehensive in its incorporation of subjective and objective usability by considering both user satisfaction and product performance, and its description of performance in terms of process (efficiency) and outcomes (effectiveness). The importance of a comprehensive analysis of usability (and usability problems) is supported by results indicating the lack of correlation usability attributes have with each other(Frøkjær, et al., 2000, Hornbæk and Law, 2007). All aspects of a usability problem must therefore be understood when rating the severity of UPs. With a more comprehensive understanding of the user's experience of usability problems, gaps between evaluator and user such as that caused by culture may be better bridged.

The findings of this study also serve to confirm the final section of the model proposed by this dissertation - that culture also affects usability measurement.

7.6 Conclusion

By comparing the severity ratings methods employed by Taiwanese evaluators with severity ratings by U.S. users, we have identified rating methods which may be more exact in cross-cultural contexts. Data-driven severity rating methods, and possibly, those based on measures of effectiveness, efficiency, and user satisfaction, given sufficient evaluator familiarity, may be the most appropriate rating method in such situations. With the greater reliability of such methods, the process of severity rating may be made easier for evaluators, and the reliability developers perceive of their results may be enhanced for the benefit of the usability profession as whole.

VIII. Improving the Cross-cultural Usability of Taiwanese-made Software Documentation with Heuristic Evaluations.

8.1 Introduction

There is a great deal of evidence of the need to improve technical documentation (Grayling, 2002, Novick and Ward, 2006), and Taiwanese-made software document is not likely to be an exception. Heuristic evaluations offer a means to make such documentation more usable. An heuristic evaluation is a usability evaluation method based on a set of usability heuristics which evaluators use to identify usability problems. Some commonly used heuristics are those developed by Molich and Nielsen (1990), which can be applied to systems generally. Other heuristic evaluations focus on the usability of a specific media or product type, such as technical documentation (DeBoard, 2004, Kantner, et al., 2002, Purho, 2002). When producing documentation for overseas users, it may be useful to know whether it is more useful to employ general usability heuristics such as those developed by Nielsen and Molich, or to apply heuristics developed specifically for technical documentation. This knowledge may be useful to those looking to improve usability in not just technical writing, but also in other domains, as well as guiding researchers who intend to develop heuristic evaluations in the future. To find whether general or domain-specific heuristic evaluations are superior in a cross-cultural context, the usability of two sets of documentation heuristics are compared, one general set of heuristics based on Molich and Nielsen's heuristics but adapted to the requirements of technical documentation (Purho, 2002), the other a set of domain-specific heuristics developed specifically for online help (Kantner, et al., 2002).

8.1.1 Related Research

Previous comparisons of general and domain-specific heuristic evaluations consistently show better results for domain-specific heuristic evaluations. Mankoff et al. (2003) compared heuristic evaluations when assessing the usability of ambient displays and found domain-specific heuristics identified more issues than general heuristics. Somervell (2004) examined the usability of large screen information

exhibits and found domain-specific heuristics were more accurate and thorough. Connell and Hammond (1999) compared Nielsen's set of general heuristics to a set of usability principles and found domain-specific results were better for expert evaluators.

These studies concluded that domain-specific heuristic evaluations were superior based on the higher number of usability problems they identified ("hits"). This result seems reasonable, given that it is likely that heuristics which focus on the product at hand require less cognitive loading, and so better support evaluators in their search for usability problems. However, comparing the number of usability problems a heuristic evaluation identifies is only reliable if these usability problems genuinely exist and are valid usability problems (Law and Hvannberg, 2004). Ascertaining the validity of usability problems that heuristic evaluations identified was managed with varying degrees of success in the studies mentioned. Law and Hvannberg point out that the study of Connell and Hammond was based on counts of "raw", unconfirmed usability problems. Somervell's study confirmed the usability problems identified by heuristic evaluations by applying a number of methods depending on the product, including reliance on the possibly biased opinion of the system developer. Mankoff assesses the validity of usability problems found by heuristic evaluations by comparing them to the results of an informal expert review. The variety of methods used to confirm usability problems found by heuristic evaluations and their potential for error suggest a more rigorous comparison of general and domain-specific heuristic evaluations is possible.

8.1.2 Issues Related to Comparing Heuristic Evaluations

Identifying the validity of usability problems is a common problem for usability evaluation in general (Gray and Salzman, 1998). The subjectivity of usability evaluation methods has led to the suggestion that a complete and accurate list of usability problems for a product is essentially unknowable (Molich and Dumas, 2008, Molich, et al., 2004). Instead of a "true" list of usability problems, Hartson et al. therefore suggest generating a “criterion of truth” (Hartson, et al., 2001), or a list of usability problems that are agreed to represent all in existence for a product. Law and Hvannberg (2004) applied a criterion of truth in their comparison of two general heuristic evaluations developed by Molich and Nielsen (1990) and Gerhardt-Powals (1996) by comparing results to those of user testing. User testing provided a list of usability problems to use as a benchmark, or criterion of truth, with which the validity of usability problems identified by heuristic evaluations could be assessed, so

supporting a more reliable comparison of the relative effectiveness of these heuristic evaluations.

Comparing the results of heuristic evaluations to those of user testing is not without its problems. User testing is a method whereby an evaluator observes and sometimes interacts with users as they use a product, usually in a laboratory situation, in order to identify usability problems. However, user testing identifies different types of usability problems from heuristic evaluations (Fu, et al., 2002, Law and Hvannberg, 2002, Tan, et al., 2009). Thus, when a heuristic evaluation identifies a usability problem which has not been confirmed by user testing (a "false positive"), or fails to identify a usability problem that has been identified by user testing (a "miss"), we cannot be certain that the heuristic evaluation is at fault. In fact, the usability problem may just have been of a type not easily identified by either heuristic evaluations or user testing.

Therefore, to measure effectiveness, in this study the use of false positives and misses is avoided, and effectiveness measures which rely on these measures, such as accuracy (hits/(hits + false positives)) and thoroughness (hits/(hits + misses)) (Cockton and Woolrych, 2002) are also not used. The number of confirmed hits instead is used, as is inter-rater agreement, which is a measure of effectiveness that does not require confirmation by user testing. Inter-rater agreement can be measured using the Any-Two Agreement method (Hertzum and Jacobsen, 2001) which measures the likelihood of any two evaluators agreeing on results. However, in this study kappa statistics were found to provide similar results to the Any-Two Agreement method and had the added benefit of providing p-values.

Other measures of comparison can be useful. Hartson et al. suggest overall usability as a measure (2001). To measure usability, we use the ISO-9241 standard definition which bases usability measurement on measures of effectiveness, efficiency, and user satisfaction (ISO, 1998). Consequently, in this study, we compare the usability of general and domain-specific heuristic evaluations by evaluating the effectiveness as indicated by the number of hits and inter-rater agreement; efficiency as shown by hits over time, and satisfaction as measured by user (evaluator) attitudes towards the heuristic evaluations.

8.1.3 Purpose and Hypothesis

This study therefore aims to evaluate whether an heuristic evaluation developed specifically for use in a domain such as technical writing is more usable (effective, efficient, and satisfying) than a general heuristic evaluation adapted for use in a domain for identifying usability problems for overseas users.

8.2 Methodology

This study examines the usability of two publically available heuristic evaluations designed for assessing the usability of online help. The help was part of a video editing application produced by Taiwanese company, Cyberlink, called MagicSports. The effectiveness, efficiency, and user satisfaction of the heuristic evaluations were evaluated by testing a group of eight technical writers, who were identified as likely users of the heuristic evaluations, on their use of the heuristic evaluations. The technical writers used the heuristic evaluations to identify usability problems in the application, the existence of which were verified by empirical user testing of the same application. Technical writers' satisfaction with the heuristic evaluations was assessed by interviewing technical writers after they used the heuristic evaluations. The effectiveness and efficiency of the heuristic evaluations were based on counts of usability problems found, evaluator reliability, and the time required to identify usability problems.

8.2.1 The Heuristic Evaluations

The heuristic evaluations were selected partly because they were publically available and likely to be used by members of the technical writing community, and also because of the important differences existing between the two. The first heuristic evaluation (GeneralHE) was suggested by Neva Purho, a technical writer at Nokia, with the support and involvement of the Society of Technical Writing (STC) community, particularly members of the STC Usability SIG mailing list, and is published on the STC web site (http://www.stcsig.org/usability/newsletter/0004-docsheuristics.html). The second heuristic evaluation (SpecificHE) was developed by Kantner et al. (2002), as employees of Tec-Ed, a user-centered design firm, and available from the company website (http://www.teced.com/PDFs/kantner_HE_of_online_doc.pdf). Only one other possible documentation-oriented HE was found by the authors, that proposed by

DeBoard (2004), but was originally thought to resemble a set of guidelines, rather than an heuristic evaluation, as defined by Gray and Salzmann (1998), and so not included in this study.

The heuristic evaluations differed mainly in the extent to which they are influenced by Nielsen's heuristics for interface evaluation. GeneralHE heuristics largely matched those of Nielsen's, differing from Nielsen's heuristics only in the way they were adapted to requirements of documentation testing, with explanations which could be described as sub-heuristics given for each heuristic, based on input from members of the STC Usability SIG mailing list. For example, the first heuristic suggests there should be a match between the documentation and the real world. Purho interprets this heuristic by stating the documentation should "speak the user's language", and "real world conventions" should be followed in order that information appears "in a natural and logical order" (Purho, 2002). For the purposes of this study, such interpretations are also counted as heuristics, suggesting a total number of 24 heuristics.

SpecificHE heuristics appear very different from Nielsen's original list. These heuristics are based on research related to usability issues in online documentation with guidelines from a number of sources, including research into hypertext, usability and design (Fillion and Boyle, 1991, Garzotto, et al., 1995, Grice and Ridgway, 1993), multimedia (Ficarra, 1997), information management (Salminen, et al., 1999) and online documentation (Mehlenbacher, 1993). Usability issues related to online help are categorized in a number of areas identified in the literature, with heuristics for each category sorted in tabular fashion into dimensions of content, structure, appearance, and dynamics. The number of heuristics is greater, with 44 identified.

Table VIII-1 Summary of GeneralHE Heuristics

Scope	Interface usability, but adapted for documentation usability.
Organization	A list of heuristics, with explanations of each heuristic including more detailed sub-heuristics. Heuristics: 1. Match between documentation and the real world 2. Match between documentation and the product 3. Purposeful documentation 4. Support for different users 5. Effective information design 6. Support for various methods for searching Information 7. Troubleshooting 8. Consistency and standards 9. Help on using documentation
Total heuristics including sub-heuristics	24

Table VIII-2 Summary of SpecificHE Heuristics

Scope	Online documentation usability.
Organization	A table of heuristic categories with heuristics for each category sorted between dimensions of content, structure, appearance, and dynamics. Categories: 1. Orientation 2. Efficiency 3. Flexibility 4. Control 5. Recognition 6. Familiarity 7. Consistency 8. Readability and aesthetics 9. Context sensitivity (help only) 10. Clarity.
Total heuristics including sub-heuristics	44

8.2.2 The Documentation

The documentation tested for usability was an online help feature in a commercially available software application from a major software corporation (Cyberlink, www.cyberlink.com). The application was a video editing product called MagicSports designed with the primary function of identifying the highlights in a sports video, such as when a goal was scored or a pitch was thrown.

8.2.3 User Testing

To set a benchmark list of usability problems, empirical usability testing was carried out on the software. Typical users from this product were identified as being English speakers, sports lovers, interested in video editing on their computer, and reasonably confident with computers. Advertisements were placed on local groups on social networking sites and local web sites, advertising for participants with the aforementioned qualities and offering a small remuneration. Eleven participants responded and ten were tested, all male, tertiary educated, employed in a range of professions, and in their twenties to their fifties. One female did respond to the advertisements, but indicated she was unlikely to use this kind of software.

Participants were given a number of tasks related to the use of the product, including setting up the product, playing a video, producing a video of game highlights, and troubleshooting. Participants were directed to use the help section to complete these tasks in order to observe how well they located information, as well as to identify problems reading and applying information. Usability problems were identified by a single evaluator based on indicators such as failure to complete, errors, delay, user dissatisfaction and were traced to usability problems by observation and user comment.

8.2.4 Heuristic Evaluations

At the same time as empirical usability testing was taking place, the heuristic evaluations were also conducted. Technical writers were used as test participants as they were likely users of the heuristic evaluations focused on online help. After advertising in a local web site in Taiwan, where the study took place, eight technical writers responded. All respondents were male, tertiary educated, native English speakers, with technical writing experience ranging from 2 to 10 years. Some had a strong background in usability, others had none. Further efforts were made to recruit female writers, but with no success, possibly because of the low number of female technical writers based in Taiwan.

The writers were asked to use GeneralHE and SpecificHE to identify the usability problems of the Help feature in the video application. In contrast to standard heuristic evaluations, writers performed the heuristic evaluations individually. Writers installed the software on their own computer and worked through each heuristic evaluation as they examined the software. Although exposure to both heuristic evaluations presented

the possibility of bias, it allowed writers to compare the heuristic evaluations, which was necessary when they were later asked to give their opinions on the heuristic evaluations. To limit bias, writers were instructed to base their responses solely on the heuristics given in each heuristic evaluation. In addition, the order in which writers worked through each heuristic evaluation varied with half the writers beginning with GeneralHE and others starting with SpecificHE. Results of the heuristic evaluations were returned by email to the authors, and shortly after, face-to-face interviews were organized with the writers where they expressed their opinions on the heuristic evaluations.

8.2.5 Data Collection and Analysis

Therefore, once user testing and the heuristic evaluations were completed, the raw data from both heuristic evaluations and user testing was passed to three additional technical writers (again male, native English speakers) respectively for classification. This was done to reduce a source of bias identified by Gray and Salzman. The accurate matching of usability problems identified by different methods requires common categories and descriptions. In so doing, researchers need to avoid bias caused by slotting usability problems identified by one method into problem categories identified by another. To remedy this potential for bias, the researchers organized additional technical writers to create a single taxonomy based on all usability problems found by all usability evaluation methods, thus avoiding the problem of fitting usability problems into categories for which they were not intended.

To create a common taxonomy of usability problems, each writer assigned categories to the usability problems available to him, and then collaboratively all three writers negotiated a framework of classification for all usability problems. With this framework, they returned to their list of usability problems and assigned them to the categories they had collectively generated. The lowest level of classification was as specific as possible, so that only the same usability problems were assigned to the same categories. Descriptions of the cause of a usability problem, its characteristics, or its solution were treated as belonging to the same usability problem. Separate instances of the same usability problem were also treated as the same. For example, user difficulty in identifying when a section ends was counted as the same problem, regardless of how many sections it occurred in.

With a common framework, it was then possible to compare the usability problems identified by each heuristic evaluation, and confirm their validity against results from user testing. Confirmed usability problems from each heuristic evaluation were summed. To assess heuristic evaluation effectiveness, average results for each heuristic evaluation were compared with paired t-testing and the association between heuristic evaluation and confirmed usability problems was measured using chi-squared, between heuristic evaluations overall and per high-level categories. To measure efficiency, the time taken to identify confirmed usability problems was compared using paired t-testing. To indicate satisfaction, writers were asked to give feedback on the heuristic evaluations.

8.3 Results

The results were mixed. Overall, there was little clear and statistically significant indication that one heuristic evaluation was more usable for writers. However, the heuristic evaluations showed a differing ability to identify usability problems in various categories. SpecificHE showed a slightly higher level of agreement among evaluators. Writers reported more satisfaction with GeneralHE.

8.3.1 Effectiveness

Results for effectiveness do not clearly show the superiority of one heuristic evaluation over the other. SpecificHE showed higher inter-rater agreement and higher unconfirmed hits. However, when hits were confirmed against user testing, neither heuristic evaluation was significantly superior, except in specific categories.

1. Reliability

As Hertzum and Jacobsen's Any-Two Agreement method showed similar results to Fleiss's Kappa, and since the latter also offered a p-value, kappa was used to calculate inter-rater agreement. For GeneralHE inter-rater agreement was 0.207 (p-value = 0.000**) and for SpecificHE inter-agreement was higher at 0.265 (p-value =0.000**).

2. Hits per Evaluator

As shown in table 3, per writer paired t-testing showed that when evaluators used SpecificHE they flagged more usability problems than when using GeneralHE, but usability problems were compared to user testing, many of these usability

problems could not be confirmed, and no significant difference between confirmed problems (p=0.632) was found.

Table VIII-3 Average Usability Problems

Heuristic Evaluations	Mean Unconfirmed Usability Problems	p-value	Mean Confirmed Usability Problems	p-value
GeneralHE	14.5	0.019**	13.12	0.632
SpecificHE	20		13.88	

3. Hits per Category

Overall there was no significant association between heuristic evaluation and the number of confirmed usability problems identified. Table 4 shows that each heuristic evaluation confirmed the existence of 39 usability problems, although the usability problems identified by each heuristic evaluation differed, and writers using SpecificHE made 115 hits on these usability problems, more than when using GeneralHE (104 hits). However, GeneralHE was more successful in identifying problems with information content, while SpecificHE helped writers identify more problems in locating information and organizing information.

Table VIII-4 Association between Heuristic Evaluation and Usability Problems Identified

UP category	Hits		Expected Hits	Measure of Association	
	GeneralHE	SpecificHE		X^2	p-value
Overall	104	115	109.5	0.731	0.392
Help Delivery Medium	8	12	10	1.16	0.281
Information Location	**27**	**52**	**39.50**	**13.442**	**0.000****
Information Content	**20**	**4**	**12**	**12.35**	**0.000****
Information Organization	**11**	**22**	**16.50**	**5.198**	**0.023***
Information Presentation	17	11	14.00	1.596	0.206
Visual Presentation	11	10	10.5	0.065	0.799
Language/ Terminology	10	4	7	2.887	0.089

* significant at alpha = 0.05

** highly significant at alpha = 0.01

4. Hits and Heuristics

Inspection of results suggests a connection between individual heuristics and the number of hits they generate.

GeneralHE was able to generate more hits in the Information Content category than SpecificHE. Inspection shows hits related to the lack of troubleshooting dominated this category. All writers agreed on this usability problem and based their finding on the same GeneralHE heuristic (no.7 on the list) which specifically mentioned the need for a troubleshooting. Nowhere did SpecificHE mention the need for troubleshooting and there was only one hit on troubleshooting in SpecificHE.

GeneralHE also performed well in identifying usability problems in navigation, albeit not as well SpecificHE. GeneralHE generated hits in the area of navigation tools, particularly in the lack of an index (6 out of 8 writers found this usability problem). All hits in this category were based on GeneralHE number 6 heuristic which gave examples of necessary navigation aids such as a table of contents, an index, and a search function. However, GeneralHE's hits in this category were significantly less than SpecificHE, which mentioned index, search, links, and menu multiple times in a number of heuristics found throughout the heuristic evaluation.

SpecificHE also did well in the Help Delivery Medium category. The main finding in this category was by six out of eight writers who found a problem switching between the help and the main application. These hits were based on the Control heuristic category in SpecificHE, and no counterpart to this heuristic was found in GeneralHE. Two writers using GeneralHE did note this usability problem, but each under a different heuristic.

8.3.2 Efficiency

Neither heuristic evaluation was shown to be more efficient to use than the other. While writers using GeneralHE could identify on average 0.359 confirmed usability problems per minute, and using SpecificHE could identify 0.551 confirmed usability problems per minute, this was not a statistically significant difference (p-value = 0.155).

8.3.3 User Satisfaction

Post-survey feedback showed that while the heuristic evaluations shared a number of problems, all evaluators bar one preferred GeneralHE. They found GeneralHE checklist format was easier to work through. SpecificHE was in the form of a table with related ideas which was found to be confusing, although one evaluator felt the table offered a more holistic understanding of the documentation's usability. The use of a table also resulted in some blank fields, which confused the evaluators and left them wondering what they should do for each field. In addition, the problem areas covered by SpecificHE were found to be too abstract. For example, categories such as "presentation" and "dynamics" required some cognitive effort to be understood.

On the other hand, there were some problems in common to both heuristic evaluations. All evaluators felt that bullets should be used more extensively, that topics could be further divided into more specific areas, and that sentences could be made less "wordy". Writers also noted that both heuristic evaluations generated a great deal of repeated ideas. They preferred the use of questions to statements and wanted some way to elicit ideas rather than yes/no responses.

8.4 Discussion

The hypothesis of this study, that an heuristic evaluation designed specifically for a domain (technical documentation) is more usable than a generalized heuristic evaluation for evaluating usability in a cross-cultural context, cannot be unreservedly confirmed. No statistically significant difference in overall effectiveness or efficiency has been observed. GeneralHE was more satisfying for writers, while SpecificHE was more effective in terms of reliability, and more accurate in two main categories, while GeneralHE was more accurate only in one category.

These results do not clearly support the research by Connell and Hammond, Mankoff et al., and Somervell. A possible reason for this discrepancy may lie in our methodology. Our results show that domain-specific heuristic evaluations generate more "raw", unconfirmed hits, which is confirmed by earlier studies. However, when these hits are confirmed by user testing, our results differ from previous studies and no clear difference between general and domain-specific heuristic evaluations can be shown.

Results suggest a number of factors worthy of consideration. The level of generality of a heuristic evaluation may not be a major factor in the deciding its effectiveness or overall usability. Also, factors such as the simpler format of GeneralHE and the use of abstract concepts by SpecificHE may influence user satisfaction, but showed no impact on effectiveness or efficiency. In fact, closer inspection of the heuristic evaluations suggested the level of detail of heuristics may have been a reason for the greater inter-rater consistency of SpecificHE, and the relative strengths of each heuristic evaluation in different categories as indicated by the number of hits.

8.4.1 Implications of this Study

The impact of detailed heuristics on effectiveness may reflect the impact of cognitive load on evaluators, as suggested by Law and Hvannberg. By removing abstraction, a heuristic evaluation allows evaluators to directly apply criteria in a simple yes/no fashion. However, this result does not imply that domain-specific heuristics are more effective, as detailed heuristics can be found in both GeneralHE and SpecificHE. Findings suggest that the level of product focus has much less effect on heuristic evaluation effectiveness than does the level of detail for each heuristic.

Before designing a heuristic evaluation replete with detailed heuristics, it is worth noting that more detailed heuristics entail greater risks. First, it may result in more false positives, although this was something which this study was not able to measure. By suggesting to an evaluator that an element is part of good document design, the evaluator may be easily led to accepting that the absence of such an element is a usability problem, even when it is not required. Another issue is the downstream utility of solutions generated by detailed heuristics. Writers gave feedback on the rigid structure of SpecificHE, suggesting that it limited their answers, and led to unimaginative problem-solving, which would in turn affect the downstream utility of evaluator suggestions. Finally, arguing for more detailed heuristics may go against the point of a heuristic evaluation which is designed to provide a set of standards or principles by which usability may be assessed. A possible solution to these issues may be to incorporate detailed heuristics under categories, or general heuristics; to include reminders to evaluators to consider detailed recommendations in light of user and task requirements; and to make space available after each heuristic category for more open-

ended responses. Such an heuristic evaluation may enhance the advantages of detailed heuristics while minimizing the drawbacks they entail.

8.4.2 Limitations of this Study

Given the use of a small number of gender-specific technical writers for this study, it may be difficult to generalize results to all technical writers who may use the heuristic evaluations in this study, let alone heuristic evaluations from a variety of domains. In light of Gray and Salzman's warnings on drawing unfounded conclusions, no such generalizations should be made without confirmation by other related studies.

8.5 Conclusion

An assessment of the usability of two heuristic evaluations, one based on heuristics developed by Nielsen and Molich and used in a variety of domains, the other focused on a specific domain, showed no clear overall difference when evaluating Taiwanese-made software documentation. No statistically significant difference was found in effectiveness as indicated by the number of hits and in efficiency as measured by hits over time. Feedback from subjects indicated greater satisfaction with the former heuristic evaluation based on its clearer format, while comparison of the number of hits showed that the latter heuristic evaluation was more effective in some categories and showed greater inter-rater agreement. These differences suggested that more detailed heuristics led to more hits and greater agreement among the writers who used the heuristic evaluations. This study therefore raises the possibility that the use of an heuristic evaluation which includes more detailed heuristics may lead to better results than otherwise.

IX.Findings

The findings of this research met the objectives of this dissertation. Culture does indeed have a major influence on usability as indicated by usability attributes effectiveness, efficiency, and user satisfaction. In addition, it impacts on how we measure usability in terms of the severity that we assign usability problems. Finally, solutions to the problem of improving cross-cultural usability have been identified, in particular, by using heuristic evaluations that are more detailed in their requirements, and severity ratings that are data-driven and based on ISO9241 usability attributes - effectiveness, efficiency, and user satisfaction.

The first part of my study examined the interaction of culture as defined by the cultural values defined by Hofstede and the World Values Survey and usability's component parts as defined by ISO9241 - effectiveness, efficiency, and user satisfaction. Culture and cultural values were found to have a significant impact on subjective and objective levels of effectiveness, efficiency, and user satisfaction, both in terms of their absolute and relative values.

The second part of my study examined usability measurement in a cross-cultural context and identified two possible methods to reduce evaluator error caused by cross-cultural differences between the evaluator or product designer and the downstream user. The third study looked at methods to identify the severity of usability problems and identified significant bias caused by cultural differences between the user and evaluator. Solutions that were found to reduce bias included ensuring that the severity ratings were based on data, rather than the subjective opinion of the evaluator, and that the ratings were based on usability attributes, effectiveness, efficiency and user satisfaction, rather than traditional criteria of frequency, impact, and persistence. The fourth study looked at heuristic evaluations as means to improve the usability of Taiwanese-produced technical documentation intended for U.S. users, and identified heuristics that were more detailed and clear in their requirements identified more usability problems, thus improving the downstream usability of the documentation.

9.1 Culture and In-Country and Between-Country Differences in the Importance of Usability Attributes

The first study into the effect of culture on the relative importance placed on usability factors, effectiveness, efficiency, and user satisfaction, found reason to support the model of cultural usability proposed in this research. This result was implied by Ford and Kotzé's Conceptual Model of Usability (2005), as well as research by Vöhringer-Kuhnt (2002) and Evers and Day (1997). Subjects from four countries were surveyed on their attitudes to the importance of the usability attributes mentioned. The study showed that significant differences in preference for usability attributes exist in some countries and between some countries. Levels of preference for usability attributes are related to one or more cultural dimensions identified by both Hofstede and the WVS. Comparison of results showed consistency with previous studies and suggest that, in most countries studied, effectiveness is most important and efficiency is more important than satisfaction. Between-country differences indicated by this study and others suggest that efficiency is more important in the Western countries studied than for the Asian countries and satisfaction is more important in the Asian countries studied than in Western countries. Australia and South Africa are exceptional for Western countries for their lower preference for effectiveness and the greater importance they place on satisfaction.

There appears to be a correlation between some cultural values and the importance place on some usability attributes. The importance of effectiveness does not appear to vary significantly across the cultures surveyed, and no relationship with cultural dimensions can be identified. However, efficiency and satisfaction appear to be influenced by a number of cultural dimensions. Correlation analysis indicated that preferences for efficiency were negatively influenced by values of power distance (acceptance of hierarchy in a society) and positively influenced by survival values. Individualism backed up by a number of values is also associated with higher preferences for efficiency. These secondary values include feminine values, uncertainty avoidance, long-term orientation, and secular-rational values. On the whole, such values relate to the achievement of individual goals, with feminine values making the exception. In contrast, preference for satisfaction is moderately correlated with self-expression values. Satisfaction is also related to combinations of values such as collective, traditional, masculine values, and short-term orientation, and reduced

levels of uncertainty avoidance. These values are related to non-task oriented goals such as status and face-saving needs, and maintaining group relationships.

These results have implications for usability evaluation and design. In international HCI practice, users, evaluators, and designers may have different understandings of usability. Designers need to keep in mind those usability attributes most values by users. Evaluators need to employ usability measurement methods that support the measurement of usability attributes, and to incorporate the preferences of users into their findings and recommendations.

9.2 Culture, User Satisfaction, and Objective and Subjective Effectiveness and Efficiency

The second study compared Taiwanese and North American users and evaluated the effect cultural differences had on usability attributes - subjective user satisfaction and subjective and objective effectiveness and efficiency. Previous studies had already noted the effect of culture on effectiveness (Ford and Gelderblom, 2003), task (Chau, et al., 2002, van Biljon and Kotzé, 2008), and product design (Ford and Kotzé, 2005). However, little research had been made into the effect of culture on usability attributes as a whole. The study showed similarities and differences between cultures. Effectiveness as measured by the number of tasks completed was similar for both cultural groups. Efficiency was measured by both the time taken and the number of mistakes made when performing a task. North American users required similar amounts of time to their Taiwanese counterparts to complete tasks, but made more errors within that time. One possible explanation for this could be a different problem-solving style between Taiwanese and North American users. It was often observed during tests that, when faced with a problem using the MP3 player, North American users sometimes became more active or even clearly frustrated, which may have been the reason for the higher number of failed attempts, or errors. The number of errors then correlated with low levels of user satisfaction and perceptions of efficiency among North American users.

This result has implications for product designers and managers in Taiwan who intend to market their products in important markets overseas such as the U.S. Designers must keep in mind the cultural differences that exist between the local and destination markets. Culture clearly affects the usability of a product with regard to not only how well the customer can use the product but just as importantly the perceptions

of the usability of the product are affected. In addition, when designing for usability, all aspects of usability need to be considered, rather than focusing on improvements to single aspects, such as the look and feel, or ease of use.

9.2.1 Culture and the Accuracy of Judgment-Driven and Data-driven Severity Rating Methods

The third study examined whether culture impacts how usability is measured, and attempted to identify methods to make usability measurement more accurate within a cross-cultural context. The impact of culture on usability measurement is shown in a number of studies. Clemmensen's Cultural Model Theory of Usability describes the interaction of culture and evaluator perceptions of usability. This happens in the way the nationality of the interviewer affects subject responses (Vatrapu and Pérez-Quiñones, 2006), and how evaluator nationality affects the evaluation process (Shi and Clemmensen, 2008). My third study confirmed that culture influences evaluators' perception of the severity of usability problems, which is a major part of the usability evaluation process, and together with the fourth study of heuristic evaluations, indicates ways to improve accuracy when measuring usability across cultures.

In the third study, four severity rating methods were compared, in order to evaluate whether any method was more accurate in supporting Taiwanese usability evaluators to rate the severity of usability problems for U.S. users. All methods confirmed the prediction of my model, that evaluators from one culture experience difficulty in accurately rating usability problems experienced by users of other cultures. However, potential solutions were found by the study. Specifically, it was found that data-driven ratings were more accurate than judgment-driven ratings, and methods based on ISO9241 usability attributes, effectiveness, efficiency, and user satisfaction were more accurate than those based on UP frequency, impact, and persistence.

Results indicate that basing severity ratings on data is clearly more reliable than relying on evaluator judgment. However, it is not clear whether ISO9241 criteria or UP frequency, impact and persistence are more reliable. Basing severity ratings on data and ISO9241 usability measures results in significantly more hits but also has lower inter-rater agreement, indicating that evaluator ratings using this method are spread more further away from correct ratings, and have a higher probability of being inaccurate. On the other hand, evaluators using data and basing their evaluation on UP frequency, impact, and persistence have better inter-rater agreement, have a higher

probability of being accurate, but achieve less hits than the ISO9241-based alternative. The greater familiarity of usability evaluators with methods based on UP frequency, impact, and persistence may have a part to play in the greater consistency of this method. If this is so, with training and increasing familiarity, improvements in inter-rater consistency and also accuracy could be achieved with ISO9241 criteria.

The study clearly has implications for severity ratings in cross-cultural contexts. A significant difference between how Taiwanese evaluators and U.S. users rate the severity of UPs has been indicated. Evaluators rating the severity of products intended for users from different cultural backgrounds from their own should consider severity rating methods which reduce the chance of bias due to culture. The use of empirical data in data-driven severity rating methods is recommended in cross-cultural contexts. In addition, the selection of criteria by which the severity of UPs is rated is also important. It is argued that ISO9241-based methods are more accurate, and that with greater exposure to this method on the part of the evaluators, greater inter-rater agreement could be achieved.

9.3 Culture and the Usability Measurement of Taiwanese-made Software Documentation Using Heuristic Evaluations

As well as identifying ways to improve severity ratings in cross-cultural situations, the use of an important usability evaluation tool, heuristic evaluations, needs to be considered. To identify ways to make heuristic evaluation more effective in cross-cultural context, I examined the factors affecting the effectiveness of two heuristic evaluations when identifying usability problems of Taiwanese-made software documentation intended for U.S. user. Overall, no statistically significant difference in overall effectiveness or efficiency was observed in the heuristic evaluations (HE1, and HE2). However, HE1 was more satisfying for evaluators, while the HE2 was more effective in terms of reliability, and more accurate in two main categories, while the HE1 was more accurate only in one category.

My research identified several factors which may influence the efficacy of an heuristic evaluation to identify usability problems in the cross-cultural context. First, the theoretical background behind an HE may therefore not be a major factor in deciding the usability of an heuristic evaluation. In fact, closer inspection of the heuristic evaluations suggest the level of detail in heuristics may have been a reason for the greater rater consistency of HE2, and the relative strengths of each HE in

different problem categories as indicated by the number of hits. Other factors reported by writers such as the simpler format of HE1 and the use of abstract concepts in HE2 did not appear to make one HE more usable than the other.

Results seem to point to the benefits of more specific heuristics. If a potential problem is mentioned directly, in specific detail, and possibly mentioned repeatedly, then it is more likely to result in a hit. The higher inter-rater agreement for HE2 suggests this is likely for all evaluators. However, greater specificity entails some risks. First, it may result in more false positives, or the mistaken identification of a usability problem. By suggesting to an evaluator that an element is part of good document design, the evaluator may be easily led to accepting that the absence of such an element is a usability problem. A possible solution to this issue may be the design of an HE which combines general and specific heuristics in a comprehensive list, along with reminders to evaluators to consider specific recommendations in light of user and task requirements, and to make space available after each heuristic category for more open-ended responses. Such an HE may enhance the advantages of specified heuristics while minimizing the drawbacks they entail, thus making the documentation more usable for overseas users.

9.4 Implications *of this Research*

The results of my research have implications for Taiwanese product designers and usability practitioners, as well as implications for usability theory. Taiwanese producers cannot assume that their products will be equally usable overseas. It has been shown that U.S. users make more errors and become more dissatisfied with a product than Taiwanese users. Markets also vary according to the mix of usability attributes they value. While generally, effectiveness is important for most users in all the markets examined, the importance of efficiency is higher for cultures that value the achievement of individual goals such as the U.S. In contrast, preference for satisfaction is more related to non-task oriented goals such as status and face-saving needs, and maintaining group relationships, such as the Philippines. Product designers need to keep in mind the differences in each market and focus their product design to better match the requirements of their users.

Differences in attitudes towards usability due to culture also affect the attitudes of usability practitioners involved in identifying and rating the severity of usability problems. Such usability evaluators need to be aware of the possibility of error due to

the cultural difference between them and the user they are testing. In order to minimize the possibility of error, when rating the severity of a usability problem, they need to rely on data-driven descriptions of the usability problems, and minimize the need to exercise judgment on the severity of a usability problem. There is also the possibility that, if they become used to basing severity ratings on ISO9241-defined usability attributes, effectiveness, efficiency and user satisfaction, they will achieve greater accuracy in their ratings. Finally, when using usability evaluation methods such as heuristic evaluations, usability evaluators need to select heuristic evaluations that are detailed in their criteria of what makes a good product. In short, a reduction of the reliance on subjective judgments by usability evaluators and greater use of detailed and data-driven criteria, is required to make better usability evaluations in a cross-cultural context.

Finally, my research proposes enhancements to existing models of culture and usability. As mentioned, the model of culturability is a synthesis of existing usability models and definitions, combined with the findings of my research . The diagram below describes the important interactions between culture, usability, usability attributes, and usability measurement. My definition of usability is based on a combination of ISO9241's definition of usability attributes - effectiveness, efficiency, and user satisfaction, with Hornbaek's Usability Aspects Model, which describes the subjective and objective aspects of these usability attributes. Finally, this usability definition is combined with the MUSIC Model, which describes the interaction of a user with features of product design, to perform a task given a specific context of use. This interaction culminates in a level of usability which can be measured based on Hornbaek's adaption of the ISO9241 definition of usability.

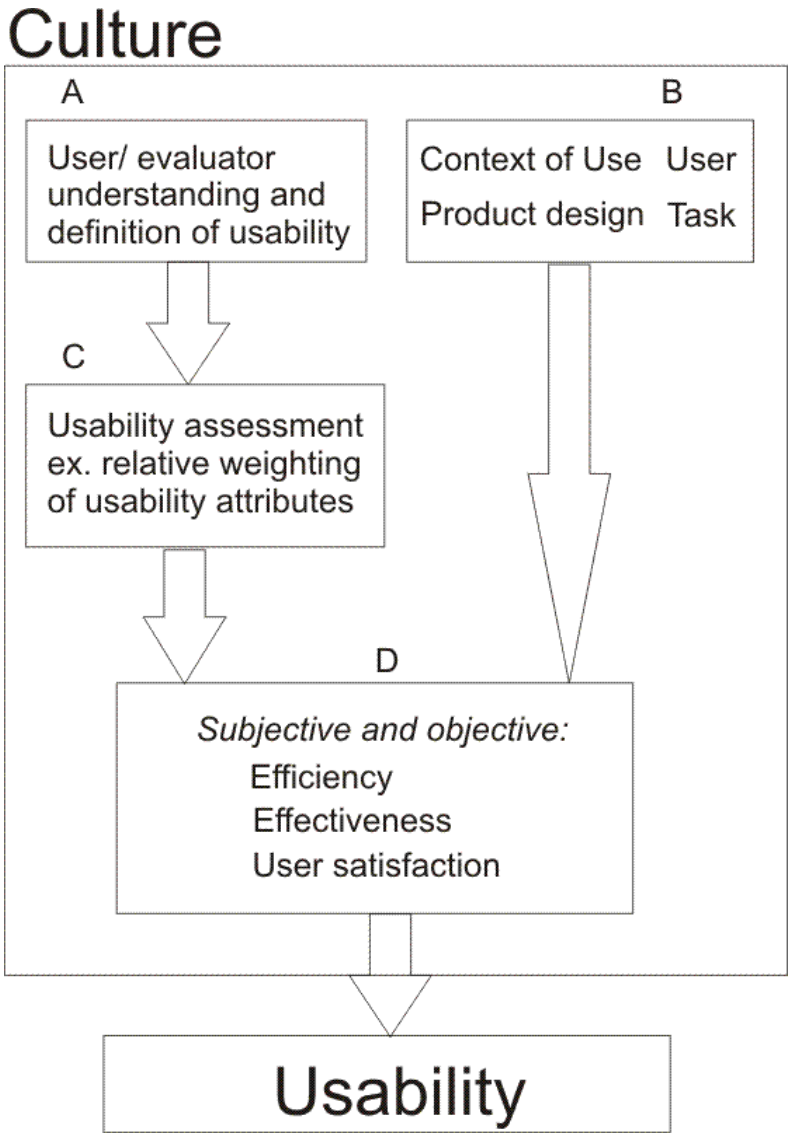


Figure IX-1 Proposed Model of Culturability

9.5 Limitations of this Research

While attempts have been made to make this research as accurate and free from error as possible, there still exist factors which may influence the accuracy of my results. Firstly, sample sizes in all studies have varied and have ranged from eight technical writers in the final study into heuristic evaluations, to several hundred respondents of different nationalities in my survey on the importance of usability factors. However, in all statistics reported, the p-value has been given. This indicates the probability a result is not random and takes into account sample size. Often p-values have been in the region of highly significant, suggesting the robustness of my results.

Secondly, while the samples on which these studies are based were selected for their ability to represent the population under consideration, limitations naturally exist. These occur in two areas - the cultural values considered in the survey on the importance of usability attributes, and in the type of products considered. Taking the latter first, in the four studies made, a number of consumer products were examined including an MP3 player, a cell phone, and software documentation. This selection does not represent the full range of products in any way. In fact, it would be impossible to test the effect of culture on usability for a full range of products. The only option is for future studies to ensure that alternative products are tested, and in this way a picture of usability is developed that may better represent the range of products types available. Similarly, the full range of cultural values also have not been examined. However, the cultural values selected may better represent those available. The countries selected represent Western and Eastern cultures, two major cultural regions of the world, and the cultural values considered include the results of two major global studies into culture, when typically in similar research only one set of results are examined (those of Hostede). By doing so, it is hoped that the impact of any inaccuracies in this research have been minimized and that results offer an acceptable level of robustness.

X. Conclusion

To conclude, this research has proposed a model of the interaction of culture and usability which bases itself on current models of culture and usability but more clearly explains the interaction of culture and usability attributes, and culture and usability measurement. Research shows that culture can affect the level of effectiveness, efficiency and satisfaction that users experience and perceive when using a product, and that also culture affects the importance users place on each of these usability attributes. Secondly, my research shows that culture affects how usability practitioners measure usability problems. However, some recommendations are made to possibly reduce the impact of culture on the accuracy of usability evaluation, thus supporting Taiwanese product designers and Taiwanese industry as whole to make the most of the opportunities presented by overseas markets.

References

1. Andre, T.S., Hartson, H.R., Belz, S.M., McCreary, F.A., "The user action framework: a reliable foundation for usability engineering support tools", International Journal of Human-Computer Studies, 54, pp. 107-136, 2001.
2. Badre, A.N., "The effects of cross cultural interface design orientation on World Wide Web user performance", GVU Tech Reports, 2000.
3. Barber, W., Badre, A., "Culturability: the merging of culture and usability", Human Factors and the Web 4th Conference, 1998.
4. Bevan, N., "The MUSIC methodology for usability measurement", 1992 SIGCHI Conference on Human Factors in Computing Systems, pp. 123-124, Monterey, California, 1992.
5. Bevan, N., "International standards for HCI and usability", International Journal of Human-Computer Studies, 55, pp. 533-552, 2001.
6. Borgman, C.L., "The user's mental model of an information retrieval system: an experiment on a prototype online catalog", International Journal of Man-Machine Studies, 24, pp. 47-64, 1986.
7. Bourges-Waldegg, P., Scrivener, S.A.R., "Meaning, the central issue in cross-cultural HCI design", Interacting with Computers, 9, pp. 287-309, 1998.
8. Capra, M.G., Smith-Jackson, T.L., 2005. Developing guidelines for describing usability problems, Technical Report #ACE/HCI-2005-002. Virginia Tech, Blacksburg.
9. Chau, P.Y.K., Cole, M., Massey, A.P., Montoya-Weiss, M.M., O'Keefe, R.M., "Cultural differences in the online behavior of consumers", Communications of the ACM, 45, pp. 138-143, 2002.
10. Choi, B., Jeon, Y., Kim, J., Lee, I., "A qualitative cross-national study of cultural influences on mobile data service design", Proceedings of CHI 2005 Portland, pp. 661-670, 2005.
11. Clemmensen, T., "Towards a theory of cultural usability: A comparison of ADA and CM-U theory", Lecture Notes in Computer Science, 5619, pp. 416-425, 2009.
12. Clemmensen, T., Hertzum, M., Hornbaek, K., Shi, Q., Yammiyavar, P., "Cultural cognition in usability evaluation", Interacting with Computers, 21, pp. 212-220, 2009.
13. Clemmensen, T., Roese, K., "An overview of a decade of journal publications about Culture and Human - Computer Interaction (HCI)", IFIP Advances in Information and Communication Technology, 316, pp. 98-112, 2010.

14. Cockton, G., Woolrych, A., "Sale must end: should discount methods be cleared off HCI's shelves?", interactions, 9, pp. 13 - 18, 2002.
15. Connell, I.W., Hammond, N.V., "Comparing usability evaluation principles with heuristics", INTERACT'99 - Human-Computer Interaction, pp. 621-629, 1999.
16. DeBoard, D., 2004. Heuristics to evaluate online help, Usability interface.
17. Evers, V., Day, D., "The role of culture in interface acceptance", IFIP TC13 International Conference on Human–Computer Interaction INTERACT'97, Chapman & Hall, Ltd., Sydney, Australia, pp. 260 - 267, 1997.
18. Ficarra, F.V.C., "Evaluation of multimedia components", Multimedia Computing and Systems '97 Ottawa, Ont., Canada, pp. 557-564, 1997.
19. Fillion, F.M., Boyle, C.D.B., "Important issues in hypertext documentation usability", 9th annual international conference on Systems documentation. ACM Chicago, pp. 59 - 66, Illinois, United States, 1991.
20. Fitzgerald, W., "Evaluation of multimedia components", Multimedia Computing and Systems '97 Proceedings IEEE, Ottawa, Ont., Canada, pp. 557-564, 2004.
21. Folmer, E., Bosch, J., "Architecting for usability", Journal of Systems and Software, 70, pp. 61-78, 2004.
22. Ford, G., Gelderblom, H., "The effects of culture on performance achieved through the use of human computer interaction", SAICSIT '03 Proceedings of the 2003 annual research conference of the South African institute of computer scientists and information technologists on Enablement through technology. ACM, pp. 218-230, 2003.
23. Ford, G., Kotzé, P., "Designing usable interfaces with cultural dimensions", Lecture Notes in Computer Science, 3585, pp. 713-726, 2005.
24. Ford, G., Kotzé, P., "Researching Culture and Usability — A Conceptual Model of Usability. In: McEwan, T., Benyon, D., and Gulliksen, J., (Eds.)", People and Computers XIX — The Bigger Picture, Springer, pp. 317-333, 2005.
25. Frandsen-Thorlacius, O., Hornbæk, K., Hertzum, M., Clemmensen, T., "Non-universal usability? A survey of how usability is understood by Chinese and Danish users", Conference on Human Factors in Computing Systems: Proceedings of the 27th international conference on human factors in computing systems ACM Press, pp. 41-50, Boston, U.S.A., 2009.
26. Frøkjær, E., Hertzum, M., Hornbæk, K., "Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated?", Proceedings of the ACM CHI 2000

Conference on Human Factors in Computing Systems ACM Press, pp. 345-352, New York, 2000.

27. Fu, L., Salvendy, G., Turley, L., "Effectiveness of user testing and heuristic evaluation as a function of performance classification", Behaviour & Information Technology, 21, pp. 137-143, 2002.
28. Garzotto, F., Mainetti, L., Paolini, P., "Hypermedia design, analysis, and evaluation issues", Communications of the ACM, 38, pp. 74-86, 1995.
29. Gerhardt-Powals, J., "Cognitive engineering principles for enhancing human-computer performance", International Journal of Human-Computer Interaction, 8, pp. 189 - 211, 1996.
30. Gray, W.D., Salzman, M.C., "Damaged merchandise? A review of experiments that compare usability testing methods", Human-Computer Interaction, 13, pp. 203-261, 1998.
31. Grayling, T., "If we build it will they come? A usability test of two browser-based embedded help systems", Technical communication, 49, 2002.
32. Grice, R.A., Ridgway, L.S., "Usability and hypermedia: Toward a set of usability criteria and measures", Technical Communication, pp. 429-437, 1993.
33. Hall, E., Beyond Culture, Anchor Press, Garden City, N.Y., 1976.
34. Han, S.H., Hwan Yun, M.H., Kwahk, J., Hong, S.W., "Usability of consumer electronic products", International Journal of Industrial Ergonomics, 28, pp. 143-151, 2001.
35. Hartson, H.R., Andre, T.S., Williges, R.C., "Criteria for evaluating usability evaluation methods", International journal of human-computer interaction, 13, pp. 373-410, 2001.
36. Hassenzahl, M., "Prioritizing usability problems: data-driven and judgment-driven severity estimates", Behaviour & Information Technology, 19, pp. 29-42, 2000.
37. Hertzum, M., "Problem prioritization in usability evaluation: From severity assessments towards impact on design", International Journal of Human Computer Interaction, 21, pp. 125-146, 2006.
38. Hertzum, M., Clemmensen, T., Hornbæk, K., Kumar, J., Shi, Q., Yammiyavar, P., "Usability constructs: A cross-cultural study of how users and developers experience their use of information systems", Lecture Notes in Computer Science, 4559, pp. 317-326, 2007.
39. Hertzum, M., Jacobsen, N.E., "The evaluator effect: A chilling fact about usability evaluation methods", International journal of human-computer interaction, 13, pp. 421-443, 2001.

40. Hofstede, G., Culture's consequences: international difference in work related values, Sage, London and Beverly Hills, 1984.
41. Hofstede, G., Bond, M.H., "The Confucius connection: From cultural roots to economic growth", Organisational Dynamics, pp. 5-21, 1998.
42. Hofstede, G., Minkov, M., Vinken, H., Values Survey Module 2008 Manual, 2008.
43. Hoft, N.L., Developing a cultural model, International users interface, John Wiley & Sons, Inc, New York, NY, USA, pp. 41 - 73, 1996.
44. Hornbæk, K., "Current practice in measuring usability: Challenges to usability studies and research", International journal of human-computer studies, 64, pp. 79-102, 2006.
45. Hornbæk, K., Law, E.L.-C., Meta-analysis of correlations among usability measures, Proceedings of the SIGCHI conference on Human factors in computing systems San Jose, California, USA, pp. 617-626, 2007.
46. ISO, 1998. ISO 9241-11, Ergonomic requirements for office work with visual display terminals (VDTs) -- Part 11: Guidance on usability. International Organization for Standardization (ISO), Geneva, Switzerland.
47. ISO, 2001. ISO 9126-1, Software Engineering - Product Quality - Part 1: Quality Model. International Organization for Standardization (ISO), Geneva, Switzerland.
48. Jacobsen, N.E., Hertzum, M., John, B., "The evaluator effect in usability studies: problem detection and severity judgments, Proceedings of the Human Factors and Ergonomics Society", Human Factors and Ergonomics Society, Chicago, U.S.A, pp. 1336-1340, 1998.
49. John, B.E., Marks, S.J., "Tracking the effectiveness of usability evaluation methods", Behaviour & Information Technology, 16, 188-202, 1997.
50. Kantner, L., Shroyer, R., Rosenbaum, S., "Structured heuristic evaluation of online documentation", Professional Communication Conference, pp. 331-342, 2002.
51. Kuroso, M., Kashimura, K., "Apparent usability vs. inherent usability: experimental analysis on the determinants of the apparent usability", Conference on Human Factors in Computing Systems, ACM, pp. 292-293, Denver, Colorado, United States, 1995.
52. Kurosu, M., Kashimura, K., "Apparent usability vs. inherent usability: experimental analysis on the determinants of the apparent usability", Conference on Human Factors in Computing Systems Denver, pp. 292 - 293, Colorado, United States, 1995.
53. Law, E.L.-C., Hvannberg, E.T., "Complementarity and convergence of heuristic evaluation and usability tests: A case study of UNIVERSAL brokerage platform", NordiCHI 2002 Aarhus, pp. 71-80, Denmark, 2002.

54. Law, E.L.-C., Hvannberg, E.T., "Analysis of strategies for improving and estimating the effectiveness of heuristic evaluation", NordiCHI 2004 Tampere, pp. 241-250, Finland, 2004.
55. Lee, I., Choi, G.W., Kim, J., Kim, S., Lee, K., Kim, D., Han, M., Park, S.Y., An, Y., "Cultural dimensions for user experience: Cross-country and cross-product analysis of users' cultural characteristics", 22nd British HCI Group Annual Conference on HCI 2008: People and Computers XXII: Culture, Creativity, Interaction British Computer Society Liverpool, pp. 3-12, United Kingdom, 2008.
56. Lund, A.M., Measuring usability with the USE questionnaire, Usability Interface, 2001.
57. Mankoff, J., Dey, A.K., Hsieh, G., Kientz, J., Lederer, S., Ames, M., "Heuristic evaluation of ambient displays", Human factors in computing systems, ACM Press, pp. 169-176, 2003.
58. Marcus, A.. "Cross-cultural user-experience design", Lecture Notes in Computer Science, 4045/2006, pp.16-24, 2006.
59. Marcus, A., Gould, E.W., "Crosscurrents - Cultural dimensions and global web user-interface design", Interactions, 7, pp. 32-46, 2000.
60. Mayhew, D.J., Mantei, M., A basic framework for cost-justifying usability engineering, In: Bias, R. G., and Mayhew, D. J., (Eds.), Cost Justifying Usability Harcourt Brace and Co., Boston, pp. 111-122, 1994.
61. McGee, M., Rich, A., Dumas, J., "Understanding the usability construct: User-perceived usability", Human Factors and Ergonomics Society Annual Meeting Human Factors and Ergonomics Society, pp. 907-911, 2004.
62. McSweeney, B., "Hofstede's model of national cultural differences and their consequences: A triumph of faith – a failure of analysis", Human relations, 55, pp. 89-118, 2002.
63. Mehlenbacher, B., "Software usability: Choosing appropriate methods for evaluating online systems and documentation", SIGDOC Conference 1993, Getting in Touch—Staying in Touch, pp. 209-222, 1993.
64. Molich, R., Dumas, J.S., "Comparative usability evaluation (CUE-4)", Behaviour & Information Technology, 27, pp. 263-281, 2008.
65. Molich, R., Ede, M.R., Kaasgaard, K., Karyukin, B., "Comparative usability evaluation", Behaviour & Information Technology, 23, pp. 65-74, 2004.
66. Molich, R., Nielsen, J., "Heuristic evaluation of user interfaces", ACM CHI'90 Seattle, WA, pp. 249-256, 1990.

67. Molich , R., Nielsen, J., "Heuristic evaluation of user interfaces", ACM CHI'90 Seattle, pp. 249-256, WA, 1990.
68. Nantel, J., Glaser, E., "The impact of language and culture on perceived website usability", Journal of Engineering and Technology Management, 25, pp. 112-122, 2008.
69. Nielsen, J., Usability 101.
70. Nielsen, J., "Finding usability problems through heuristic evaluation", ACM CHI'92 Monterey, pp. 676-680, CA., 1992.
71. Nielsen, J., "Reliability of severity estimates for usability problems found by heuristic evaluation", Conference on Human Factors in Computing System, pp. 676-680, Monterey, California, 1992.
72. Nielsen, J., "Usability Engineering", Academic Press, New York, 1993.
73. Novick, D.G., Ward, K., "What users say they want in documentation", 24th Annual ACM International Conference on Design of Communication, pp. 84-91, Myrtle Beach, SC, 2006.
74. O'Keefe, R.M., Cole, M., Chau, P.Y.K., Massey, A., Montoya-Weiss, M., Perry, M., "From the user interface to the consumer interface: results from a global experiment", International Journal of Human-Computer Studies, 53, pp. 611-628, 2000.
75. Olie, R., "The 'culture' factor in personnel and organization policies", In: A., H., and London, V. R. J., (Eds.), International Human Resource Management: An integrated approach. Sage pp. 124-143, 1995.
76. Purho, N., "Heuristic Inspections for Documentation – 10 Recommended Documentation Heuristics", Usability Interface, 2002.
77. Salminen, A., Tiitinen, P., Lyytikäinen, V., "Usability evaluation of a structured document archive", Proceedings of the 32nd Hawaii International Conference on System Sciences, pp. 1-10, 1999.
78. Sauro, J., Lewis, J.R., "Correlations among prototypical usability metrics: evidence for the construct of usability", Conference on Human Factors in Computing Systems, ACM, pp. 1609-1618, Boston, MA, USA, 2009.
79. Schwartz, S.H., "Mapping and interpreting cultural differences around the world", In: Vinken, H., Soeters, J., and Ester, P., (Eds.), Comparing cultures, Dimensions of culture in a comparative perspective. Brill., Leiden, The Netherlands, pp. 43-73, 2004.
80. Shackel, B., "Usability—context, framework, definition, design and evaluation", Human factors for Informatics Usability, pp. 21 - 37, 1991.

81. Shen, S.T., Woolley, M., Prior, S., "Towards culture-centered design", Interacting with Computers, 18, pp. 820-852, 2006.
82. Shi, Q., Clemmensen, T., "Communication patterns and usability problem finding in cross-cultural thinking aloud usability testing", Conference on Human Factors in Computing Systems, pp. 2811-2816, Florence, Italy, 2008.
83. Singh, N.. "Culture and the World Wide Web: a cross-cultural analysis of web sites from France, Germany and USA", American Marketing Association, 14, pp. 30-31, 2003.
84. Straub, D., Loch, K., Evaristo, R., Karahanna, E., Srite, M., "Toward a theory-based measurement of culture", Journal of Global Information Management, 10, pp. 13–23, 2002.
85. Tan, W.-s., Liu, D., Bishu, R., "Web evaluation: Heuristic evaluation vs. user testing", International journal of industrial ergonomics, 39, pp. 321-327, 2009.
86. Tractinsky, N., "Aesthetics and apparent usability: empirically assessing cultural and methodological issues", SIGCHI Conference on Human Factors in Computing Systems. ACM Press, pp. 115–122, Atlanta, Georgia, United States, 1997.
87. Triandis, H.C., The Analysis of Subjective Culture, Wiley, New York, N.Y, 1972.
88. Trompenaars, F., Riding the Waves of Culture: Understanding Diversity in Global Business, Irwin Professional Publishing, New York, 1994.
89. van Biljon, J., Kotzé, P., "Cultural factors in a mobile phone adoption and usage model", Journal of Universal Computer Science, 14, pp. 2650-2679, 2008.
90. Vatrapu, R., Pérez-Quiñones, M., Culture and usability evaluation: The effects of culture in structured interviews, Journal of Usability Studies, 1, 156-170, 2006.
91. Virzi, R.A., "Refining the test phase of usability evaluation: How many subjects is enough?", Human Factors, 34, pp. 457-468, 1992.
92. Vöhringer-Kuhnt, T., "The influence of culture on usability", Freie Universität Berlin, Masters Thesis, 2002.
93. Walker, M.A., Fromer, J., Mestel, C., Hindle, D., "What can I say?: evaluating a spoken language interface to Email", CHI 98 Conference on Human Factors in Computing Systems. ACM, pp. 582-589, 1998.
94. Winter, S., Wagner, S., Deissenboeck , F., "A comprehensive model of usability", Lecture Notes In Computer Science, Berlin, Heidelberg, pp. 106 - 122, 2008.
95. Witold, A., Suryn , W., Khelifi , A., Rilling, J., Seffah, A., Robert, F., "Consolidating the ISO usability models", 11th International Software Quality Management Conference, Springer, Glasgow, Scotland, UK. , 2003.

96. Woolrych, A., Cockton, G., "Why and when five test users aren't enough", In: Vanderdonckt, J., Blandford, A., and Derycke, A., (Eds.), Proceedings of IHM-HCI 2001 Conference, pp. 105- 108, Cépadèus, Toulouse, France, 2001.

Appendix 1

Heuristic Inspections for Documentation – 10 Recommended Documentation Heuristics

We all are familiar with Jakob Nielsen's heuristics for evaluating the usability of interfaces. When I was conducting a study on documentation usability, I started wondering if there existed a similar set of heuristics for evaluating the usability of documentation. The natural place to pose such a question was the STC Usability SIG mailing list. The response was that there was no heuristics set available although someone had tried to open the discussion in the mailing list some time ago. An answer, which led to the list of heuristics presented below, was something along the line "Well, now that you asked, why don't you put the heuristics together" and so I did.

I combined all kinds of general ideas about what is good documentation and made a list of heuristics that I sent for comments to the usability mailing list. After making some modifications based on the comments, I came up with the following ten statements:

1. Match between documentation and the real world

   The documentation should speak the users' language, with words, phrases, and concepts familiar to the user, rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order.

2. Match between documentation and the product

   The forms, screens, manuals, and online helps system should match so that the same terminology is used in all of them. This may contradict with "Match between the documentation and real world" if the interface uses strange terminology.

3. Purposeful documentation

   If the documentation set contains several documents, the purpose of each type of document should be clear, as well as the intended use. The media of the documentation must be purposeful so that users get what they need. For example, people working on a rooftop installing some hardware would not necessarily be delighted with nice multimedia CD-ROMs but prefer a laminated quick reference card.

4. Support for different users

   The documentation should support users with different levels of knowledge on the domain as well as those assigned different tasks in the domain. Any unnecessary

information for a specific user must be hidden from other users or be easily overlooked. Quick reference information for expert users should be available.

5. Effective information design

   Information must be presented in a way that it is easily found and understood by the users. Short lines and paragraphs are easier to read. Graphics, tables, and lists are easy to scan and read, and appropriately used to support the information need the user has. Unnecessary graphics only slow the reading and the download time of web-based documentation. Write instructions in imperative form and address the user directly using active sentences.

6. Support for various methods for searching Information

   Documentation should support people with different strategies for finding information: some search through the table of contents, some use the index, some browse, and some use searches (in electronic documentation). The index should contain users' own terminology as well as system terms, terms from international standards, and those used by competitors. The layout of documentation should support browsing so that beginnings of new chapters and important warnings and notes are easily picked up.

7. Task orientation

   Instructional documentation should be structured around the users' job tasks, that is, tasks that are independent of the tools used. The job tasks remain the same although the tools may change. For example, the job task "baking bread" remains the same although the baker may do it all by hand or using latest state-of-the-art tools. This reduces the need to restructure the documentation when the product is changed. The tasks should be approximately at the same level of granularity throughout the documentation

8. Troubleshooting

   The documentation should contain a troubleshooting section giving users guidance for common problem situations and how to analyze rare situations. All documentation related to errors must be easily accessible.

9. Consistency and standards

   Users should not have to wonder whether different words, situations, or actions mean the same thing. If the product has several documents, they should be consistent in their structure and the information in different documents should be designed so that no unnecessary overlapping exists. Follow platform conventions when creating the help system. Be sure that the terminology is consistent throughout the documentation suite.

10. Help on using documentation

If the documentation set is large, provide instructions on intended use, and how it is going to be updated (if separate updates are delivered).

Heuristic Statements

Of course, heuristic statements are only generalizations and never replace careful planning and user-centered design processes. However, they can be used as a checklist on matters that need to be considered when designing documentation. One way to use them is to derive more specific checklists from the heuristics and add company and product specific statements to be used in planning and evaluation. If you have any comments on the heuristics, I would be glad if you can post them to the STC Usability SIG mailing list for discussion.

Appendix 2

Usability Heuristics for Evaluating Online Documentation

The usability heuristics presented here:

- Enable a quick and thorough inspection of online documentation to identify problems that may hinder the user's ability to access and navigate the information.
- Focus on online help and online manuals, the two forms of online documentation that are the most popular today.

The heuristics address the issues discussed earlier in this paper. A step-by-step process for applying the heuristics appears later in this paper.

Criteria Category	**Evaluation Dimensions**			
	Structure	**Presentation**	**Dynamics**	**Content (not the focus)**
1. Orientation	The user: • Knows how to get to all information about a topic. • Knows his/her place within the information structure. • Can determine where a search result is in the document structure.	Primary related topics are easy to distinguish from tangential related topics.	The user: • Can get back to previous location. • Can hold current place while checking other information.	
2. Efficiency	Frequently sought information is located at high levels of the structure. Within headings, the most important words come first. Layered information and branching provide minimalist paths through information.	Related-information links are positioned where they are most easily noticed.	The user can navigate quickly to the last item, first item, any item, and selected item. Commonly performed operations require no more than 2 clicks. The user can easily identify how to access online help or the online manual.	
3. Flexibility		The same information can appear in different views depending on context.	Several paths are provided to the same information: context sensitivity, index with alternative terms, contents that reflects user's place, related-topics linking, maps. Simple and advanced searching are available.	
4. Control	The user can choose how many levels of the structure to view.	The user: • Can customize window size, contents pane size, and text size. • Can move information to keep product in view.	The user: • Can easily annotate information. • Can easily place bookmarks. • Can easily print needed	

			information.	
5. Recognition	Titles can be recognized without the context of surrounding information. Link labels clearly indicate their destination. Search results show how each result differs from the others. Index keywords are easy to recognize and distinguish.	Symbols and color are used to reinforce information structure.	Navigation aids are clearly labeled. Links look like links. Search results show term searched for.	
6. Familiarity	Titles represent real-world concepts and tasks.	Symbols do not conflict with other common uses, such as international symbols.	Navigation takes advantage of people‘s experience with books.	New terms are defined within 1 click at every occurrence.
7. Consistency	Headings at the same level in a section use parallel structure to help users identify the information they cover.	Presentation of information on various pages is consistent so that differences between pages are meaningful.	Navigation follows the same pattern throughout the documentation.	Words have the same meaning from one use to another.
8. Readability and aesthetics		Information is presented in a size and color that is easy to read. Different levels of headings are easy to distinguish. Online documentation does not clash with the design of the product screen. Information is pleasing to the eye.		
9. Context-sensitivity (help only)	Context sensitive topics are logically positioned within the information structure. If the current topic does not present the information needed, the supporting information is no more than 1 or 2 clicks away.	Context-sensitive help windows do not cover up the work area. Related-topic links are easy to identify.	The user can easily identify how to access context-sensitive help.	Content answers user‘s expected questions, whether through task support (suitability) or reference information or conceptual information or other.
10. Clarity	Information pieces can be read in any order, or a guide to the order is provided.			Reused information retains meaning in any context. Meaning holds no matter how the user combines the information pieces.

Appendix 3

Draft USE Questionnaire

The first step in identifying potential items for the questionnaire was to collect a large pool of items to test. The items were collected from previous internal studies, from the literature, and from brainstorming. The list was then massaged to eliminate or reword items that could not be applied across the hardware, software, documentation, and service domains. One goal was to make the items as simply worded as possible, and as general as possible. But they should also have face validity to the people completing the questionnaire. As rounds of testing progressed, standard psychometric techniques were used to weed out additional items that appeared to be too idiosyncratic or to improve items through ongoing tweaking of the wording. In general, the items contributing to each scale were of approximately equal weight, the Chronbach's Alphas were very high, and for the most part the items appeared to tap slightly different aspects of the dimensions being measured. Work does need to continue to ensure each item is tapping something slightly different about a given factor, and a lot of work is required to find reliable negative versions of the items.

The questionnaires were constructed as 7-point Likert rating scales. Users were asked to rate agreement with the statements, raging from strongly disagree to strongly agree. Various forms of the questionnaires were used to evaluate user attitudes towards a variety of consumer products (mostly hardware products), towards internal software applications (e.g., e-mail and time reporting), and towards voice-mail systems. Most of the testing occured using products with which the users had had significant experience, but some testing was conducted when the only experience the user had was during the traditional laboratory usability testing. An example of a prototypical questionnaire is available. The actual questionnaires included additional items begin tested for inclusion, and items intended to measure usage, loyalty, and branding. One interesting side note is that Marketing found the items worked well as part of telephone interviews to understand customer attitudes about products that had been in the field for some time, and this suggested an approach to linking more traditional market research with usability work in the laboratory. The items have also been useful in assessing user perceptions of an ISP and of Web sites.

Factor analyses following each study suggested that users were evaluating the products primarily using three dimensions, Usefulness, Satisfaction, and Ease of Use. Evidence of other dimensions was found, but these three served to most effectively discriminate between interfaces. Partial correlations calculated using scales derived for these dimensions suggested

that Ease of Use and Usefulness influence one another, such that improvements in Ease of Use improve ratings of Usefulness and vice versa. While both drive Satisfaction, however, Usefulness is relatively less important when the systems are internal systems that users are required to use and users are more variable in their Usefulness ratings when they have had only limited exposure to a product. As expected from the literature, Satisfaction was strongly related to the usage (actual or predicted). For internal systems, the items contributing to Ease of Use for other products actually could be separated into two factors, Ease of Learning and Ease of Use (which were obviously highly correlated).

The items that appeared across tests for the 3 factors plus Ease of Learning are listed below. The items in italics loaded relatively less strongly on the factors.

- Usefulness
    1. It helps me be more effective.
    2. It helps me be more productive.
    3. It is useful.
    4. It gives me more control over the activities in my life.
    5. It makes the tings I want to accomplish easier to get done.
    6. *It saves me time when I use it.*
    7. *It meets my needs.*
    8. *It does everything I would expect it to do.*
- Ease of Use
    1. It is easy to use.
    2. It is simple to use.
    3. It is user friendly.
    4. It requires the fewest steps possible to accomplish what I want to do with it.
    5. *It is flexible.*
    6. *Using it is effortless.*
    7. *I can use it without written instructions.*
    8. *I don't notice any inconsistencies as I use it.*
    9. *Both occasional and regular users would like it.*
    10. *I can recover from mistakes quickly and easily.*
    11. *I can use it successfully every time.*
- Ease of Learning
    1. I learned to use it quickly.
    2. I easily remember how to use it.
    3. It is easy to learn to use it.
    4. *I quickly became skillful with it.*
- Satisfaction

1. I am satisfied with it.
2. I would recommend it to a friend.
3. It is fun to use.
4. It works the way I want it to work.
5. It is wonderful.
6. *I feel I need to have it.*
7. *It is pleasant to use.*

Work to refine the items and the scales continues. There is some evidence that for Web sites and certain Consumer products there is an additional dimension of fun or aesthetics associated with making a product compelling.

If you would like to collaborate to evolve an instrument that will be available in the public domain, please contact me. If you try some of these items out as part of your testing, I would be very interested in whatever results you can share.

Appendix 4

An example of a prototypical questionnaire

Instructions:

Based on your experience with similar products, please rate your agreement with the following statements about how you feel in general when using _____.
Just circle or X out the level of agreement that applies (where 1 means strongly disagree, 4 means neither disagree nor agree, and 7 means strongly agree; and NA means it doesn't apply), as in the example.

Strongly Disagree 1---2---3---X---5---6---7 Strongly Agree NA
**********Statements about the product to use for ratings.**********
Both occasional and regular users would like it.
Strongly Disagree 1---2---3---4---5---6---7 Strongly Agree NA

I am satisfied with it.
Strongly Disagree 1---2---3---4---5---6---7 Strongly Agree NA

It is simple to use.
Strongly Disagree 1---2---3---4---5---6---7 Strongly Agree NA

It is wonderful.
Strongly Disagree 1---2---3---4---5---6---7 Strongly Agree NA

It is useful.
Strongly Disagree 1---2---3---4---5---6---7 Strongly Agree NA

I don't notice any inconsistencies as I use it.
Strongly Disagree 1---2---3---4---5---6---7 Strongly Agree NA

It is fun to use.
Strongly Disagree 1---2---3---4---5---6---7 Strongly Agree NA

It does everything I would expect it to do.
Strongly Disagree 1---2---3---4---5---6---7 Strongly Agree NA

It requires the fewest steps possible to accomplish what I want to do with it.
Strongly Disagree 1---2---3---4---5---6---7 Strongly Agree NA

It is user friendly.

Strongly Disagree 1---2---3---4---5---6---7 Strongly Agree NA

It would give me more control over activities in my life.
Strongly Disagree 1---2---3---4---5---6---7 Strongly Agree NA

I feel I need to have it.
Strongly Disagree 1---2---3---4---5---6---7 Strongly Agree NA

I can use it successfully every time.
Strongly Disagree 1---2---3---4---5---6---7 Strongly Agree NA

It is flexible.
Strongly Disagree 1---2---3---4---5---6---7 Strongly Agree NA

I learned to use it quickly.
Strongly Disagree 1---2---3---4---5---6---7 Strongly Agree NA

It would make the things I want to accomplish easier to get done.
Strongly Disagree 1---2---3---4---5---6---7 Strongly Agree NA

I can recover from mistakes quickly and easily.
Strongly Disagree 1---2---3---4---5---6---7 Strongly Agree NA

It would help me be more effective.
Strongly Disagree 1---2---3---4---5---6---7 Strongly Agree NA

Using it is effortless.
Strongly Disagree 1---2---3---4---5---6---7 Strongly Agree NA

I can use it without written instructions.
Strongly Disagree 1---2---3---4---5---6---7 Strongly Agree NA

It meets my needs.
Strongly Disagree 1---2---3---4---5---6---7 Strongly Agree NA

It works the way I want it to work.
Strongly Disagree 1---2---3---4---5---6---7 Strongly Agree NA

It would save me time when I use it.
Strongly Disagree 1---2---3---4---5---6---7 Strongly Agree NA

I easily remember how to use it.
Strongly Disagree 1---2---3---4---5---6---7 Strongly Agree NA

I quickly could become skillful with it.

Strongly Disagree 1---2---3---4---5---6---7 Strongly Agree NA

It is pleasant to use.
Strongly Disagree 1---2---3---4---5---6---7 Strongly Agree NA

It is easy to use.
Strongly Disagree 1---2---3---4---5---6---7 Strongly Agree NA

I would recommend it to a friend.
Strongly Disagree 1---2---3---4---5---6---7 Strongly Agree NA

It would help me be more productive.
Strongly Disagree 1---2---3---4---5---6---7 Strongly Agree NA

It is easy to learn to use it.
Strongly Disagree 1---2---3---4---5---6---7 Strongly Agree NA