Development of Usability and Context-of-Use Taxonomies, Integration with Techniques for the Study of Usability and Application to Real-World Intelligent Systems

Author: David Alonso Ríos

Doctoral thesis, UDC 2015

Supervisors: Eduardo Mosqueira Rey
Vicente Moret Bonillo

Department of Computer Science. Doctorate in Computer Science[1]

UNIVERSIDADE DA CORUÑA

[1] Program regulated by RD 778/1998

Eduardo Mosqueira Rey, Profesor Titular del Departamento de Computación de la Universidad de A Coruña y

Vicente Moret Bonillo, Profesor Titular del Departamento de Computación de la Universidad de A Coruña

CERTIFICAN QUE:

La tesis doctoral titulada "Development of Usability and Context-of-Use Taxonomies, Integration with Techniques for the Study of Usability and Application to Real-World Intelligent Systems" ha sido realizada por D. David Alonso Ríos bajo nuestra dirección en el Departamento de Computación de la Universidad de A Coruña y constituye la Tesis que presenta para optar al Grado de DOCTOR en Informática con Mención Internacional.

En A Coruña a 2 de marzo de 2015.

MOSQUEIRA REY EDUARDO - DNI 32805695J

Firmado digitalmente por MOSQUEIRA REY EDUARDO - DNI 32805695J
Nombre de reconocimiento (DN): c=ES, o=UNIVERSIDAD DE A CORUÑA, ou=certificado electrónico de empleado público, ou=COMPUTACION, ou=210, serialNumber=32805695J, sn=MOSQUEIRA REY, givenName=EDUARDO, cn=MOSQUEIRA REY EDUARDO - DNI 32805695J
Fecha: 2015.03.31 12:26:24 +02'00'

Fdo. Eduardo Mosqueira Rey

NOMBRE MORET BONILLO VICENTE - NIF 32746092A

Firmado digitalmente por NOMBRE MORET BONILLO VICENTE - NIF 32746092A
Nombre de reconocimiento (DN): c=ES, o=FNMT, ou=FNMT Clase 2 CA, ou=500170212, cn=NOMBRE MORET BONILLO VICENTE - NIF 32746092A
Fecha: 2015.03.31 13:34:11 +02'00'

Fdo. Vicente Moret Bonillo

To my mother, who encouraged me to do this.

And in memory of my father, who never saw its completion.

Acknowledgments

I would like to thank all the people who have offered me their support and advice throughout all these years.

First of all, I would like to thank my PhD supervisors, Eduardo and Vicente, without whom this thesis would not have been possible.

I would also like to thank the other co-authors of the research papers I wrote for this thesis, namely, Ana and the staff at Vienna's Institut für Computertechnik (Dr. Raneburger, Dr. Popp, Dr. Falb, and Prof. Dr. Kaindl – the latter of whom also arranged my research stay in Vienna). And also all the people I collaborated with in research projects, especially the members of the Galileo Speed Warning project. Anyone involved in the field of academic publishing, such as editors, reviewers, and so on, was also indispensable. As were all the authors I cited.

I would also like to thank other people who gave me the initial impetus to pursue a PhD: my mother – who suggested the possibility to me before I had even finished my degree – and the people I was working for or working with as a computer programmer when I joined my PhD program, especially Nieves, Mon, and Santi.

I am also indebted to the people who, at certain points, helped me secure the economic stability I needed in order to carry out this research freely, especially Amparo and Bertha.

I would also like to thank all the other members of the LIDIA group, both present and past. Instead of giving a long list of names I would simply like to thank them all, especially those who were involved with usability in one way or another: Elena, Edu, and Diego.

Of course, I would also like to thank all my family and friends.

However, I feel no gratitude towards those who are hellbent on destroying Spain's higher education system. Not that they would care anyway. This work exists *in spite* of them.

Abstract

A major obstacle to the implantation of User-Centered Design in the real world is the fact that no precise definition of the concept of usability exists that is widely accepted and applied in practice. Generally speaking, the literature tends to define usability in overly brief and ambiguous terms and to describe its application in informal terms. The same criticisms can be leveled at the concept of context of use, to which usability is always relative. As a consequence of these drawbacks, ad hoc techniques predominate in usability study methodologies. This thesis proposes detailed taxonomies for the concepts of usability and the context of use. The taxonomies are organized hierarchically and contain precise descriptions of their attributes and subattributes. In order to illustrate the practical usefulness of the taxonomies, this thesis describes and discusses how the taxonomies were integrated into the development life cycle of two real-world projects in the field of Intelligent Systems, namely, an Intelligent Speed Adaptation device and an automatic generator of user interfaces. At a specific point of each project, a usability study was conducted, in which the taxonomies were used to structure and guide well-known usability activities such as usability requirements analysis, heuristic evaluation, and subjective analysis.

Keywords: Usability, Context of Use, Taxonomies, Heuristic Evaluation, Usability Testing, User Questionnaires, User Interfaces, Intelligent Systems.

Resumen

Un importante obstáculo para la implantación del Diseño Centrado en el Usuario en el mundo real es que no existe una definición del concepto de usabilidad que sea precisa y comúnmente aceptada y aplicada en la práctica. La literatura tiende a definir la usabilidad en términos demasiado escuetos y ambiguos, y a describir su aplicación en términos informales. Idénticas críticas pueden hacerse al concepto del contexto de uso, al cual la usabilidad es siempre relativa. Como consecuencia de esto, las técnicas ad hoc predominan en las metodologías de usabilidad. Esta tesis propone taxonomías detalladas para los conceptos de usabilidad y contexto de uso. Las taxonomías están organizadas jerárquicamente y contienen descripciones precisas de sus atributos y subatributos. Para ilustrar la utilidad práctica de las taxonomías, se describe cómo fueron integradas en el ciclo de vida de desarrollo de dos productos reales en el campo de los Sistemas Inteligentes. Concretamente, un dispositivo de Adaptación Inteligente de la Velocidad y un generador automático de interfaces de usuario. En un punto específico de cada proyecto se realizó un estudio de usabilidad, usando las taxonomías para estructurar y guiar actividades de usabilidad como el análisis de los requisitos de usabilidad, la evaluación heurística y el análisis subjetivo.

Palabras clave: Usabilidad, Contexto de uso, Taxonomías, Evaluación heurística, Pruebas de usabilidad, Cuestionarios de usuario, Interfaces de usuario, Sistemas Inteligentes.

Resumo

Un importante obstáculo para a implantación do Deseño Centrado no Usuario no mundo real é que non existe unha definición do concepto de usabilidade que sexa precisa e comunmente aceptada e aplicada na práctica. A literatura tende a definir a usabilidade en termos demasiado concisos e ambiguos, e a describir a súa aplicación en termos informais. Idénticas críticas poden facerse ao concepto do contexto de uso, ao cal a usabilidade é sempre relativa. Como consecuencia disto, as técnicas ad hoc predominan nas metodoloxías de usabilidade. Esta tese propón taxonomías detalladas para os conceptos de usabilidade e contexto de uso. As taxonomías están organizadas xerárquicamente e conteñen descricións precisas dos seus atributos e subatributos. Para ilustrar a utilidade práctica das taxonomías, descríbese como foron integradas no ciclo de vida de desenvolvemento de dous produtos reais no campo dos Sistemas Intelixentes. Concretamente, un dispositivo de Adaptación Intelixente da Velocidade e un xerador automático de interfaces de usuario. Nun punto específico de cada proxecto realizouse un estudo de usabilidade, usando as taxonomías para estruturar e guiar actividades de usabilidade como a análise dos requisitos de usabilidade, a avaliación heurística e a análise subxectiva.

Palabras clave: Usabilidade, Contexto de uso, Taxonomías, Avaliación heurística, Probas de usabilidade, Cuestionarios de usuario, Interfaces de usuario, Sistemas Intelixentes.

Summary in Spanish / Resumen en castellano

La Ingeniería de la Usabilidad es una disciplina relativamente reciente pero establecida. El interés por la misma se ve justificado por el hecho de que la usabilidad de un producto afecta a diferentes aspectos: la calidad del producto, la utilización del mismo, la productividad alcanzable o el correcto desarrollo del flujo de trabajo. La usabilidad es un aspecto a cuidar a lo largo de todo el ciclo de vida de un producto, y eso incluye las etapas de desarrollo, mantenimiento, comercialización y soporte.

La usabilidad es una cualidad deseable pero al mismo tiempo compleja y difícil de garantizar. Existen diversos motivos para ello: es difícil de definir, medir y evaluar; económicamente hablando, es infravalorada en relación a otros aspectos más "prioritarios" y "visibles"; en la práctica, tiende a quedar finalmente un tanto desconectada del resto del ciclo de vida de un proyecto.

En el ámbito académico existe ya una abundante literatura sobre usabilidad, de carácter tanto teórico como práctico. En este sentido, los campos de investigación más relevantes para esta tesis son los relacionados, por un lado, con la definición y el modelado del concepto de *usabilidad* (y el concepto asociado del *contexto de uso*) y, por el otro, con la metodología de los *estudios de usabilidad* en el mundo real.

En el ámbito práctico existen numerosas *técnicas* (o métodos) para estudiar, analizar y evaluar la usabilidad de un producto. Una forma tradicional de integrar dichas técnicas en el ciclo de vida de desarrollo del producto es a través del diseño iterativo. De esta forma, los aspectos de usabilidad (y los planes sobre cómo mejorarlos) son identificados a lo largo de sucesivas iteraciones. Por ejemplo, al comienzo de un proyecto los expertos en usabilidad pueden reunir información y requisitos de usabilidad mediante entrevistas, listas de verificación, etc. Esto puede dar lugar a un prototipo inicial de un producto, el cual puede ser probado por usuarios representativos en un contexto de laboratorio. Como resultado, se obtendrían nuevos datos de usabilidad que servirían de entrada al diseño de una versión del producto totalmente funcional que pueda ser probada por usuarios finales.

Las técnicas de usabilidad existentes son de muy diversos tipos y pueden ser clasificadas de la siguiente forma [13]:

1. **Técnicas expertas** (o heurísticas), que son llevadas a cabo por profesionales que identifican los puntos fuertes y débiles del producto con respecto a la usabilidad (por ejemplo, listas de verificación y evaluaciones heurísticas).

2. **Técnicas subjetivas**, que obtienen opiniones personales de los usuarios sobre la usabilidad del producto (por ejemplo, cuestionarios y entrevistas).

3. **Técnicas empíricas**, que obtienen datos objetivos acerca de lo bien que utilizan realmente el sistema los usuarios (por ejemplo, medidas de tiempo y de número de errores).

Es quizá en el ámbito teórico donde podemos encontrar los mayores obstáculos. A día de hoy, no puede decirse que exista un consenso real respecto al significado exacto del término *usabilidad*. Existen, sin embargo, ciertos modelos de usabilidad, como el

estándar ISO 9241-11:1998, que son mayoritariamente considerados como referencia. ISO 9241-11 define la usabilidad como el grado hasta el cual un producto puede ser utilizado por usuarios específicos para alcanzar metas específicas con *efectividad*, *eficiencia* y *satisfacción*, dentro de un contexto de uso específico [87]. ISO-9241-11 también proporciona definiciones individuales para los atributos que componen la usabilidad:

1. **Efectividad**. La precisión y completitud con las cuales los usuarios alcanzan metas específicas.

2. **Eficiencia**. Los recursos invertidos en relación con la precisión y completitud con las cuales los usuarios alcanzan metas específicas.

3. **Satisfacción**. Ausencia de incomodidades y actitudes positivas hacia el uso de un producto.

ISO-9241-11 no continúa definiendo la usabilidad más allá de este punto, sino que describe el concepto en más detalle de un modo informal, ofreciendo una lista parcial de ejemplos.

Otros importantes modelos de usabilidad que pueden encontrarse en la literatura, y que han servido de referencia para el trabajo realizado en esta tesis, son los de ISO/IEC 9126-1:2001 [90], Nielsen [129], Preece et al. [146], Quesenbery [147], Abran et al. [11] y Seffah et al. [167].

En la práctica, los ingenieros de usabilidad tienden a encontrar la mayoría de los modelos de usabilidad de la literatura demasiado limitados para resultar útiles a la hora de identificar problemas de usabilidad más específicos. Estas limitaciones, unidas a la falta de consenso entre los diferentes modelos de usabilidad existente, llevan a muchos investigadores a construir sus propios modelos de usabilidad, generalmente de un modo ad hoc y enfocados hacia un campo de aplicación específico.

Esta tesis propone una taxonomía de usabilidad genérica y aplicable a cualquier tipo de sistema y que tiene como propósito sintetizar y refinar los principales modelos de usabilidad de la literatura con el objetivo de construir un nuevo modelo de usabilidad detallado y preciso. Dentro de la literatura, dicha taxonomía encaja dentro de una rama llamada “modelos expandidos de usabilidad”[111].

Más concretamente, los objetivos a la hora de construir la taxonomía de usabilidad fueron:

- Hacer que la taxonomía fuese lo más completa posible, cubriendo todos los diversos atributos incluidos en los principales modelos de la literatura, pero evitando contradicciones y repeticiones.

- Seguir fielmente los modelos de usabilidad de la literatura, dando prioridad a a una síntesis y mejora de lo existente en vez de a una reinterpretación radical del concepto de la usabilidad.

- Estructurar la taxonomía jerárquicamente en varios niveles de detalle, yendo de lo general a lo particular (los modelos de usabilidad de la literatura tienden a constar de un único nivel).

- Plantear la taxonomía como un modelo de usabilidad de propósito general, aplicable a cualquier tipo de producto (a diferencia de la mayoría de los modelos de la literatura, que están pensados sólo para sistemas informáticos).

- Proporcionar definiciones para cada uno de los atributos y subatributos de la taxonomía, que es algo de lo que generalmente carecen los modelos de usabilidad de la literatura.

La taxonomía de usabilidad propuesta [16] consta de los siguientes atributos:

1. **Cognoscibilidad**, subdividida en *claridad*, *consistencia*, *rememorabilidad* y *ayuda*. Cada uno de estos cuatro subatributos está a su vez dividido en más subniveles de atributos.

2. **Operatividad**, subdividida en *completitud*, *precisión*, *universalidad* y *flexibilidad*, que a su vez se dividen en más subatributos.

3. **Eficiencia**, con subatributos agrupados en las siguientes categorías: *eficiencia en cuanto a esfuerzo humano*, *eficiencia en cuanto a tiempo de ejecución de tareas*, *eficiencia en cuanto a recursos ocupados* y *eficiencia en cuanto a costes económicos.*

4. **Robustez**, subdividida en *robustez ante errores internos*, *robustez ante usos indebidos*, *robustez ante abusos por terceros* y *robustez ante problemas en el entorno.*

5. **Seguridad**, con subatributos agrupados en las categorías de *seguridad del usuario*, *seguridad de terceros* y *seguridad del entorno.*

6. **Satisfacción subjetiva**, con subatributos agrupados en las categorías de *interés* y *estética.*

Esta tesis propone asimismo una nueva taxonomía para el concepto de contexto de uso, con objetivos análogos a los descritos para la taxonomía de usabilidad. Como señala la definición de usabilidad del ISO 9241-11 citada anteriormente, la usabilidad de un producto no es una cualidad intrínseca al mismo, sino relativa a las características del contexto en el cual es utilizado. ISO 9241-11 define el contexto de uso en base a las características de los *usuarios*, las *tareas*, el *equipo* y el *entorno* [87].

Nuevamente, la literatura sobre el término es analizada con el objetivo de construir un modelo más detallado y que sintetice definiciones diversas evitando contradicciones y repeticiones. La taxonomía del contexto de uso propuesta [17] consta de los siguientes atributos:

1. **Usuario**, subdividido en *rol*, *experiencia*, *formación*, *actitud hacia el sistema*, *características físicas* y *características cognitivas*. Cada uno de estos seis subatributos está a su vez dividido en más subniveles de atributos.

2. **Tarea**, subdividida en los atributos de *elección del uso del sistema*, *complejidad*, *características temporales*, *exigencias*, *controlabilidad del flujo de trabajo*, *seguridad* y *criticidad*, que a su vez pueden estar subdivididos en más atributos.

3. **Entorno**, subdividido en *entorno físico*, *entorno social* y *entorno técnico*, a su vez subdivididos en varios niveles jerárquicos de subatributos.

Otra carencia significativa de la literatura sobre la usabilidad es la relativa escasez de publicaciones que describan en detalle estudios de usabilidad reales. El principal motivo para ello es que los resultados de usabilidad suelen mantenerse en secreto, ya que las compañías prefieren no divulgar abiertamente los defectos de sus productos. Eso lleva a que los artículos científicos sobre usabilidad aplicada tiendan a centrarse más en cuestiones metodológicas que en resultados concretos [111].

Por el contrario, esta tesis complementa las cuestiones teóricas y metodológicas con una descripción detallada de los resultados y las lecciones aprendidas al llevar a cabo dos estudios de usabilidad reales asociados a proyectos de investigación internacionales centrados en el área de Sistemas Inteligentes.

El primer estudio de usabilidad descrito en esta tesis tuvo lugar dentro del proyecto de investigación Galileo Speed Warning [2]. Dicho proyecto involucró a compañías como Mapflow Ltd. y la compañía tecnológica holandesa Technolution B.V. El objetivo del proyecto era buscar nuevas soluciones al control automatizado de los límites de velocidad en vehículos. La novedad de la propuesta radica en ofrecer recompensas a los usuarios en vez de aplicar penalizaciones. Para ello, se instala en los vehículos de los usuarios un dispositivo de Adaptación Inteligente de la Velocidad llamado CARAT counter. Si se respetan adecuadamente los límites de velocidad, el CARAT counter acumula puntos que pueden ser posteriormente canjeados por el usuario. El estudio de usabilidad del CARAT counter está planteado en base a las taxonomías de usabilidad y del contexto de uso y comprende diversas actividades de la Ingeniería de la Usabilidad que proporcionan perspectivas complementarias sobre el producto: análisis de requisitos de usabilidad, evaluación heurística y análisis subjetivo (cuestionarios de usuario) [14].

El segundo estudio de usabilidad fue realizado durante una estancia en la Universidad Técnica de Viena (Austria) y en él se evalúa comparativamente la usabilidad de diferentes interfaces generados automáticamente. Más concretamente, se analizan diseños alternativos de páginas web que han sido creados con un generador automatizado y multidispositivo desarrollado por investigadores del Institut für Computertechnik de Viena. Dicha herramienta generadora crea interfaces multiplataforma (en el caso de este estudio de usabilidad, se trata de páginas web codificadas en HTML, CSS y Javascript) en base a un modelo abstracto de la interacción con el usuario. La herramienta utiliza técnicas propias de la Inteligencia Artificial –como las búsquedas en espacios de soluciones– para decidir la mejor forma de distribuir los elementos de la interfaz (por ejemplo, para evitar el *scrolling* en dispositivos pequeños). Nuevamente, esta tesis describe en detalle el estudio de usabilidad correspondiente, que consiste en evaluar la usabilidad de las diferentes distribuciones obtenidas (con *scroll* horizontal, con *scroll* vertical, con pestañas) para varias páginas web de idéntico contenido. Las actividades de la Ingeniería de la Usabilidad llevadas a cabo fueron evaluación heurística, medidas de rendimiento y análisis subjetivo (cuestionarios y entrevistas) [15][153][154].

Ambos estudios de usabilidad están estructurados de forma similar, combinando técnicas heurísticas y subjetivas y, en el caso del segundo estudio, también empíricas. Antes de comenzar con el estudio de usabilidad, es necesario primero documentar las características del sistema que va a ser examinado. En lo que a esta tesis respecta, ello implica dividir el sistema en las partes que lo componen e identificar los atributos de las mismas que sean relevantes para el estudio de la usabilidad, obteniendo como resultado una lista de aspectos a ser evaluados posteriormente empleando la taxonomía de usabilidad. Otro paso previo a la realización de los estudios de usabilidad es caracterizar también el contexto de uso, tarea que es realizada basándose directamente en la taxonomía del contexto de uso propuesta en esta tesis.

Pero no sólo la usabilidad y el contexto de uso pueden ser conceptualizados en taxonomías genéricas. Aunque cada sistema a analizar posee características propias que lo distinguen de los demás, algunos elementos de la interfaz [69] y algunas tareas [41] son lo bastante comunes como para ser generalizados a diferentes sistemas. Esta tesis muestra mediante ejemplos –correspondientes a los estudios de usabilidad descritos en ella– cómo puede realizarse esto, aunque la elaboración de taxonomías genéricas y completas para describir los elementos de las interfaces y las tareas quedan fuera de los objetivos de esta tesis.

Desde un punto de vista metodológico, el aspecto más destacable de los dos estudios de usabilidad descritos no radica en las técnicas utilizadas –que son comúnmente aplicadas en muchos otros estudios– sino en cómo las taxonomías han sido integradas dentro de dichas técnicas. Aunque cualquier estudio de usabilidad está inherentemente basado en algún modelo de usabilidad, ya sea un estándar de la literatura o un modelo ad hoc, dichos modelos tienden a ser poco detallados, lo que lleva en la práctica a depender demasiado de guías de usabilidad o del "sentido común" para suplir las carencias de los modelos. El nivel de detalle de las taxonomías tiene como propósito facilitar el estudio de la usabilidad en este sentido.

En el estudio de usabilidad del CARAT counter, las taxonomías facilitaron la identificación de problemas en las siguientes categorías: legibilidad y consistencia de los elementos de la pantalla, intuitividad en los mensajes de error y las tareas, universalidad respecto a las unidades de medida de la velocidad y el idioma de la interfaz, precisión del GPS, seguridad del usuario, robustez respecto a abusos y a errores internos y, por último, el atractivo estético.

Por su parte, el estudio de usabilidad de las interfaces generadas automáticamente identificó problemas en aspectos como: claridad (de significado, de funcionamiento y de estructura), consistencia (de terminología, de aspecto), rememorabilidad, completitud, accesibilidad, universalidad cultural (de lenguaje, de formatos de fechas), flexibilidad, eficiencia, robustez ante el uso inadecuado, confidencialidad y estética.

Aunque ambos estudios de usabilidad están estructurados de modo similar, existen diferencias significativas que merecen ser descritas. En primer lugar, mientras que el CARAT counter consiste en una combinación muy específica de hardware y software, las interfaces generadas automáticamente eran páginas web multiplataforma visualizadas en dispositivos móviles de propósito general (smartphones y tablets). En segundo lugar, mientras que el estudio de usabilidad del CARAT counter analizaba la usabilidad del mismo en términos "absolutos", el estudio de las interfaces generadas automáticamente fue de naturaleza comparativa, examinando distribuciones alternativas de los mismos elementos de una misma aplicación informática. En tercer lugar, el estudio de la usabilidad de las interfaces generadas automáticamente estaba centrado en un subconjunto reducido de atributos de usabilidad, ya que las interfaces eran idénticas en todos los sentidos excepto en lo relativo a la disposición de sus elementos. En cuarto lugar, este último estudio de usabilidad obtuvo no sólo datos cualitativos (por ejemplo, sobre las opiniones de los usuarios) sino también datos cuantitativos, incluyendo medidas de rendimiento objetivas (tiempo, número de errores) y estadísticamente significativas, que fueron obtenidas a partir de anotar grabaciones en vídeo de 60 pruebas realizadas con usuarios.

Ambos estudios de usabilidad están centrados en las llamadas *metas formativas*, que buscan detectar problemas de usabilidad y proponer soluciones. Mientras que el nivel de detalle de las taxonomías propuestas en esta tesis ha ayudado en la consecución de estas metas (por ejemplo, en las evaluaciones heurísticas de los dos estudios de usabilidad), existen otros tipos de metas en la Ingeniería de la Usabilidad. Las llamadas *metas sumativas* tienen como objetivo obtener evaluaciones generales, no específicas. Por ejemplo, puntuaciones numéricas. La cuestión de la adecuación de la taxonomía de usabilidad a las metas sumativas ha quedado fuera de los objetivos de esta tesis. Éste es un tema complejo, ya que por el momento los investigadores no han llegado a un consenso ni siquiera respecto a si el modelo de usabilidad del ISO 9241-11 puede ser utilizado para metas sumativas; más concretamente, si tiene sentido agregar los clásicos atributos de *efectividad*, *eficiencia* y *satisfacción* en una única puntuación de usabilidad.

Otra parte importante de esta tesis está dedicada a investigar la validez de los supuestos teóricos en los que se basa la misma. Más concretamente, se compara el enfoque propuesto (es decir, uno basado en taxonomías detalladas y de propósito general) con su opuesto (es decir, modelos de usabilidad ad hoc, no detallados, y creados ex profeso para un campo de aplicación específico). Lógicamente, no tiene por qué existir un modelo ad hoc para cada campo de aplicación que podamos considerar pero, afortunadamente, existe uno para el campo de sistemas de información situados en el interior de vehículos [72], lo que permitió establecer una comparación basada en los resultados obtenidos previamente para el CARAT counter. Dicha comparación no tiene más objetivo que establecer las bases para una discusión más formal de la expresividad de los modelos de usabilidad y, más concretamente, de los problemas ocasionados por la falta de expresividad en este sentido. Si estudiamos un nuevo modelo de usabilidad y lo comparamos con un modelo de referencia como puede ser un estándar de la literatura, podemos clasificar los posibles problemas de expresividad del nuevo modelo de la siguiente forma [22]:

1. **Sobrecarga de constructos**. Un concepto del nuevo modelo corresponde a más de un concepto del modelo de referencia.
2. **Exceso de constructos**. Un concepto del nuevo modelo no corresponde a ningún concepto del modelo de referencia.
3. **Redundancia de constructos**. Un concepto del modelo de referencia corresponde a más de un concepto del nuevo modelo.
4. **Incompletitud de constructos**. Hay un concepto en el modelo de referencia que no corresponde a ningún concepto del nuevo modelo.

Esta clasificación permite analizar un modelo de usabilidad –o comparar dos modelos– y detectar problemas de ambigüedad, completitud o redundancia. Adicionalmente, también puede argumentarse que modelos creados ex profeso para un campo de aplicación específico carecen de la propiedad de ser generalizados a otras áreas, la cual suele ser una propiedad deseable en el ámbito de la Ingeniería.

Por último, las contribuciones más relevantes de esta tesis pueden ser resumidas de la siguiente manera:

- Se analizan las diferentes formas en que la usabilidad y el contexto de uso han sido definidos en la literatura y se investigan los problemas y las consecuencias de esta multiplicidad de definiciones.
- Se proponen taxonomías genéricas, detalladas y jerárquicamente organizadas para clasificar los atributos que componen, por un lado, la usabilidad y, por el otro, el contexto de uso.
- Se propone un enfoque para llevar a cabo estudios de usabilidad basados en la integración de dichas taxonomías en el ciclo de vida de desarrollo de un producto.
- Se documenta detalladamente cómo las taxonomías han sido integradas en la práctica en sendos estudios de usabilidad que han tenido lugar dentro de proyectos de investigación internacionales.
- Se presenta un marco teórico de discusión de la validez y la expresividad de un modelo de usabilidad dado basándose en establecer una comparación con un modelo de referencia.

Summary in Galician / Resumo en galego

A Enxeñaría da Usabilidade é unha disciplina relativamente recente pero establecida. O interese pola mesma vese xustificado polo feito de que a usabilidade dun produto afecta a diferentes aspectos: a calidade do produto, a utilización do mesmo, a produtividade alcanzable ou o correcto desenvolvemento do fluxo de traballo. A usabilidade é un aspecto a coidar ao longo de todo o ciclo de vida dun produto, e iso inclúe as etapas de desenvolvemento, mantemento, comercialización e soporte.

A usabilidade é unha calidade desexable pero ao mesmo tempo complexa e difícil de garantir. Existen diversos motivos para iso: é difícil de definir, medir e avaliar; economicamente falando, é infravalorada en relación a outros aspectos máis "prioritarios" e "visibles"; na práctica, tende a quedar finalmente un tanto desconectada do resto do ciclo de vida dun proxecto.

No ámbito académico existe xa unha abundante literatura sobre usabilidade, de carácter tanto teórico como práctico. Neste sentido, os campos de investigación máis relevantes para esta tese son os relacionados, por unha banda, coa definición e o modelado do concepto de *usabilidade* (e o concepto asociado do *contexto de uso*) e, polo outro, coa metodoloxía dos *estudos de usabilidade* no mundo real.

No ámbito práctico existen numerosas *técnicas* (ou métodos) para estudar, analizar e avaliar a usabilidade dun produto. Unha forma tradicional de integrar as devanditas técnicas no ciclo de vida de desenvolvemento do produto é a través do deseño iterativo. Desta forma, os aspectos de usabilidade (e os plans sobre como melloralos) son identificados ao longo de sucesivas iteracións. Por exemplo, ao comezo dun proxecto os expertos en usabilidade poden reunir información e requisitos de usabilidade mediante entrevistas, listas de verificación, etc. Isto pode dar lugar a un prototipo inicial dun produto, o cal pode ser probado por usuarios representativos nun contexto de laboratorio. Como resultado, obteríanse novos datos de usabilidade que servirían de entrada ao deseño dunha versión do produto totalmente funcional que poida ser probada por usuarios finais.

As técnicas de usabilidade existentes son de moi diversos tipos e poden ser clasificadas da seguinte forma [13]:

1. **Técnicas expertas** (ou heurísticas), que son levadas a cabo por profesionais que identifican os puntos fortes e débiles do produto con respecto á usabilidade (por exemplo, listas de verificación e avaliacións heurísticas).

2. **Técnicas subxectivas**, que obteñen opinións persoais dos usuarios sobre a usabilidade do produto (por exemplo, cuestionarios e entrevistas).

3. **Técnicas empíricas**, que obteñen datos obxectivos acerca do ben que utilizan realmente o sistema os usuarios (por exemplo, medidas de tempo e de número de erros).

É quizais no ámbito teórico onde podemos atopar os maiores obstáculos. A día de hoxe, non pode dicirse que exista un consenso real respecto ao significado exacto do termo *usabilidade*. Existen, con todo, certos modelos de usabilidade, como o estándar ISO 9241-11:1998, que son maioritariamente considerados como referencia. ISO 9241-11 define a usabilidade como o grao até o cal un produto pode ser utilizado por usuarios específicos para alcanzar metas específicas con *efectividade*, *eficiencia* e *satisfacción*, dentro dun contexto de uso específico [87]. ISO-9241-11 tamén proporciona definicións individuais para os atributos que compoñen a usabilidade:

1. **Efectividade**. A precisión e completitude coas cales os usuarios alcanzan metas específicas.

2. **Eficiencia**. Os recursos investidos en relación coa precisión e completitude coas cales os usuarios alcanzan metas específicas.

3. **Satisfacción**. Ausencia de incomodidades e actitudes positivas cara ao uso dun produto.

ISO-9241-11 non continúa definindo a usabilidade máis aló deste punto, senón que describe o concepto en máis detalle dun modo informal, ofrecendo unha lista parcial de exemplos.

Outros importantes modelos de usabilidade que poden atoparse na literatura, e que serviron de referencia para o traballo realizado nesta tese, son os de ISO/IEC 9126-1:2001 [90], Nielsen [129], Preece et al. [146], Quesenbery [147], Abran et al. [11] e Seffah et al. [167].

Na práctica, os enxeñeiros de usabilidade tenden a atopar a maioría dos modelos de usabilidade da literatura demasiado limitados para resultar útiles á hora de identificar problemas de usabilidade máis específicos. Estas limitacións, unidas á falta de consenso entre os diferentes modelos de usabilidade existente, levan a moitos investigadores a construír os seus propios modelos de usabilidade, xeralmente dun modo ad hoc e enfocados cara a un campo de aplicación específico.

Esta tese propón unha taxonomía de usabilidade xenérica e aplicable a calquera tipo de sistema e que ten como propósito sintetizar e refinar os principais modelos de usabilidade da literatura co obxectivo de construír un novo modelo de usabilidade detallado e preciso. Dentro da literatura, a devandita taxonomía encaixa dentro dunha rama chamada "modelos expandidos de usabilidade"[111].

Máis concretamente, os obxectivos á hora de construír a taxonomía de usabilidade foron:

- Facer que a taxonomía fose o máis completa posible, cubrindo todos os diversos atributos incluídos nos principais modelos da literatura, pero evitando contradicións e repeticións.

- Seguir fielmente os modelos de usabilidade da literatura, dando prioridade á unha síntese e mellora do existente no canto de a unha reinterpretación radical do concepto da usabilidade.

- Estruturar a taxonomía xerárquicamente en varios niveis de detalle, indo do xeral ao particular (os modelos de usabilidade da literatura tenden a constar dun único nivel).

- Formular a taxonomía como un modelo de usabilidade de propósito xeral, aplicable a calquera tipo de produto (a diferenza da maioría dos modelos da literatura, que están pensados só para sistemas informáticos).
- Proporcionar definicións para cada un dos atributos e subatributos da taxonomía, que é algo do que xeralmente carecen os modelos de usabilidade da literatura.

A taxonomía de usabilidade proposta [16] consta dos seguintes atributos:

1. **Coñecibilidade**, subdividida en *claridade*, *consistencia*, *memorabilidade* e *axuda*. Cada un destes catro subatributos está á súa vez dividido en máis subniveis de atributos.
2. **Operatividade**, subdividida en *completitude*, *precisión*, *universalidade* e *flexibilidade*, que á súa vez se dividen en máis subatributos.
3. **Eficiencia**, con subatributos agrupados nas seguintes categorías: *eficiencia en canto ao esforzo humano*, *eficiencia en canto ao tempo de execución de tarefas*, *eficiencia en canto a recursos ocupados* e *eficiencia en canto a custos económicos*.
4. **Robustez**, subdividida en *robustez ante erros internos*, *robustez ante usos indebidos*, *robustez ante abusos por terceiros* e *robustez ante problemas na contorna*.
5. **Seguridade**, con subatributos agrupados nas categorías de *seguridade do usuario*, *seguridade de terceiros* e *seguridade da contorna*.
6. **Satisfacción subxectiva**, con subatributos agrupados nas categorías de *interese* e *estética*.

Esta tese propón así mesmo unha nova taxonomía para o concepto de contexto de uso, con obxectivos análogos aos descritos para a taxonomía de usabilidade. Como sinala a definición de usabilidade do ISO 9241-11 citada anteriormente, a usabilidade dun produto non é unha calidade intrínseca ao mesmo, senón relativa ás características do contexto no cal é utilizado. ISO 9241-11 define o contexto de uso en base ás características dos *usuarios*, as *tarefas*, o *equipo* e a *contorna* [87].

Novamente, a literatura sobre o termo é analizada co obxectivo de construír un modelo máis detallado e que sintetice definicións diversas evitando contradicións e repeticións. A taxonomía do contexto de uso proposta [17] consta dos seguintes atributos:

1. **Usuario**, subdividido en *rol*, *experiencia*, *formación*, *actitude cara ao sistema*, *características físicas* e *características cognitivas*. Cada un deste seis subatributos está á súa vez dividido en máis subniveis de atributos.

2. **Tarefa**, subdividida nos atributos de *elección do uso do sistema*, *complexidade*, *características temporais*, *esixencias*, *controlabilidade do fluxo de traballo*, *seguridade* e *criticidade*, que á súa vez poden estar subdivididos en máis atributos.

3. **Contorna**, subdividida en *contorna física*, *contorna social* e *contorna técnica*, á súa vez subdivididas en varios niveis xerárquicos de subatributos.

Outra carencia significativa da literatura sobre a usabilidade é a relativa escaseza de publicacións que describan en detalle estudos de usabilidade reais. O principal motivo para iso é que os resultados de usabilidade adoitan manterse en segredo, xa que as compañías prefiren non divulgar abertamente os defectos dos seus produtos. Iso leva a que os artigos científicos sobre usabilidade aplicada tendan a centrarse máis en cuestións metodolóxicas que en resultados concretos [111].

Pola contra, esta tese complementa as cuestións teóricas e metodolóxicas cunha descrición detallada dos resultados e as leccións aprendidas ao levar a cabo dous estudos de usabilidade reais asociados a proxectos de investigación internacionais centrados na área de Sistemas Intelixentes.

O primeiro estudo de usabilidade descrito nesta tese tivo lugar dentro do proxecto de investigación Galileo Speed Warning [2]. O devandito proxecto involucrou a compañías como Mapflow Ltd. e a compañía tecnolóxica holandesa Technolution B.V. O obxectivo do proxecto era buscar novas solucións ao control automatizado dos límites de velocidade en vehículos. A novidade da proposta radica en ofrecer recompensas aos usuarios no canto de aplicar penalizacións. Para iso, instálase nos vehículos dos usuarios un dispositivo de Adaptación Intelixente da Velocidade chamado CARAT counter. Se respéctanse axeitadamente os límites de velocidade, o CARAT counter acumula puntos que poden ser posteriormente trocados polo usuario. O estudo de usabilidade do CARAT counter está formulado en base ás taxonomías de usabilidade e do contexto de uso e comprende diversas actividades da Enxeñaría da Usabilidade que proporcionan perspectivas complementarias sobre o produto: análise de requisitos de usabilidade, avaliación heurística e análise subxectiva (cuestionarios de usuario) [14].

O segundo estudo de usabilidade foi realizado durante unha estancia na Universidade Técnica de Viena (Austria) e nel avalíase comparativamente a usabilidade de diferentes interfaces xerados automaticamente. Máis concretamente, analízanse deseños alternativos de páxinas web que foron creados cun xerador automatizado e multidispositivo desenvolvido por investigadores do Institut für Computertechnik de Viena. A devandita ferramenta xeradora crea interfaces multiplataforma (no caso deste estudo de usabilidade, trátase de páxinas web codificadas en HTML, CSS e Javascript) en base a un modelo abstracto da interacción co usuario. A ferramenta utiliza técnicas propias da Intelixencia Artificial –como as procuras en espazos de solucións– para decidir a mellor forma de distribuír os elementos da interfaz (por exemplo, para evitar o *scrolling* en dispositivos pequenos). Novamente, esta tese describe en detalle o estudo de usabilidade correspondente, que consiste en avaliar a usabilidade das diferentes distribucións obtidas (con *scroll* horizontal, con *scroll* vertical, con pestanas) para varias páxinas web de idéntico contido. As actividades da Enxeñaría da Usabilidade levadas a cabo foron avaliación heurística, medidas de rendemento e análise subxectiva (cuestionarios e entrevistas) [15][153][154].

Ambos estudos de usabilidade están estruturados de forma similar, combinando técnicas heurísticas e subxectivas e, no caso do segundo estudo, tamén empíricas. Antes de comezar co estudo de usabilidade, é necesario primeiro documentar as características do sistema que vai ser examinado. No que a esta tese respecta, iso implica dividir o sistema nas partes que o compoñen e identificar os atributos das mesmas que sexan relevantes para o estudo da usabilidade, obtendo como resultado unha lista de aspectos a ser avaliados posteriormente empregando a taxonomía de usabilidade. Outro paso previo á realización dos estudos de usabilidade é caracterizar tamén o contexto de uso, tarefa que é realizada baseándose directamente na taxonomía do contexto de uso proposta nesta tese.

Pero non só a usabilidade e o contexto de uso poden ser conceptualizados en taxonomías xenéricas. Aínda que cada sistema a analizar posúe características propias que o distinguen dos demais, algúns elementos da interface [69] e algunhas tarefas [41] son o bastante comúns como para ser xeneralizados a diferentes sistemas. Esta tese mostra mediante exemplos –correspondentes aos estudos de usabilidade descritos nela– como pode realizarse isto, aínda que a elaboración de taxonomías xenéricas e completas para describir os elementos das interfaces e as tarefas quedan fóra dos obxectivos desta tese.

Dende un punto de vista metodolóxico, o aspecto máis destacable dos dous estudos de usabilidade descritos non radica nas técnicas utilizadas –que son comunmente aplicadas en moitos outros estudos– senón en como as taxonomías foron integradas dentro de ditas técnicas. Aínda que calquera estudo de usabilidade está inherentemente baseado nalgún modelo de usabilidade, xa sexa un estándar da literatura ou un modelo ad hoc, os devanditos modelos tenden a ser pouco detallados, o que leva na práctica a depender demasiado de guías de usabilidade ou do "sentido común" para suplir as carencias dos modelos. O nivel de detalle das taxonomías ten como propósito facilitar o estudo da usabilidade neste sentido.

No estudo de usabilidade do CARAT counter, as taxonomías facilitaron a identificación de problemas nas seguintes categorías: lexibilidade e consistencia dos elementos da pantalla, intuitividade nas mensaxes de erro e as tarefas, universalidade respecto das unidades de medida da velocidade e o idioma da interface, precisión do GPS, seguridade do usuario, robustez respecto de abusos e erros internos e, por último, o atractivo estético.

Pola súa banda, o estudo de usabilidade das interfaces xeradas automaticamente identificou problemas en aspectos como: claridade (de significado, de funcionamento e de estrutura), consistencia (de terminoloxía, de aspecto), memorabilidade, completitude, accesibilidade, universalidade cultural (de linguaxe, de formatos de datas), flexibilidade, eficiencia, robustez ante o uso inadecuado, confidencialidade e estética.

Aínda que ambos estudos de usabilidade están estruturados de modo similar, existen diferenzas significativas que merecen ser descritas. En primeiro lugar, mentres que o CARAT counter consiste nunha combinación moi específica de hardware e software, as interfaces xeradas automaticamente eran páxinas web multiplataforma visualizadas en dispositivos móbiles de propósito xeral (smartphones e tablets). En segundo lugar, mentres que o estudo de usabilidade do CARAT counter analizaba a usabilidade do mesmo en termos "absolutos", o estudo das interfaces xeradas automaticamente foi de natureza comparativa, examinando distribucións alternativas dos mesmos elementos dunha mesma aplicación informática. En terceiro lugar, o estudo da usabilidade das interfaces xeradas automaticamente estaba centrado nun subconxunto reducido de atributos de usabilidade, xa que as interfaces eran idénticas en todos os sentidos excepto no relativo á disposición dos seus elementos. En cuarto lugar, este último estudo de usabilidade obtivo non só datos cualitativos (por exemplo, sobre as opinións dos usuarios) senón tamén datos cuantitativos, incluíndo medidas de rendemento obxectivas (tempo, número de erros) e estatisticamente significativas, que foron obtidas a partir de anotar gravacións en vídeo de 60 probas realizadas con usuarios.

Ambos estudos de usabilidade están centrados nas chamadas *metas formativas*, que buscan detectar problemas de usabilidade e propor solucións. Mentres que o nivel de detalle das taxonomías propostas nesta tese axudou na consecución destas metas (por exemplo, nas avaliacións heurísticas dos dous estudos de usabilidade), existen outros tipos de metas na Enxeñaría da Usabilidade. As chamadas *metas sumativas* teñen como obxectivo obter avaliacións xerais, non específicas. Por exemplo, puntuacións numéricas. A cuestión da adecuación da taxonomía de usabilidade ás metas sumativas quedou fóra dos obxectivos desta tese. Este é un tema complexo, xa que polo momento os investigadores non chegaron a un consenso nin sequera respecto de se o modelo de usabilidade do ISO 9241-11 pode ser utilizado para metas sumativas; máis concretamente, se ten sentido agregar os clásicos atributos de *efectividade*, *eficiencia* e *satisfacción* nunha única puntuación de usabilidade.

Outra parte importante desta tese está dedicada a investigar a validez dos supostos teóricos nos que se basea a mesma. Máis concretamente, compárase o enfoque proposto (é dicir, un baseado en taxonomías detalladas e de propósito xeral) co seu oposto (é dicir, modelos de usabilidade ad hoc, non detallados, e creados ex profeso para un campo de aplicación específico). Loxicamente, non ten por que existir un modelo ad hoc para cada campo de aplicación que podamos considerar pero, afortunadamente, existe un para o campo de sistemas de información situados no interior de vehículos [72], o que permitiu establecer unha comparación baseada nos resultados obtidos previamente para o CARAT counter. A devandita comparación non ten máis obxectivo que establecer as bases para unha discusión máis formal da expresividade dos modelos de usabilidade e, máis concretamente, dos problemas ocasionados pola falta de expresividade neste sentido. Se estudamos un novo modelo de usabilidade e comparámolo cun modelo de referencia como pode ser un estándar da literatura, podemos clasificar os posibles problemas de expresividade do novo modelo da seguinte forma [22]:

1. **Sobrecarga de construtos**. Un concepto do novo modelo corresponde a máis dun concepto do modelo de referencia.

2. **Exceso de construtos**. Un concepto do novo modelo non corresponde a ningún concepto do modelo de referencia.

3. **Redundancia de construtos**. Un concepto do modelo de referencia corresponde a máis dun concepto do novo modelo.

4. **Incompletitud de construtos**. Hai un concepto no modelo de referencia que non corresponde a ningún concepto do novo modelo.

Esta clasificación permite analizar un modelo de usabilidade –ou comparar dous modelos– e detectar problemas de ambigüidade, completitude ou redundancia. Adicionalmente, tamén pode argumentarse que modelos creados ex profeso para un campo de aplicación específico carecen da propiedade de ser xeneralizados a outras áreas, a cal adoita ser unha propiedade desexable no ámbito da Enxeñaría.

Por último, as contribucións máis relevantes desta tese poden ser resumidas da seguinte maneira:

- Analízanse as diferentes formas en que a usabilidade e o contexto de uso foron definidos na literatura e investíganse os problemas e as consecuencias desta multiplicidade de definicións.

- Proponse taxonomías xenéricas, detalladas e xerarquicamente organizadas para clasificar os atributos que compoñen, por unha banda, a usabilidade e, polo outro, o contexto de uso.

- Proponse un enfoque para levar a cabo estudos de usabilidade baseados na integración das devanditas taxonomías no ciclo de vida de desenvolvemento dun produto.

- Documéntase detalladamente como as taxonomías foron integradas na práctica en senllos estudos de usabilidade que tiveron lugar dentro de proxectos de investigación internacionais.

- Preséntase un marco teórico de discusión da validez e a expresividade dun modelo de usabilidade dado baseándose en establecer unha comparación cun modelo de referencia.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The concept of *usability* derives from the term *user friendly*, defined as "an expression used to describe computer systems which are designed to be simple to use by untrained users, by means of self-explanatory or self-evident interaction between user and computer" [47].

With time, the term *user friendly* came to be criticized as having "acquired a host of undesirably vague and subjective connotations" [30, p. 651], among them those of being "unnecessarily anthropomorphic" [129, p. 23] and suggesting that "users' needs can be described along a single dimension" [129, p. 23]. The concept of *usability* was then coined in order to overcome the limitations of the term *user friendly*.

So far, several definitions of usability and the attributes that characterize it have been proposed. However, these definitions tend to be brief and informal, and neither HCI (Human-Computer Interaction) researchers nor standards bodies have really achieved consensus on its exact meaning [11]. The best-known definition of usability at present would probably be the one in the ISO-9241-11 standard [87], which describes usability in terms of *effectiveness*, *efficiency*, and *satisfaction*.

Planning for usability is an essential part of the design and development of products and involves different types of activities during the life cycle of a product, such as identifying requirements, conducting user tests, obtaining performance measurements, and collecting subjective data. In turn, this produces both quantitative and qualitative information. Engineering for usability also requires to produce documentation in which the context of use is identified, problems are described, and targets are established [87].

The impact of usability on a product can be described along several dimensions, as follows [13][98][129][168]:

- **Actual use**. Usability affects the accomplishment of tasks. If usability is deficient, users will not feel motivated to use a product beyond what is strictly necessary and will never feel compelled to take full advantage of its possibilities.

- **Work management**. Inadequate usability reduces productivity and interrupts the work flow. In the worst-case scenario, it might actually be preferable to not use the product at all. Low usability in a product also makes training users more difficult.

- **Development**. Usability has a considerable impact on the characteristics and the quality of a product. Paying attention to usability from the start of the development life cycle also helps to avoid wasting valuable resources afterwards.

- **Maintenance**. Good usability is not a fixed property. A constantly evolving system – which is what much of today's software is by default – must seek to maintain good usability with each successive iteration. From another point of view, taking good care of usability from the beginning of the development of the product, and anticipating the needs of the users, also saves maintenance effort later.

- **Commercialization**. Low usability has a negative impact on the image and credibility of a product and its associated brand. Usability is also commonly used as a sales pitch, which is a big change compared to what happened over 20 years ago, when computers were more technologically limited and the number of functionalities was what mattered the most.

- **Support**. A non-intuitive, difficult to use product makes it necessary to dedicate more resources to user support.

From a socio-cultural point of view, usability becomes increasingly important with the rise of the so-called "information society". An ever-growing spectrum of users, characterized by great diversity (in terms of age, social class, technological expertise, physical characteristics, etc.) are now using complex computers and applications every day. Increasingly, this situation has been recognized with examples like the Section 508 of the government of the United States of America [10], which "applies to Federal agencies and which many people believe defines accessibility for Federal Web sites" [83, p. 39], or the 90/270/EEC Council Directive of the European Union [8], which "has been implemented in national law in member countries" [120, p. 2] and "is primarily concerned with the physical ergonomics of workstations, but makes some interesting demands concerning usability".

At the same time, technological progress has brought about a significant change of mentality in users, encouraging us to become more demanding and to get what we want with the least effort. As a consequence, we are quick to become impatient with bad usability.

However, achieving good usability is not easy. As we all know, even popular tools and applications can be found lacking in this regard. Some authors [123][129][131][168] have offered concrete reasons for this:

- Usability has always been difficult to define, measure, and evaluate, which makes it difficult to set goals, establish requirements, and communicate concepts.

- For the same reasons, it becomes difficult to justify in pragmatic terms the additional investment required to pursue good usability. Instead, it is very common to give higher priority to other aspects that are considered, in the short term, more “urgent” and “tangible”, whose results are more visible and immediate. This contrasts with the benefits of improving usability, which tend to be cumulative and are more easily perceived later, when the product is in actual use. In practice, usability problems are often fixed in an ad hoc way.

- Traditionally, the study of usability has been somewhat disconnected from the development life cycle of a product. Software engineers and usability professionals often see things from different – and even hard to reconcile – perspectives, especially in terms of priorities and implementation.

- User support is also treated separately in terms of budget. Nielsen [131] argued that the costs of improving support centers are often assigned to an account that is different from the one assigned to the costs of improving usability. This can be a problem, as both things are clearly not independent. Rather, the need for support is often a consequence of bad usability.

1.1 Background

Usability Engineering is a young but thriving field of research, and this thesis takes place in the context of a significant body of literature. Briefly, the three topics in the usability literature that are of most relevance to this thesis are:

1. **Usability**. That is, the meaning of the concept of usability itself.

2. **Context of use**. Usability is always relative to the context in which a product is used.

3. **Techniques for the study of usability**. Currently, there exist many different types of techniques for assessing the usability of a product. A usability study is always a small selection of some of them.

The research conducted for this thesis fits into the usability literature as part of what Lewis [111] calls “expanded models of usability”, which include the models by Bevan [29], Seffah et al. [167], Winter et al. [186], and the usability taxonomy developed for this thesis [16].

The motivations of these researchers for expanding the concept of usability are manifold. For example, Seffah et al. argue that:

> “Although there are many individual methods for evaluating usability, they are not well integrated into a single conceptual framework that facilitate their usage by developers who are not trained in the field of HCI. This is true in part because there are now several different standards

> [...] or conceptual models [...] for usability, and not all of these standards or models describe the same operational definitions and measures." [167, p. 160]

Winter et al. identify another limitation of the existing models of usability, as "they do not decompose the attributes and criteria to a level that is suitable for actually assessing them for a system" [186, p. 106]. This "constrains the use of these models as the basis for analyses".

Even though the usability definitions proposed in standards are widely accepted as valid, their vagueness restricts practitioners to dealing with *problem types*, not specific problems. As Gray and Salzman [65] state:

> "knowing what problem types an interface has is not really useful for developers. Developers need to know the specific problem (e.g., a problem with an item in a particular menu) and not the general one (e.g., 'there are menu problems' or 'speak the users' language')." [65, p.242]

This idea is borne out in practice by many usability engineers. For example, after examining ISO 9241-11, Hu and Chang conclude that if only "effectiveness, efficiency, and satisfaction are measured, the different specific aspects of usability problems related to a goal-task's human-tool interaction process cannot be reflected in the final usability evaluation so that the final usability evaluation appears to be too abstract and empty" [81, p.135].

The most recent group of standards proposed by ISO, namely, ISO FDIS 9241-210 [88], have done little to solve these problems, as their definition of usability remains the same, which has led researchers, such as Conger, to conclude that "all of the standards are generic, non-specific, and oriented toward a process for involving users in the development of interfaces. This approach, while useful, ignores the characteristics of usability and, as a result, is too abstract to guarantee any usability outcomes" [49, p.16].

This situation inspires usability practitioners to propose new usability definitions, often accompanied by their own methodologies. Some of the proposed solutions are intended as general-purpose and applicable to any kind of system (which is the approach taken in this thesis), whereas others are context-specific in the sense of being expressly tailored for a specific field of application. Recent examples of the latter approach can be found in the fields of biometrics [59], in-vehicle information systems [72], web sites [81], digital libraries [95], mobile applications [139], and open source software [156]. Context-specific conceptualizations of usability are typically constructed by borrowing concepts from different usability definitions, metrics, and so on, in an ad hoc way. This ad hoc approach, however, can easily lead to oversights, redundancy, or contradictions.

Unfortunately, it is difficult to get a clear sense of the actual, practical usefulness of the different usability definitions and methodologies employed by usability practitioners. As pointed out by Lewis:

> "Much of industrial usability engineering work is confidential. Companies are reluctant to expose the usability blemishes of their products and services to the public, preferring instead to keep them 'in house' as they track them and seek to eliminate or reduce their impact on users. Practitioners have been much freer to publish the results of methodological investigations, exposing and discussing methodological controversies of significant importance to the development of the field of usability engineering." [111, p. 663]

Therefore, usability researchers should not limit ourselves to proposing and discussing new models and methodologies; it is also essential to document in print how these models and methodologies are put to use in real-world usability studies.

1.2 Scope and Goals

The goals of this doctoral thesis are as follows:

- To study the different ways in which usability and the context of use have been defined in the literature, and to investigate the problems with this multiplicity of definitions.
- To create comprehensive and general-purpose taxonomies of attributes for the concepts of usability and the context of use, with the aim of synthesizing and refining the most prominent definitions for both terms.
- To propose a generic taxonomy-based approach to the study of usability as an alternative to context-specific (or ad hoc) approaches.
- To illustrate how this taxonomy-based approach can be integrated with selected techniques for the study of usability in specific moments of the life cycle of a product.
- To document how the approach has been actually used in real-world international research projects.

The conceptualization of usability in this doctoral thesis aims for both inclusiveness and relevance. Instead of taking sides on the many debates that occur in the usability field, or trying to reinvent the concept, the approach taken is to try and encompass all the currently relevant interpretations of the term.

While the taxonomies are meant to be applicable to any kind of product, the systems to which they will be applied belong to the field of Intelligent Systems, which is one of the main interests of our research group. More specifically, the usability studies will be focused on either the user interfaces of the systems or on the user interfaces that are generated by them. The usefulness of the taxonomies will be tested by how they are actually put into practice in order to identify and solve real-world problems. The general validity of the approach will also be discussed, but a properly empirical demonstration of its validity falls outside the scope of this thesis.

It should be also kept in mind that usability is neither an axiomatic system from which mathematical proofs can be derived nor a natural phenomenon than can be empirically tested against discrete things that exist in the world. Rather, it is a collectively-constructed concept whose ultimate goal is to allow engineers to understand and predict real-world problems in Human-Machine Interaction, which is a highly subjective area.

1.3 Structure of the Document

Chapter 2 (*Description of the Domain*) provides further background on the central topics of this research, which have been already identified in Section 1.1. That is, the concept of *usability*, the concept of *context of use*, and the *techniques* for the study of usability.

Chapter 3 (*A Taxonomy of Usability Attributes*) presents a new usability taxonomy that was specifically developed for this thesis and aims to be a synthesis and refinement of well-known definitions from the literature.

Chapter 4 (*A Taxonomy of Context-of-Use Attributes*) does the same for the concept of context of use.

Chapters 5 (*A Usability Study of an Intelligent Speed Adaptation Device*) and 6 (*A Usability Study of Automatically-Generated User Interfaces*) describe two practical, real-world applications of the usability and context-of-use taxonomies. The field of application is Intelligent Systems and the objects of study are, on the one hand, a prototype of an Intelligent Speed Adaptation device known as CARAT counter and, on the other hand, alternative HTML designs created by an automated multi-device UI generator. The specific usability activities include requirements analysis, heuristic evaluation, performance measurements, and subjective analysis (user questionnaires and interviews).

Chapter 7 (*Validity of the Approach*) explores and discusses the validity of the overall approach taken in this thesis. This will be carried out mainly by means of a comparison with the opposite approach, with the ultimate goal of laying the groundwork for a more formal discussion of expressiveness in general, applied to the area of usability models.

Chapter 8 (*Discussion*) addresses in depth several topics that have been raised in the previous chapters regarding the scope and the methodology of this thesis.

Finally, Chapter 9 (*Conclusions and Future Work*) summarizes the main findings of this thesis and suggests new lines of research.

1.4 Recommended Reading

What follows is a short list of recommended texts that help to put into context the research conducted for this doctoral thesis. The texts are chosen for their relevance to this thesis, more than for their historical importance or their authors' notoriety. Nevertheless, some of the authors are recognized authorities in usability and their referenced texts are regarded as important contributions to the literature. Some texts are reviews of the state of the art, some have goals in common with this thesis, and some are examples of opposite approaches. Many of them are repeatedly referenced throughout different chapters of this thesis.

- J. R. Lewis. Usability: Lessons learned. . . and yet to be learned. *International Journal of Human-Computer Interaction*, 30(9):663–684, 2014.

  A review of the current state of longstanding controversies in the field of usability.

- M. Y. Ivory and M. A. Hearst. The state of the art in automating usability evaluation of user interfaces. *ACM Computing Surveys (CSUR)*, 33(4):470–516, 2001.

  A review of the state of the art of usability techniques, classified according to their degree of automation.

- *ISO 9241-11:1998 - Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability*, 1998.

  The most widely-referenced standard about usability.

- A. Seffah, M. Donyaee, R. B. Kline, and H. K. Padda. Usability measurement and metrics: A consolidated model. *Software Quality Journal*, 14:159–178, 2006.

  One of the most comprehensive usability models published prior to this research.

- S. Winter, S. Wagner, and F. Deissenboeck. A comprehensive model of usability. In *Engineering interactive systems*, pages 106–122. Springer, 2008.

  Another comprehensive usability model, published contemporaneously with this research.

- D. Alonso-Ríos, A. Vázquez-García, E. Mosqueira-Rey, and V. Moret-Bonillo. Usability: A critical analysis and a taxonomy. *International Journal of Human-Computer Interaction*, 26(1):53–74, 2009.

  Introduces the usability taxonomy.

- D. Alonso-Ríos, A. Vázquez-García, E. Mosqueira-Rey, and V. Moret-Bonillo. A context-of-use taxonomy for usability studies. *International Journal of Human-Computer Interaction*, 26(10):941–970, 2010.

  Describes the context-of-use taxonomy.

- D. Alonso-Ríos, E. Mosqueira-Rey, and V. Moret-Bonillo. A taxonomy-based usability study of an Intelligent Speed Adaptation device. *International Journal of Human-Computer Interaction*, 30(7):58–603, 2014.

  Describes in detail the usability study of the CARAT counter and outlines the validity of the taxonomy-based approach.

- D. Raneburger. *Interactive model-driven generation of graphical user interfaces for multiple devices*. PhD thesis, Institute of Computer Technology, Vienna University of Technology, 2014.

  Describes the automated user interface generator.

- C. Harvey, N. A. Stanton, C. A. Pickering, M. McDonald, and P. Zheng. Context of use as a factor in determining the usability of in-vehicle devices. *Theoretical issues in ergonomics science*, 12(4):318–338, 2011.

  An example of the opposite approach to constructing usability models.

Chapter 2

Description of the Domain

2.1 Definitions of Usability

The first attempt at a definition of *usability* is generally attributed to Shackel in 1981 [170]. Expanding on the concept of *ease of use*, Shackel's conception of usability "is not only conceived of as ease of use, but also equally involves efficacy i.e., effectiveness in terms of measures of (human) performance" [171, p. 24]. Shackel's formal definition of usability, applied to systems or equipment, is as follows:

> "the capability in human functional terms to be used easily and effectively by the specified range of users, given specified training and user support, to fulfil the specified range of tasks, within the specified range of environmental scenarios." [171, p. 24]

In a more convenient shortened form, Shackel also proposes the following definition:

> "the capability to be used by humans easily and effectively." [171, p. 24]

Shackel also proposes a brief list of usability criteria, namely [171]:

1. **Effectiveness**.
2. **Learnability**.
3. **Flexibility**.
4. **Attitude**.

Nielsen [129][135] is another pioneering and influential author in the field of usability. Nielsen does not provide a definition as such [62], but proposes instead a list of usability attributes and provides definitions for them, as follows [129, p. 26]:

1. **Learnability**. The system should be easy to learn so that the user can rapidly start getting some work done with the system.

2. **Efficiency**. The system should be efficient to use, so that once the user has learned the system, a high level of productivity is possible.

3. **Memorability**. The system should be easy to remember, so that the casual user is able to return to the system after some period of not having used it, without having to learn everything all over again.

4. **Errors**. The system should have a low error rate, so that users make few errors during the use of the system, and so that if they do make errors they can easily recover from them. Further, catastrophic errors must not occur.

5. **Satisfaction**. The system should be pleasant to use, so that users are subjectively satisfied when using it; they like it.

As can be seen, these two early models of usability attributes by Shackel and Nielsen are quite different in form, although not necessarily in content, as there is much semantic overlap between them. For example, attitude (Shackel) and satisfaction (Nielsen) are meant to represent subjective criteria. Other attributes, like effectiveness (Shackel) and errors (Nielsen) are clearly related to each other.

Other important authors who contributed to the early history of usability are Norman, whose "work on usability is perhaps better viewed as a contribution to the philosophy of usability, rather than as a definitive list of usability criteria" [72, p. 5] and Shneiderman, whose "golden rules of dialog design" overlap significantly with the Nielsen and Shackel models and are, in any case, closer to design guidelines than usability attributes [72].

Probably the best-known definitions of usability are the ones proposed by the International Organization for Standardization (ISO), particularly in the following standards:

- **ISO 9241-11**:1998, Ergonomic requirements for office work with visual display terminals (VDTs), Part 11: Guidance on usability [87].

- **ISO/IEC 9126-1**:2001, Software engineering, product quality, Part 1: Quality model [90].

These two standards are the result of several iterations and revisions, during which their definitions of usability have eventually changed. What follows are simply the ones in current use.

ISO 9241-11 [87] defines usability as:

> "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use."

More specifically, the attributes that characterize usability are defined by ISO 9241-11 as follows:

1. **Effectiveness**. The accuracy and completeness with which users achieve specified goals.

2. **Efficiency**. The resources expended in relation to the accuracy and completeness with which users achieve specified goals.

3. **Satisfaction**. Freedom from discomfort and positive attitudes towards the use of the product.

ISO 9241-11 does not define usability further, but illustrates the concept using a non-exhaustive list of examples. The Annex B of said document, thus, refers to several measures of usability, with effectiveness measured in terms of the "percentage of goals achieved, percentage of users successfully completing tasks and average accuracy of completed tasks," efficiency measured in terms of the "time to complete a task, tasks completed per unit time and monetary cost of performing the task," and finally, with satisfaction measured in terms of a "rating scale for satisfaction, frequency of discretionary use and frequency of complaints."

ISO 9241-11 was recently replaced by the more comprehensive standard ISO FDIS 9241-210 [88], but its definition of usability is basically the same, simply substituting "product" with "system, product, or service". Its classification of sub-attributes remain the same.

The other relevant ISO standard is ISO/IEC 9126-1 [90], which classifies usability as one of the components representing *internal and external software quality*. The attributes of *usability* are further defined by ISO/IEC 9126-1 as follows:

1. **Understandability**. The capability of the software product to enable the user to understand whether the software is suitable and how it can be used for particular tasks and conditions of use.

2. **Learnability**. The capability of the software product to enable the user to learn its application.

3. **Operability**. The capability of the software product to enable the user to operate and control it.

4. **Attractiveness**. The capability of the software product to be attractive to the user.

5. **Usability compliance**. The capability of the software product to adhere to standards, conventions, style guides or regulations relating to usability.

As can be seen, both usability standards have little in common, especially at the relatively superficial level of detail at which they are described. As mentioned, *usability* is actually considered an attribute of *quality* in ISO/IEC 9126-1, along with other five characteristics. The ISO/IEC 9126-1 *quality* model consists of:

1. **Functionality**. The capability of the software product to provide functions which meet stated and implied needs when the software is used under specified conditions.

2. **Reliability**. The capability of the software product to maintain a specified level of performance when used under specified conditions.

3. **Usability**. The capability of the software product to be understood, learned, used and attractive to the user, when used under specified conditions.

4. **Efficiency**. The capability of the software product to provide appropriate performance, relative to the amount of resources used, under stated conditions.

5. **Maintainability**. The capability of the software product to be modified. Modifications may include corrections, improvements or adaptation of the software to changes in environment, and in requirements and functional specifications.

6. **Portability**. The capability of the software product to be transferred from one environment to another.

ISO/IEC 9126-1 also includes a separate but related model for *quality in use*, which is defined as "the user's view of quality" [90, p. 12]. It is broken down into *effectiveness*, *productivity*, *safety*, and *satisfaction*. This concept overlaps significantly with usability and brings to mind the title of a 1995 paper by Nigel Bevan: "Usability is quality of use" [26]. It should be noted that ISO/IEC 9126-1 acknowledges that "usability is defined in ISO 9241-11 in a similar way to the definition of quality in use in this part of ISO/IEC 9126. Quality in use may be influenced by any of the quality characteristics, and is thus broader than usability" [90, p. 12].

The ISO/IEC 9126 quality model was recently replaced by ISO/IEC 25010:2011 [89]. The latter differs from ISO ISO/IEC 9126 in that it does not feature usability as an attribute of its *quality* model. Instead, usability is considered a subset of its new *quality in use* model.

The ISO/IEC 25010 *quality* model consists of:

1. **Functional suitability**.

2. **Reliability**.

3. **Operability**.

4. **Performance efficiency**.

5. **Security**.

6. **Compatibility**.

7. **Maintainability**.

8. **Transferability**.

And the ISO/IEC 25010 *quality in use* model consists of:

1. **Effectiveness**. Accuracy and completeness with which users achieve specified goals.

2. **Efficiency**. Resources expended in relation to the accuracy and completeness with which users achieve goals.

3. **Satisfaction**. Degree to which user needs are satisfied when a product or system is used in a specified context of use. This attribute is subdivided into *usefulness*, *trust*, *pleasure*, and *comfort*.

4. **Freedom from risk**. Degree to which a product or system mitigates the potential risk to economic status, human life, health, or the environment. This attribute is subdivided into *economic risk mitigation*, *health and safety risk mitigation*, and *environmental risk mitigation*.

5. **Context coverage**. Degree to which a product or system can be used with effectiveness, efficiency, freedom from risk and satisfaction in both specified contexts of use and in contexts beyond those initially explicitly identified. This attribute is subdivided into *context completeness* and *flexibility*.

Usability, then, "is defined as a subset of quality in use consisting of effectiveness, efficiency and satisfaction, for consistency with its established meaning" [89], which finally brings usability back in line with its traditional (and competing) definition from ISO 9241-11, therefore ending what has been the source of much confusion over the years.

Another standards body, the Institute of Electrical and Electronics Engineers (IEEE), proposes as a definition for usability "the ease with which a user can learn to operate, prepare inputs for and interpret outputs of a system or component" [150, p. 80].

Other recent additions to the literature that are also relevant to this doctoral thesis are as follows:

- **Preece et al.** developed an initial usability model that includes *safety*, *effectiveness*, *efficiency*, and *enjoyableness* [145]. Subsequently, they proposed a new classification composed of *learnability*, *throughput*, *flexibility*, and *attitude* [146].

- **Quesenbery** lists the attributes of a usable product as *effectiveness*, *efficiency*, *engagement*, *error tolerance*, and *ease of learning*. [147][148][149].

- **Abran et al.** proposed extending [11] the ISO 9241-11 definition by adding two further attributes, namely, *learnability* (which had already been adopted by the IEEE [150], ISO/IEC 9126-1 [90], and Nielsen [129]) and *security*.

- **Seffah et al.**, also taking ISO 9241-11 as the starting point, propose the most complete usability model of all these, which they call the QUIM model [167]. The model defines a total of ten usability factors (*efficiency*, *effectiveness*, *productivity*, *satisfaction*, *learnability*, *safety*, *trustfulness*, *accessibility*, *universality*, and *usefulness*) that are associated with 26 measurable usability criteria. The criterion of *privacy*, for example, is associated with trustfulness, universality, and usefulness (given that it measures aspects associated with each of these factors).

It is also worth mentioning that a new term named *user experience* has recently come to the fore. This term is related to usability but is not necessarily synonymous with it. Once more, many definitions have been proposed for this term, which, according to Bevan [27], fall into three groups:

1. An elaboration of the satisfaction component of usability.
2. Distinct from usability, which has a historical emphasis on user performance.
3. An umbrella term for all the user's perceptions and responses, whether measured subjectively or objectively.

2.2 Definitions of the Context of Use

Usability is not absolute but relative. In 1991, Bevan et al. [30] identified three views of usability:

1. **The product-oriented view**: that usability can be measured in terms of the ergonomic attributes of the product.
2. **The user-oriented view**: that usability can be measured in terms of the mental effort and attitude of the user.
3. **The user performance view**: that usability can be measured by examining how the user interacts with the product, with particular emphasis on either ease-of-use (how easy the product is to use) or acceptability (whether the product will be used in the real world).

To these, Bevan et al. added the following view:

4. **The contextually-oriented view**: that usability of a product is a function of the particular user or class of users being studied, the task they perform, and environment in which they work.

This contextual view was an important part of the ISO 9241-11 [87] definition of usability quoted in Section 2.1, as the standard explicitly refers to effectiveness, efficiency, and satisfaction "in a specified context of use".

This means that a product or system that may be usable in one *context of use* may not be usable in another context characterized by different users, activities, technologies, or environments. For example, a text editor requires different quality attributes depending on its intended purpose, which could range from producing letters to writing programming code [25]. Another example is how the usability of an application varies depending on the personal characteristics of its users, as demonstrated decades ago in studies by Dillon and Song [52] and Mack et al. [118]. Consequently, it seems clear that the results of a usability study of a product in one specific context of use cannot be directly generalized to other contexts [24].

Traditionally, recognition of the concept of context of use was limited to the principles deriving from Human-Computer Interaction. For example, the literature cited environmental factors such as safety and noise and light levels, and it drew attention to the importance of understanding the characteristics of the users and the nature of the expected work to be accomplished [64]. Nonetheless, this was still a somewhat simplified vision of the context of use that failed to clarify precisely which characteristics should be studied.

Based on previous definitions by Schilit et al. [165] and Pascoe [141], Dey et al. [51] proposed the following definition of context:

> "Context is any information that can be used to characterise the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves." [51, p. 5]

Thus, the idea of context explicitly encompasses the characteristics of the user, the application, and things external to both.

As with the concept of usability, researchers and standardization bodies eventually proposed detailed context-of-use models consisting in comprehensive enumerations of attributes. The most relevant ones for the purposes of this thesis are described below.

The ISO 9241-11 [87] context-of-use model is probably the best-known in the usability literature. It identifies the first-level attributes as *users*, *tasks*, *equipment*, and *environment*. These attributes are further broken down into several subattributes (which are in turn sometimes grouped into categories), as follows:

1. **Users** is subdivided into three categories: *user types*, *skills and knowledge*, and *personal attributes*. *User types* is classified into two types; *skills and knowledge* is subdvided into nine attributes; and *personal attributes* into seven.

2. **Tasks** is not subdivided into categories like the other attributes, but is simply associated with ten attributes (i.e., *breakdown*, *name*, *frequency of use*, *duration*, *flexibility*, *physical and mental demands*, *dependencies*, *output*, *risk resulting from error*, and *safety critical demands*).

3. **Equipment** is broken down into two categories: *basic description* and *specification*. The former has three attributes and the latter has five.

4. **Environment** is the most complex context-of-use attribute. It consists of 31 attributes at the lowest level, which are hierarchically grouped into categories, as follows. *Environment* is subdivided into *organizational environment* (characterized in turn by *structure*, *attitudes and culture*, and *job design*), *technical environment* (characterized by *configuration*, which in turn is subdivided into *hardware*, *software*, and *reference materials*), and *physical environment* (characterized by *workplace conditions*, *workplace design*, and *workplace safety*).

ISO 9241-11 indicates, however, that its classification merely constitutes a set of examples and is not to be rigidly applied. This standard is a revision of the 1992 version, and its context-of-use taxonomy is explicitly based on two intermediate models by Bevan and Macleod [31] and Thomas and Bevan [179], which are described next.

The Bevan and Macleod [31] definition refers to *users*, *tasks*, *equipment*, and *environment* as first-level attributes (i.e., the same as ISO 9241-11 [87]). *User* is broken down into *personal details*, *skills and knowledge*, and *personal attributes.* The rest of the attributes are subdivided into the same categories as ISO 9241-11 [87]), that is, *equipment* is broken down into *basic description* and *specification*; *environment* is broken down into *organizational environment* (*structure*, *attitudes and culture*, and *job design*), *technical environment* (*configuration*), *and physical environment* (*workplace conditions*, *workplace design*, and *workplace safety*). This classification is, moreover, based on that by Maissel, Dillon, Maguire, Rengger, and Sweeny [122].

The Thomas and Bevan context-of-use model [179] refers to *users*, *task characteristics*, *organizational environment*, *technical environment*, and *physical environment*, with their main subattributes as follows: for the *user* attribute, *user types*, *skills and knowledge*, *physical attributes*, *mental attributes*, *job characteristics*, and *lists of tasks*; for the *organizational environment* attribute, *structure*, *attitudes and culture*, and *worker control*; for the *technical environment* attribute, *hardware*, *software*, and *reference materials*; and finally, for the *physical environment* attribute, *environmental conditions*, *workplace design*, and *health and safety.*

Another definition of the context of use, given by Kirakowski and Cierlick [103] refers to *users*, *tasks*, and *environment*, with subattributes as follows: for the *user* attribute, *role*, *skills and knowledge*, *physical attributes*, *attitude and motivation*, and *job characteristics*; for the *task* attribute, *execution*, *flow*, *demand on users*, and *safety*; for the *environment* attribute, *social environment*, *organizational environment*, *technical environment*, and *physical environment.*

Finally, the definition by Maguire [121] also refers to attributes and subattributes for *users* (*name*, *experience*, *knowledge*, and *skills*, and *personal attributes*), *tasks*, and *environment* (*technical environment*, broken down into *hardware*, *software*, *network*, and *reference materials*; *physical environment*, broken down into *workplace conditions*, *workplace design*, and *workplace safety*; and *organizational environment*, broken down into *structure*, *attitudes and culture*, and *job design*).

These models have been used in real context-of-use studies. For example, one such study was conducted by Träskbäck and Haller [180], who used the ISO 9241-11 model to analyze the context of use for a training application for oil refineries.

Other researchers, however, have been focused on constructing ad hoc context-of-use models for specific types of software systems. There is a great deal of interest, for example, in studying context-of-use components for mobile systems, as demonstrated by the classification by Coursaris and Kim [50], who distinguished between *user*, *task*, *environment*, and *technology*. In the field of audiovisual consumer electronic products, Kwahk and Han [107] proposed a taxonomy that highlights, on a first level, *user*, *product*, *user activity*, and *environment*. Estrella, Popescu-Belis, and Underwood [57], meanwhile, classified the context of use for automated translation systems in terms of *characteristics of the translation task*, *input characteristics*, and *user characteristics*.

In sum, most of the definitions of context of use in the literature suffer from the same kinds of problems as the definitions for usability. They tend to propose attribute definitions that are overly brief or ambiguous. In fact, definitions are not even provided in some cases – for example, in Bevan and Macleod [31] and ISO 9241-11 [87]. The hierarchies in existing context-of-use models, furthermore, tend to have too few layers of attributes, making it impossible to describe context in sufficient detail. Another problem is that practically all the context-of-use models establish sets of attributes that are relevant to information technology (IT) products but are incomplete or unsuitable for the study of other types of systems.

2.3 Techniques for the Study of Usability

According to ISO 9241-11, there are three potential ways in which the usability of a product could be measured, namely [87, p. 35]:

- **By analysis of the features of the product**, required for a particular context of use. Usability could be measured by assessing the product features required for usability in a particular context.

- **By analysis of the process of interaction**. Usability could be measured by modeling the interaction between a user carrying out a task with a product. However, current analytic approaches do not give very precise estimates of usability. As the interaction is a dynamic process in the human brain, it cannot be studied directly.

- **By analyzing the effectiveness and efficiency** which results from use of the product in a particular context and measuring the satisfaction of the users of the product. These are direct measures of the components of usability. If a product is more usable in a particular context, usability measures will be better.

At present, there are many *techniques* (or methods) for studying, analyzing, and evaluating usability. A common way of integrating such techniques in the development life cycle is through iterative design. Thus, usability characteristics are identified and improvements are suggested during successive iterations. For example, at the beginning of a project usability experts might gather usability information and requirements though interviews, ergonomics checklists, and so on. This can lead to an initial prototype of a product, which can then be tested by users in a laboratory setting. In turn, this gives more usability information that can be used to design a fully operational first version of the product that can be empirically tested by actual users [13].

Usability techniques are characterized by great diversity, which makes sense given that usability is such a multifaceted concept. These techniques are distinguished by the type of activities they involve and the aspects of usability they address. Different techniques may obtain different types of results, so it is recommended to use several complementary techniques during a usability study. It is also advisable to obtain a varied and reasonably high amount of results, for which it is recommended to have a sufficient number of participants and to ensure that the usability analysis is conducted in a systematic way.

Adelman and Riedel's conceptual framework [13] includes a classification of 13 usability evaluation methods, which they group into three categories:

1. **Expert evaluation methods**, which "determine what is good and bad about the system from a usability perspective" [13, p. 235]. These include ergonomic checklists, interface surveys, architecture (or navigation) evaluation, the GOMS model, and the Seven Stages of Action model.

2. **Subjective evaluation methods**, which obtain "users' opinions about the usability of evolving prototypes and operational systems" [13, p. 242]. These include user feedback, user diaries, teaching how a system works, and questionnaires.

3. **Empirical evaluation methods**, which "obtain objective data about how well people can actually use a system" [13, p. 251]. These include objective performance measures, usability testing, experiments, and logging actual use.

Ivory and Hearst later proposed a more comprehensively taxonomy of techniques, which they group into five classes [92, p. 473]:

1. **Inspection**. An evaluator uses a set of criteria or heuristics to identify potential usability problems in an interface.

2. **Inquiry**. Users provide feedback on an interface via interviews, surveys, and the like.

3. **Testing**. An evaluator observes users interacting with an interface (i.e., completing tasks) to determine usability problems.

4. **Analytical Modeling**. An evaluator employs user and interface models to generate usability predictions.

5. **Simulation**. An evaluator employs user and interface models to mimic a user interacting with an interface and report the results of this interaction (e.g., simulated activities, errors, and other quantitative measures).

The first three classes essentially match the three categories in Adelman and Riedel's conceptual framework (see above). According to Ivory and Hearst, the first three classes are "appropriate for formative (i.e., identifying specific usability problems) and summative (i.e., obtaining general assessments of usability) purposes" [92, p. 473]. On the other hand the last two classes concern themselves with the construction and automation of user models and interface models.

Ivory and Hearst also further subdivide these five categories into 35 individual techniques, which are listed in tables 2.1 through 2.5. This should serve to illustrate the wide range of techniques currently available to usability practitioners.

Table 2.1: Inspection Techniques

Guideline Review	The expert checks guideline conformance.
Cognitive Walkthrough	The expert simulates the user's problem solving.
Pluralistic Walkthrough	Multiple people conduct cognitive walkthrough.
Heuristic Evaluation	The expert identifies violations of heuristics.
Perspective-Based Inspection	The expert conducts a narrowly focused heuristic evaluation.
Feature Inspection	The expert evaluates product features.
Formal Usability Inspection	The expert conducts formal heuristic evaluation.
Consistency Inspection	The expert checks consistency across products.
Standards Inspection	The expert checks for standards compliance.

Table 2.2: Inquiry Techniques

Contextual Inquiry	The interviewer questions users in their environment.
Field Observation	The interviewer observes system use in user's environment.
Focus Groups	Multiple users participate in a discussion session.
Interviews	One user participates in a discussion session.
Surveys	An interviewer asks the user specific questions.
Questionnaires	The user provides answers to specific questions.
Self-Reporting Logs	The user records UI operations.
Screen Snapshots	The user captures UI screens.
User Feedback	The user submits comments.

Practical considerations demand that a given usability study will employ only a small subset of the usability techniques available to practitioners. The work described in this doctoral thesis made use of techniques in all three of the Ademal and Riedel [13] categories. The actual implementation of the usability techniques in the respective usability studies will be described in their own chapters, but a brief overview is given below.

Table 2.3: Testing Techniques

Thinking-Aloud Protocol	The user talks during the test.
Question-Asking Protocol	The tester asks the user questions.
Shadowing Method	An expert explains user actions to the tester.
Coaching Method	The user can ask an expert questions.
Teaching Method	An expert user teaches a novice user.
Codiscovery Learning	Two users collaborate.
Performance Measurement	The tester records usage data during the test.
Log File Analysis	The tester analyzes usage data.
Retrospective Testing	The tester reviews a videotape with the user.
Remote Testing	The tester and the user are not co-located during the test.

Table 2.4: Analytical Modeling Techniques

GOMS Analysis	Predict execution and learning time.
UIDE Analysis	Conduct GOMS analysis within a UIDE.
Cognitive Task Analysis	Predict usability problems.
Task-Environment Analysis	Assess mapping of user's goals into UI tasks.
Knowledge Analysis	Predict learnability.
Design Analysis	Assess design complexity.
Programmable User Models	Write a program that acts like a user.

The main expert (or inspection) technique employed in the two usability studies described in this thesis is heuristic evaluation [136][130], which was first introduced by Nielsen and Molich and is currently one of the most actively used and researched usability methods [76]. In heuristic evaluation, one or more usability specialists examine the usability of a product guided by a set of heuristics, such as guidelines or good practices. Many different heuristics have been proposed over time, but there is still little scientific evidence on the preferability of one over the others in terms of effectiveness, efficiency, or inter-evaluator reliability [84]. Heuristic evaluation remains mostly informal, but it has the advantages of being quick and inexpensive. According to Dumas, "Nielsen's work on heuristic evaluation freed expert reviews from the burden of the hundreds of guidelines that characterized earlier inspections" [55, p. 55]. Heuristic evaluation works best when the results of several evaluators are aggregated, as these typically only find some of the actual usability problems [128]. Of course, during the heuristic evaluations the systems were operated in ways that simulate typical user tasks – although not exhaustively – and as a consequence the heuristic evaluations had elements of other expert techniques.

Subjective (or inquiry) techniques, were also an important part of the usability studies described in this thesis, as these techniques collect information on the actual opinions of the users. Subjective techniques, often performed together with user testing, require feedback from the users in the form of surveys, questionnaires, interviews, and so on [92]. Questionnaires and interviews are perhaps the most-widely used forms of feedback, and they were the ones used for the work presented in this thesis.

Table 2.5: Simulation Techniques

Information Processor Modeling	Mimic user interaction.
Petri Net Modeling	Mimic user interaction from usage data.
Genetic Algorithm Modeling	Mimic novice user interaction.
Information Scent Modeling	Mimic Web site navigation.

The last group of usability techniques under consideration, that is, the ones called empirical (or testing), provide direct information on system use and actual (as opposed to predicted) problems [77]. During usability testing, participants operate the system or a prototype in order to complete a predetermined set of tasks while the testers record the results. The testers then use the results to evaluate the suitability of the system to task completion, as well as other measures like completion time and number of errors [92]. Actual user testing was very important in one of our usability studies, as one of the intended goals was to obtain quantitative measures of realistic task completion (in a laboratory setting). According to the testing techniques in Table 2.3, this means focusing on the technique known as performance measurement. Other types of testing techniques, particularly those in which the flow is interrupted in some way to talk or collaborate, may not be suitable for collecting realistic measures, and are best reserved for usability studies of a different nature.

Other researchers have classified usability techniques according to different criteria. For example, Rohrer [162] has recently addressed the question of "what to do when" by proposing a three-dimensional framework of user-experience methods along the following axes:

1. Attitudinal vs. behavioral.
2. Qualitative vs. quantitative.
3. Context of Use.

The *attitudinal vs. behavioral* dimension contrasts "what people say" versus "what people do". Subjective questionnaires would fall into the former category, whereas comparative testings that present "changes to a site's design to random samples of site visitors, but attempts to hold all else constant" [162] – which was basically the domain of one of our usability studies – would fall into the latter category.

The *qualitative vs. quantitative* dimension addresses whether the data about behaviors or attitudes is generated via direct observation or indirect measurement. This thesis covers both ends of this spectrum, as the actions of the users are, at specific points, observed in person (and later discussed with them in interviews) or measured through computer tools (and later analyzed statistically).

The *context of use* dimension is related to how and whether participants in a study are using the product or service in question. Rohrer draws the following distinctions:

1. Natural or near-natural use of the product.
2. Scripted use of the product.
3. Not using the product during the study.
4. A hybrid of the above.

The two examples of usability testings described in this thesis fall into opposite categories of this dimension. One falls into the *natural* category, as users had to simply incorporate the system under study into their daily routine. The other falls into the *scripted* category in the sense that users were given very specific instructions regarding the information that had to be entered.

A different type of classification was proposed by Whitefield et al. [185]. Their taxonomy classifies evaluation methods into four categories according to whether they involve *real end users* or *representative users*, and whether they evaluate a *real system* or a *partial version*.

Bastien and Scapin [23] and Sears [166] have defined *validity*, *thoroughness*, and *reliability* as criteria for comparing different evaluation techniques. The latter provides definitions and formulas for these criteria, as follows [166, pp. 214–215]:

1. **Validity**. A technique is valid if evaluators are capable of focusing on issues that are pertinent. Validity can be measured as the ratio of "real" usability problems identified to all issues identified as usability problems.
2. **Thoroughness**. A technique is thorough if evaluators are capable of evaluating all aspects of an interface. A natural measure of thoroughness is the ratio of real problems that are identified to the number of problems that exist in the system
3. **Reliability**. Reliability implies that similar results should be obtained under similar conditions. This can be measured using the ratio of the standard deviation of the number of problems found to the average number of problems found.

These metrics, however, have been criticized on the grounds that it is generally impossible to determine whether all usability problems have been actually identified in a particular test or whether the problem set for assessing the effectiveness of a given technique is complete [112].

Chapter 3

A Taxonomy of Usability Attributes

3.1 Introduction

This chapter presents a usability taxonomy that was specifically developed for this thesis. The main motivations for creating a usability taxonomy were the relative vagueness of the models that exist in the literature and the lack of consensus on the precise meaning of the term (see Section 2.1). The aim of the taxonomy is to be a synthesis and refinement of well-known definitions of the term from the literature while avoiding ambiguity and redundancy. The main contributions of the taxonomy lie in the way the attributes in the literature are merged, split, renamed, structured, and given depth as a result of the synthesis.

The chapter is structured as follows: (1) The construction process is outlined; (2) the taxonomy is described in detail by specifying the full hierarchy of subattributes and providing definitions for each term; (3) the taxonomy is compared with other usability models in the literature in terms of the attributes that are covered.

3.2 The Construction Process

The first step in building the usability taxonomy consisted in examining the usability models in the literature (see Section 2.1) and choosing a small subset of relevant and non-redundant models. The criteria for selection were based on acceptance by practitioners (e.g., standards) and the introduction of new attributes that became widely accepted thereafter. A balance had to be struck between, on the one hand, searching for models that added something new to traditional definitions of usability and, on the other hand, trying not to deviate too heavily from them.

The usability models that were chosen for constructing the usability taxonomy at the time were as follows:

- **ISO 9241-11** (1998) [87].

- **ISO/IEC 9126-1** (2001) [90].
- **Nielsen** (1993) [129].
- **Preece et al.** (1994) [146].
- **Quesenbery** (2001) [147].
- **Abran et al.** (2003) [11].
- **Seffah et al.** (2006) [167].

All these models were described in Section 2.1. Note that, as discussed in that section, ISO 9241-11 and ISO/IEC 9126-1 have been recently replaced by ISO FDIS 9241-210 [88] and ISO/IEC 25010 [89], respectively. These two new standards were released after this research had been conducted and the usability taxonomy [16] had been published. Nevertheless, the current definitions of usability in both standards are virtually identical to the traditional ISO 9241-11 one, which makes the references to ISO 9241-11 in this chapter still relevant. This also means that the other old definition of usability, namely, the one in ISO/IEC 9126-1, has been basically superseded, but the references to it are retained in this chapter for the sake of consistency with the published results of this research.

The pros and cons of the models in the literature were weighed and their attributes were structured and given precise definitions during several refinement cycles. All this involved many meetings among the authors of the taxonomy [16] during several weeks. Completeness (i.e., retaining as many attributes as possible from the literature) was prioritized, whereas redundancy and overlap between meanings were to be avoided. While no formal methodology was followed for achieving this, the general principles are similar to the ones conceptualized afterwards in Section 7.5.

More specifically, the first step in the construction process was to list all the attributes described in existing usability models. Given the lack of consensus in the usability field, existing models are clearly divergent. Furthermore, when models do overlap, they tend to do so only partially and unevenly, with different terms used to designate the same attribute or with the same term used to describe different concepts. Given the ambiguity of the definitions offered by the models in the literature, it is also difficult to match concepts in different models. This preliminary study nevertheless served as the groundwork for our own synthesis. The purpose of this synthesis is to cover the usability aspects of any type of system and to be adaptable to different contexts of use (i.e., not only IT systems). As a result, some of the attributes in the literature were directly used, while other attributes were redefined. Furthermore, new attributes that had not been covered as such by any of the existing models were also added. Finally, for each attribute a precise definition of its meaning was coined.

Once the attributes had been defined, they were structured and ordered in such a way as to populate a first level in the taxonomy with generic attributes that could be progressively refined to obtain more specific subattributes that would populate subsequent taxonomic levels. It was important that the criteria used to structure the taxonomy permitted to group related attributes together and separate unrelated attributes as much as possible. Another important issue was to pay special attention to avoiding redundancy in the different attributes. By preventing overlapping, attributes can be evaluated in isolation from other attributes, thereby simplifying the study of usability, ensuring greater thoroughness, and avoiding contradiction.

An initial analysis round was followed by a series of revisions and progressive refinements of the taxonomy that consisted of adding new attributes, regrouping attributes in more suitable categories, eliminating redundant attributes, and developing more appropriate terms for certain attributes.

The taxonomy resulting from this process of construction has the first-level usability attributes depicted in Figure 3.1.

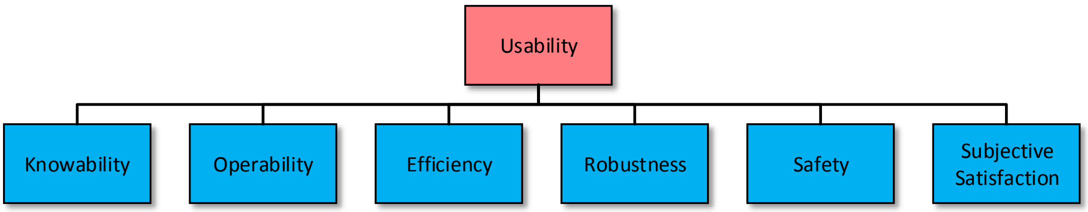


Figure 3.1: Usability attributes

3.3 Description of the Taxonomy

3.3.1 Knowability

Knowability is defined as the property by means of which the user can understand, learn, and remember how to use the system. This attribute has subattributes as follows (see Figure 3.2):

- **Clarity**, defined as the ease with which the system can be perceived by the mind and the senses. A distinction is drawn between three kinds of clarity:
  - *Clarity of the elements*, classified in turn in terms of *formal* clarity (capacity of the system to facilitate perception of individual system elements through the senses) and *conceptual* clarity (capacity of the system to facilitate comprehension of the meaning of the system elements).
  - *Clarity of the structure*, divided in turn into *formal* clarity (property of the system in terms of having its elements organized in a way that enables them to be perceived with clarity) and *conceptual* clarity (property of the system in terms of having its elements organized in a way that enables their meaning to be easily understood).

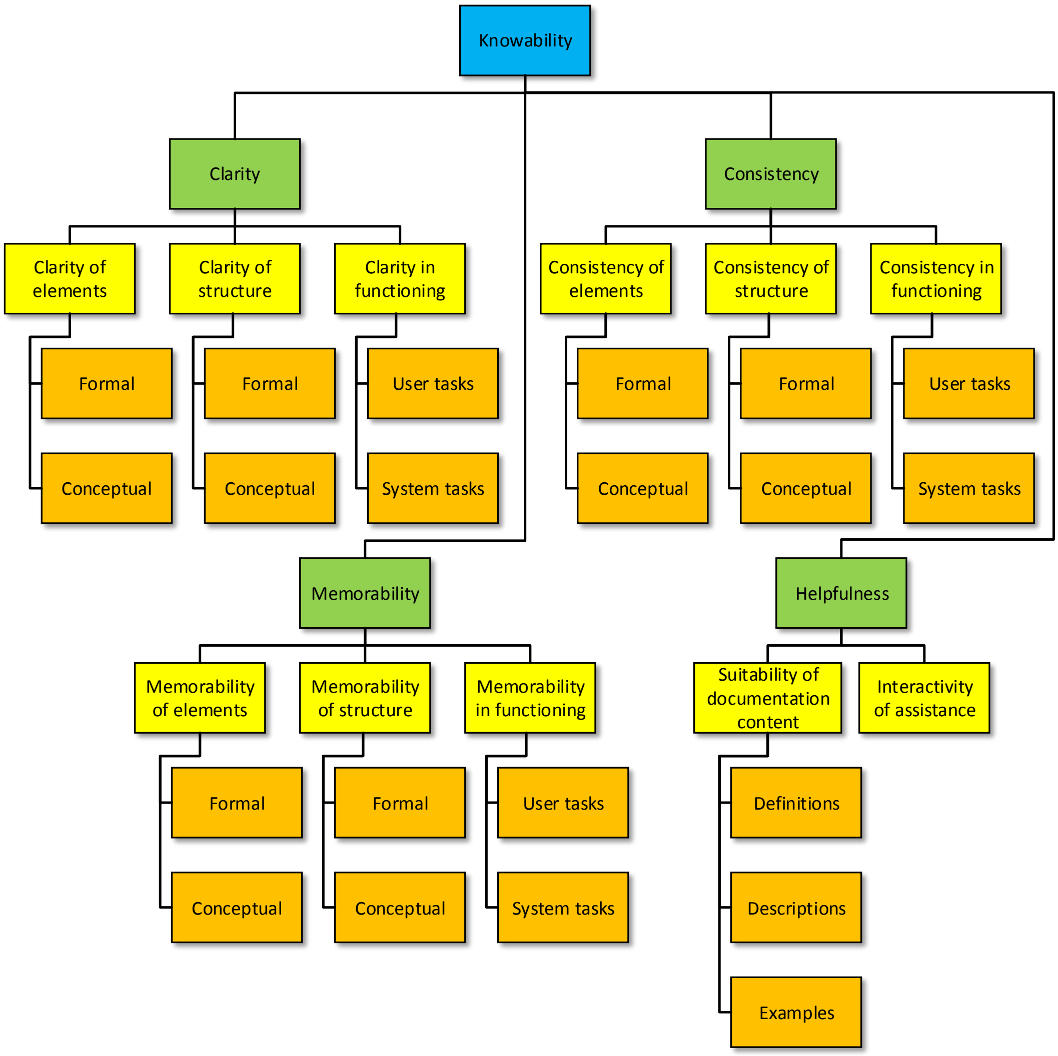


Figure 3.2: Knowability subattributes

  - *Clarity in functioning*, referring to both the way *user tasks* are performed and the way *system tasks* are automatically executed.

- **Consistency**, defined as system uniformity and coherence. It is subdivided in a similar way to clarity.

- **Memorability**, defined as the property of the system that enables the user to remember the elements and the functionality of the system. This attribute, like clarity and consistency, is also referred to in terms of individual *elements*, *structure*, and *functioning*.

- **Helpfulness**, defined as the means provided by the system to help users when they cannot infer or remember how to use the system. For this attribute a distinction is drawn between two aspects:

  - *Suitability of documentation content*, that is, content should be useful and adequate, bearing in mind that it includes *definitions*, *descriptions*, and *examples*.
  - *Interactivity of assistance*, that is, the extent to which the help provided by the system responds to the actions of the user.

The term *knowability* does not feature as an attribute in previously existing usability models in the literature. However, some models describe attributes that partially reflect related criteria, such as *learnability* (Abran et al. [11], ISO/IEC 9126-1 [90], Nielsen [129], Preece et al. [146], Quesenbery [147], Seffah et al. [167]), *memorability* (Nielsen [129]), and *understandability* (ISO/IEC 9126-1 [90]).

Furthermore, most of these models describe the attributes at a superficial level. Understandability, for example, is defined as the "capability of the software product to enable the user to understand whether the software is suitable, and how it can be used for particular tasks and conditions of use" [90, p. 9], but no distinction is drawn between different types of understandability, nor is there an indication of the specific characteristics that contribute to understandability. This contrasts with the taxonomy presented in this chapter, which offers detailed descriptions of the factors that determine the knowability of a system.

3.3.2 Operability

Operability is defined as the capacity of the system to provide users with the necessary functionalities and to permit users with different needs to adapt and use the system. It is divided into the following subattributes (see Figure 3.3):

- **Completeness**, defined as the capacity of the system to provide the functionalities necessary to implement the tasks intended by the user.
- **Precision**, defined as the capacity of the system to perform tasks correctly.
- **Universality**, defined as the extent to which the system can be used by all kinds of users. It is broken down as follows:
  - *Accessibility*, defined as the extent to which the system can be used by all kinds of users regardless of any physical or psychic characteristic they may have (e.g., disabilities, limitations, age, etc.). This attribute is subdivided into others in accordance with specific characteristics (*visual*, *auditory*, *vocal*, *motor*, and *cognitive*).
  - *Cultural universality*, defined as the extent to which users from different cultural backgrounds can use the system. In this sense, two features are identified, namely, *language* and other *cultural conventions* (use of symbols, measurement units, numeric formats, etc.).
- **Flexibility**, defined as the capacity of the system to adapt itself and to be adapted to different user preferences and needs. It has two distinct aspects:

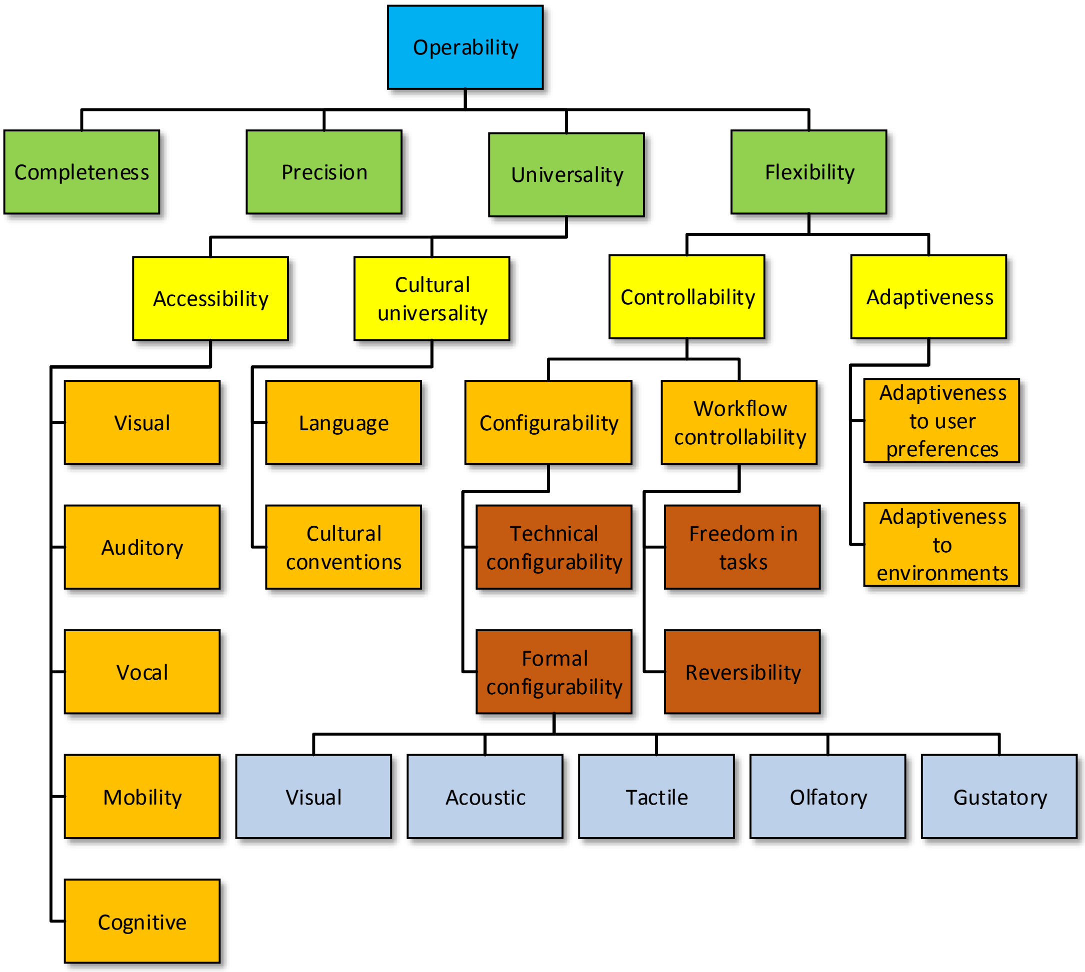


Figure 3.3: Operability subattributes

- *Controllability*, defined as the capacity of the system to permit users to choose the most appropriate way to use the system. A distinction is drawn between two subattributes:

  * *Configurability*, defined as the capacity of the system to permit users to personalize the system, with a distinction drawn between the configurability of *technical* aspects and of *formal* aspects.

  * *Workflow controllability*, defined as the capacity of the system to permit users to control tasks as they are implemented. This attribute includes *freedom in tasks* (i.e., controllability over the steps to be followed, allowing alternative approaches to performing tasks) and task *reversibility* (i.e., the system allows users to reverse actions).

- *Adaptiveness*, defined as the capacity of the system to adapt itself to *user preferences* and to different types of *environments*.

Most of the existing usability models do not include an attribute that is defined in equivalent terms as our *operability* concept. ISO/IEC 9126-1 [90] does refer to operability but describes it as related to aspects of *suitability* (coinciding partially with our completeness attribute), *error tolerance* (reflected in our robustness attribute, see Section 3.3.4), and *controllability.*

Our completeness and precision attributes are very similar to those that comprise the ISO 9241-11 [87] definition of *effectiveness* (i.e., the accuracy and completeness with which users achieve specified goals), which was subsequently adopted by Quesenbery [147], Abran et al. [11], and Seffah et al. [167].

The QUIM model by Seffah et al. [167] is the only usability model that includes attributes related to accessibility and cultural universality (although these authors simply refer to *universality*). Our taxonomy describes both these concepts in greater detail, including new kinds of disabilities, such as speech and cognitive handicaps, and specifying distinct types of cultural universality. Since both these attributes share the notion that a system should be capable of use by all potential users, they have been grouped under the universality attribute.

Preece et al. [146] referred to *flexibility*, briefly defined as "the extent to which the system can accommodate changes to the tasks and environments beyond those first specified." Our taxonomy is more precise in that it explicitly distinguishes between the system's own capacity to adapt itself and its capacity to be adapted.

3.3.3 Efficiency

Efficiency is the capacity of the system to produce appropriate results in return for the resources that are invested. The taxonomy for the branch of usability referring to efficiency (see Figure 3.4) reflects four subattributes:

- **Efficiency in human effort**, referring to the capacity of the system to produce appropriate results in return for the *physical* or *mental* effort that the user invests.
- **Efficiency in task execution time**, referring to the time invested by the *user* in performing actions and the time taken by the *system* to respond.
- **Efficiency in tied-up resources**, both *material* and *human.*
- **Efficiency in economic costs**, which refers to different types of expenses, namely, the cost of the *system* itself, *human resource* costs, the cost of the *equipment* that is required to work with the system, and the cost of *consumables.*

Although the term *efficiency* features in most of the usability models in the literature, our understanding of this concept is somewhat different. Nielsen described efficiency of use as referring to the level of performance achieved by expert users (users who have learned to use the system) and measured as "the time it takes users

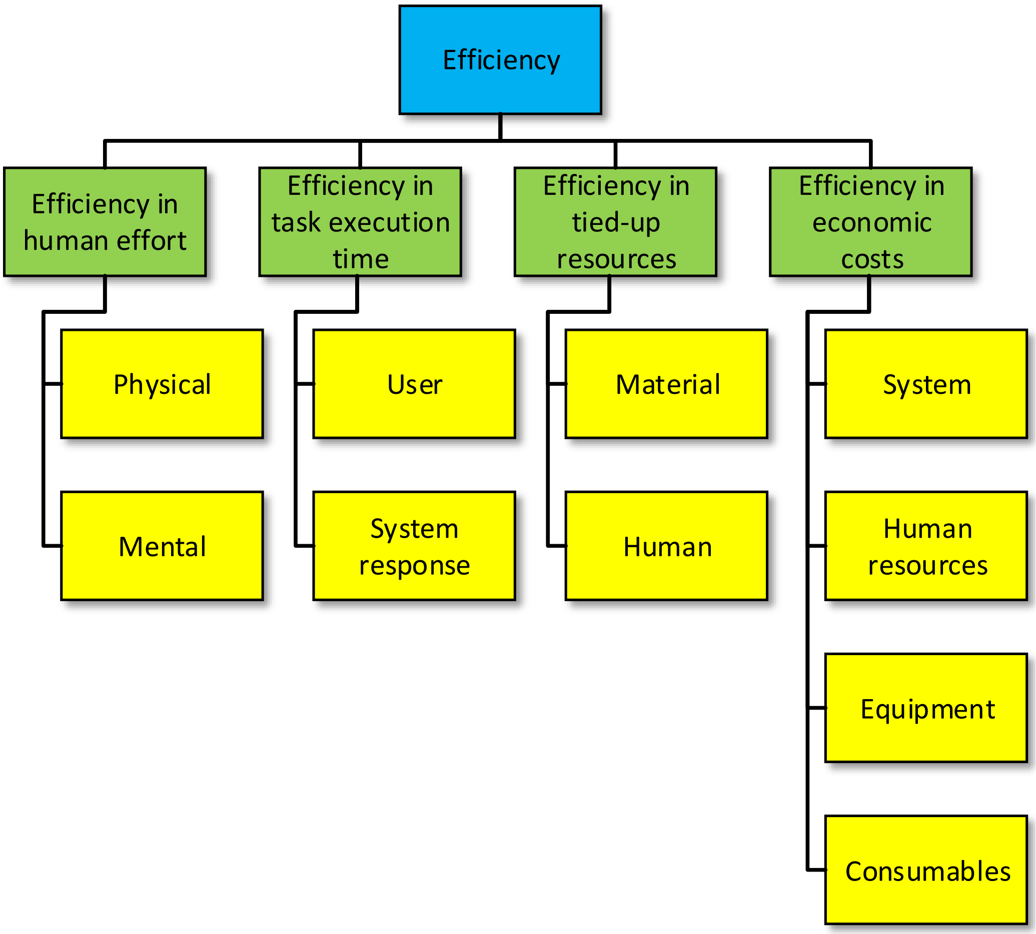


Figure 3.4: Efficiency subattributes

to perform some typical test tasks" [129, p. 31]. The usability model by Preece et al. [146], which also refers to users with experience, includes the attribute *throughput* but also considers new measures such as *tasks accomplished* and *errors made*. A key problem with these models is that they take into account only expert users, and thus ignore the fact that usability depends on the context of use. This means that a usability study of a system should consider the skills and knowledge of its intended users, which might not be proficient in its use.

ISO 9241-11 [87] presents the most complete conceptualization of efficiency. Even though the actual definition of efficiency provided by ISO 9241-11 is brief, Annex B offers some examples of measures of efficiency that include *task execution time*, *mental* and *physical effort*, *materials used*, and *financial cost*. Our definition of efficiency reflects all the aspects mentioned in ISO 9241-11 but extends it further by adding new subattributes.

3.3.4 Robustness

Robustness is defined as the capacity of the system to resist error and adverse situations. It is broken down into subattributes as follows (see Figure 3.5):

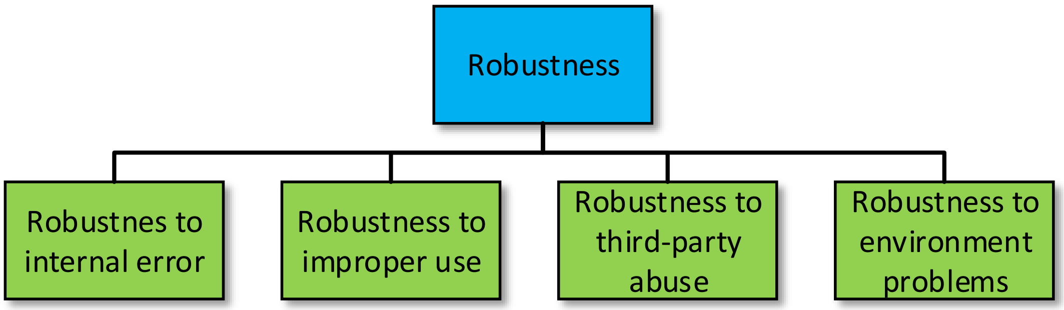


Figure 3.5: Robustness subattributes

- **Robustness to internal error.**
- **Robustness to improper use.**
- **Robustness to third-party abuse.**
- **Robustness to environment problems.**

None of the previous usability models in the literature include robustness as a main attribute. Nielsen [129] referred to *errors* in terms of the user making few errors when using the system and being able to easily recover from them. He drew a distinction between catastrophic errors and errors that just slow down the user's transaction rate (pointing out that the latter should be included in his efficiency of use attribute). The Quesenbery [147] model includes the attribute *error tolerance*, analogous to Nielsen's *errors* but taking into account errors that are not caused by the user (e.g., system failures).

The *robustness* attribute in this taxonomy differs somewhat from the attributes described by these other authors. First, different sources of error are specified; second, the meaning of the term is restricted to the system's capacity to resist adverse situations; and, finally, the need for the system to be able to recover by itself is also taken into account. Note also that in the *error tolerance* attribute, the Quesenbery model includes the capacity of the system to reverse actions. The taxonomy in this chapter, however, reflects this reversal capacity as an element of operability (see Section 3.3.2).

Other usability models also reflect robustness as an element in other attributes. ISO/IEC 9126-1 [90], for example, includes *error tolerance* in the *operability* attribute. Our taxonomy, however, proposes that robustness needs to be dealt with as a separate attribute, given that it covers aspects that go beyond the mere operation of the system, such as robustness against abuse by third parties and against problems in the environment.

3.3.5 Safety

Safety is defined as the capacity to avoid risk and damage derived from the use of the system. It is broken down into the following subattributes (see Figure 3.6):

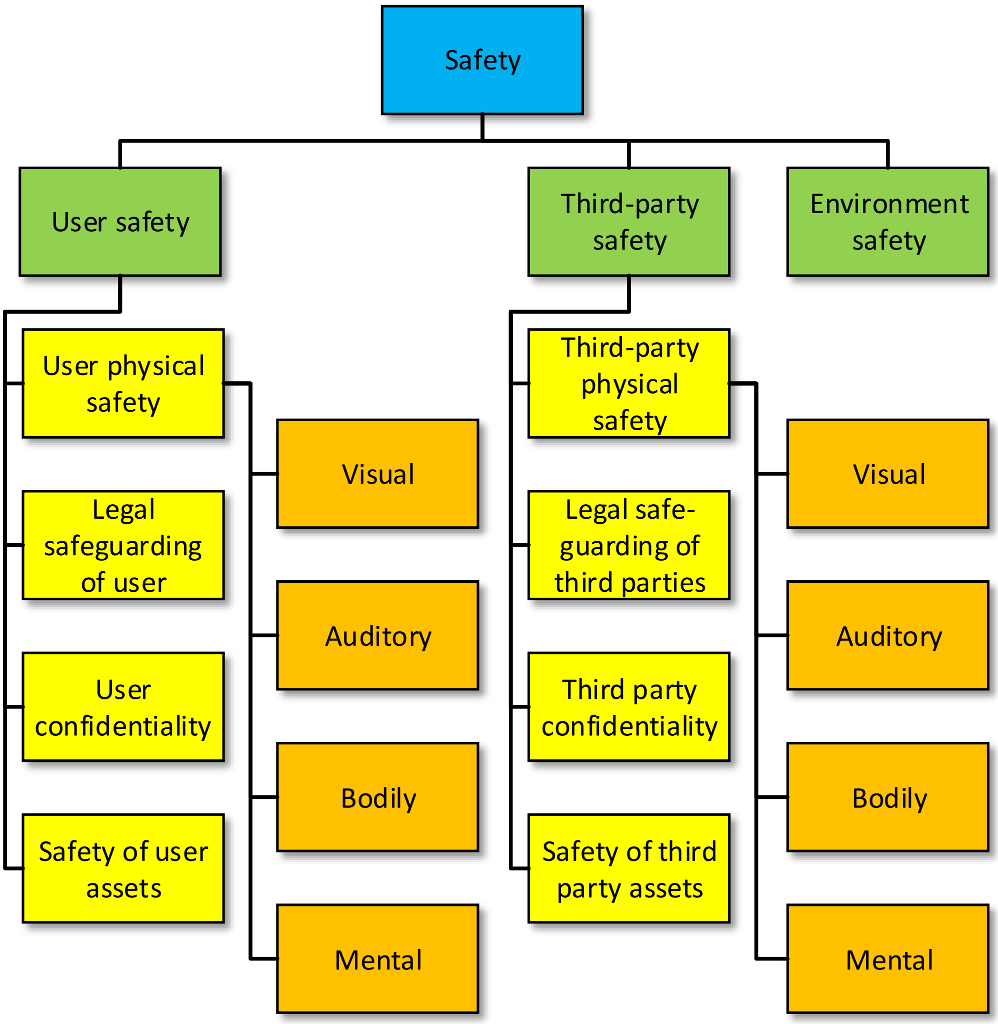


Figure 3.6: Safety subattributes

- **User safety**, defined as the capacity to avoid risk and damage to the user when the system is in use. Specifying risk or damage in more detail, a distinction is made between notions such as *physical safety*, *legal safeguarding*, *confidentiality*, and the *safety of the material assets* of the user.

- **Third-party safety**, defined as the capacity of avoiding risk and damage to individuals other than the user when the system is in use. This attribute is broken down analogously to user safety.
- **Environment safety**, defined as the capacity of the system to avoid risk and damage to the environment when being used.

Safety does not appear in most of the previously existing usability models in the literature. Although Abran et al. [11] and Seffah et al. [167] referred to the concepts of *security* and *safety*, respectively, these concepts remain only defined briefly. Furthermore, only physical harm or damage to people and resources is mentioned in their descriptions of specific types of safety. On the other hand, this taxonomy broadens safety to also include the capacity of the system to avoid any breach of the law or of the confidentiality rights of the user or of other individuals.

This comparison would not be complete without a mention to the *quality in use* model in the ISO/IEC 25010 standard [89]. Even though this standard is too recent to have taken part in the construction of the taxonomy, it is worth pointing out that it includes a first-level attribute named *freedom from risk*. This attribute is in turn subdivided into *economic risk mitigation*, *health and safety risk mitigation*, and *environmental risk mitigation*, which makes it somewhat similar to this taxonomy's *safety* attribute.

3.3.6 Subjective Satisfaction

Subjective satisfaction is the capacity of the system to produce feelings of pleasure and interest in users. It consists of two subattributes (see Figure 3.7):

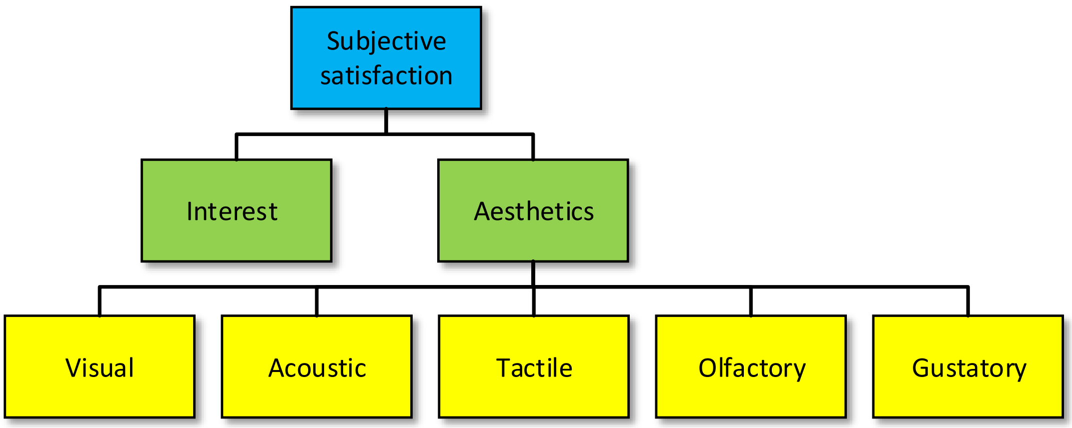


Figure 3.7: Subjective satisfaction subattributes

- **Interest**, defined as the capacity of the system to capture and maintain the attention and intellectual curiosity of the user.

- **Aesthetics**, defined as the capacity of the system to please its user in sensorial terms. Depending on the type of sensation, this attribute can be subdivided into *visual*, *acoustic*, *tactile*, *olfactory* and *gustatory* aesthetics.

All the usability models in the literature describe attributes that fully or partially reflect this concept. Preece et al. [146] defined the attribute *attitude* as the capacity of the system to produce positive feelings in the user. ISO/IEC 9126-1 [90] referred to the attribute *attractiveness* as the capacity of the system to be aesthetically pleasing to the user (for example, in terms of color use and graphic design). ISO 9241-11 [87], Nielsen [129], and Seffah et al. [167] gave their preferred term, *satisfaction*, a broader meaning by considering other subjective sensations such as, for example, the absence of discomfort when using the system and the capacity of the system to fulfill the aims of the user. Quesenbery [147] referred to the attribute *engagement* in a similar way, and mentioned aspects such as user interface interaction style (gamelike simulation, menu command, etc.).

In the taxonomy presented in this chapter, *subjective satisfaction* differs from the abovementioned attributes in two ways. First, although the freedom from discomfort is considered by some authors to comprise part of subjective satisfaction, it was consciously excluded from this taxonomy. Absence of discomfort is related to the absence of risk and damage, and is thus already covered by the *physical safety* attribute. Its inclusion here could therefore create redundancy in the taxonomy. Second, satisfaction is viewed as composed of two distinct concepts: satisfaction from an intellectual perspective and satisfaction from a sensorial perspective. The subdivision of satisfaction into these two dimensions is necessary in order to be able to generalize the concept to all types of systems. As can be appreciated from the literature (which focuses on IT systems), it is only possible to explore more specific kinds of satisfaction when the domain is restricted to a particular type of system.

3.4 Comparative Summary of the Taxonomies

Taking the first-level attributes in the taxonomy presented in this chapter as a baseline, Table 3.1 represents correspondences between the attributes described in the different usability models in the literature. It should be pointed out that in most cases there is no full equivalence between concepts, even though several models use the same or very similar terms to refer to similar attributes. The differences arise for a range of reasons: the meaning of attributes is not identical, attributes may correspond to different levels of granularity, and what is rated an attribute in one model may be a metric or guideline in another model. Finally, an added difficulty arises in the fact that terms are often defined ambiguously.

A number of terms feature in most of the models, namely, *satisfaction*, *efficiency*, *learnability*, and *effectiveness*. The first two concepts are present in the taxonomy; the second two are also included but form part of the *knowability* and *operability* concepts, respectively, both of which are more complete.

Other attributes that appear in several models, but with less frequency, are

Table 3.1: First-Level Usability Attributes from our Taxonomy Featuring in Other Models

	Attributes in other models						
Proposed attributes	*ISO 9241-11 (1998)*	*ISO/IEC 9126-1 (2001)*	*Nielsen (1993)*	*Preece et al. (1994)*	*Quesenbery (2001)*	*Abran et al. (2003)*	*Seffah et al. (2006)*
Knowability		Understandability; learnability	Learnability; memorability	Learnability	Ease of learning	Learnability	Learnability
Operability	Effectiveness	Operability		Flexibility	Effectiveness	Effectiveness	Effectiveness; usefulness; accessibility; universality
Efficiency	Efficiency		Efficiency	Throughput	Efficiency	Efficiency	Efficiency; productivity
Robustness			Errors		Error tolerance		
Safety						Security	Safety
Subjective satisfaction	Satisfaction	Attractiveness	Satisfaction	Attitude	Engagement	Satisfaction	Satisfaction
		Usability compliance					

robustness (or the related concept, *errors*) and *safety* (or *security*). These attributes, incidentally, have not been included in the two ISO standards.

ISO/IEC 9126-1 [90] includes, however, the *usability compliance* attribute. This does not feature as an attribute in the taxonomy, given that it consists of aspects that are covered by a range of standards, conventions, style guides, and regulations that already form part of attributes – such as *accessibility*, *cultural universality*, *consistency*, and so forth – already included in the taxonomy. In other words, the creation of a new attribute for this concept is not justified.

This chapter ends with some tables that list the attributes described in the models just analyzed that explicitly refer to second-level concepts in our taxonomy, that is, the subattributes of *knowability* (Table 3.2), *operability* (Table 3.3), *efficiency* (Table 3.4), *robustness* (Table 3.5), *safety* (Table 3.6), and *subjective satisfaction* (Table 3.7). From these tables, two things become more apparent at this level of detail: first, the lack of consensus among the different models, and second, the new concepts covered by the taxonomy that are not explicitly included in other usability models.

Table 3.2: Knowability Subattributes from our Taxonomy Featuring in Other Models

	Attributes in other models						
Proposed attributes	*ISO 9241-11 (1998)*	*ISO/IEC 9126-1 (2001)*	*Nielsen (1993)*	*Preece et al. (1994)*	*Quesenbery (2001)*	*Abran et al. (2003)*	*Seffah et al. (2006)*
Clarity					Ease of learning		Learnability
Consistency					Ease of learning		Learnability
Memorability			Memorability				
Helpfulness		Understandability			Ease of learning		Learnability

Table 3.3: Operability Subattributes from our Taxonomy Featuring in Other Models

	Attributes in other models						
Proposed attributes	*ISO 9241-11 (1998)*	*ISO/IEC 9126-1 (2001)*	*Nielsen (1993)*	*Preece et al. (1994)*	*Quesenbery (2001)*	*Abran et al. (2003)*	*Seffah et al. (2006)*
Completeness	Effectiveness	Operability			Effectiveness	Effectiveness	Effectiveness
Precision	Effectiveness				Effectiveness	Effectiveness	Effectiveness; usefulness
Universality							Accesibility; universality
Flexibility		Operability		Flexibility			Effectiveness; trustfulness; usefulness

Table 3.4: Efficiency Subattributes from our Taxonomy Featuring in Other Models

	Attributes in other models						
Proposed attributes	*ISO 9241-11 (1998)*	*ISO/IEC 9126-1 (2001)*	*Nielsen (1993)*	*Preece et al. (1994)*	*Quesenbery (2001)*	*Abran et al. (2003)*	*Seffah et al. (2006)*
Efficiency in human effort	Efficiency					Efficiency	Efficiency; usefulness
Efficiency in task execution time	Efficiency		Efficiency	Throughput	Efficiency	Efficiency	Efficiency; productivity; usefulness
Efficiency in tied up resources							Efficiency; productivity; usefulness
Efficiency in economic costs	Efficiency					Efficiency	

Table 3.5: Robustness Subattributes from our Taxonomy Featuring in Other Models

	Attributes in other models						
Proposed attributes	*ISO 9241-11 (1998)*	*ISO/IEC 9126-1 (2001)*	*Nielsen (1993)*	*Preece et al. (1994)*	*Quesenbery (2001)*	*Abran et al. (2003)*	*Seffah et al. (2006)*
Robustness to internal error					Error tolerance		
Robustness to improper use		Operability	Errors		Error tolerance		
Robustness to third party abuse							Trustfulness
Robustness to environment problems							

Table 3.6: Safety Subattributes from our Taxonomy Featuring in Other Models

Proposed attributes	*Attributes in other models*						
	ISO 9241-11 (1998)	*ISO/IEC 9126-1 (2001)*	*Nielsen (1993)*	*Preece et al. (1994)*	*Quesenbery (2001)*	*Abran et al. (2003)*	*Seffah et al. (2006)*
User safety							Safety; trustfulness
Third party safety							Safety; trustfulness
Environment safety							

Table 3.7: Subjective Satisfaction Subattributes from our Taxonomy Featuring in Other Models

Proposed attributes	*Attributes in other models*						
	ISO 9241-11 (1998)	*ISO/IEC 9126-1 (2001)*	*Nielsen (1993)*	*Preece et al. (1994)*	*Quesenbery (2001)*	*Abran et al. (2003)*	*Seffah et al. (2006)*
Interest	Satisfaction		Satisfaction	Attitude	Engagement	Satisfaction	Satisfaction
Esthetics	Satisfaction	Attractiveness	Satisfaction		Engagement	Satisfaction	Satisfaction

Chapter 4

A Taxonomy of Context-of-Use Attributes

4.1 Introduction

Analogously to the usability taxonomy introduced in Chapter 3, this chapter presents a context-of-use taxonomy that was specifically developed for this thesis. The context-of-use taxonomy complements the usability taxonomy in the sense that usability is, by definition [87], always relative to a specified context of use. Once more, the relative vagueness of the context-of-use models in the literature and the lack of consensus on the precise meaning of the term (see Section 2.2) were the motivations for creating a synthesis and refinement of the term.

The chapter is structured as follows: (1) The construction process is outlined; (2) the taxonomy is described in detail by specifying the full hierarchy of subattributes and providing definitions for each term; (3) the taxonomy is compared with other context-of-use models in the literature in terms of the attributes that are covered.

4.2 The Construction Process

The construction process of the context-of-use taxonomy was basically similar to that of the usability taxonomy described in Chapter 3, and took place in two main phases. In the first phase an exhaustive analysis of existing context-of-use models was conducted. Other publications that referred to the importance of certain context-of-use factors were also analyzed, even if they did not refer explicitly to the context of use [71][158]. A point was made, however, that our conceptualization of context of use had to be strongly connected to our conceptualization of usability, which meant adhering to the definition of context of use in usability standards such as ISO 9241-11 and related publications.

The context-of-use models that were chosen for constructing the taxonomy were as follows:

- **Bevan and Macleod** (1994) [31].
- **Thomas and Bevan** (1996) [179].
- **ISO 9241-11** (1998) [87].
- **Kirakowski and Cierlick** (1999) [103].
- **Maguire** (2001) [121].

These models are described in detail in Section 2.2. As with the usability taxonomy, the main criteria for inclusion were relevance (particularly with respect to ISO standards, and also including some authors who were also explicitly involved in ISO standards, as explained in Section 2.2) and non-redundancy (i.e., adding something new to traditional definitions).

These context-of-use models laid the groundwork for compiling an initial set of attributes for our taxonomy. Attributes were then successively reviewed as to refine them further, whether by adding attributes, removing attributes, or restructuring hierarchical relationships between them. As with the construction process of the usability taxonomy, no methodology was explicitly spelled out, but there was a clear focus in prioritizing completeness while avoiding redundancy and overlap between meanings.

Of the new attributes added, these were necessary to ensure that the taxonomy reflected an adequate level of detail. We thus separated out aspects that, although similar, had different impacts on usability; we added or, in some cases, modified other attributes to ensure that we complied with our aim of developing a taxonomy that was suitable for any kind of product or system. Context of use, like usability, is a concept that affects all types of products and systems, yet the models in the literature focus almost exclusively on IT systems.

Because we wished to include only attributes that were relevant to usability, it was decided to dispense with some attributes. For example, the attributes list initially included the attribute *user age*, which is widely accepted as a relevant attribute in the literature. We decided to exclude this attribute, however, because we are of the opinion that it is not age but the physical and mental characteristics accompanying age that determine the need to attribute greater importance to certain usability attributes. Similarly, other attributes were excluded to eliminate redundancy within the taxonomy and to minimize the possibility that any one attribute might reflect or contradict another attribute.

Finally, the reason for restructuring the hierarchy was that successive revisions revealed criteria for creating more appropriate categories that would enable related attributes to be grouped together in isolation from unrelated attributes.

4.3 Description of the Taxonomy

The taxonomy resulting from this construction process has the first-level attributes depicted in Figure 4.1. At this broad level, the attributes are those commonly accepted in the literature on usability, namely, the system *users*, the *tasks* performed by the users, and the *environments* in which the system is used. They are described in detail next.

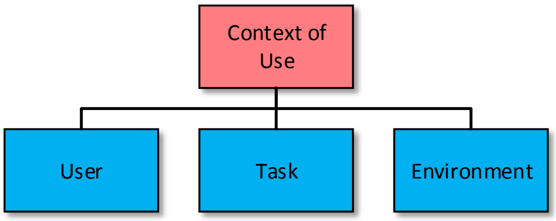


Figure 4.1: Context-of-use attributes

4.3.1 User

A *user* is a person who interacts directly or indirectly with the system. The user attributes that are relevant to usability (Figure 4.2) are listed as follows:

- **Role** is about how the user functions in the interaction with the system, whether in a *direct* operation role (performing typical system tasks), in a *support role* (carrying out maintenance or installation tasks, etc.), in a *monitoring role* (supervising the work of system users), or in an *indirect role* (as a user affected by the interaction of another user with the system, because, e.g., the former provides inputs to or needs outputs from the system).
- **Experience** refers to the practical skills and knowledge of the user in relation to the system, with a distinction drawn between experience *with the system* and experience *with similar systems*.
- **Education** is the knowledge the user has acquired not through using the system but through formal education or from the social, cultural, or organizational environment. This attribute is further broken down as follows:
  - *Educational background* refers to general (rather than system-related) knowledge, acquired through instruction or study.
  - *Knowledge of system domain* is the familiarity with the field to which the system pertains.
  - *Knowledge of system language* refers to the ability of the user to understand the linguistic system of communication used by the system (which can be verbal, speech, tactile, etc.).

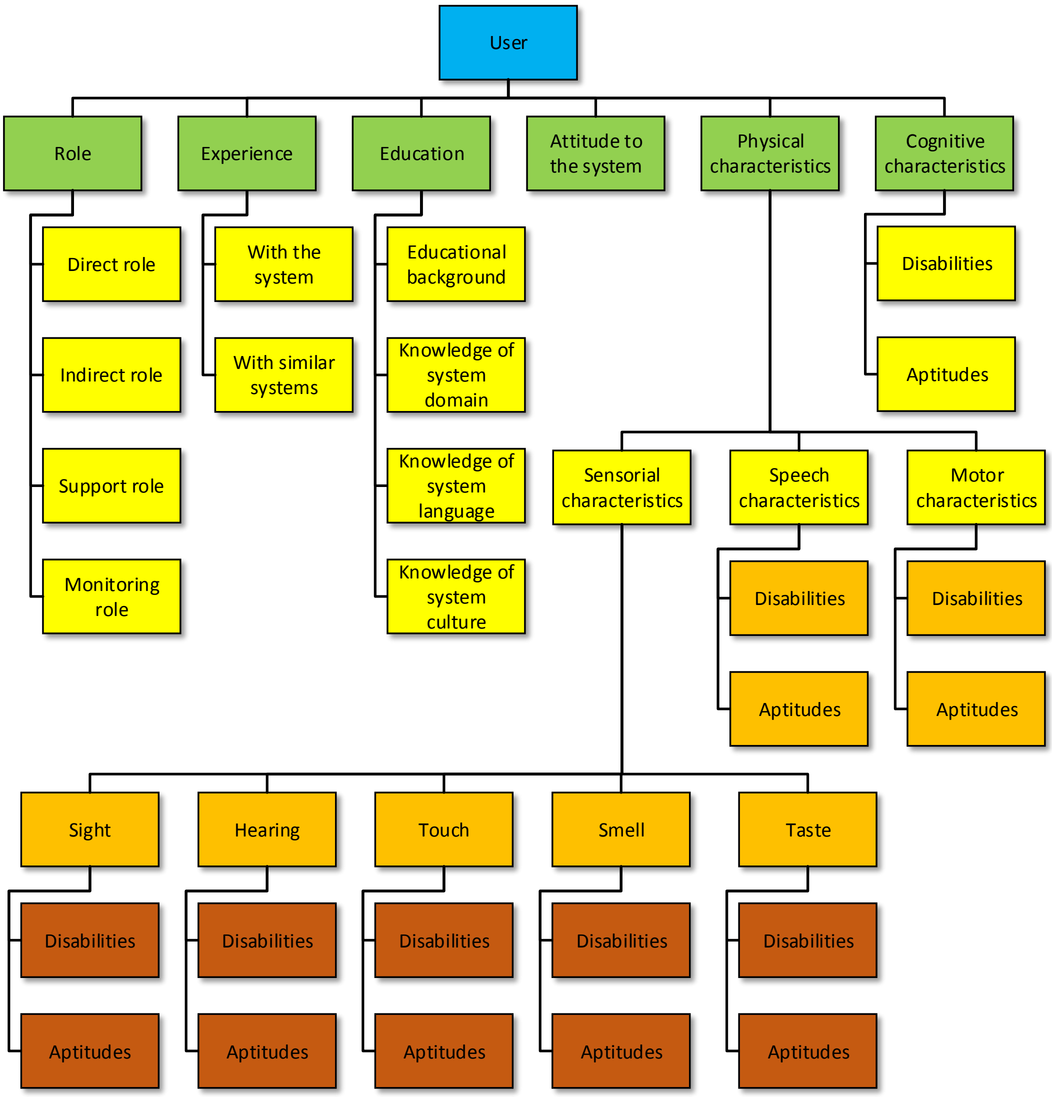


Figure 4.2: User subattributes

- *Knowledge of system culture* is the ability of the user to understand the cultural conventions used by the system (e.g., symbols, formats, metaphors).

- **Attitude to the system** refers to the feelings or emotions experienced by the user when operating the system.

- **Physical characteristics** describes the characteristics of the user's body that have an impact on usability, namely, *sensorial characteristics*, *speech characteristics*, and *motor characteristics*. Each characteristic is described in terms of *disabilities* (functional limitations of the user) and *aptitudes* (optimal performance of functions by the user). By drawing these distinctions, our taxonomy reflects specific situations in which the same physical characteristic simultaneously presents a disability and an aptitude, such as, e.g., a user who is insensitive to certain colors but who has great visual acuity. The *physical characteristics* attribute is divided into the following categories:

- *Sensorial characteristics* draws a distinction between the five senses (with each further broken down into disabilities and aptitudes), to wit:
  - *Sight* (disabilities include *complete* or *partial loss of vision*, *insensitivity to certain colors*, and *color blindness*).
  - *Hearing* (disabilities include *complete* or *partial loss of hearing*).
  - *Touch* (disabilities include *complete* or *partial loss of feeling*).
  - *Smell* (disabilities include *complete* or *partial loss of smell*, and *insensitivity to certain smells*).
  - *Taste* (disabilities include *complete* or *partial loss of taste*, and *insensitivity to certain tastes*).
- *Speech characteristics*, similarly subdivided into:
  - *Disabilities*, including *complete loss of speech*, *speech disorder* (referring to problems with volume, tone, resonance, etc.), *articulation disorder* (incorrect pronunciation of sounds), and *fluency disorder* (a lack of uniformity or continuity in speech).
  - *Aptitudes*, covering aspects such as *voice quality*, *quality of articulation*, and *fluency.*
- *Motor characteristics*, also subdivided into:
  - *Disabilities*, referring to muscular or skeletal disorders, or problems with central nervous system structures, that restrict user movements and can be classified as follows: *complete* or *partial loss of movement*, *loss of control over movements*, and *altered body shape* (whether in terms of posture or size). In addition, each of these disabilities can be further classified in more specific terms according to the body part affected (right or left arm, hand, leg or foot, trunk, neck, etc.).
  - *Aptitudes*, broken down further into *physical strength*, *physical endurance*, *physical speed*, *physical flexibility*, and *physical coordination*

• **Cognitive characteristics** refers to the mental characteristics of the user, with a distinction also drawn between:

- *Disabilities*, broken down into *reasoning*, *memory*, *dyslexia*, *dyscalculia*, *attention capacity*, *mental health*, and *epilepsy.*
- *Aptitudes*, broken down into *conceptual reasoning*, *memory*, *linguistic*, *mathematical*, and *attention capacity aptitudes.*

Although many of our first-level user attributes are included in the models from the literature, they are neither structured in the same way nor described in terms of the same set of subattributes. As one example, the models in the literature include a broad-based attribute referring to user *experience*, *knowledge*, and *skills.* We, however, draw a distinction between experience in using the system (*experience*) and experience and knowledge that will help the user yet do not arise directly from using the system (*education*). Some authors refer to *experience with similar systems* but, rather than give this attribute the status of an independent attribute, it is confusingly considered as part of product experience [103][179]. However, we consider this aspect to have a differentiating effect on usability and so consider it to warrant treatment as an attribute in its own right.Within *education*, furthermore, we include attributes that have been included in none or only some of the other models, such as *educational background*, *knowledge of system domain* (not included in any model), and *knowledge of system culture* (only partially mentioned in the model by Thomas & Bevan [179]). Because these aspects affect the ease with which a user will learn to use a system, they need to be included in a context-of-use taxonomy.

All the context-of-use models in the literature consider physical attributes to be important in determining how a user will operate a system. Some authors specify physical attributes by means of a subattribute referring to *limitations and disabilities* [179], other authors distinguish between *limitations* and *capabilities* [31][87], whereas other authors prefer to treat *limitations and capabilities* as a single attribute [103][121]. Our taxonomy takes a more refined view of the physical attributes of the user, first by distinguishing between two specific issues that have different usability outcomes, namely, the capacity to perceive the system (*sensorial characteristics*) and the capacity to communicate commands to the system by means of physical actions (*speech characteristics* and *motor characteristics*). For each physical attribute, therefore, we distinguish between *disabilities* that could prevent the user from perceiving or executing actions in the system, and *aptitudes* that determine effective use of the system. Our taxonomy also differs in that we describe specific types of disabilities and aptitudes. Furthermore, there is a consensus in the literature in regard to classifying *age* and *gender* as physical or personal attributes, with some authors claiming that it is more difficult for older people to learn to use a system. We have chosen to exclude age and gender from our taxonomy, however, because they do not by themselves affect usability. What really concerns us here are characteristics such as educational background and mental and physical aptitudes, which are not necessarily derived from age or gender and are thus better covered by independent taxonomy attributes.

We applied a similar approach to defining subattributes for *cognitive characteristics* – some models exclude these altogether [103] or simply consider *ability* [31][179][87]. We considered it necessary, moreover, to draw a distinction again between *disabilities* and *aptitudes*, and furthermore, to further specify types of mental functions, given that *reasoning*, *memory*, and so forth, all have different effects on usability.

Finally, the models in the literature tend to combine *attitude toward the system* with personal or mental attributes. We consider these separately, however, because what is referred to is not an intrinsic property of the user's mind but an initial user reaction to a particular system.

4.3.2 Task

A task is a piece of work that the user carries out by interacting with the system. Figure 4.3 represents the branch of our taxonomy covering task attributes, described as follows:

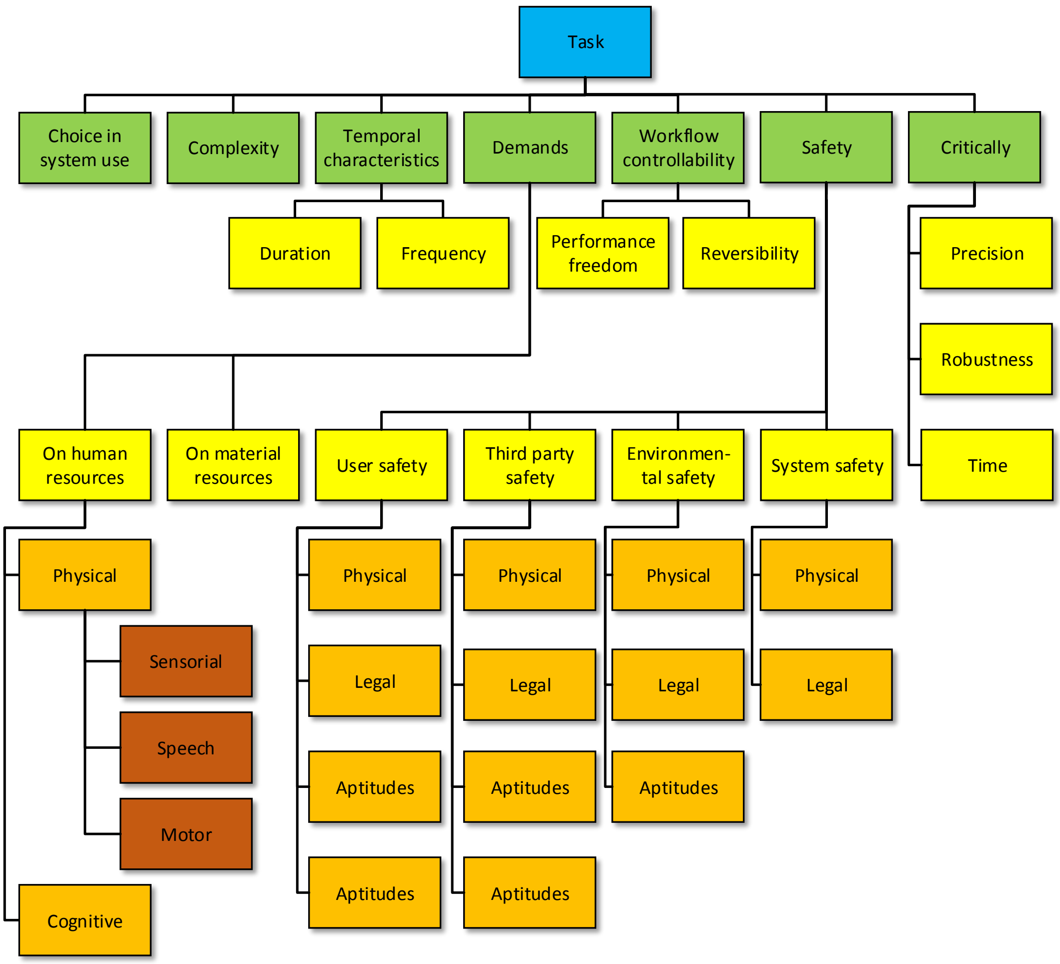


Figure 4.3: Task subattributes

- **Choice in system use** is the extent to which users may choose whether or not to use the system to complete the task.
- **Complexity** is defined as the degree to which completion of the task is difficult for the user.
- **Temporal characteristics** as an attribute includes both task *duration* and task *frequency*.
- **Demands** refers to the resources necessary to complete the task successfully. For this attribute we propose two subattributes to distinguish between demands *on human resources* and demands *on material resources*:

- *Demands on human resources*:
    - *Physical demands* refers to requirements related to the following groups of characteristics: *sensorial* (*sight*, *hearing*, *touch*, *smell*, and *taste*), *speech* (*voice quality*, *articulation*, and *fluency*), and *motor* (*strength*, *endurance*, *speed*, *flexibility*, and *coordination*).
    - *Cognitive demands* covers *reasoning*, *memory*, *linguistic*, *mathematical*, and *attention capacity* demands.
- *Demands on material resources* refers to the physical resources (equipment, consumables, etc.) necessary for completing the task.

- **Workflow controllability** refers to the extent that the task can be controlled by the user during implementation. Within this attribute a distinction is drawn between:
    - *Performance freedom*, which refers to the extent to which there are alternative ways to complete the task.
    - *Reversibility*, which refers to the possibility for undoing actions and returning to a previous state.

- **Safety** refers to the degree to which the task as implemented does not cause damage or risk. This can affect *users*, *third parties*, the *environment*, or the *system* itself. Each type of safety risk is associated with a different subattribute:
    - *Physical safety.* Depending on the circumstances, a distinction is drawn between *user physical safety*, *third-party physical safety*, *environmental physical safety*, and *system physical safety.*
    - *Legal safety* is the extent to which task performance does not incur legal problems. A distinction is drawn between *user legal safety*, *third-party legal safety*, and *environmental legal safety.*
    - *Confidentiality* refers to the degree to which implementation of the task does not incur any risk of unauthorized access to personal or organizational data. Depending on who owns the information, a distinction is drawn between *user confidentiality*, *third-party confidentiality*, and *environmental confidentiality.*
    - *Safety of material assets* refers to the extent that implementation of a task does not damage property, with a distinction drawn between *user property safety* and *third-party property safety.*

- **Criticality** refers to the extent to which performance of the task is decisive. Depending on the aspect of the task in question, this attribute can be broken down into:
    - *Precision criticality* refers to the level of accuracy required in the task.
    - *Robustness criticality* refers to the importance of the task being resistant to error and adverse circumstances.

- *Time criticality* refers to the speed with which the task is required to be completed.

Many of the existing models do not reflect the attributes that we propose for *task*. For example, *choice in system use* only appears in the models by Thomas and Bevan [179] and Kirakowski and Cierlick [103], and *complexity* only appears in the Kirakowski and Cierlick model. Furthermore, although all the models have an attribute similar to *demands*, our definition is more exhaustive; first, we specify several types of *physical* and *cognitive demands on human resources* and, second, we refer to *demands on material resources*.

Workflow controllability does not correspond directly with any attributes in the existing models. Only one of its components, namely, *performance freedom*, is similar to the *task flexibility* attribute included in all the models and described as the extent to which the user can decide on how to sequence actions and whether any of these actions are governed by a time constraint [103]. Our attribute refers to the fact that the user can both choose actions from a set of alternatives and choose the order of actions. Our second *workflow controllability* component, that is, *reversibility*, does not feature in any of the other models.

Although the *safety* attribute features in most of the models, it does not always have the same meaning. In Thomas and Bevan [179], it refers to the safety of the task in relation to the health of the user and other persons, whereas in Kirakowski and Cierlick [103] the meaning is broadened to also cover the safety of the informational environment. ISO 9241-11 [87] and Maguire [121] both include the attribute *safety-critical demands*, but fail to explain the scope of this attribute. We propose, in addition to *physical safety*, further concepts such as *legal safety*, *confidentiality*, and *safety of material assets*. We also consider safety to refer not only to persons, but also to the system itself and to the environment over and above the informational environment.

Only some of the models include attributes or subattributes that correspond, even partially, to the *criticality* subattributes. Kirakowski and Cierlick [103] proposed *degree of precision required in output*, which is very similar to our *precision criticality*. The models by Bevan and Macleod [31], ISO 9241-11 [87], and Maguire [121] include *risk resulting from error*, and the models by Thomas and Bevan [179] and by Kirakowski and Cierlick include *criticality of the task output*, defined as the importance of a task being completed. Our *robustness criticality* attribute is different from both attributes, however. It differs from *risk resulting from error* in that it reflects the fact that a task should be resistant both to error and to any adverse circumstances in the environment, and it differs from *criticality of the task output* in that, besides indicating that a task is completed, it also determines that task integrity is maintained. Finally, Kirakowski and Cierlick proposed the attribute *task postponement*, referring to the possibility for postponing the termination of a task. Our *time criticality* attribute, however, reflects not only the fact that a task may be postponed but also the extent to which postponement may have negative repercussions.

4.3.3 Environment

The *environment* consists of the external factors that affect the use of the system. For the *environment* attribute (Figure 4.4) we distinguish between the *physical environment* (the surroundings and space in which the user operates the system), the *social environment* (the people with whom the user interacts and who effect the user's interaction with the system), and the *technical environment* (the technical equipment and infrastructures that support the functioning of the system).

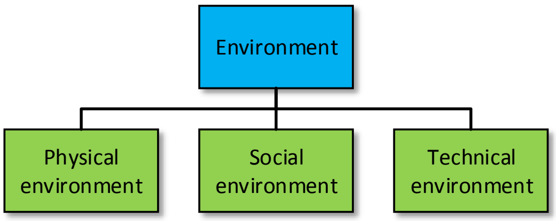


Figure 4.4: Environment subattributes

Physical environment

The attributes that refer to the physical environment (see Figure 4.5) are described as follows:

- **Sensorial conditions** refers to the characteristics of the physical environment that affect the sensorial perceptions of the user. For this attribute a distinction is drawn between:
  - *Sensorial quality* is the extent to which the sensorial conditions are appropriate for using the system. This subattribute, in turn, is divided into *visual quality*, *auditory quality*, *tactile quality*, *olfactory quality*, and *gustatory quality*, reflecting the degree to which the environment enables satisfactory perception in terms of each of these senses.
  - *Sensorial stability* refers to the extent to which sensorial conditions do not change frequently. This is also broken down according to the five senses into *visual stability*, *auditory stability*, *tactile stability*, *olfactory stability*, and *gustatory stability*.

- **Atmospheric conditions** refers to the characteristics of the air. A distinction is drawn between:
  - *Atmospheric quality* is the extent to which the properties of the air are acceptable for working with the system. This is broken down into air *purity*, *humidity*, *temperature*, and *air movement*.

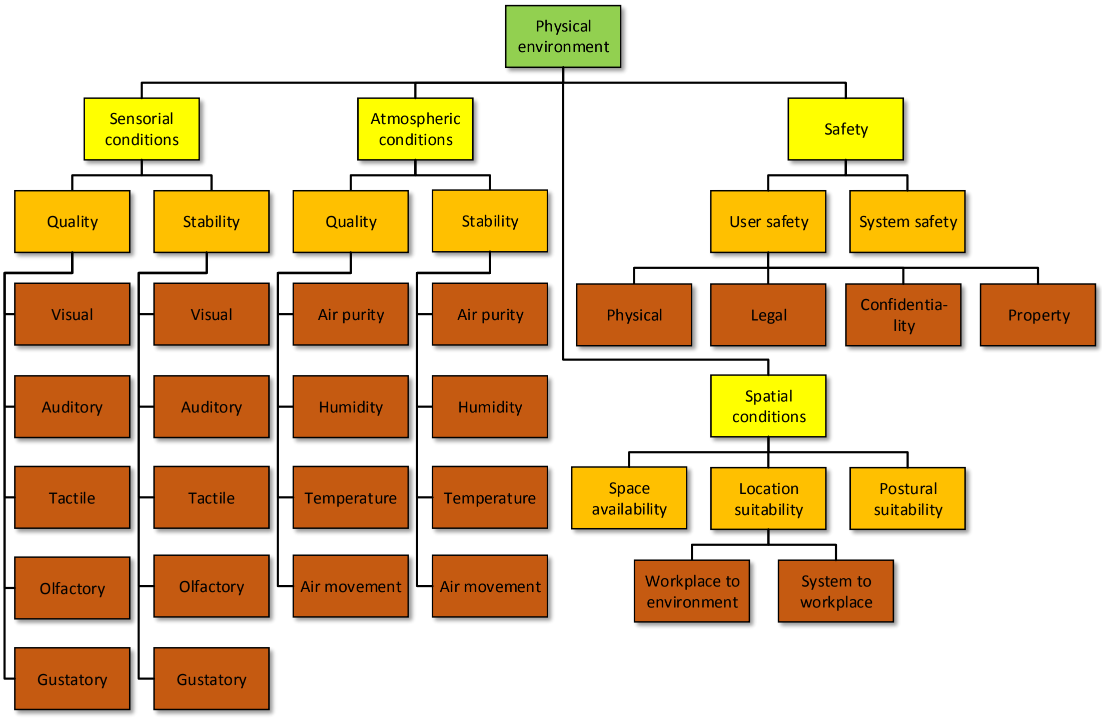


Figure 4.5: Physical environment subattributes

- *Atmospheric stability* is the extent to which atmospheric quality does not change frequently. This attribute is further broken down into *air purity stability*, *humidity stability*, *temperature stability*, and *air movement stability*.

- **Spatial conditions** refers to physical location within the environment in which the system is used. A distinction is drawn between:
  - *Space availability* refers to having the space necessary to operate the system.
  - *Location suitability* is further broken down into *workplace suitability in relation to the environment* and *system suitability in relation to the workplace*.
  - *Postural suitability* refers to the degree to which the work space permits the user's body to work in an adequate position.

- **Safety** refers to the degree to which the physical environment does not cause damage or risk. Depending on the circumstances, a distinction is drawn between the following subattributes:
  - *User safety* refers to *physical safety*, *legal safety*, *confidentiality*, and *property safety*.

  - *System safety* refers to the degree to which the environment does not cause damage to the system.

All the models in the literature include attributes for *auditory environment* and *visual environment*. Our taxonomy, however, is more comprehensive in that it includes a general attribute for *sensorial conditions* that is subdivided on the basis of the five senses.

Most models include an attribute for *atmospheric conditions*, and all of them include a separate attribute for *thermal environment*. However, these attributes are defined sketchily and only Thomas and Bevan [179] tried to provide examples of specific atmospheric conditions. We decided then to add subattributes for *air purity*, *humidity*, and *air movement*, and to include temperature in the same group for the sake of clarity.

All the models in the literature include an attribute for *environmental instability*. However, they do not define it, and therefore it is impossible to know exactly what they mean. Again, only Thomas and Bevan [179] tried to provide examples (in this case, related only to vibration and motion). We aimed to be more comprehensive and specific in terms of usability and thought it necessary to define a *stability* attribute for every subattribute of *sensorial* and *atmospheric conditions*.

For the *spatial conditions* attribute, whereas other models propose workplace design aspects such as furniture,we restrict ourselves again to the properties of these attributes that interest us from a usability perspective, namely, *space availability*, *location suitability*, and *postural suitability*. Also, even though all models include the *location* attribute, they actually mean different things. Only Maguire [121] covered, to some extent, our two meanings: *the workplace in relation to the environment*, and *the system in relation to the workplace*.

Finally, all the models in the literature include a *workplace safety* attribute, referring exclusively to risks to health. Our concept of safety, however, is broader, covering as it does both users and the system itself. Furthermore, as far as users are concerned, it is not limited to just *physical safety* but also includes *legal safety*, *confidentiality*, and *user property safety*.

Social environment

The *social environment* is described by means of the following attributes (see Figure 4.6):

- **Work relations** refers to interactions between members of the organization within which the system is used. This attribute has the following subattributes:
  - *Team work* is defined as the extent to which the user interacts with other persons while working.
  - *Human support* is the extent to which the user can count on help from other persons in the organization.

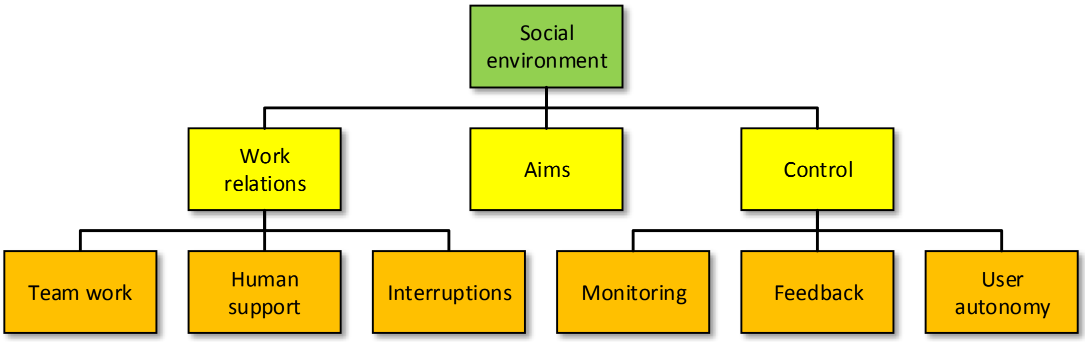


Figure 4.6: Social environment subattributes

- *Interruptions* refers to the degree to which the work of a user is interrupted by other persons in the organization.

- **Aims** refers to the intentions of the organization in regard to user interactions with the system.
- **Control** is the degree to which the organization checks and controls the work of the user so as to ensure a suitable level of productivity and quality. This can be broken down further as follows:
  - *Monitoring* refers to supervision of the user's work by the organization.
  - *Feedback* refers to information provided to the user in relation to the work with the system.
  - *User autonomy* refers to the freedom granted to the user in terms of how to implement tasks.

The subattributes that we propose for *work relations* are broadly accepted by the models in the literature, although there is a difference in how they are grouped – in some cases, under the attribute structure of the *social environment* or *structure of the organizational environment* (Bevan & Macleod [31], ISO 9241-11 [87], Maguire [121], Thomas & Bevan [179]) and, in other cases, directly under *social* or *organizational environment* (Kirakowski & Cierlick [103]).

Likewise, the subattributes for *user control* are organized in a different way. In some models, they are grouped under *job design* (Bevan & Macleod [31],ISO 9241-11 [87], Maguire [121]). In others, *autonomy* is considered to be a subattribute of *task* rather than of *social environment* (Kirakowski & Cierlick [103]).

Although the attribute *aims* appears in existing models, it generally differs from our concept, however, in that it has a broader meaning and is not specific to aims associated with interaction with the system (Thomas & Bevan [179]).

Finally, it should be mentioned that we have dispensed again with some attributes that appear in the literature but are not directly related to usability, such as *remuneration*.

Technical environment

Finally, the *technical environment* is described by means of the following attributes (see Figure 4.7):

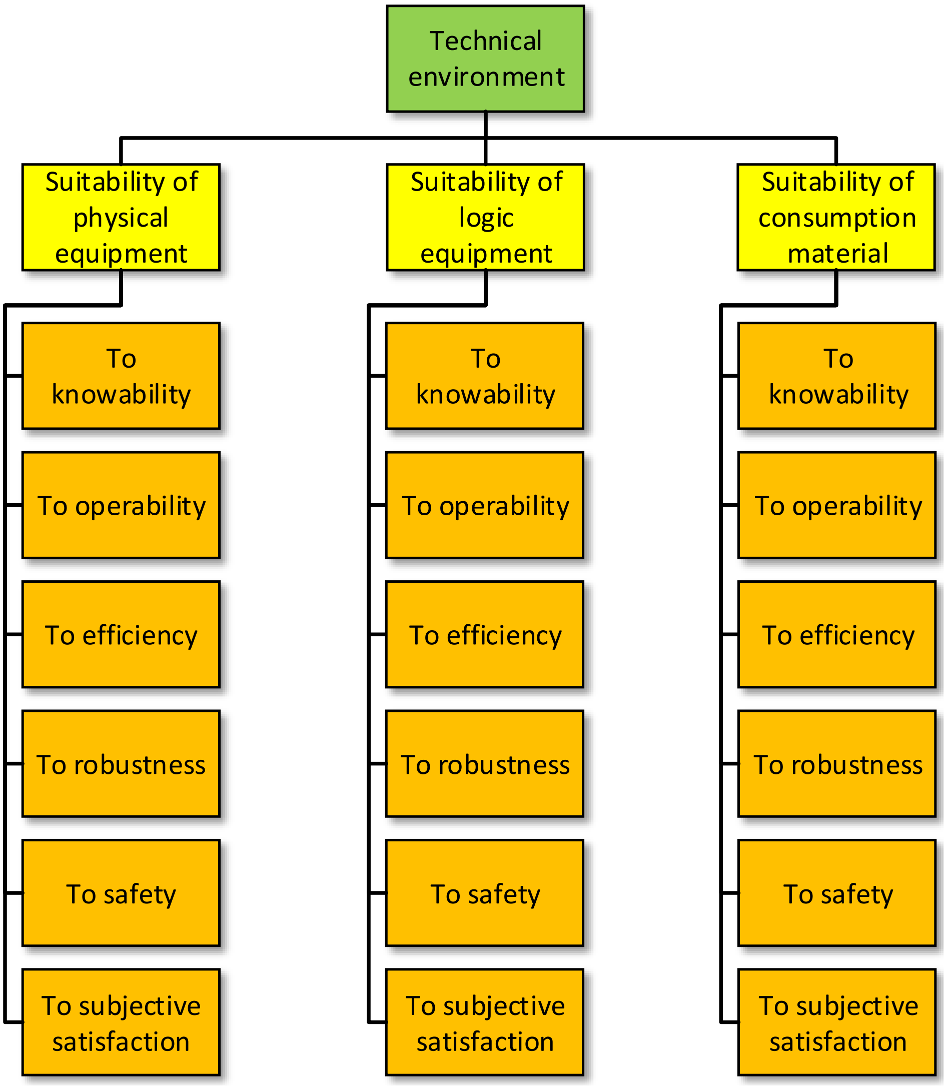


Figure 4.7: Technical environment subattributes

- **Suitability of physical equipment to usability** refers to the extent to which the physical equipment supporting the functioning of the system enables the system to be usable. Based on the usability taxonomy [16] presented in this thesis, this attribute can be divided into the following subattributes:
  - *Suitability of physical equipment to knowability.* Knowability is defined as the property by means of which the user can understand, learn, and remember how to use the system.

  - *Suitability of physical equipment to operability.* Operability is defined as the capacity of the system to provide users with the necessary functionalities and to permit users with different needs to adapt and use the system.
  - *Suitability of physical equipment to efficiency.* Efficiency is the capacity of the system to produce results in line with the invested resources.
  - *Suitability of physical equipment to robustness.* Robustness is defined as the capacity of the system to resist error and adverse situations.
  - *Suitability of physical equipment to safety.* Safety is defined as the capacity to avoid risk and damage when the system is in use.
  - *Suitability of physical equipment to subjective satisfaction.* Subjective satisfaction is the capacity of the system to produce feelings of satisfaction and interest in users.

- **Suitability of logic equipment to usability**, referring to the extent to which the logic (nonphysical) equipment supporting the functioning of the system enables it to be usable. Exactly as with the physical equipment, a distinction is drawn between suitability to the *knowability*, *operability*, *efficiency*, *robustness*, *safety*, and *subjective satisfaction* aspects of usability.

- **Suitability of consumption materials to usability**, referring to the extent to which the materials used during the functioning of the system enable it to be usable. Exactly as with the physical and logic equipment, this attribute is broken down according to the *knowability*, *operability*, *efficiency*, *robustness*, *safety*, and *subjective satisfaction* aspects of usability

The main difference between the technical environment attribute in our taxonomy and in the literature lies in the fact that existing models focus almost exclusively on IT systems, as indicated by the inclusion of subattributes such as support *hardware* and *software* (to be found in all the models), *reference materials* (Bevan & Macleod [31], ISO 9241-11 [87], Maguire [121], Thomas & Bevan [179]), and *type of network* (Kirakowski & Cierlick [103], Maguire [121]). Our taxonomy proposes generic attributes related to the *suitability of physical equipment to usability* and the *suitability of logic equipment to usability*, which makes the taxonomy applicable to any kind of system and not just computer systems. We also include a new component, namely, *suitability of consumption materials to usability*, as consumption materials could also feasibly affect the performance of a system.

Unlike the other context-of-use models, our taxonomy does not endeavor to provide a technical description of the elements in the environment but merely to reflect their influence on the usability of the system. We have therefore proposed only attributes that reflect the suitability of these elements in terms of the usability of the system.

4.4 Comparative Summary of the Taxonomies

Tables 4.1, 4.2, 4.3, 4.4 and 4.5 depict the *user* attributes proposed for our taxonomy and their correspondence with the attributes given in the different models in the literature. First-level attributes are indicated in capital letters. At this point it is appropriate to point out that because meanings are formally different, there is no such thing as absolute equivalence between attributes, even when identified using the same or similar terms. Furthermore, some authors' models lack definitions, and this makes it even more difficult to identify correspondences.

Certain attributes feature in all the existing models, namely, *role*, *skills and knowledge* (which we divide into *education* and *experience*), *physical characteristics* (included with *personal characteristics* in some models), and *attitude to the system* (also sometimes included in *personal characteristics*). *Cognitive characteristics* is an attribute that appears in all the models except that by Kirakowski and Cierlick [103]. Finally, the models developed in ISO 9241-11 [87] and by Kirakowski and Cierlick propose a *job characteristics* attribute for organizational settings, which we, however, exclude. This is because we consider that these characteristics are properties not of the user but of the social environment.

Tables 4.6, 4.7 and 4.8, describe correspondences between the *task* attribute for our taxonomy compared to other models. All the existing models agree in proposing subattributes for *task duration* and *frequency*, although we differ in that we group these under *temporal characteristics*. All the models include the attributes *physical and cognitive demands*, *task flexibility*, *safety in relation to the user's health*, and *risk resulting from error* (or *output criticality*). In this regard, our taxonomy is more comprehensive, as not only have we proposed new components as subattributes to refine previously described attributes, we also have grouped together associated attributes.

The attributes *choice in system use* and *complexity* appear much less frequently in the models from the literature.

We have excluded *task* attributes that are included in other taxonomies, either because we consider them to be irrelevant to usability studies (for example, *task dependencies*, *linked tasks* or *task breakdown*) or because we view them as part of the environment (e.g., *user autonomy*, which we considered to depend on the *social environment* attribute).

Tables 4.9, 4.10, 4.11, 4.12, and 4.13 describe correspondences between *environment* attributes. For the *physical environment*, there is unanimity in distinguishing between *conditions*, *workplace design*, and *safety*, and, for the *social environment*, between *structure* (in our taxonomy, *work relations*), *attitudes* and *culture* (in our taxonomy, *aims*), and *job design* (in our taxonomy, *control*). The main differences between models are to be found in the subattributes, which sometimes refer to different concepts even when identified by means of similar terms.

There is less consensus in regard to *technical environment*. All the models in the literature describe technical components that are necessary for the functioning

Table 4.1: User Subattributes from our Taxonomy Featuring in Other Models (part I: Role & Experience)

Proposed attributes	Attributes in other models				
	Bevan and Macleod (1994)	Thomas and Bevan (1996)	ISO 9241-11 (1998)	Kirakowski and Cierlick (1999)	Maguire (2001)
ROLE	PERSONAL DETAILS	USER TYPES	USER TYPES	USER ROLE	USER NAME
Operation role	User types	User types	Primary	-	User types
Support role	-	-	-	-	-
Monitoring role	-	-	-	-	-
Indirect role	Secondary users	Secondary users	Secondary and indirect	-	User types
-	-	-	-	-	User role
EXPERIENCE	SKILLS AND KNOWLEDGE	SKILLS AND KNOWLEDGE	SKILLS AND KNOWLEDGE	SKILLS AND KNOWLEDGE	EXPERIENCE, KNOWLEDGE, AND SKILLS
Experience with the system	Product experience System knowledge	Product experience	Product skill / knowledge System skill / knowledge	Product experience	Product experience
-	Training	Training	Level of training	Training or experience in product use	Training
Experience with similar systems	-	Product experience	-	Product experience	Related experience

Table 4.2: User Subattributes from our Taxonomy Featuring in Other Models (part II: Education)

Proposed attributes	Attributes in other models				
	Bevan and Macleod (1994)	Thomas and Bevan (1996)	ISO 9241-11 (1998)	Kirakowski and Cierlick (1999)	Maguire (2001)
EDUCATION	SKILLS AND KNOWLEDGE	SKILLS AND KNOWLEDGE	SKILLS AND KNOWLEDGE	SKILLS AND KNOWLEDGE	EXPERIENCE, KNOWLEDGE, AND SKILLS
Educational background	General knowledge	-	General knowledge	Education level	-
Knowledge of system domain	-	-	-	-	-
Knowledge of system language	Linguistic ability	Linguistic ability	Language skills	Linguistic ability	Language skills
Knowledge of system culture	-	Background knowledge	-	-	-
-	Task experience	Training and experience in the business processes	Task experience	Training or experience in the business processes	Task knowledge
-	Qualifications	Qualifications	Qualifications	Qualifications for job	Qualifications
-	Keyboard and input skills	Input skills	Input device skills	Input skills	Input device skills
-	Organizational experience	-	Organizational experience	-	Organizational knowledge
-	-	-	-	General computer experience	-

Table 4.3: User Subattributes from our Taxonomy Featuring in Other Models (part III: Attitude to the System)

Proposed attributes	Attributes in other models				
	Bevan and Macleod (1994)	Thomas and Bevan (1996)	ISO 9241-11 (1998)	Kirakowski and Cierlick (1999)	Maguire (2001)
ATTITUDE TO THE SYSTEM	PERSONAL ATTRIBUTES	MENTAL ATTRIBUTES	PERSONAL ATTRIBUTES	ATTITUDE AND MOTIVATION	PERSONAL ATTRIBUTES
-	Motivation	Motivations	Motivation	Level of motivation to use system	Attitude and motivation
-	Attitude	-	Attitude	Attitude to the product	Attitude and motivation
-	-	-	-	Attitude to the job / task	-
-	-	-	-	Attitude to information technology	-
-	-	-	-	Attitude to employing organization	-

Table 4.4: User Subattributes from our Taxonomy Featuring in Other Models (part IV: Physical Characteristics)

Proposed attributes	Attributes in other models				
	Bevan and Macleod (1994)	Thomas and Bevan (1996)	ISO 9241-11 (1998)	Kirakowski and Cierlick (1999)	Maguire (2001)
PHYSICAL CHARACTER-ISTICS	PERSONAL ATTRIBUTES	PHYSICAL ATTRIBUTES	PERSONAL ATTRIBUTES	PHYSICAL ATTRIBUTES	PERSONAL ATTRIBUTES
Sensorial characteristics	Physical capabilities Physical limitations and disabilities	Physical limitations and disabilities	Physical capabilities Physical limitations and disabilities	Physical characteristics	Physical capabilities and limitations
Speech characteristics	Physical capabilities Physical limitations and disabilities	Physical limitations and disabilities	-	Physical characteristics	Physical capabilities and limitations
Motor characteristics	Physical capabilities Physical limitations and disabilities	Physical limitations and disabilities	-	-	-
-	Age	Age	Age	Age	Age, gender
-	Gender	Gender	Gender	Gender	Age, gender

Table 4.5: User Subattributes from our Taxonomy Featuring in Other Models (part V: Cognitive Characteristics)

Proposed attributes	Attributes in other models				
	Bevan and Macleod (1994)	Thomas and Bevan (1996)	ISO 9241-11 (1998)	Kirakowski and Cierlick (1999)	Maguire (2001)
COGNITIVE CHARACTERISTICS	PERSONAL ATTRIBUTES	MENTAL ATTRIBUTES	PERSONAL ATTRIBUTES	-	PERSONAL ATTRIBUTES
Cognitive disabilities	-	Intellectual ability	-	-	Cognitive capabilities and limitations
Cognitive aptitudes	Intellectual ability	Intellectual ability	Intellectual ability	-	Cognitive capabilities and limitations
-	-	JOB CHARACTERISTICS	-	JOB CHARACTERISTICS	-
-	-	Job function	-	Job function	-
-	-	Job history	-	Job history	-
-	-	Hours of work or operation	-	Hours of work or operation	-
-	-	Job flexibility	-	Job flexibility	-
-	-	-	-	Work autonomy	-

Table 4.6: Task Subattributes from our Taxonomy Featuring in Other Models (part I)

	Attributes in other models				
Proposed attributes	Bevan and Macleod (1994)	Thomas and Bevan (1996)	ISO 9241-11 (1998)	Kirakowski and Cierlick (1999)	Maguire (2001)
CHOICE IN SYSTEM USE	-	CHOICE	-	EXECUTION	-
-	-	-	-	Degree of choice in use of system to carry out task	-
COMPLEXITY	-	-	-	TASK DEMAND ON USERS	-
-	-	-	-	Complexity as perceived by user	-
TEMPORAL CHARACTERISTICS	-	-	-	TASK DEMAND ON USERS	-
Task duration	DURATION	DURATION	DURATION	Duration	DURATION
Task frequency	FREQUENCY	FREQUENCY	FREQUENCY	Frequency	FREQUENCY
DEMANDS	-	-	-	TASK DEMAND ON USERS	-
Demands on human resources	PHYSICAL AND MENTAL DEMANDS	PHYSICAL AND MENTAL DEMANDS	PHYSICAL AND MENTAL DEMANDS	Physical and mental demands	PHYSICAL AND MENTAL DEMANDS
Demands on material resources	-	-	-	-	-

Table 4.7: Task Subattributes from our Taxonomy Featuring in Other Models (part II)

Proposed attributes	Attributes in other models: Bevan and Macleod (1994)	Thomas and Bevan (1996)	ISO 9241-11 (1998)	Kirakowski and Cierlick (1999)	Maguire (2001)
WORKFLOW CONTROLLABILITY	-	-	-	TASK DEMAND ON USERS	
Performance freedom	FLEXIBILITY	FLEXIBILITY	FLEXIBILITY	Flexibility	FLEXIBILITY
Reversibility	-	-	-	-	-
SAFETY	-	-	-	SAFETY Safe to operator Safe to secondary users Implications for immediate informational environment Implications for wider informational environment	-
Physical safety	-	SAFETY	SAFETY CRITICAL DEMANDS	-	SAFETY CRITICAL DEMANDS
Legal safety	-	-	-	-	-
Confidentiality	-	-	-	-	-
Safety of material assets	-	-	-	-	-

Table 4.8: Task Subattributes from our Taxonomy Featuring in Other Models (part III)

Proposed attributes	Attributes in other models: Bevan and Macleod (1994)	Thomas and Bevan (1996)	ISO 9241-11 (1998)	Kirakowski and Cierlick (1999)	Maguire (2001)
CRITICALITY	-	-	-	EXECUTION	-
Precision criticality	-	-	-	Degree of precision required in output	-
Robustness criticality	RISK RESULTING FROM ERROR	CRITICALITY OF THE TASK OUTPUT	RISK RESULTING FROM ERROR	Criticality of the task output	RISK RESULTING FROM ERROR
Time criticality	-	-	-	FLOW Postponement	-
-	-	-	-	FLOW	-
-	-	-	-	Input	-
-	OUTPUT	OUTPUT	OUTPUT	Output	OUTPUT
-	-	SIDE EFFECTS	-	Side effects	-
-	DEPENDENCIES	DEPENDENCIES	DEPENDENCIES	Dependencies	DEPENDENCIES
-	-	LINKED TASKS	-	Linked tasks	-
-	GOAL	GOAL	GOAL	Goal	GOAL
-	NAME	-	NAME	Name	NAME
-	BREAKDOWN	-	BREAKDOWN	-	STEPS
-	FREQUENCY OF EVENTS	-	FREQUENCY OF EVENTS	-	-
-	-	-	-	EXECUTION	-
-	-	-	-	Autonomy of user in completing task	-
-	-	-	-	Other execution constraints	-

of IT systems; we, on the other hand, refer to attributes that are applicable to any kind of system.

Finally, Table 4.14 describes the *equipment* subattributes proposed in the models by Bevan & Macleod [31] and ISO 9241-11 [87]. In our taxonomy (as in other models) *equipment* is excluded, given that we consider this aspect of context of use to be adequately reflected in the *technical environment* attribute.

Table 4.9: Physical Environment Subattributes from our Taxonomy Featuring in Other Models (part I: Sensorial Conditions & Atmospheric Conditions)

Proposed attributes	Attributes in other models: Bevan and Macleod (1994)	Thomas and Bevan (1996)	ISO 9241-11 (1998)	Kirakowski and Cierlick (1999)	Maguire (2001)
SENSORIAL CONDITIONS	CONDITIONS	CONDITIONS	CONDITIONS	-	CONDITIONS
Sensorial quality	Auditory environment Visual environment	Auditory environment Visual environment	Auditory environment Visual environment	AUDITORY ENVIRONMENT VISUAL ENVIRONMENT	Auditory environment Visual environment
Sensorial stability	Environmental instability	Environmental instability	Environmental instability	ENVIRONMENTAL INSTABILITY	Environmental instability
ATMOSPHERIC CONDITIONS	CONDITIONS	CONDITIONS	CONDITIONS	-	CONDITIONS
Atmospheric quality	Atmospheric conditions Thermal environment	Atmospheric conditions Thermal environment	Atmospheric conditions Thermal environment	THERMAL ENVIRONMENT	Atmospheric conditions Thermal environment
Atmospheric stability	Environmental instability	Environmental instability	Environmental instability	ENVIRONMENTAL INSTABILITY	Environmental instability

Table 4.10: Physical Environment Subattributes from our Taxonomy Featuring in Other Models (part II: Spatial Conditions & Safety)

Proposed attributes	Attributes in other models				
	Bevan and Macleod (1994)	Thomas and Bevan (1996)	ISO 9241-11 (1998)	Kirakowski and Cierlick (1999)	Maguire (2001)
SPATIAL CONDITIONS	DESIGN	DESIGN	DESIGN	-	DESIGN
Space availability	Space and furniture	Space and furniture	Space and furniture	Amount of available space	Space and furniture
Location suitability	Location	Location	Location	Location	Location
Postural suitability	User posture	User posture	User posture	Posture required of user	User posture
-	Space and furniture	Space and furniture	Space and furniture	Necessary furniture / resources	Space and furniture
SAFETY	SAFETY	HEALTH AND SAFETY	HEALTH AND SAFETY	-	SAFETY
User safety	Health hazards	Health hazards	Health hazards	Health hazards	Health hazards
-	Protective clothing and equipment	Protective clothing and equipment	Protective clothing and equipment	Protective clothing needed	Protective clothing and equipment
System safety	-	-	-	-	-

Table 4.11: Social Environment Subattributes from our Taxonomy Featuring in Other Models (part I: Work Relations)

Proposed attributes	Attributes in other models: Bevan and Macleod (1994)	Thomas and Bevan (1996)	ISO 9241-11 (1998)	Kirakowski and Cierlick (1999)	Maguire (2001)
WORK RELATIONS	STRUCTURE	STRUCTURE	STRUCTURE	-	STRUCTURE
Team work	Group working	Group working	Group working	MULTI / SINGLE USER ENVIRONMENT	Group working
Human support	Assistance	Assistance	Assistance	ASSISTANCE	Assistance
Interruptions	Interruptions	Interruptions	Interruptions	INTERRUPTIONS	Interruptions
-	Management structure	Management structure	Management structure	-	Management structure
-	Communications structure	Communications structure	Communications structure	MODE OF COMMUNICATION	Communications structure
-	Work practices	-	Work practices	-	Work practices
-	Remuneration	-	-	-	Salary or payment
-	-	-	-	-	JOB DESIGN
-	Job function	-	Job function	-	Job function
-	Hours of work	-	Hours of work	-	Hours of work

Table 4.12: Social Environment Subattributes from our Taxonomy Featuring in Other Models (part II: Aims & Control)

	Attributes in other models				
Proposed attributes	Bevan and Macleod (1994)	Thomas and Bevan (1996)	ISO 9241-11 (1998)	Kirakowski and Cierlick (1999)	Maguire (2001)
AIMS	ATTITUDES AND CULTURE	ATTITUDES AND CULTURE	ATTITUDES AND CULTURE	-	ATTITUDES AND CULTURE
-	Policy on use of computers	IT Policy	Policy on use of computers	POLICY	Policy on computer use
-	Organizational aims	Organizational aims	Organizational aims	AIMS	Organizational aims
-	Industrial relations	Industrial relations	Industrial relations	-	Industrial relations
-	-	-	-	CULTURE	-
-	-	-	-	PROCEDURES	-
CONTROL	JOB DESIGN	WORKER CONTROL	JOB DESIGN	-	JOB DESIGN
Monitoring	Performance monitoring	Performance monitoring	Performance monitoring	USER MONITORING IN PROGRESS	Performance monitoring
Feedback	Performance feedback	Performance feedback	Performance feedback	FEEDBACK ON JOB GIVEN	Performance feedback
User autonomy	Autonomy	-	Autonomy	-	Autonomy
-	Pacing	Pacing	Pacing	-	Pacing
-	Job flexibility	-	Job flexibility	-	Job flexibility
-	Discretion	-	Discretion	-	Discretion

Table 4.13: Technical Environment Subattributes from our Taxonomy Featuring in Other Models

	Attributes in other models				
Proposed attributes	Bevan and Macleod (1994)	Thomas and Bevan (1996)	ISO 9241-11 (1998)	Kirakowski and Cierlick (1999)	Maguire (2001)
-	CONFIGURATION	-	-	-	-
SUITABILITY OF PHYSICAL EQUIPMENT TO USABILITY	Hardware	HARDWARE	HARDWARE	HARDWARE REQUIRED	HARDWARE
SUITABILITY OF LOGIC EQUIPMENT TO USABILITY	Software	SOFTWARE	SOFTWARE	(SUPPORTING) SOFTWARE REQUIRED	SOFTWARE
SUITABILITY OF CONSUMPTION MATERIAL TO USABILITY	-	-	-	-	-
-	Reference materials	REFERENCE MATERIALS	REFERENCE MATERIALS	-	REFERENCE MATERIALS
-	-	-	-	STANDALONE / NETWORKED	-
-	-	-	-	ADDITIONAL HARDWARE/ SOFTWARE RESOURCES REQUIRED	OTHER EQUIPMENT
-	-	-	-	TYPE OF NETWORK CONNECTION REQUIRED	NETWORK

Table 4.14: Equipment Subattributes in Other Models

Proposed attributes	Attributes in other models				
	Bevan and Macleod (1994)	Thomas and Bevan (1996)	ISO 9241-11 (1998)	Kirakowski and Cierlick (1999)	Maguire (2001)
-	BASIC DESCRIPTION	-	BASIC DESCRIPTION	-	-
-	Product identification	-	Product identification	-	-
-	Product description	-	Product description	-	-
-	Main application areas	-	Main application areas	-	-
-	Major functions	-	Major functions	-	-
-	SPECIFICATION	-	SPECIFICATION	-	-
-	Hardware	-	Hardware	-	-
-	Software	-	Software	-	-
-	Materials	-	Materials	-	-
-	-	-	Services	-	-
-	Other Items	-	Other Items	-	-

Chapter 5

A Usability Study of an Intelligent Speed Adaptation Device

5.1 Introduction

This chapter describes a practical, real-world application of the taxonomies described in Chapters 3 and 4. As explained in Section 2.1, usability studies are often based on ad hoc definitions of usability. That is, "ad hoc" in the sense of "created for a specific purpose at a particular time". These kinds of traditional usability studies can be difficult to generalize, might require a steep learning curve, and there is always the danger of being inconsistent with the concept of usability as defined in standards and the literature. The approach taken in this thesis, on the other hand, starts from comprehensive, general-purpose, and hierarchically structured taxonomies that can be later applied to a specific product.

This chapter illustrates this approach through a usability study [14] of a prototype of an Intelligent Speed Adaptation (ISA) device called CARAT (Continuous Assessment of Road ATtributes) counter. The development of this ISA device (ISAD) and the usability study are part of an international research project called Galileo Speed Warning (GSW), which proposes a reward-based alternative to traditional speed enforcement systems that are typically punitive [1, pp. 4-5].

The methodology in this usability study involves a selection of widely-used usability techniques (see Section 2.3 for an overview of common usability techniques). The focus was on, firstly, inspecting a preliminary version of the CARAT counter at the earliest stages of its development, and, secondly, conducting some usability tests with actual users and inquiring them later about their opinions. The entire process was supported by the two taxonomies presented in this thesis.

The specific usability activities that were carried out are as follows:

1. **Usability requirements analysis**. An informal list of usability requirements was produced based on the most preliminary design of the CARAT counter.
2. **Heuristic evaluation** of the first working prototype of the CARAT counter.

3. **User tests**, in which several participants tried the first working prototype during their daily driving routine, followed by a **subjective analysis** of the opinions of the users after testing the prototype. All the usability information on the tests was gathered through **questionnaires**.

The bulk of this chapter is structured as follows: (1) Some background is provided by describing the domain of application (i.e., ISA for in-vehicle systems); (2) the scope of the GSW project is outlined; (3) the system under study (i.e., the CARAT counter) is described; (4) the methodology for the usability study is explained; (5) characterizations of the system and its context of use are provided; (6) the usability activities that comprise the study (i.e., usability requirements analysis, heuristic evaluation, and subjective analysis) are described in detail; and (7) the results of the usability study are analyzed.

5.2 Background

ISA is a technology for intelligent in-vehicle transportation systems that warns the driver when speeding, discourages the driver from speeding, or prevents the driver from exceeding the speed limit [159]. Because vehicle speed has a direct impact on road safety, and studies have long suggested that reduced speed results in fewer accidents [18][60], there is a growing interest in the potential of ISA systems.

ISA systems can be classified according to three attributes:

1. **Intervention**. The degree of intervention classifies ISA systems into *passive*, *intermediate*, and *active*[1] [140]. A passive system simply warns the drivers that they are exceeding the speed limit, typically by means of a simple acoustic or visual warning. On the opposite end of the spectrum, active systems intervene and automatically correct the vehicle's speed to comply with the speed limit. It is also possible to have an intermediate level of intervention, which uses haptic feedback to prevent the user from exceeding the speed limit.

2. **Freedom**. According to the degree of freedom the user has when following and using the ISA recommendations, ISA systems can be classified into *voluntary* and *mandatory* [45]. In voluntary systems the user has the freedom of following or not following the ISA warnings. On the other hand, in mandatory systems the ISAD cannot be disconnected and usually limits the vehicle's maximum speed to the ISA recommendations.

3. **Adaptability**. The adaptability to legal speed limits is also taken into account [113]. In this regard an ISA device can be *fixed* or *variable*. In fixed systems the ISAD compares the vehicle's speed with the fixed legal speed limit of the road on which it is traveling. On the other hand, in a variable system other driving characteristics are taken into account in order to establish the speed

[1]Other common names for these three attributes are *informative*, *supportive*, and *mandatory*, respectively [106].

limit. For example, weather conditions, driving near a school, or cornering speed.

In spite of the obvious advantages of ISA systems [124], there is no widespread use of such products so far. This is related to two factors: technological reasons and consumer acceptance reasons.

From the technological point of view, ISA systems need accurate maps and accurate positioning systems in order to work properly. There has been a great deal of research in this area lately, but not enough to ensure good performance.

As for acceptance, it should be borne in mind that successful implementation of ISA would ultimately rely upon the attitude of the general public [46]. There are many privacy, freedom, and usability concerns that prevent users from adopting ISA technology. According to the Organisation for Economic Co-operation and Development (OECD) [9], speed management is a government task, and the European governments would realize important economic benefits for their citizens if they decided to encourage and eventually require them to install ISADs in their cars. As a first step, governments could promote the industry's efforts by supporting additional research and standardization and encourage its private use by introducing tax cuts as incentives to installing ISA and becoming first customers of ISA technology (e.g., in public transport systems).

Regarding the usability of ISADs, one could mention the Swedish large-scale trial involving ISA in urban areas performed between 1999 and 2002 [32]. Several thousand vehicles were equipped with voluntary, supportive, and informative systems to keep drivers from exceeding the speed limit. One of the main conclusions of the study was that the "systems must be improved to become more attractive" [32, p. 2]. So usability is a very important feature to consider in ISA systems in order to make them more appealing, easier to use, and safer.

5.3 The GSW Project

The GSW Project [2] took place between January 2009 and July 2010, and was partially funded by the European 7th Framework Programme. The project's coordinator was Jonathan Guard, of Mapflow Ltd.[2], nowadays a LexisNexis company. Two partners contributed to the project, namely, the Dutch company Technolution B.V.[3] and the University of A Coruña through researchers from the LIDIA group[4]. The participation of the author of this thesis, specifically, lies in the design and implementation of the usability study described in this chapter.

The main goal of the GSW project is to test the acceptability of a novel approach to ISA. Typically, the default strategy for reducing speed in roads has been to use punitive fines. For example, the Danish project Spar på Farten was based on Pay As

[2]http://www.mapflow.com/
[3]http://www.technolution.eu/en/
[4]http://lidiagroup.org/

You Drive principles, which means that, besides giving a warning when the driver is speeding, the ISA equipment also gives penalty points, which reduce a promised bonus of 30% on the insurance rate [108]. In contrast, the GSW project follows a "carrots rather than sticks" approach "based on the concept of 'reward' schemes in other industry sectors (for example, air miles, Nectar points, retail loyalty card schemes)" [2]. This is put into practice through an on-board ISAD that monitors good driving (i.e., the CARAT counter).

The GSW website lists these as the core objectives for the project [2]:

- "To design and develop an innovative new product which allows good driving behaviour to be rewarded."
- "To partner with other SMEs in the design and development, and also in the future commercialisation of the product."

Also according to the GSW website, the final outcome of the project can be summarized thus:

> "The key result is the design and development of a new innovative solution for rewarding drivers for good driving behaviour. This is a new and exciting solution that will speed up the adoption of ISA and create new opportunities for interested parties, such as insurance companies, local authorities and large organisations to reward good driving behaviour." [2]

5.4 The ISAD under Study

The CARAT counter is a passive, voluntary, and fixed ISAD that maintains a count of "CARATs" (i.e., points) that increases as the driver keeps within specified speed limit thresholds. It is also possible to review how many CARATs have been collected and how many could have been collected through better compliance with the limits.

The ISAD is designed to be placed in the vehicle where it can be seen by the driver without causing distraction. The device lights up green when the vehicle is within the speed limits, which also increases the CARAT count. The device turns yellow, and no CARATs are accumulated when the limit is slightly exceeded. Otherwise, it turns red, making it more difficult to earn CARATs.

At the end of the trip, the driver can see the earned CARATs and the "ideal" number of CARATs that could have been accumulated for "perfect" driving. The earned CARATs can then be exchanged for real-world benefits such as reduced insurance premiums, loyalty points, and so on. Furthermore, if users decide to make such information available, they can participate in good-driving competitions, statistical information gathering, and various other activities.

Two prototypes of the CARAT counter were developed during the project, namely, the *concept prototype* and the *field prototype*.

The concept prototype (see Figure 5.1) was done on paper and consists in a small USB device with a display.

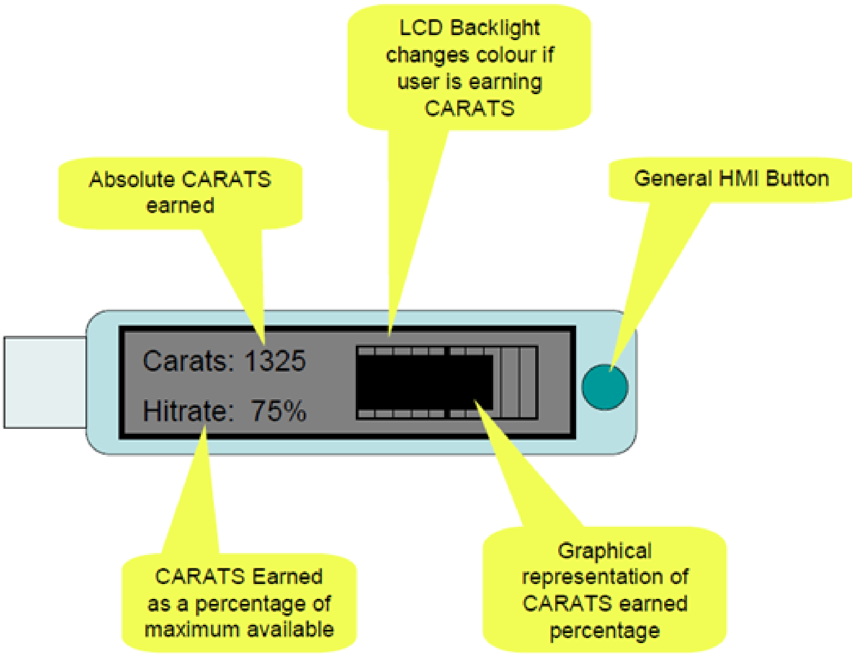


Figure 5.1: Concept prototype of the CARAT counter, including the USB connector, the earned CARATs (shown as an absolute value, a numerical percentage, and a percentage bar), and the "general HMI button."

The field prototype (see Figure 5.2) was subsequently developed in order to test an actual working prototype of the CARAT counter. This is a more complex in-vehicle information system with a display, a USB stick for storing CARATs, and a cigarette lighter plug. The display has been completely redesigned for this prototype. The left side of the display now shows the earned CARATs accompanied by the "ideal" number of CARATs that could have been earned. Also on the left side of the display are a text describing the driving behavior, and the "end-of-journey button" (which the driver needs to push before stopping the engine). On the right side of the display, the current speed limit is displayed through a speed limit sign, and overall compliance with the speed limits is represented by means of a smiley.

As seen in Figure 5.3, the degree of compliance with the speed limit is represented by a green happy face, a yellow neutral face, or a red sad face. When the degree of compliance is unknown, a gray face with a question mark is displayed.

The text describing the driving behavior informs of the driver's progress using the values described in Table 5.1:

At that point in the development life cycle of the CARAT counter, the behavior was determined purely by compliance with speed limits. Future improvements of the algorithm were also contemplated in the design, however, which would include other driving actions like tailgating or cornering speed.

Figure 5.2: Field prototype of the CARAT counter, including the external shell with two buttons and a speaker, the display, the USB stick, the USB connector, and the cigarette lighter plug.

Under unexpected situations, the CARAT counter displays the error alerts described in Table 5.2 and depicted in Figure 5.4.

5.5 Methodology for the Usability Study

What distinguishes the usability methodology in this study from the more familiar "ad hoc" usability methodologies is that the former takes a more analytic approach based on, firstly, comprehensive and generic taxonomies [16] [17] and, secondly, a detailed decomposition of the system under study into its constituent parts. Because the taxonomies are general-purpose and the system is project-specific, the first step in the usability study is to provide a detailed characterization of the system and its context of use. This characterization is then used to "instantiate" the generic

Figure 5.3: Compliance with the speed limit represented through smileys.

Table 5.1: Driving Behavior

Displayed Text	**Description**
Good	In compliance with speed limits
Bad	Not in compliance with speed limits
Improving	Current behavior is “good”, but was previously “bad”
Neutral	No CARATs are being earned due to GPS or map errors

taxonomies with the specific characteristics of both the system and the context of use. In contrast with the more abstract and vague nature of traditional usability definitions, the taxonomies presented in this thesis try to help to achieve this goal by going from the general to the particular at several levels of detail.

Once this “instantiation” has taken place, specific usability techniques can be applied to obtain usability information on distinct aspects of the system and from different points of view.

The usability study of the CARAT counter comprises in fact the study of the two prototypes of the CARAT counter described in Section 5.4. More specifically:

- For the concept prototype (see Figure 5.1), a usability requirements analysis was conducted.
- For the field prototype (see Figure 5.2), a heuristic evaluation was performed and some user tests were conducted, followed by a subjective analysis of the opinions of the users.

Usability requirements analysis consists in identifying the usability needs and possible limitations of the concept prototype. The goal is to provide a set of recommendations to be used as an input to the design of the first working prototype of the product (i.e., the field prototype).

This is followed by a usability assessment of the actual product. As mentioned, the techniques that were selected are as follows:

1. **Heuristic evaluation**, which is based on the judgment of usability specialists who analyze the product in order to find strengths and weaknesses.
2. **User testing**, which consists in having real users performing tasks with the actual product. In this study, several participants tried the first working prototype of the CARAT counter during their daily driving routine. No data

Table 5.2: Error Alerts

Name	Description
No CARAT stick	The USB stick has not been plugged in.
No CARAT functionality	The USB stick has been plugged in, but it is not valid.
No GPS	The GPS position could not been retrieved and, consequently, the speed limit could not be determined.
Off map	The vehicle is outside the geographic boundary of the map, or the map does not contain the road or its speed limit.

Figure 5.4: CARAT counter error alerts. Clockwise from top left: "no CARAT stick", "no CARAT functionality", "no GPS", "off map".

collection techniques were employed during these tests, however, so as not to distract the users.

3. **Subjective analysis**, which interprets the opinions of the users on the usability of the product. This was done by giving each user a questionnaire after having tested the CARAT counter.

The specific usability techniques selected for in this study are widely used and have been proved to yield useful results, so they were considered appropriate and comprehensive enough at this stage of the development life cycle. According to the classification of usability techniques in Section 2.3, the selected techniques cover the categories of *inspection*, *inquiry*, and *testing*. As many more usability techniques exist, different techniques might have been used and, certainly, different techniques could be used at later stages, some of which could also be supported by the taxonomies. For example, the taxonomies could be used to structure user interviews or field observations in the same way they are used to structure user questionnaires or heuristic evaluations.

5.6 Characterizations of the System and the Context of Use

The first step in the usability study is to characterize the system (i.e., the CARAT counter). This involves breaking down the system into its constituent parts and identifying their usability-relevant attributes. As a result, this characterization provides a checklist of the aspects to be assessed through the usability taxonomy.

The characterization was done for the two versions of the CARAT counter described in Section 5.4, namely, the concept prototype (see Figure 5.1) and the field prototype (see Figure 5.2). However, in order to avoid redundancy, this section provides a brief outline of the process that is basically applicable to either design.

The process begins by removing the branches of the usability taxonomy that are not relevant to the CARAT counter. For example, of all the *formal* aspects, which appear in *knowability*, *operability*, *safety*, and *aesthetics*, only the *visual* and, in some cases, *acoustic* or *tactile* aspects are relevant. Some attributes related to *structure* (from *knowability*) were also removed because the structure of the physical product is such that their analysis is not necessary.

The next step is to instantiate specific attributes of the usability taxonomy with the components whose usability must be examined, always from the point of view of the user. Some examples, applied to the field prototype of the CARAT counter, are as follows:

1. **The visual elements of the CARAT counter** that appear in the *knowability*, *operability*, and *subjective satisfaction* attributes can be divided into *physical elements* (e.g., physical buttons, speaker, USB stick, USB connector, and cigarette lighter plug) and *displayed elements* (e.g., text and graphics shown on the display of the device).

2. **The functioning of the CARAT counter**, which is a subattribute of *knowability*, is further subdivided into two attributes in the taxonomy: *user tasks* (i.e., the actions that the user must perform in order to operate the CARAT counter) and *system tasks* (i.e., those functions that do not need the input of the user). *User tasks* can be instantiated with activities like pushing the end-of- journey button and removing the USB stick. *System tasks* include, for example, recording and displaying information.

3. **Cultural universality**, which is a subattribute of *operability*, is subdivided in the taxonomy into *universality of language* and *cultural conventions*. Language is instantiated as *displayed text*, and *cultural conventions* include the

measurement units and the directionality of the interface (i.e., whether it is read from left to right or the other way around).

4. **The environment**, which appears in the *operability*, *robustness*, and *safety* attributes, can be subdivided into *physical environment* and *technological environment*, both widely-accepted context-of-use attributes.

5. **The third parties** (from the *safety* attribute) are instantiated in this case with the passengers who go along with the driver.

6. **Confidentiality** (from the *safety* attribute) includes confidentiality of the personal details of the user and of the use of the product, such as the roads traveled.

This breakdown of the system gives rise to what could be in fact considered taxonomies of elements. For example, the taxonomy of *visual elements* is divided in *physical elements* (Figure 5.5) and *displayed elements* (Figure 5.6).

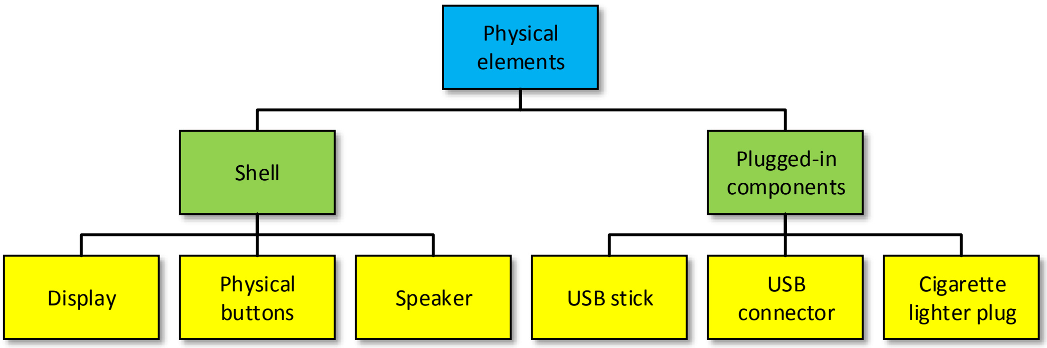


Figure 5.5: Taxonomy of the physical elements of the CARAT counter

Many of the elements characterized above are entirely dependent on the specific characteristics of the system under study (i.e., the CARAT counter), which makes this activity partially ad hoc. The elements that characterize the system were agreed on by observing the design of the system and studying the purpose and the design of the system as described in the project's documentation. Some components, however, might be generalized into generic taxonomies of objects with shared properties, and related research has already been conducted on this topic. For example, Han et al. [69] propose a framework for human interface elements that divides a consumer electronics product into *physical components* (e.g., display, button, and body) and *logical components* (e.g., menu, metaphor, and message). Each of these components possesses a set of usability-relevant properties, which they classify into three categories: *individual* (e.g., shape, size, color), *integration* (e.g., structure, ordering, layout), and *interaction* (e.g., response, feedback, range of motion). Following an analogous approach, a tentative list of usability-relevant attributes was produced for the *displayed elements* of the CARAT counter (see Table 5.3 for some examples).

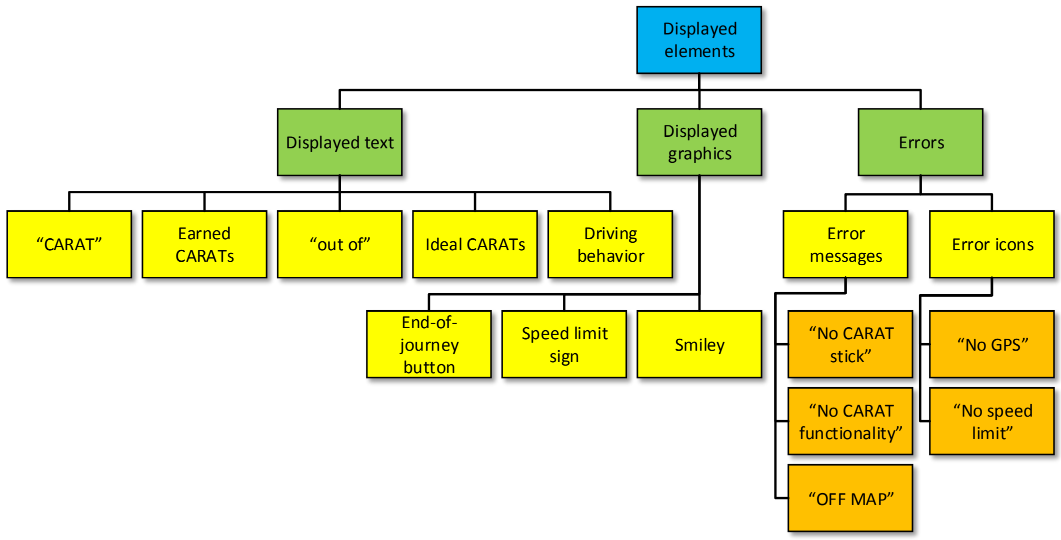


Figure 5.6: Taxonomy of the displayed elements of the CARAT counter

Similarly, while some tasks may be specific to the system under study, other tasks might be so common as to be included in generic taxonomies of tasks. For example, Byrne et al. [41] created a "taskonomy" of WWW use that consists of *use information*, *locate on page*, *go to page*, *provide information*, *configure browser*, and *react to environment*. This "taskonomy" has been used by automated tools such as Web TANGO [91].

In addition to characterizing the system it is also necessary to specify the context of use, as the usability of a system is always relative to it. For example, the context of use determines the user's perception of the usability of the system and makes certain usability attributes more desirable. A brief summary of the relevant context-of-use attributes for this usability study, including how they are instantiated and their predicted impact on usability, would be as follows:

1. **User knowledge of system domain** and, similarly but to a lesser extent, *user experience with similar systems*. In this study, *system* means ISADs. Previous knowledge and experience on the part of the user, or the lack thereof, have a significant impact on knowability.

2. **User knowledge of system language**. In the CARAT counter, the system's language is English. If the intended audience lacks proficiency in that language, knowability and universality become particularly critical.

3. **User attitude to the system**. This attribute has an impact on the user's perception of the usability of the system, especially regarding user safety and subjective satisfaction.

4. **Sensorial disabilities and aptitudes of the user** (*sight*, *hearing*). This affects usability attributes like clarity and safety.

Table 5.3: Examples of Usability-Relevant Attributes for the Components of the CARAT Counter

Component	**Attributes**	**Subattributes**
Displayed text (visual)	Font	Font family, weight, style, decoration, capitalization
	Size	-
	Orientation	-
	Color	Hue, saturation, lightness
	Content	-
Displayed graphics (visual)	Shape	-
	Size	Width, height
	Color	Hue, saturation, lightness
Displayed graphics (conceptual)	Shape	Literal meaning, symbolism
	Color	Symbolism
Display structure (visual)	Ordering	-
	Positioning	Alignment, spacing
	Color contrast	Hue, saturation, lightness
...		

5. **Task demands on the user**. The most sensitive task here is the system task of displaying information while the user is driving, which puts significant *sensorial* (*sight*, *hearing*) and *cognitive* (*memory*, *attention*) demands on the user. Excessive demands on human effort would have a negative impact on perceived efficiency, user safety, and subjective satisfaction.

6. **Task safety**. Two tasks are particularly sensitive here. Firstly, displaying information while the user is driving can affect the physical safety of the user and the passengers. Second, recording driving information is directly related to user confidentiality. Obviously, the inherent safety of the tasks has an impact on the usability of the whole system, as safety is a subattribute of usability.

7. **Sensorial conditions of the physical environment** (*visual*, *auditory*), which are subdivided into the *quality* and the *stability* of said conditions. In this study, this is instantiated with specific real-world problems such as luminosity, glare, noise, and so on. This is related to usability attributes such as knowability, efficiency, and safety.

8. **Spatial conditions of the physical environment** (*suitability of the location of the system*, *postural suitability*). This is instantiated as the placement of the CARAT counter and the posture of the user while looking at the CARAT counter. This has an impact on knowability, efficiency, subjective satisfaction, and, in the worst-case scenario, safety.

9. **Social environment** (*human support*, *interruptions*). This is instantiated into several aspects: the presence of passengers who notify the driver of important information displayed on the CARAT counter, distractions caused

by the passengers, the traffic, and so on. This affects efficiency, safety, and subjective satisfaction.

5.7 Usability Requirements Analysis

The initial usability requirements of the CARAT counter establish a series of basic principles for the design and a list of usability aspects that should be taken into account before carrying out a detailed design of the product. The input to this activity was the concept prototype of the CARAT counter (see Figure 5.1).

The usability taxonomy guides the process in a straightforward manner, as the usability attributes are instantiated with the particular characteristics of the product. For example, visual clarity of the elements is instantiated as visual clarity of the physical elements, the displayed text, and the displayed graphics, each one with its relevant characteristics like font, size, and so on (see Table 5.3).

Incidentally, this systematic and structured approach fits well with the technique of formulating requirements as “boilerplates”, which contributes to standardizing the language and making requirements more reusable [82]. In this technique, requirements are constructed by completing placeholders in ready-made sentences. For example the boilerplate “<the component> shall have a <usability attribute> <component attribute>” could be instantiated as “the displayed text shall have a legible font”.

For more technical considerations, the requirements can be refined by consulting documents such as usability guidelines. The guidelines can be generic or domain-specific. Generic guidelines include the ones by Smith and Mosier [175], Microsoft Corporation [7], and Apple Inc. [5][3]. As for guidelines specific to driver information systems, some were proposed by Dingus et al. [53], Green, et al. [66], Stevens et al. [177], and Japan Automobile Manufacturers Association [19]. Also consulted was the list of common usability problems in automotive navigation systems that was compiled by Nowakowski et al. [138].

The usability requirements serve as an input to the design of the field prototype of the CARAT counter. Once the field prototype has been designed, the next step is to perform a heuristic evaluation in order to analyze its strong and weak points in terms of usability.

5.8 Heuristic Evaluation

In heuristic evaluation (see Section 2.3), usability specialists examine a product and assess its usability guided by a set of heuristics, such as guidelines or good practices. In a way, the methodology presented in this thesis aims to make this informal task a bit more systematic and more connected to accepted definitions of usability, as it provides a checklist of common usability aspects to be analyzed and facilitates a shared understanding of usability aspects among the participants. For example, the attributes and subattributes in Table 5.3 point the specialist to the color contrast between adjacent elements of the display in terms of hue, saturation, and lightness (incidentally, these three attributes can be measured and subsequently evaluated according to color guidelines such as [117]).

Based on the usability taxonomy applied to the field prototype of the CARAT counter (see Figure 5.2), the usability of each instantiated attribute was assessed. The assessments were focused on detecting usability problems (i.e., *formative* goals), not on obtaining numerical usability ratings (i.e., *summative* goals). Keevil [99] and Lynch et al. [116], for instance, have conducted some research on the latter type of goals, but this was not the aim of this usability study.

Table 5.4 shows examples of the heuristic evaluation of the visual clarity of the physical and displayed elements. As can be seen, Table 5.4 focuses only on a single usability subattribute applied to a few interface elements. The full heuristic evaluation repeated the same process for all the relevant attributes in the usability taxonomy and all the components of the CARAT counter. A full report was subsequently produced for the GSW project, which presented the conclusions of the heuristic evaluation and offered suggestions for improvement.

As a summary, the heuristic evaluation identified the following strong points of the CARAT counter:

1. The two most important elements on the display while the user is driving, namely, the speed limit sign and the smiley, are clearly visible.

2. Color contrast on the right side of the display (i.e., light colors on a dark background) is appropriate and follows the widely-accepted guideline of negative contrast[5].

3. The general purpose of the CARAT counter is intuitive and the way the display is structured is simple and readable.

4. Even though the color of the smiley has an associated meaning, the different faces make it easily understandable to users with color blindness (see Figure 5.7).

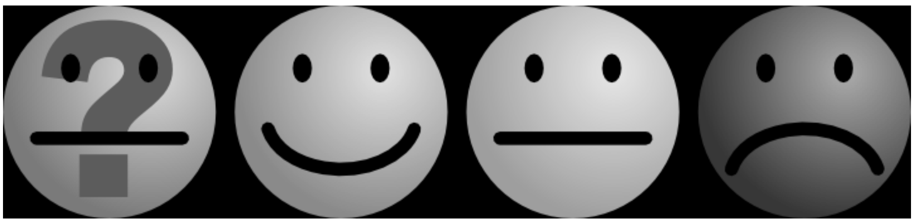


Figure 5.7: Color is not essential in order to understand the meaning of the smiley.

[5] "In cars, the guiding principle is to minimize glare. Since there are more pixels for the background than the text (foreground), using a dark background will minimize the luminous output, and consequently minimize glare from the display" [66, p. 36].

Table 5.4: Examples of the Heuristic Evaluation of the CARAT Counter

Usability attribute		**Heuristic Evaluation**
Knowability		
Clarity		
Visual clarity of the physical elements		
Visual clarity of the physical buttons	+	Visibility is good
Visual clarity of the speaker	+	Visibility is good
Visual clarity of the USB stick	+	Visibility is good
Visual clarity of the USB connector	+	Visibility is good
Visual clarity of the cigarette lighter plug	+	Visibility is good
Visual clarity of the displayed text		
Visual clarity of the "CARAT" text	–	Size is small
Visual clarity of the earned CARATs	–	Size is small
Visual clarity of the "out of" text	–	Size is very small
Visual clarity of the "ideal" CARATs	–	Size is small
Visual clarity of the driving behavior text	–	Size is small
Visual clarity of the "no CARAT stick" error message	–	Size is small
Visual clarity of the "no CARAT functionality" error message	–	Size is small
Visual clarity of the "OFF MAP" error message	+	Text size is appropriate
Visual clarity of the displayed graphics		
Visual clarity of the end-of-journey button	–	Size is too small
Visual clarity of the speed limit sign	+	Size is appropriate
Visual clarity of the smiley	+	Size is appropriate
Visual clarity of the "no GPS" icon	+	Size is appropriate
Visual clarity of the "no speed limit" icon	+	Size is appropriate
...		

Note. The plus sign indicates a positive assessment and the minus sign a negative one.

5. The CARAT counter does not risk the legal safety of the driver or the passengers, as it does not record information about the position or the speed of the car at a given moment, something that could be used as evidence in a legal process. Great care has also been taken in order to avoid compromising the confidentiality of its use.

The main usability problems detected in the heuristic evaluation were as follows:

1. Some elements on the left side of the display tend to be small and have low contrast (e.g., gray or red on blue).

2. The purpose of some elements is not intuitive, such as (a) the two buttons on the left side of the CARAT counter; (b) "ideal" CARATs, which by themselves say little to the user and would be better replaced by a percentage or some kind of graph; (c) the "no CARAT functionality" error message, which should be replaced by a more descriptive sentence; (d) the end-of-journey button, whose purpose is difficult to infer, and which does not really look like a standard button due to its color palette.

3. There is an inconsistency in how the driving behavior text and the smiley seem to represent the same concept, namely, success, but actually do it in different ways. The smiley is associated only with the speed limit, whereas the behavior text is associated with driving performance in general. This distinction is not intuitive and can even lead to visual messages that may seem contradictory at first (as shown in Figure 5.8).

Figure 5.8: Seemingly contradictory messages: Neutral driving behavior and a happy smiley.

4. Another problem with the behavior text is the vagueness with which it describes the progress of the driving behavior of the user. It should be replaced with more meaningful text messages or with self-explanatory graphics, such

as arrows. The latter option would also contribute to solving the general universality problem of having English-language messages whose meaning may be needlessly unintelligible to non-English speakers.

5. There is a certain inconsistency in the use of text and images to convey information. For example, the use of purely textual error messages like "OFF MAP" is not consistent with other error alerts that have a graphical representation, such as the "no GPS" error icon (see Figure 5.4). Graphical representation is also generally preferable over text.

6. The speed limit sign does not indicate the speed measurement units (i.e., mph or km/h). Moreover, it supports only km/h.

7. The prototype has no audio, which is a recommended feature for similar devices[6].

8. The CARAT counter is too dependent on the availability of a GPS signal that is not precise enough. The use of more accurate signals such as EGNOS might solve some of these problems.

As a result of the heuristic analysis, a new interface structure was suggested. Figure 5.9 shows the wireframe of the display, which was divided in three parts: percentage of points earned in the current trip, driving behavior trend, and current speed limit conformance.

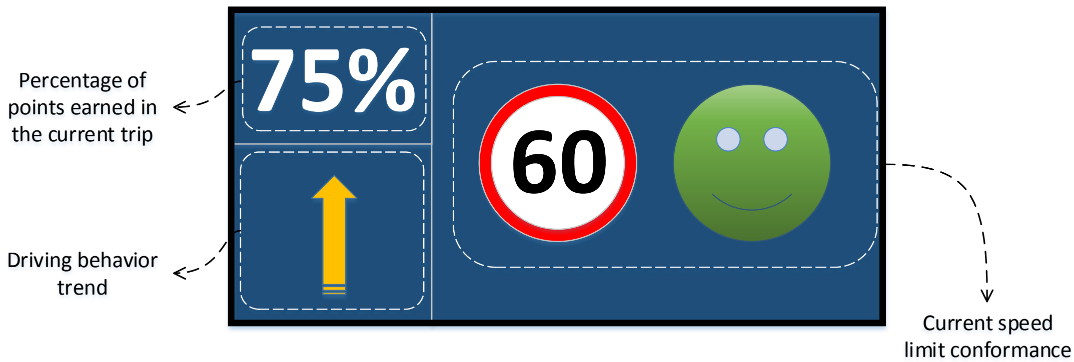


Figure 5.9: Wireframe of the suggested display after the heuristic analysis.

The driving behavior trend is represented in Figure 5.10. There are four icons that represent the four possible trends in the user's driving behavior, namely, "good", "bad", "improving", and "neutral" (see Table 5.1).

[6]For example, [58] demonstrated the benefits of duplicating visual information through sound to remedy some of the difficulties in reading textual information on systems with small-screen displays that are used when in motion.

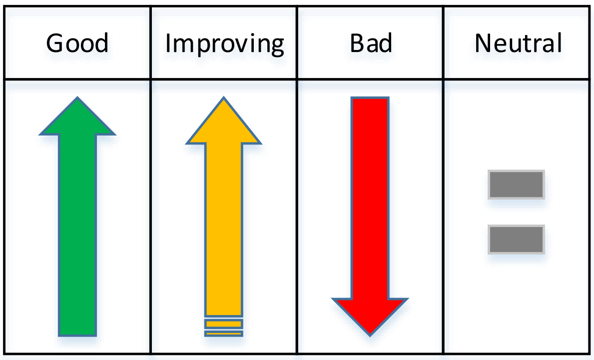


Figure 5.10: Icons representing the driving behavior and their corresponding meaning.

5.9 Subjective Analysis

The next step in the usability study consists in performing tests of the first working prototype of the CARAT counter (i.e., the field prototype) with real users in actual driving situations, and subsequently inquire them about their opinions.

The users had no previous hands-on experience with the prototype, but its basic functioning had been described to them. The users were not assigned a fixed set of predefined tasks and were told instead to use the prototype during their daily work routine. The goal was to perform realistic and natural tasks that did not interfere with the users' jobs.

After testing the prototype, the users were asked to answer a usability questionnaire. The questionnaire included 65 items inquiring about the usability of the prototype. The items were specific questions that follow almost directly from the usability taxonomy, with two caveats: (a) the number of questions must seem reasonable to the users, and (b) some questions must be slightly rephrased in order to make them as clear and self-explanatory as possible to the participants. The questionnaire covered all the first-level attributes in the usability taxonomy (see Figure 5.11). However, not all the lower-level subattributes are covered, as there are some aspects that the user does not necessarily know about (e.g., economic costs) and some system functionalities had not yet been implemented.

Some examples of the items included in the questionnaire, and their correspondences with the usability attributes, are shown in Table 5.5.

The rest of the questionnaire follows the same structure. Except for the final questionnaire item, which inquires the user about the overall usability of the CARAT counter, all the items in the questionnaire are phrased as positive statements for which the user could indicate a degree of agreement. The answers are on a Likert scale, from 1 to 5. The associated meaning of these values is shown in Tables 5.6

Table 5.5: Excerpts from the Subjective Questionnaire of the CARAT Counter

Usability attribute	**Questionnaire item**
Knowability	
Clarity	
Visual clarity of the displayed text	The displayed text is easy to read.
Visual clarity of the displayed graphics	The displayed graphics are easy to see.
Clarity of the meaning of the physical components	The purpose of the physical components is clear.
Clarity of the meaning of the displayed text	The meaning of the displayed text is clear.
Clarity of the meaning of the displayed graphics	The meaning of the displayed graphics is clear.
Clarity of the meaning of the structure of the displayed elements	The purpose of the division of the display into zones is clear.
...	
Consistency	
Visual consistency of the displayed text	The appearance of the displayed text remains consistent throughout.
Visual consistency of the displayed graphics	The appearance of the displayed graphics remains consistent throughout.
Consistency in functioning of user tasks	Similar user actions consistently follow similar steps.
...	
Memorability	
Memorability of the meaning of the displayed text	The meaning of the displayed text is easy to remember.
Memorability of the meaning of the displayed graphics	The meaning of the displayed graphics is easy to remember.
Memorability of the functioning of the user tasks	The actions that the user must perform are easy to remember.
Memorability of the functioning of the system tasks	The actions performed automatically by the CARAT counter are easy to remember.
...	

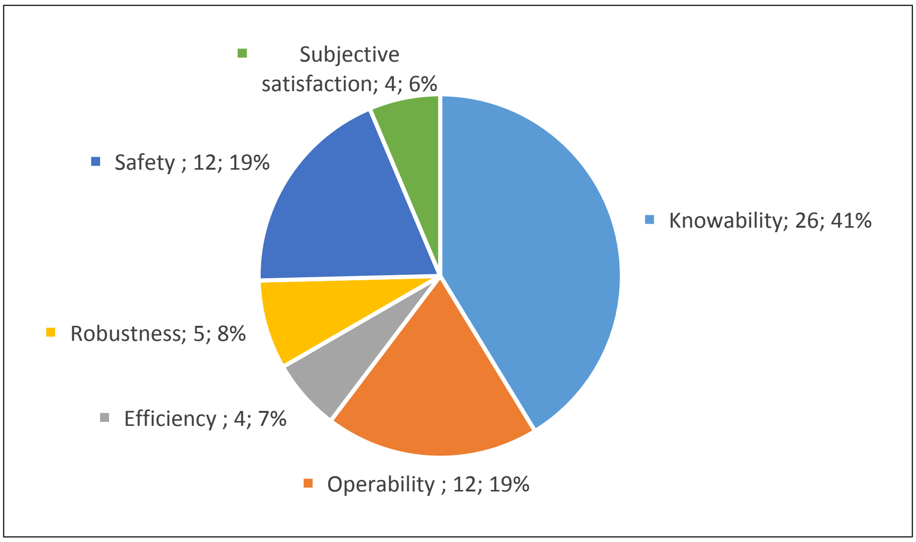


Figure 5.11: Distribution of the questionnaire items for each first-level usability attribute, indicating the number of items and the percentage.

and 5.7 (the latter corresponding to the final questionnaire item). The questionnaire also has a text box for adding comments or observations.

Table 5.6: Possible Answers to the Items in the Usability Questionnaire

Value	**Meaning**
1	Strongly disagree
2	Disagree
3	Neutral
4	Agree
5	Strongly agree
DK/NA	Don't Know/Not Applicable

Note. This applies to all items except the final "overall usability" item.

The subjective analysis was conducted in 2010 with four actual users. Nielsen [132] recommends as a rule of thumb[7] a maximum of five users for nonquantitative studies like this, although, obviously, this depends on the particular characteristics of the study and the subject sample [34].

All the participants in the study were male and from the United Kingdom but were reasonably diverse in all other respects. Their year of birth varied between 1963 and 1983 (the average age at the time was 35). The level of computer expertise varied from low to high. Their annual mileage was between 1,000 and 6,000, except for one user who had 60,000. Only one driver had been given speeding fines or penalty points (3 points at the time).

A summary of the ratings given by the users, and grouped for each second-level usability attribute, is shown in Figure 5.12 (it should be kept in mind again that not every criterion of the usability taxonomy was applicable).

[7]See also the early research by Nielsen and Landauer [133] and Virzi [182].

Table 5.7: Possible Answers to the "Overall Usability" Item in the Usability Questionnaire

Value	Meaning
1	Very bad
2	Bad
3	Adequate
4	Good
5	Very good

The aggregated ratings shown in Figure 5.12 were obtained through the technique known as Multi-Attribute Utility Theory (MAUT) [13]. MAUT analysis is a formal subjective multicriterion analysis technique, used in usability environments to assess the utility of systems or alternatives that have more than one evaluable attribute. The procedure for a MAUT analysis is as follows:

1. Specification of the evaluation criteria and attributes
2. Weighting of these criteria and attributes according to their relative importance.
3. Testing of how the system complies with each of the defined attributes.
4. Creation of utility functions that will convert the above mentioned scores into utility measures.
5. Integration of the utility values obtained for each attribute into a single measure.
6. Sensitivity analysis.

Due to the lack of information on the context of use at this point – which would provide further information on which usability aspects are more relevant to the users – it was decided to use linear utility functions and give equal weight to each criterion in order to avoid bias. If reaching a consensus on the relative weights of the taxonomy criteria is deemed necessary, techniques exist for this purpose, such as Analytic Hierarchy Process [163]. But this was out of the scope of this usability study, as the ultimate goal was not to come up with an overall usability score but to detect strong and weak points in the usability of the CARAT counter. Note that, as shown in Figure 5.12, the aggregated usability score is 3.17 and the overall usability rating given by the users is 2.75. The percentage difference[8] of the two values is 14.19%.

Based on the ratings and the observations made in the comments boxes, the strong points of the CARAT counter would be as follows:

1. **Clarity**. The displayed elements are legible, the functioning is intuitive, and the external components are big, tactile, and easy to identify.

[8]Calculated as the difference divided by the average.

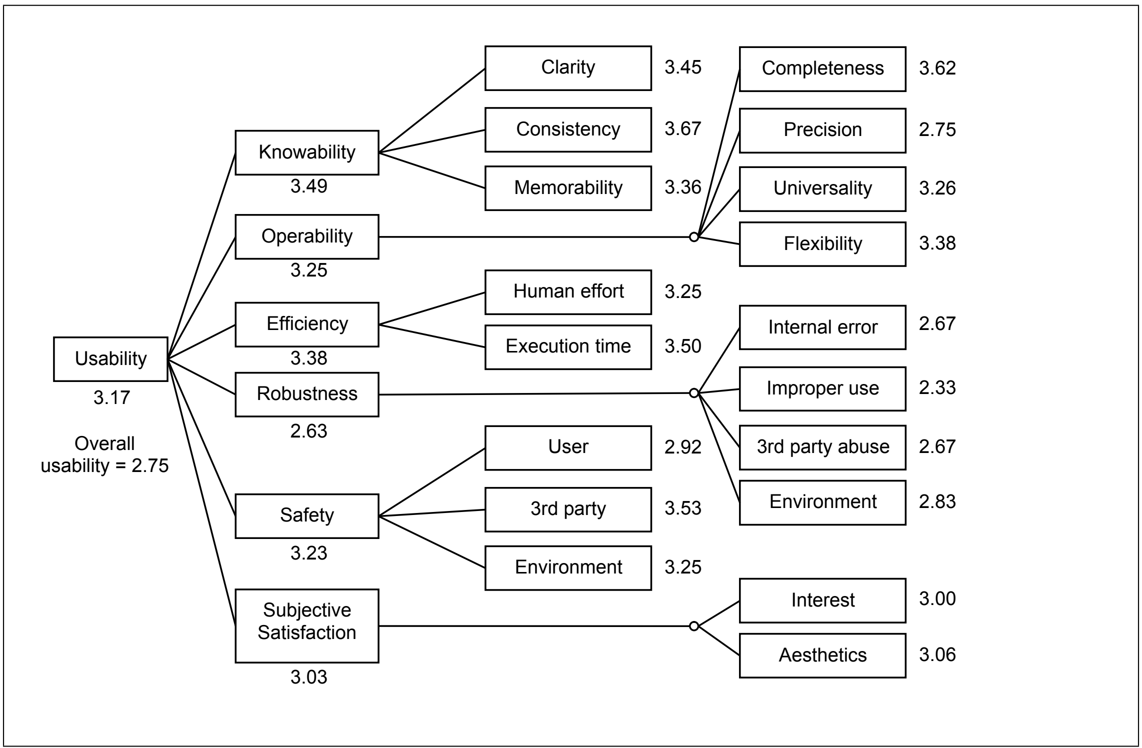


Figure 5.12: Average ratings given by the users in the subjective questionnaires, on a scale from 1 (*lowest*) to 5 (*highest*). *Note.* The "overall usability" rating is an independent item in the questionnaire and is rated separately by the users.

2. **Consistency** of design, which contributes to making the CARAT counter easy to understand.

3. **Completeness**. The CARAT counter includes all the functionalities that the users need.

4. **Efficiency**, particularly in task execution time.

5. **Confidentiality** in the use of the product.

6. **Legal safety**.

The main criticisms expressed by the users are as follows:

1. **Lack of precision** in the reception of the GPS signal.

2. **Lack of robustness to internal error**. A user specifically remarked that the counting of CARATs sometimes seems to stop working.

3. **Lack of robustness** of the USB connector (*to improper use*).

4. **Bodily safety**. The CARAT counter can be distracting to the point of being a safety issue, at least for some users.

5. **Mental safety**. The CARAT counter made some users feel emotionally uncomfortable.

6. **Aesthetics**. The external shell of the CARAT counter is aesthetically unappealing.

5.10 Summary of the Results and Suggestions for Improvement

The overall results of the usability study suggest that the usability of the current prototype of the CARAT counter is adequate in general, but with some significant flaws that need to be improved upon in the next stage of development of the product. Most of the problems found, however, are due to the fact that the device was just a prototype used to test interest and adoption for stakeholders. Nevertheless, these problems must be taken into account in the following design and implementation phases.

Taking together the heuristic evaluation and the subjective analysis, the most significant results of the usability study are summarized in Tables 5.8 and 5.9. The results are also classified according to the attributes and subattributes in the usability taxonomy.

It should be borne in mind that the classification of correspondences has been simplified for Tables 5.8 and 5.9 for conciseness. For example, the usability result "text on left side of display is small" would actually correspond to "visual clarity *of the size of the displayed text*" (which is even closer to the actual usability result). This is obtained as follows:

- The result is initially derived from the usability taxonomy subattribute *visual clarity of elements*, which is in turn a subattribute of *clarity*, which is a subattribute of *knowability*.

- The word "elements" in "visual clarity of elements" is instantiated with the distinct elements identified in the characterization of the CARAT counter (see Table 5.6). In this case, the element being analyzed is the *displayed text*.

- Each of the elements identified in the characterization of the CARAT counter has certain attributes whose usability must be examined (see Table 5.3). In this case, *displayed text* was identified as having the usability-relevant attribute of *size*.

In addition to detecting usability problems, it is also necessary to suggest concrete redesign proposals [79]. After assessing the strong and weak points of the CARAT counter, the most significant suggestions for improvement would be as follows:

Table 5.8: Results of the Usability Study of the CARAT Counter (CC), Classified according to the Usability Taxonomy by Alonso-Ríos et al. (part I)

	Usability Results for the CC	**Usability subattributes**
	Knowability	
1.	Text on left side of display is small	Visual clarity of elements (size)
2.	Graphics on right side of display are big	
3.	Text on left side of display has bad contrast	Visual clarity of structure (color contrast)
4.	Graphics on right side of display have good contrast	
5.	End-of-journey button has bad contrast	
6.	Display is structurally simple	Visual clarity of structure (positioning)
7.	External buttons have unclear purpose	Conceptual clarity of elements
8.	End-of-journey button does not look like a button	
9.	Speed limit sign does not indicate units	
10.	Error messages have unclear meaning	
11.	Pushing end-of-journey button before stopping engine is not intuitive	Clarity in functioning of user tasks
12.	Errors are arbitrarily represented as text messages or graphics	Visual consistency of elements
13.	Messages of behavior text and smiley sometimes seem contradictory	Conceptual consistency of elements
	Operability	
14.	CC includes necessary functionalities	Completeness
15.	GPS signal is not precise enough	Precision
16.	Smiley has alternatives to color for users with color blindness	Visual accessibility (color)
17.	Text is English only	Cultural universality of language
18.	Speed limit sign is km/h only	Cultural universality of conventions
	Efficiency	
19.	CC does not demand much physical effort	Efficiency in human physical effort
20.	CC does not demand much time	Efficiency in task execution time

Note. Where relevant, instantiated elements are shown in parentheses.

Table 5.9: Results of the Usability Study of the CARAT Counter (CC), Classified according to the Usability Taxonomy by Alonso-Ríos et al. (part II)

	Usability Results for the CC	**Usability subattributes**
	Robustness	
21.	Counting seems to stop working	Robustness to internal error
22.	USB is vulnerable to abuse	Robustness to improper use
	Safety	
23.	CC may be distracting to the point of being dangerous	User bodily safety
24.	CC may cause emotional discomfort	User mental safety
25.	CC does not compromise legal safety	Legal safeguarding of user
26.	CC does not compromise confidentiality	User confidentiality
	Subjective satisfaction	
27.	External shell is unattractive	Visual aesthetics

- The size of the text on the left side of the display should be maximized and good contrast should be sought in all the displayed elements.
- The end-of-journey button must be more visible and its purpose must be made clearer.
- More consistency in the visual representation of errors should be sought.
- The speed limit sign needs to specify the speed measurement units (mph, km/h).
- The CARAT counter should support other languages besides English. Graphical representations should also be prioritized.
- The precision problems need to be solved, particularly the GPS issues. In this regard, using the EGNOS technology can improve the reliability and accuracy of the signal.
- Robustness to abuse of the USB connection needs to be improved.
- It is very important to avoid distracting the users. This is not due to the complexity of the display, as the CARAT counter is simple[9]. Rather, the problem seems to lie in the demands the information places on the attention of the users.
- The external shell should be redesigned to make it more appealing, and the two buttons on the external shell ought to be removed.

It may also be advisable to include audio in the device. Because sound effects can easily go unheard, voice messages might be generally preferable. Moreover, synthesized speech may sound artificial, whereas digitized prerecorded voice messages usually sound natural and acceptable to the driver [40].

[9]For example, it follows the recommendation that "five or fewer information items (consisting of simple phrases, icons, sign graphics, etc.) should be presented at one time visually" [114, p. 3].

Chapter 6

A Usability Study of Automatically-Generated User Interfaces

6.1 Introduction

This chapter describes a usability study conducted from November 2011 to March 2012 in Vienna's University of Technology. The study was a collaboration between the LIDIA group through the author of this thesis and the staff of the Institut für Computertechnik (ICT) of Vienna's University of Technology. The work was done under the direction of Prof. Hermann Kaindl and the guidance of the supervisors of this thesis. The author of this thesis stayed in Vienna for three months, during which meetings took place regularly. This allowed him to work closely, and for an extended period of time, with the developers of the system under study. After that, the collaboration and further processing of the results continued via email and video chat.

The system under study is a set of user interfaces (UIs) composed of automatically generated web pages written in HTML (HyperText Markup Language), CSS (Cascading Style Sheets), and JavaScript. Starting from a high-level model of the interface, an automated multi-device UI generator [152] is able to produce different versions of the same UI. The generation process can be customized according to certain parameters, and the resulting UIs can be also modified manually. Although the resulting UIs are typically sets of web pages, other technologies might be employed (for example, earlier versions of the tool produced Java Swing interfaces). The hardware on which the UIs were tested consisted of touch devices, namely, smartphones and tablets.

Several types of usability activities were employed in this study:

1. **Heuristic evaluation**, prior to actual user involvement.

2. **User tests**, complemented with **questionnaires**, **interviews**, and **video recordings**.
3. **Performance measurements**, which were directly extracted from the video recordings and subsequently subjected to statistical analysis.

The general methodology for this study was analogous to the one followed for the usability study of the CARAT Counter (see Chapter 5). That is, both studies begin with a characterization of the system and its context of use, which is then followed by a combination of inspection, inquiry, and testing usability techniques. In fact, some of the techniques – such as heuristic evaluation, usability testing, and user questionnaires – are common to both studies. There are, however, some significant differences that allowed the author of this thesis to explore different aspects of usability engineering, as will be shown throughout this chapter and further discussed in Chapter 8.

The bulk of this chapter is structured as follows: (1) Some background is provided by reviewing the relevant literature for the user tests; (2) the automated UI generator is introduced; (3) the UIs whose usability is under study are described; (4) characterizations of the UIs and the context of use are provided; (5) the heuristic evaluation is described in detail; (6) the methodology for the user tests is explained; and (7) the results of the tests are analyzed.

6.2 Background

Small mobile devices like the ones used in this usability study pose significant challenges to usability researchers, especially with regard to the input method and the availability of screen real estate. The input method for these devices is touch-based, which is markedly different from the WIMP (Window, Icon, Menu, Pointer) interfaces that have traditionally dominated Human-Computer Interaction. Consequently, this requires rethinking many of the previously held assumptions in the field. The other significant factor from a usability standpoint is space availability – if the contents do not fit into the display, designers are forced to choose between partitioning the contents in some way (e.g., in tabs) or using scrolling, which would require swipe gestures on the part of the users. Questions like these make mobile interaction a relatively unexplored and thriving area of research, and a brief overview of the relevant literature for this particular usability study is given below.

Balagtas-Fernandez et al. [21] conducted a study evaluating different UI designs and input methods for touch-based mobile phones. The study was focused on three characteristics of the UIs: layout (i.e., scrolling vs. tabbed), input (i.e., keyboard vs. tapping through a modal dialog), and menu access (device menu vs. context menu). Two UIs were studied, which differed in all three characteristics. The users were assigned three predefined tasks, each of which was meant to focus on one of the three characteristics mentioned previously. However, as the UIs in that study are so radically different from one another, it is hard to prove that only the specific characteristic being investigated accounts for differences in performance, and that there is no influence from other characteristics. In order to avoid this, the usability study in this chapter is focused on UIs that differ only in their layout and are otherwise identical.

Recently, Lasch et al. [109] investigated touch-screen scrolling techniques and their effect on car driver distraction. Their results indicate that swiping was less distracting while driving than traditional scroll buttons or kinetic scrolling. The latter method uses the same gesture as swiping, but it accelerates the menu – the menu stops automatically after deceleration or if the user taps on the screen. Their research, then, investigates three different ways of scrolling in a context where distraction is a safety risk. In contrast, the primary concern of the usability study in this chapter is to investigate the relative usability of different layouts in a context where the user is focused on a specific task. Nevertheless, the conclusions of both usability studies complement each other and contribute to the literature on the topic from different angles.

Other comparable usability studies have been performed in the past but, for example, only with desktop devices [70][20][142], old cellphones [37], or PDAs [100], so their results are not necessarily valid for current smartphones. Early usability studies for touch devices (e.g., PDAs or smartphones) take for granted that the user does not like to scroll and, therefore, are primarily focused on investigating strategies to avoid or minimize scrolling [93][97][48]. For instance, Jones et al. [94] found that users preferred vertical over two-dimensional scrolling on small screens. The results in this chapter extend these findings for current touch devices through the comparison of several different layouts tailored for touch devices.

The specific type of device on which these kinds of usability studies are performed is of particular importance. Suzuki et al. [178] demonstrated that familiarity with a given device may have a significant influence on the users' perceptions of usability and correlated time-on-task. In contrast, the UIs under study in this chapter are generic HTML pages rendered on Web browsers, which helps to minimize this effect and, consequently, the differences in background among the users.

6.3 The Automated Multi-Device UI Generator

The UI generator [152] developed by the ICT is an intelligent system that automatically[1] generates user interfaces from Discourse-Based Communication Models. These kinds of models are device-independent and are intended to specify high-level communicative interactions between a user and an application. The generation process is broken down into several steps [143], which are depicted in Figure 6.1 and explained below.

Firstly, the Communication Model that represents the interaction is defined by

[1] Or, using Raneburger's [152] terminology, "(semi-)automatically", in the sense that "additional input from the designer is possible and in general required to achieve a satisfactory level of usability for the resulting GUI, but not mandatory". [152, p. 47]

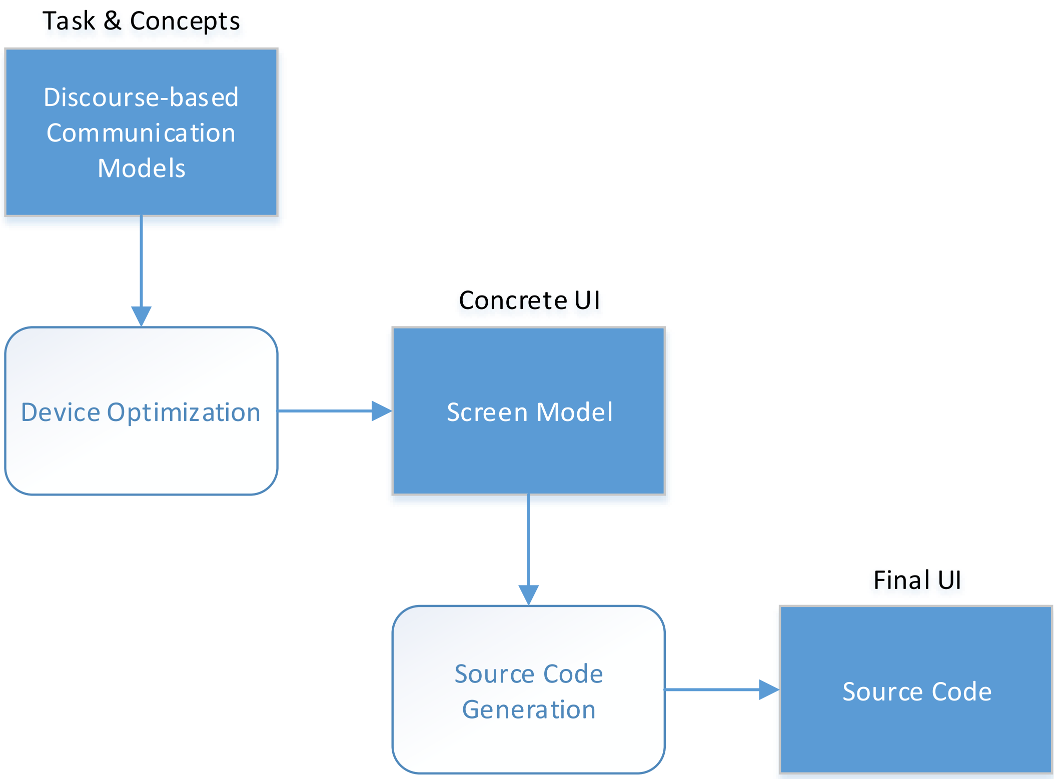


Figure 6.1: UI transformation process (adapted from [143]).

using a purpose-built editor based on the Eclipse Modeling Framework[2]. Figure 6.2 shows a screenshot of the editor, using a bike rental Communication Model as an example.

The generator tool receives a Discourse-Based Communication Model like this as an input and then performs a device-optimization step to tailor the Communication Model to specific platforms with the aim of avoiding issues such as unnecessary scrolling. In order to achieve this, it is necessary to provide information on constraints such as screen size and resolution.

As a result, the generator tool produces a Screen Model that corresponds to the Concrete User Interface Level of the CAMELEON [42][43][44] Reference Framework. This Screen Model is then converted into source code, for which a distinction is made between the dynamic and the structural parts of the model. For the dynamic part, a UI Controller is generated by using Xpand[3]. For the structural part, Apache Velocity[4] templates for HTML pages are generated.

The run-time environment on which the source code is interpreted is based on

[2]http://www.eclipse.org/modeling/emf/
[3]http://wiki.eclipse.org/Xpand
[4]http://velocity.apache.org/

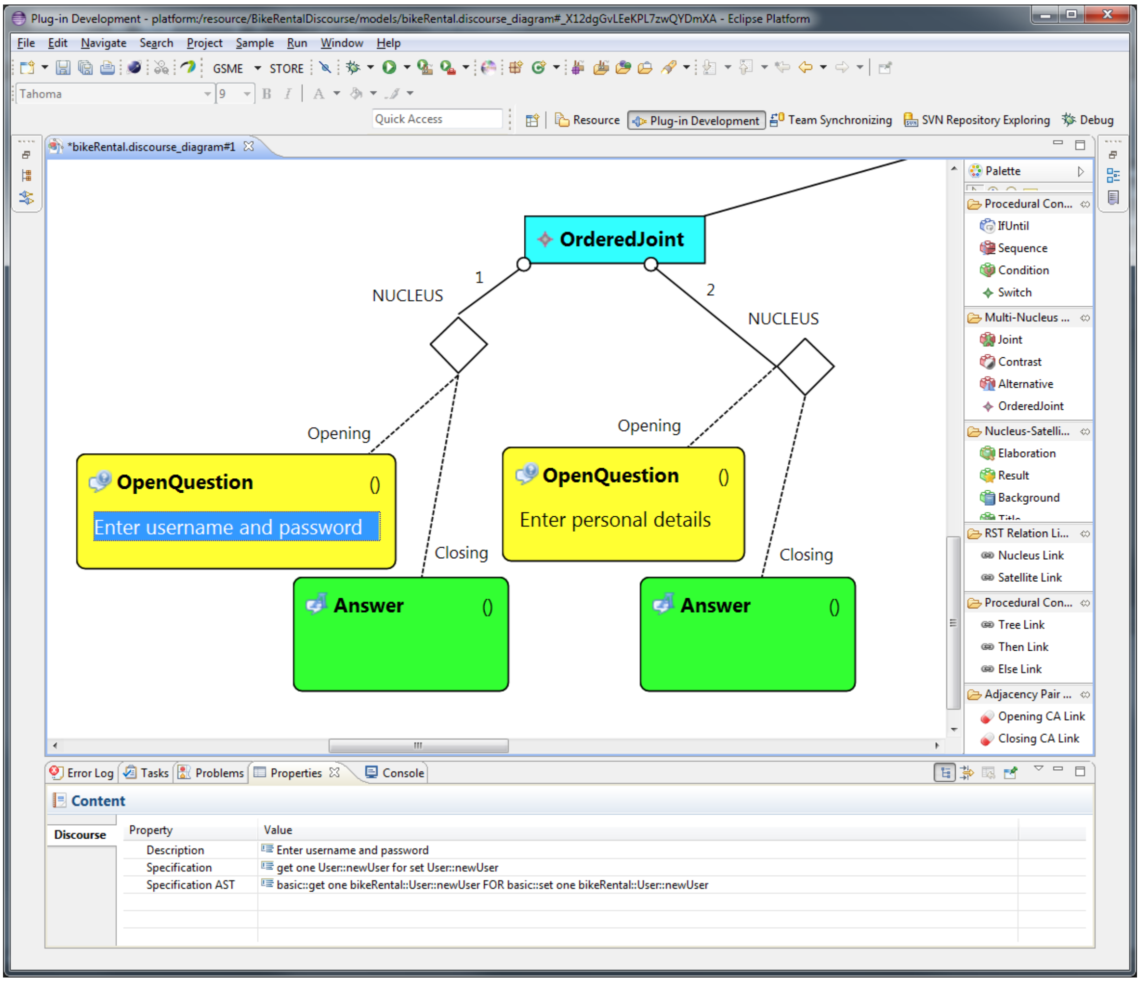


Figure 6.2: Communication Model editor (reprinted from [143]).

the Model-View-Controller (MVC) pattern [105], with the generated UI Controller acting as the Controller in the MVC pattern and the Velocity templates as a main part of the View component. During runtime, a Java-based web server such as Apache Tomcat[5] or Jetty[6] receives HTTP requests. The UI Controller then decides which page to present. The corresponding Velocity template is dynamically filled with the values received from the application logic in order to create the HTML page. It is also possible to perform a manual customization of the generated UIs.

The automatic generation process uses Artificial Intelligence techniques to determine the best way to arrange the elements of the UI [152]. The possible UIs constitute a search space. As an example, Figure 6.3 shows a search tree in which the leaves are the resulting UIs and the nodes are partially constructed UIs that are built step by step starting from the root. In Figure 6.3, two pieces of information are indicated for each edge: the *Communication Model pattern* that is matched and the *transformation rule* that is executed. In some cases, more than one transformation

[5]http://tomcat.apache.org/
[6]http://www.eclipse.org/jetty/

rule is applicable, which results in branches. For example, in Figure 6.3, starting from the root, the only rule that can be executed is *R1*. Similarly, the only rule that can be executed next is *R4*. From then on, two rules can be executed for each of the subsequent nodes, which results in 8 different UIs.

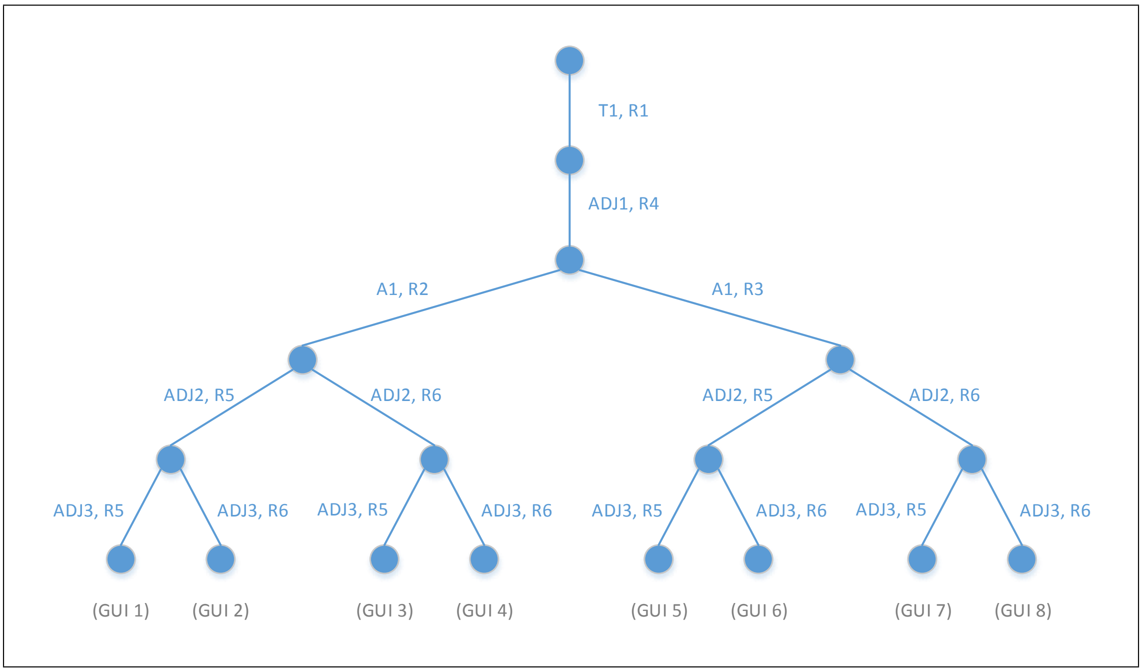


Figure 6.3: An example of a search tree for the UI generation problem (adapted from [152]).

The search algorithm is based on the branch and bound algorithm. The cost is calculated based on the following optimization objectives:

1. Maximum use of the available space.
2. Minimum number of navigation clicks.
3. Minimum scrolling (excluding list widgets).

The parameters of the cost function can be modified by the designer in order to have more control over how the UIs are tailored.

6.4 The UIs under Study

The usability study that is described next is focused on a simplified but realistic model of a real-world application for flight booking. As described in Section 6.3, the UIs are deployed as a web application running on a web server. Each UI is a slightly different – and automatically generated – version of the same set of web pages. The difference lies in the type of layout into which the UI elements have been arranged.

Table 6.1: Types of Layout under Study

Abbreviation	**Name**
T-UI	Tabbed UI
H-UI	Horizontal UI
V-UI	Vertical UI

The distinct types of layout under study will be referred to as T-UI, H-UI, and V-UI (see Table 6.1), and are described in detail below.

On T-UI, the layout has been tailored to split into tabs any screen where the contents do not fit into the display area. After entering all the corresponding data in the current tab panel, the user needs to switch to the next panel by tapping on the associated tab. This layout is intended to make the most of the space limitations of small mobile devices.

On H-UI, the layout has not been tailored for any particular device. Given the way HTML pages are rendered, this means that the contents are likely to "spill over" to the right if they do not fit into the display area. In that case, navigating through the UI would require sideways scrolling on the part of the users.

On V-UI, on the other hand, the layout has been expressly tailored to expand vertically if the contents do not fit into the display area. This is another type of layout tailored for small mobiles devices with significant space limitations.

Technically, T-UI and H-UI would also expand vertically if a given page happened to be big enough but, as the literature on usability demonstrates (see Section 6.2), this would be an unusually awkward layout, so the chosen scenario was intentionally designed to avoid it. That is to say, whenever scroll was required in this usability study, it was always one-dimensional.

Regardless of layout, the flight booking scenario consists of four individual screens that are listed in Table 6.2, along with the type of information that the user needs to enter on them.

Table 6.2: Screens of the Flight Booking Scenario

Name	**Information to be entered**
Screen 1	Origin airport Destination airport Departure date
Screen 2	Flights available on departure date
Screen 3	Billing information
Finish screen	-

Figures 6.4 to 6.7 show screenshots of the four screens for T-UI on an iPod Touch. As the contents of screens 1 and 3 exceed the space limitations of the device, T-UI divides both screens into tabs. The contents of screen 2 and the finish screen, on the other hand, fit perfectly into the display, which means that no tailoring is necessary.

In fact, these two screens will be identical for all the different layouts throughout this usability study.

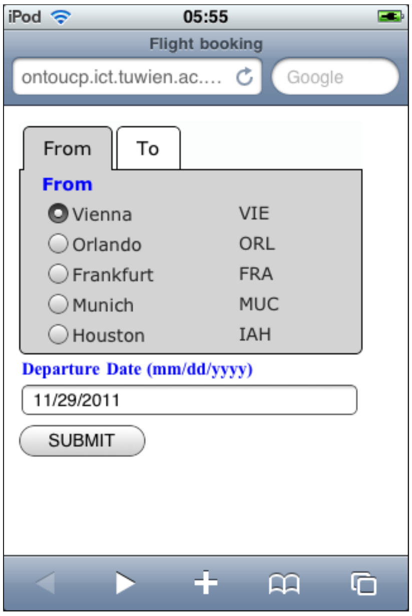


Figure 6.4: Screen 1

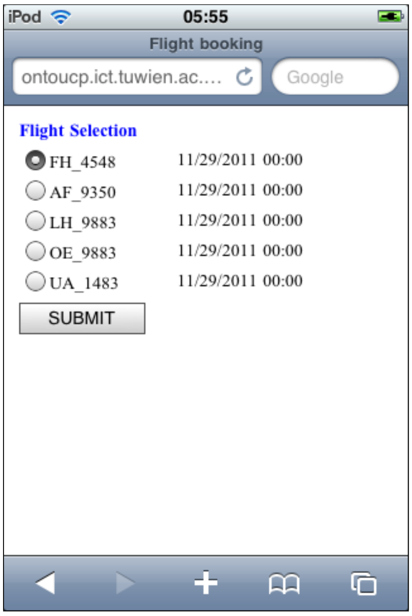


Figure 6.5: Screen 2

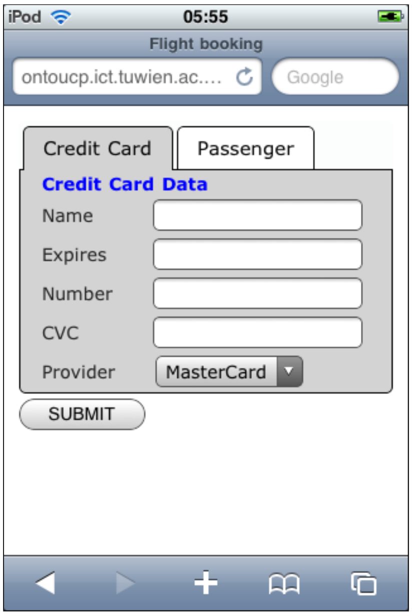


Figure 6.6: Screen 3

Figure 6.7: Finish screen

Screens 1 to 3 are composed of text and basic HTML controls. For example, Screen 1 consists of a tab panel with two tabs and radio buttons inside for choosing the origin and the destination, an editable text field for the departure date, and a "submit" button to advance to the next screen. Screen 3 is structured similarly but includes a drop-down list to choose the credit card provider. And so on. All the information to be entered in all screens is mandatory, so if the user fails to provide it, a small dialog box in the form of an error alert will pop up. Up to this point in the development life cycle, error checking is limited to ensuring that the fields are not left blank. That is to say, no other types of validation are performed.

To these screens one could also add the web server error pages for HTTP response status codes such as “500 Internal Server Error” or “502 Proxy Error”, although at this point those pages were simply the default error pages of the web server being used (in this case, Jetty).

6.5 Characterizations of the System and the Context of Use

The components of the UI are mainly HTML 4.01 elements. The fact that HTML is a standardized language composed of a relatively small number of predefined components – which are described in the HTML 4.01 specification [151] – contributed to simplifying the task. Some parts of the UI were not made up of pure HTML, however, but of JavaScript code and CSS. Complementing HTML with JavaScript and CSS is a common practice, as the three technologies fit together more or less seamlessly in a way that is typically transparent to the end user. It should be also mentioned that a significant part of the terminology used in this section for naming UI elements is taken directly from the established terminologies of the HTML, JavaScript, and CSS specifications, rather than inventing new names for them.

The system characterization that was done for this usability study is very different from the system characterization done for the CARAT Counter (see Chapter 5), which is to be expected, as the characteristics of the system can differ greatly for one usability study to another. Nevertheless, the characterization done for this study has the benefit of being much more generalizable and reusable, as the elements of the HTML specification are shared by many other languages for designing user interfaces. The analysis performed in order to achieve this characterization can be considered a first step towards defining a taxonomy of UI elements for computer applications in general. Such a taxonomy would be a substantial undertaking in itself, and is thus reserved for future work.

Table 6.3 shows the usability-relevant attributes of the UI components only in the broadest sense. That is, these attributes are basically applicable to any component in the UI, whereas the more specialized attributes of specific components (e.g., dialog boxes) will be described separately below. A distinction is made in Table 6.3 between the individual UI elements themselves and the different types of relations that exist between these elements. As in the usability taxonomy [16] described in Chapter 3, the individual elements are examined from two perspectives: *formal* (which for this type of UI means “visual”) and *conceptual*. The relations between elements, on the other hand, are part of what is considered *structure* in the taxonomy.

Next, the specific types of elements that make up the UIs will be examined individually and in detail. Some typical HTML elements, such as checkboxes, concealed

Table 6.3: Usability-Relevant Attributes for the Global Components of the UIs

Component	**Attributes**	**Subattributes**
Individual elements (visual)	Shape	-
	Size	Width
		Height
	Orientation	-
	Color	Hue
		Saturation
		Lightness
Individual elements (conceptual)	Shape	Literal meaning
		Symbolism
	Color	Symbolism
Relations between individual elements	Ordering	-
	Focus	-
	Positioning	Horizontal alignment
		Vertical alignment
		Horizontal spacing
		Vertical spacing
	Color contrast	Hue
		Saturation
		Lightness

passwords, or two-dimensional text areas, were not part of the UIs under study, so they are omitted from this characterization. Similarly, the fact that the study is focused on touch devices means that important HTML elements such as the caret and the tab indexes are also left out, as they are mainly relevant to keyboard and mouse interaction. Table 6.4 offers a summary and a classification of the UI elements under study.

Table 6.5 lists the attributes for text qua text. That is, the essential attributes shared by any type of text regardless of the type of control in which it happens to appear in the UI. The terminology for the text attributes has been borrowed from the CSS 2.1 specification [35], as HTML is somewhat limited in this regard (e.g., it makes no distinction between *font style* and *font weight*) and CSS has become the de facto standard for advanced typesetting on web pages.

Containers (see Table 6.6) are a common way to group collections of individual UI elements, to the point that containers are considered UI components in their own right and merit a separate characterization. The containers relevant to this usability study are the browser window and the tab panels. Since tabs are not part of the HTML 4.01 standard, the tabs in the UIs were implemented with CSS. Other types of containers may exist in the source code of the pages, such as borderless tables (often used for formatting purposes) or forms, but these were left off this usability study if they were completely transparent to the user, as this makes them inconsequential from a usability standpoint. It should be also noted that Table 6.6 only lists the attributes that specifically distinguish the containers from other UI

Table 6.4: UI Elements

Type	Element
Text	Text in non-editable labels
	Text in editable text fields
	Text in buttons
	...
Container	Screen
	Tab panel
Dialog box	Alert
Controls	Drop-down list
	Button
	Radio button
	Text field

Table 6.5: Usability-Relevant Attributes for the UI Text

Attribute	Subattributes	Example values
Font	Font family	Serif, Sans-Serif, Condensed, Monospace...
	Font weight	Normal, bold, light...
	Font style	Normal, italic...
	Decoration	Underline...
	Capitalization	Uppercase, lowercase...
Content	-	-

components. That is, it would be necessary to add to them the attributes previously listed in Table 6.3 for global UI elements or groups of elements (e.g., *size*, *color*, *color contrasts* between the elements contained inside, etc.). Likewise, containers can have all kinds of individual UI elements inside, such as controls and text, each one with its own specific characteristics.

Table 6.6: Usability-Relevant Attributes for the Containers

Container	Attributes
Screen	Title
Tab panels	Number of tabs
	Number of rows of tabs
	Panel size (fixed or variable)

Table 6.7 shows the usability-relevant attributes for the dialog boxes. The only type of dialog box in the UIs under study is an alert with a brief description and an “OK” button. Technically, these dialog boxes are not HTML elements – they are invoked via the JavaScript `alert()` function, passing the text to be shown as a parameter. The dialog boxes in the UIs under study correspond to validation errors that are displayed when the user fails to enter required information. An important usability consideration in this way of implementing alerts is that standard JavaScript does not allow to set a value for the *title* of the dialog box. This has always been intentional, and the motivation is to protect users from security threats such as

phishing. Whatever ends up being displayed as a title will depend on the particular browser being used, so designers should focus on making the content of the message reasonably self-explanatory.

Table 6.7: Usability-Relevant Attributes for the Dialog Boxes

Dialog box	**Attributes**
Alerts	Title
	Message

Table 6.8 shows the usability-relevant attributes for what are typically called controls, that is, UI elements that allow to manipulate information through direct interaction.

Table 6.8: Usability-Relevant Attributes for the Controls

Control	**Attributes**
Controls in general	Enabled
Input controls in general	Required
Drop-down lists	Editable
	Number of options
	Maximum displayed number of options
	Value
Buttons	Keyboard shortcut
Radio buttons	Label
Text fields (one-dimensional)	Maximum length
	Value

Some points about the controls are worth clarifying:

- As this characterization is always done from the point of view of the users, no distinction has been made between HTML's so-called "submit" controls and customized HTML buttons that superficially look identical and are in fact programmed to function in the same way (by using the JavaScript `submit()` function). Choosing between one implementation or the other is purely a programming decision that is generally of no interest to the users – assuming they notice it at all.

- Even though drop-down lists are part of the HTML 4.01 standard, *editable* drop-down lists (also known as combo boxes) are not.

- If the maximum displayed number of options of a drop-down list is exceeded, a scroll bar is shown.

- The labels attached to radio buttons and text fields can be clickable, in which case the caret or focus would be placed on the associated control (though, as mentioned above, the concepts of caret and focus are not very relevant for touch devices).

Once the characterization of the system under study has been completed, the next step is to characterize the context of use (i.e., the *users*, *tasks*, and *environments*). As with the components of the system, the exact elements of the context of use depend on the problem in question, but most of the characteristics of the context of use are common to many problems, which is why they were formalized in the context-of-use taxonomy [17] described in Chapter 4. Compared to the usability study of the CARAT Counter, the context of use was broader in some respects and more specific in others. On the one hand, the general-purpose and platform-independent nature of the automatically-generated UIs makes the context of use very unspecific compared to more specialized applications. In fact, the UIs were mainly a proof of concept and there was no actual target audience at this point. On the other hand, characteristics such as the limitations of the hardware pose significant usability challenges.

For the purpose of this usability study, the UIs are operated only by one user at a time. The most relevant user characteristics are:

1. **Experience with similar systems** and **knowledge of the system's domain and language**. The application assumes a certain familiarity with the *technology* being used (e.g., the hardware and web browser), as well as the *domain of the application* (i.e., flight booking) and the *language* in which it is implemented (i.e., English).

2. **Physical characteristics**. In particular, *aptitudes* and *disabilities* related to *sight* (as the only type of feedback is visual), *hand movements*, and *touch* (especially in terms of *precision*). *Cognitive* aspects such as *memory* are less relevant here. Nevertheless, the basic requirements for all these characteristics are not very demanding, as the application is simple and the UI elements are reasonably sized.

The tasks – in the broadest sense – are shown in Table 6.9. These types of tasks can have very dissimilar context-of-use characteristics, particularly in terms of complexity, duration, frequency, and workflow controllability.

Table 6.9: Tasks

Type	**Name**
User tasks	Navigation Data input
System tasks	Data validation Refreshing the UI after the user's input

The identified tasks can be described as follows:

1. **Navigation**. This includes moving between different screens and being able to keep track of where one is.

2. **Data input**. The user provides information that is required to fill the forms. This includes entering text, tapping on radio buttons, and so on.

3. **Data validation**. In this usability study, validation means that the UI checks that required fields are not left blank. If the user fails to do this, an error dialog box is shown.

4. **Refreshing the UI after the user's input**. In response to some interaction from the user (e.g., selecting an item from a drop-down list), the system must update the contents of the display. A small amount of time may pass to allow the user to notice changes.

The environment can be subdivided into three categories: *physical*, *social*, and *technical*. The characteristics of the environment to take into account in usability studies like this are as follows:

1. **Physical environment**. The *sensorial*, *atmospheric*, and *spatial conditions* are particularly important. Mobile applications can be operated in all kinds of physical environments, with very variable characteristics. Even though the user tests in this usability study were conducted in a controlled laboratory environment (because booking a flight is meant to be done carefully and attentively), different types of applications could be used in other types of environments.

2. **Social environment**. The versatility of mobile devices means that there could be a great diversity of social factors at play, from *interruptions* to other people's *support* or *feedback*.

3. **Technical environment**. The fact that the UI depends on specific types of technology means that these can become a significant bottleneck. For example, the display area limitations of smartphones – which in the context-of-use taxonomy would be classified under *suitability of the physical equipment to knowability* – are so significant that finding the best way to tailor the UIs to the available display area was one of the main goals of the usability study. The technical environment for this study is characterized as shown in Table 6.10.

Table 6.10: Technical Environment

Type	**Name**
Physical equipment	Hardware
Logical equipment	Operating system Browser
Consumption materials	Battery or power Internet connection

6.6 Heuristic Evaluation

The heuristic evaluation in this usability study is mainly aimed at improving the preliminary design of the UIs. The ultimate goal is to ensure that the subsequent usability tests are reasonably realistic, in the sense that the users' attention is not unduly diverted from the questions being investigated towards irrelevant usability problems. Therefore, some of the conclusions of the heuristic evaluation refer to issues specific to the flight booking scenario, while other conclusions refer to the automatically-generated UIs in general and were in fact used to improve the UI generator. The hardware used for the heuristic evaluation was the same as the one for the user tests, namely, an 320×480px iPod Touch (which for all intents and purposes is basically an iPhone without phone capability) and an Android-based Motorola Xoom tablet. Due to the simplicity of the UIs and the assistance provided by the taxonomies, the heuristic evaluation was performed solely by the author of this thesis.

6.6.1 Heuristic Evaluation of the Components

The heuristic evaluation was performed by instantiating the usability taxonomy with the characterization of the system and its context of use (the latter two were described in Section 6.5). As an example, Table 6.11 shows the heuristic evaluation for the most general components of the UIs, with the plus sign indicating a positive assessment and the minus sign a negative one. The first column of Table 6.11 shows the relevant attributes of the usability taxonomy (with some attributes, such as *confidentiality* and *legal safeguarding*, being omitted from the analysis due to the fact that the application is a proof-of-concept), whereas the second column shows the heuristic evaluation for them. Where relevant, this evaluation references the attributes of the global components previously listed in Table 6.3 (e.g., *size*, *color*, *meaning*, and *horizontal and vertical spacing*), as well as the characteristics of the context of use, particularly the *technological environment* shown in Table 6.10 (e.g., *hardware*, *operating system*, *browser*, and *Internet connection*).

This methodology was then repeated for the more specific UI components discussed in Section 6.5. It should be kept in mind again that some of the characteristics of low-level UI components are unique to them, whereas other characteristics are common to all UI components. So, for example, in order to evaluate the tab panels one should pay attention not only to the attributes in Table 6.6 but to the more general ones in Table 6.3 as well (such as *size*, *color*, etc.).

For brevity, the analysis of the more specific UI components that follows will be succinct and mainly concerned with the aspects that are of particular interest for this usability study.

Focusing first on the containers, the heuristic evaluation yielded the following information:

1. The UI title is too short and identical for all screens, namely, "Flight booking", which does not help to clarify what the current screen is.

2. The tab structure (as a reference, see Figure 6.8 for two screenshots of screen 3, one for each tab) is not cluttered, which means that it is easy to read. There is a maximum of two tabs and one row of tabs on each screen.

Table 6.11: Heuristic Evaluation of the Global Components of the UIs

Usability attribute		**Heuristic Evaluation**
		Knowability
Visual clarity of elements	+	Size is good enough.
	+	Color is adequate.
	–	Occasionally, some UI elements may not be visible, depending on layout.
Conceptual clarity of elements	+	Meaning is clear due to standardized technologies.
Visual consistency of elements	+	Visual consistency is generally good due to standardized technologies.
	=	The shape of some UI elements may vary between different platforms or browsers.
	–	Horizontal and vertical spacing between UI elements is occasionally inconsistent between different screens and also depending on orientation (portrait or landscape).
Conceptual consistency of elements	+	Meaning is consistent due to standardized technologies.
Memorability of elements	+	Memorability is good due to standardized technologies.
Helpfulness	–	No interactive help is provided, just alerts.
		Operability
Completeness	+	HTML, JavaScript, and CSS are enough for the desired UI elements.
Configurability	–	Application is not configurable.
Adaptiveness	+	Adaptiveness to different hardware/OS environments is guaranteed due to platform-independent technologies.
		Robustness
Robustness to internal error	+	Robustness to internal error is enhanced due to standardized technologies.
Robustness to environment problems	–	UI is vulnerable to Internet connection problems.
	–	UI is vulnerable to browser problems.
		Subjective satisfaction
Visual aesthetics	=	UI has a simple and unadorned look.

3. Width, spacing, and horizontal alignment are not necessarily consistent between tab panels. The height of the tab panels can also vary greatly depending on the number of controls inside the tab panel, and with it the placement of the submit button, as can be clearly seen in Figure 6.8. However, there is a trade-off between consistency and efficiency in this case, as the fact that the submit button is always close to the other controls means that less effort is required from the user.

4. The background color is not very consistent throughout the application. It is 100% white in general but 83% white (i.e., light gray) for the tab panels, which on the iPod Touch fill up most of the available space. These sudden changes in background color could be distracting for some users.

5. Tab panels have the same meaning throughout the application, namely, a ordered sequence of pages with forms that need to be filled. This consistency of purpose is a good thing, as tab panels can be used for other ends. An example of this can be found in the "advanced configuration" option of other applications, in which tabs are used for grouping a great number of (unordered, and perhaps unrelated) options simply because they cannot be easily fitted into the screen.

6. The "Credit Card Data" tab and the "Passenger Data" tab (see Figure 6.8) could be merged. From an effectiveness standpoint, it would demand fewer actions on the part of the user. From a conceptual standpoint, it would make sense, because the concepts are somewhat related. "Name" appears on both, and one needs to be an adult to have a credit card. However, there is a trade-off between these benefits and the loss of screen real estate.

7. It is always clear on which tab the focus is placed.

8. Color contrasts inside the tab panels could be higher. For example, in Figure 6.8, the pure blue of some labels ("Credit Card Data" and "Passenger Data") clashes with the gray in the tab panel. Using higher contrast schemes could make the text more legible and the UI arguably more aesthetically pleasing.

Figure 6.8: Screen 3, both tabs

After discussing items number 4 and 8 with the developers of the UI generator, it was decided to increase the lightness of the tab panels to 86% white (i,e, a lighter shade of gray). This improved the consistency between screens while still communicating clearly the fact that this is a container and where the focus is.

The usability assessment of the dialog boxes was as follows:

1. The title of all the dialog boxes is just the URL, which may be confusing for some users. As discussed before, this is simply a security feature of Safari, and its purpose is to indicate which site is displaying the alert. There are workarounds for this, but they are neither standard nor trivial. After discussing it with the developers of the UI generator, they decided to save this recommendation for the next version.

2. Dialog boxes tend to have very brief and not very descriptive messages, such as "Please select from" to remind the user that it is necessary to choose the origin airport (which, in the initial version of the UIs was simply labeled as "From"). Some messages are also too browser-centric and might be unclear to some users or too close to computer jargon. For example, the use of "submit" and "form" in the "Are you sure you want to submit this form again?" message. These comments motivated some changes in the UI in order to make the messages more self-explanatory.

3. There are some inconsistencies in terminology between the dialog boxes and the rest of the UI. For example, the "Please enter credit card owner" message box refers to the "Name" label in screen 3, and there is no "Owner" field as such.

The assessment for the UI controls was as follows:

1. It is not clear in advance which fields are required and which ones are not.

2. Non-editable drop-down lists, which are the ones used in the UIs, are less flexible than combo boxes, which are editable. Combo boxes are not standard HTML, however, and their use would require customization, which might lead to a new set of usability problems.

3. As suggested previously, the "Submit" text in the buttons is perhaps too close to computer jargon. It is better replaced with a text that describes specific actions, which was the course of action finally taken by the developers of the UIs.

4. Text fields representing dates might be improved by adding calendar controls. However, these are also non-standard, which would go against the simplicity and the benefits of using standard HTML.

5. The "expires" field in screen 3 does not specify which date format to use. This was another suggestion that was easily fixed by the developers.

The usability assessment for the text elements of the UIs yielded the following results:

1. The simultaneous use of serif and sans serif fonts seems inconsistent or arbitrary. Moreover, some users could also consider it less aesthetically appealing than a more uniform style. Given that serif fonts are generally considered more difficult to read at small sizes, it was decided to make all fonts sans serif.

2. The "From" and "To" labels in screen 1 could be substituted with more descriptive terms like "Origin" and "Destination".

3. If the "Name" label on the "credit card" tab of screen 3 refers to the client name as it literally appears on the credit card (which is an important distinction in e-commerce), the label could be changed to something less ambiguous like "Name on Credit Card".

4. The meaning of the "CVC" label may not be clear to the user. Moreover, there are many different terms for this code, and it is a standard procedure on many websites to include some kind of "what's this?" button or tooltip. (The label was finally changed to "Validation Code".)

5. The text is in English only. Moreover, other types of elements, such as icons, could be used to make the application less language-dependent.

6.6.2 Heuristic Evaluation of the Tasks

The last part of the heuristic evaluation is to assess usability from the point of view of the tasks (see Table 6.9). Similarly to the heuristic evaluation of the components of the system discussed so far, the usability of the tasks is assessed by following the usability taxonomy. Thus, each task is evaluated in terms of *clarity*, *consistency*, and so on. The difference between evaluating tasks and UI components is that the latter are more "static" and can be analyzed more or less by looking at the UI, whereas the former are more "dynamic" and the problems are best detected during hands-on interaction. However, there is some conceptual overlap between UI components and tasks, of course. For instance, since input controls are a type of control – and hence a type of system component – the heuristic evaluation of the controls may have some overlap with the heuristic evaluation of the task of data input, which is performed using said controls.

The most significant results for the heuristic evaluation of the user task of navigation are as follows:

1. Going back and forth between screens can be difficult, assuming some users would want to do that (e.g., in order to double-check the data). There are no "back" or "forward" buttons besides the ones in the browser, which, as a quick test shows, can be very unpredictable in combination with HTML forms. These browser buttons could be replaced by ones controlled by the application.
2. Users have no way of keeping track of where they are. As mentioned, all screens share the same title, namely, "Flight booking". Some kind of breadcrumb trail would make navigation more clear and reduce the mental workload of the user. (This suggestion motivated the developers of the UIs to change the title of each screen in two ways: they made the titles more descriptive and appended an "(x/y)" progress indicator.)
3. Users need to go all the way up to the top of the screen in order to switch tabs. The possibility of adding some kind of "next tab" button at the end is worth considering, although this would take up valuable screen real estate.
4. There are no actions available to the user on the finish Screen. This is inconsistent with the other screens and may even confuse some users. A "home" button could be added.

The other user task is data input. The main results of the heuristic evaluation are as follows:

1. Tapping on a text field always brings up the default keyboard, when for numeric fields it would be preferable to bring up the numeric keyboard. Moreover, the default keyboard reappears when a space is entered in the "credit card number" field. These problems are very device- or OS-specific and can be difficult to control, however.
2. A missing feature, perhaps, is that there is no easy way to clear all the fields in a screen (what in HTML is called a "reset" button). There is no way to easily clear an individual field, either. Solutions exist for this, but they would make the UI more complicated or add clutter to the screen, so there is definitely a trade-off in this case. This is certainly a well-known issue in small mobile devices.

3. Regarding freedom in tasks, a strong point is that data can be entered in any order within a screen. The screens themselves, however, are ordered in a predetermined sequence that cannot be ignored, as the application would not submit a form if any of the required fields is blank.

4. Efficiency would be improved by automatically setting the focus on the first field that needs to be filled. As mentioned, the concept of focus is significantly less important in tactile displays as it is in WIMP interfaces, but this would still be useful in the sense of moving the focus to the relevant tab panel after validation errors.

5. Time and effort would also be saved by implementing the standard technique of allowing specific types of input controls (e.g., radio buttons and text fields) to be selected when the labels attached to them are tapped. This is particularly important for radio buttons, as they are rather small and selecting them requires precision on the part of the user.

As for the system task of data validation, the heuristic evaluation identified the following issues:

1. The UI does not perform any type of validation besides checking that the fields are not blank. Some additional validations can be suggested: (1) checking that the origin airport is not the same as the destination; (2) validating dates and numeric formats; (3) checking the number of characters of the "CVC".

2. It would be advisable to take the standard precautions against hacking (validating parameters, preventing SQL injection, etc.), but this would require server-side validation, which is not currently supported by the UI generator tool.

Finally, a strange usability problem was found for the system task of refreshing the UI after the user's input:

1. At certain points, the corners of some buttons change from rounded to square, for no apparent reason. This was certainly unintentional and probably a browser bug. An example of this problem can be clearly seen in Figure 6.8.

As has been mentioned in passing, some of the suggested changes were immediately implemented, whereas other, more complex, modifications were saved for future iterations of the UI generator. Figure 6.9 shows an example of the modifications for screen 1, with the old version on the left and the new version on the right. Notwithstanding some modifications done specifically for the user tests, such as increasing the number of cities, the most noticeable changes in Figure 6.9 are:

1. The title of screen 1 has been changed from "Flight booking" (initially shared by all screens) to "Origin and destination selection (1/4)", which is more descriptive and specifies the number of steps.

2. The "From" and "To" labels have been changed to "Origin" and "Destination" in order to make them less vague and more consistent with the terminology in the new screen title, and flight booking in general.

3. The font family of the "Departure Date" label has been changed from serif to sans serif to make it consistent with the rest of the UI.

4. The text in the button has been changed from "SUBMIT" to "Search flights", which is less "jargony" and more descriptive of its specific purpose.

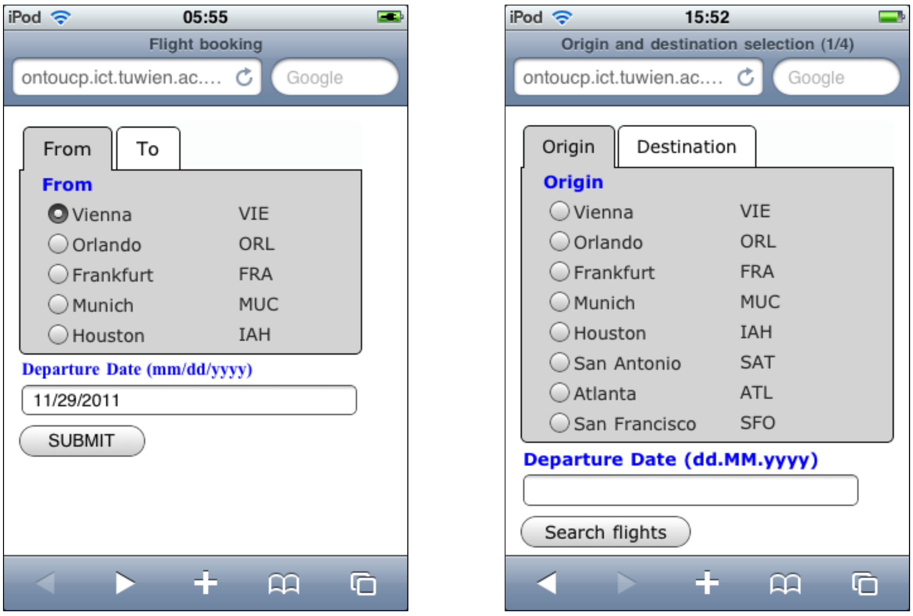


Figure 6.9: Changes in Screen 1

6.7 User Tests

The purpose of the user tests was to study the *comparative* usability of the different UIs of the application (i.e., T-UI, H-UI, and V-UI), rather than their *absolute* usability. Each user was asked to perform the same task on two different UIs in order to decide which layout was more usable. All UI combinations were tested on an iPod Touch (see Figure 6.10), and a Motorola Xoom tablet was additionally used in some tests (see Figure 6.11). More specifically, it was decided not to test V-UI on the Motorola Xoom, as it was considered that a vertical layout was not advisable for this type of device.

Each user operated the same kind of device for the two assigned taks. Similarly, the same layout was maintained on each task from screen 1 to the finish screen. That is, there were no changes of layout in the middle of a task.

In this usability study, the user tests are complemented with empirical and subjective techniques. The former include performance measurements of task completion times and error rates, whereas the latter include questionnaires and interviews.

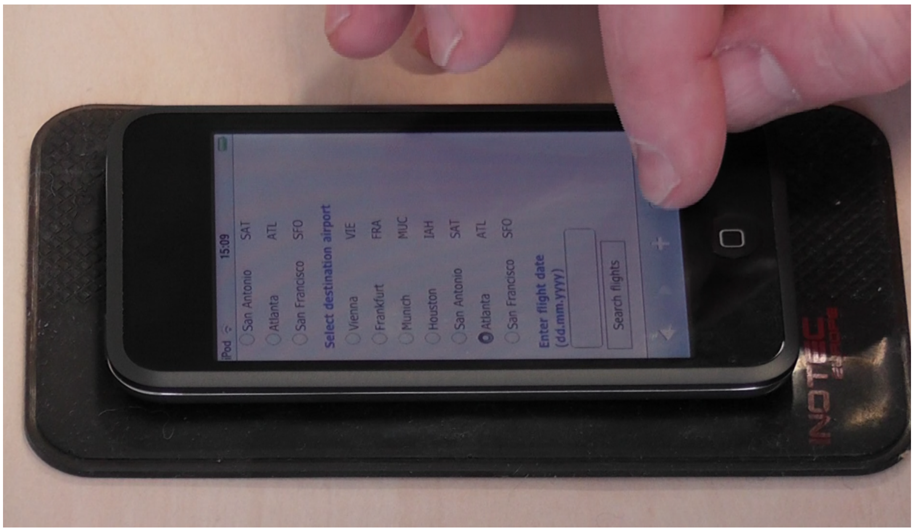


Figure 6.10: Testing V-UI on an iPod Touch

This is a fairly standard approach (used in studies such as [37] and [93], for instance), and has the benefit of yielding both quantitative and qualitative results. Moreover, studies have shown that personal preference and performance tend to be associated in general [134], but not always [67][119].

The planning and the content of the tests were discussed at length by the author of this doctoral thesis and the ICT staff during several meetings. Once everything had been agreed on, the tests were conducted by the author of this thesis with additional support by ICT staff (specifically, Drs. Raneburger and Popp) at certain points. The tests took place at the ICT itself and the language used was English.

6.7.1 Setup

The tests are divided into three test runs, as follows:

- **Run 1: T-UI vs. H-UI**. The tabbed layout and the horizontal layout are compared. This is the only test run in which both the iPod Touch and the Motorola Xoom tablet were used. H-UI is too large to fit on the iPod Touch screen, so for that device H-UI can be considered a horizontally-scrolling layout.

- **Run 2: V-UI vs. H-UI**. Vertically- and horizontally-scrolling layouts are compared on an iPod Touch.

- **Run 3: V-UI vs. T-UI**. The vertically-scrolling layout and the tabbed layout are compared on an iPod Touch.

The iPod Touch was intended to be used in portrait mode, whereas the Motorola Xoom was meant for landscape mode. Both the iPod Touch and the Motorola Xoom

Figure 6.11: Testing H-UI on a Motorola Xoom

were affixed to a table, as this ensures that all participants operate the device in the same orientation and also facilitates the task of recording to video.

As mentioned, the small size of mobile devices means that some screens in the application might not fit into the display area. This is what happened when testing H-UI and V-UI on the iPod Touch, as some components in screens 1 and 3 ended up being hidden from view. Depending on how a given screen is rendered by the browser, the components could be hidden only in part or completely. The latter situation (see Figure 6.12) is the worst-case scenario, as the user might not initially notice that there is something on the other side of the screen.

When the user does not notice that something is hidden from view, the user will probably tap the submit button without having entered some required information and the application will show an error alert informing that a particular field has been left blank, as shown in Figure 6.13.

In order to reach the other side of the screen, the user needs to make a swiping gesture. In these tests, swiping is always one-dimensional, that is, the user needs to swipe sideways on H-UI and downwards on V-UI. When the user presses and drags the finger on the screen, the iPod Touch shows a small scroll bar indicating the position relative to the entire screen, as shown in Figure 6.14. This behavior varies depending on the browser and the hardware, however.

While space limitations were a significant bottleneck on the iPod Touch, the same cannot be said of the Motorola Xoom tablet, as all the screens of all the UIs fit without any problems into the display.

A total of 60 users participated in the usability study. The number of participants on each run varied based on the amount of people necessary to obtain significant results. There were 20 users for run 1, which were divided into 10 users for the iPod Touch and 10 for the Motorola Xoom. There were also 10 users for run 2. Run 3

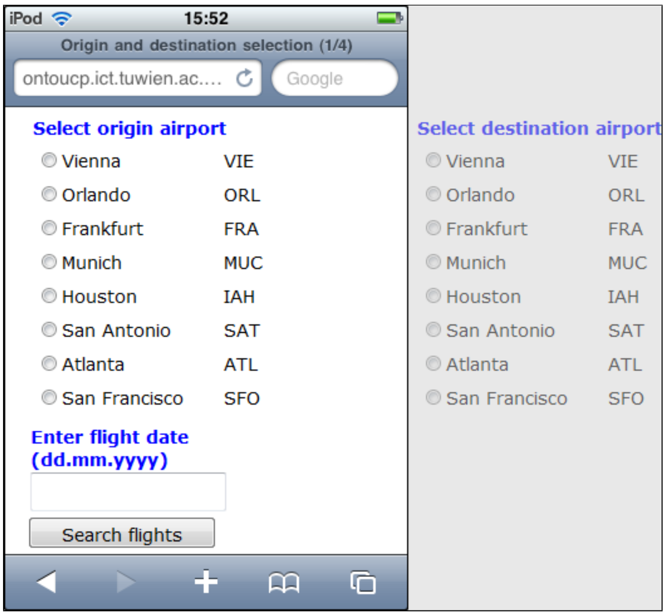


Figure 6.12: H-UI on an iPod Touch: the information does not fit into the display

had 30 users, as this was by far the most difficult comparison. To avoid bias, the participants of each run were split into two groups that differed in the order in which the two layouts were tested.

Table 6.12 shows a summary of the setup for each run, including the UIs, devices, and number of users.

Table 6.12: Test Runs

Run number	UIs	Devices	Number of users
1	T-UI, H-UI	iPod Touch	10
		Motorola Xoom	10
2	V-UI, H-UI	iPod Touch	10
3	V-UI, T-UI	iPod Touch	30

The results of the three runs were published in [154], [15], and [153], respectively.

At the beginning of each test, the participants were informed about the content of the usability study and the procedure. The participants were also informed that the goal of the experiment was to test the UIs, not their personal skills. The users were then asked to fill in some background information about themselves (see Table 6.13), which was derived from the characterization of the context of use described in Section 6.5. Finally, the users were asked to give their consent to filming their hands operating the device and recording their voices.

The participants were given a sheet of paper with instructions for completing the assigned task. The information to be entered was stated as unambiguously as possible and in the same order in which it was meant to be entered. This way, the focus of the test is diverted from the particularities of the application and the data,

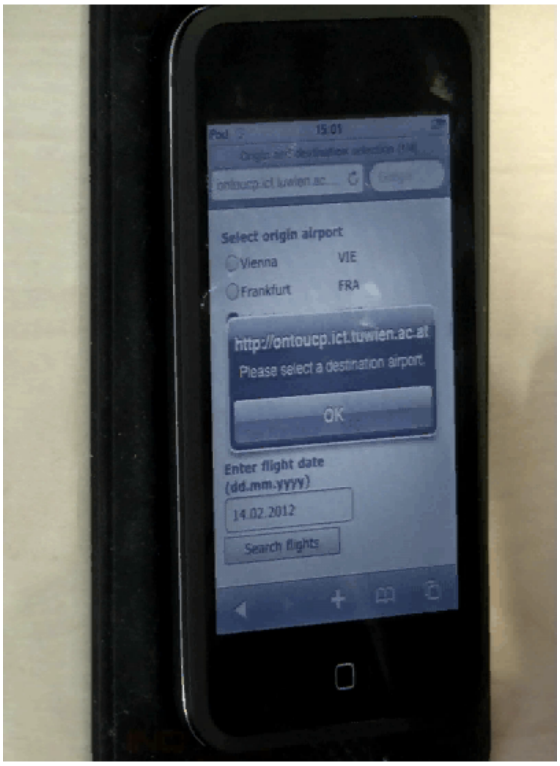


Figure 6.13: An error alert appears when the user hits submit before entering all the required information

so that users can better concentrate on the differences in usability between layouts. The sheet of paper that was given to the users describes the following scenario:

> Imagine it is Tuesday 14/02/2012, 11:55am and your boss Mr. Huber tells you to book a flight ticket for his wife as quickly as possible. Mrs. Huber is already waiting at the airport! Book a flight from Munich to Atlanta on 02/14/2012 at 1pm for Mrs. Anna Huber (age 47). Pay for it using her husband's (Max Huber's) VISA Credit Card with the number: 1258 8569 7532 1569 (validation code: 354) and the expiration date 12/14.

The sheet of paper was left on the table where it could be easily seen by the participants, so that they did not have to memorize anything. The need to perform the task quickly and realistically was also emphasized to them. Users were not allowed to ask for help if they got "stuck" – they were asked to act as they would do using a real-world application on their own.

While testing the UIs, the participants were recorded on a high-definition digital camera mounted on a tripod, and the recordings were saved for subsequent video annotation. The high resolution of the recordings allowed to discern clearly all the displayed elements in the screen, which facilitated the extraction of precise performance measurements.

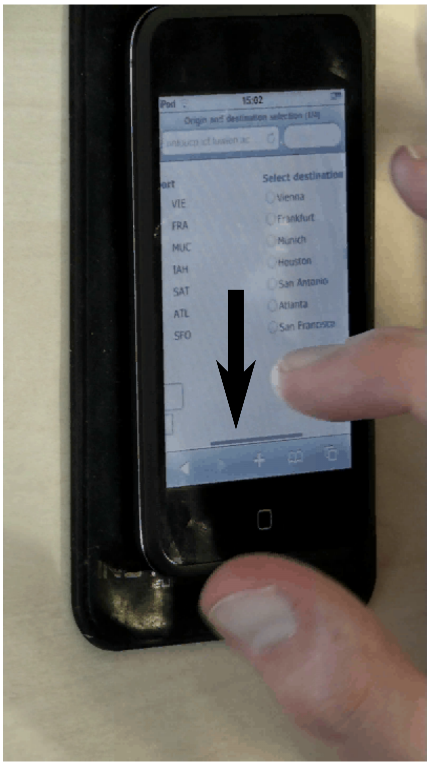


Figure 6.14: A small scroll bar appears at the bottom of the screen when the user presses and drags his finger

After the users had tested the two UIs, they were asked to fill two questionnaires on paper. Once the questionnaires had been filled, the users were asked if they had additional comments besides what they had already stated in writing. If a user wanted to add something, the author of this thesis conducted an informal interview, generally no longer than 15 minutes.

For each user, all the activities described in this section (i.e., testing, questionnaires, and interview) took place during the same session and without interruptions, which typically lasted around half an hour in total.

6.7.2 Background Information on the Users

The information on the background of the users who participated in the tests is described below. For the sake of conciseness, this information is for all the test runs, that is, the statistics correspond to the 60 users, without distinguishing between runs. Almost all users were recruited at the ICT, and consisted mostly of Electrical Engineering students, along with a few members of the faculty's staff. The gender distribution was similar to that of the students of the faculty, as only four of the 60 participants were women.

The users' proficiency in English is shown in Figure 6.15. 19 users stated that their English level was "high", 38 chose "medium", and only 3 indicated "low".

Table 6.13: Background Information on the Users

Question	Possible answers
Year of birth	Write-in.
How would you rate your English level?	Low. Medium. High.
Do you have any previous experience with smartphones, tablets or similar devices?	I have never used any of them. I have used such devices, but I have never owned any of them. I am a regular user.
Did you ever book a flight on the Internet before?	I never booked a flight on the Internet. I booked a few flights. My last booking was in... I regularly book flights on the Internet.

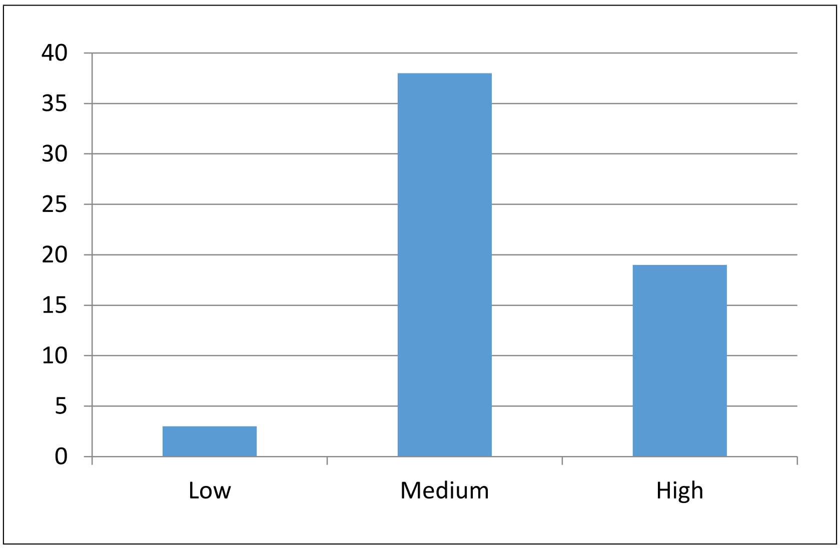


Figure 6.15: User background: English level

The users' response to the question "Do you have any previous experience with smartphones, tablets, or similar devices?" is shown in Figure 6.16. 41 users reported being regular users, 17 said that they had used similar devices before but had never owned one, and only 2 answered that they had never used a device like that before.

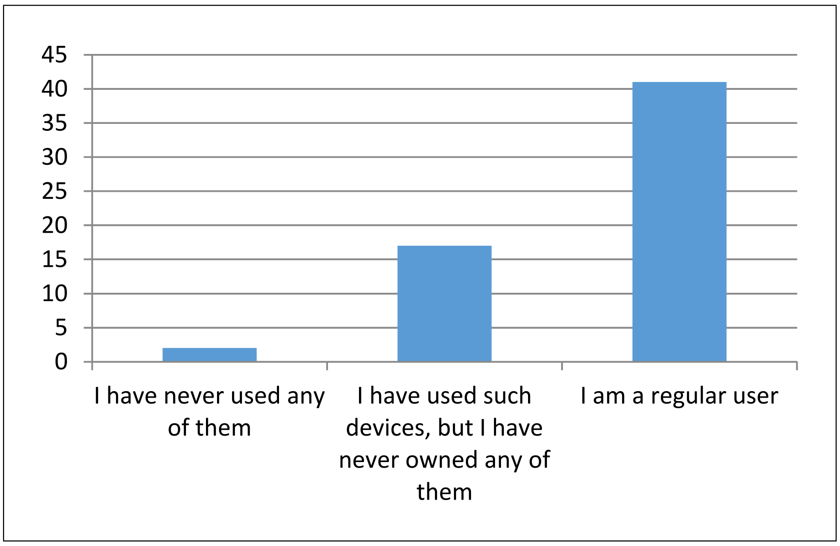


Figure 6.16: User background: Previous experience with devices

The users' previous experience with booking flights on the Internet is shown in Figure 6.17. 14 users reported booking flights on the Internet regularly, 20 answered that they had booked a few flights before, and 26 said that they had never booked a flight on the Internet previously.

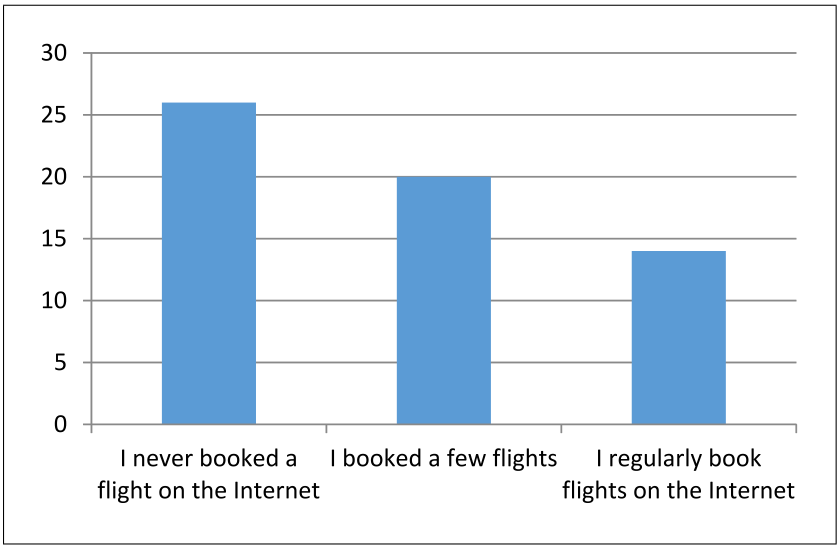


Figure 6.17: User background: Previous experience with booking flights on the Internet

The average year of birth of the users was 1986. Given that the tests were conducted in early 2012, this means that the average age of the users would be around 26. Not all users indicated their year of birth, however: 11 of the 60 users left this field blank.

6.7.3 Subjective Data

After hands-on testing of the UIs, the participants were given two questionnaires: the first one will be referred to as the *usability questionnaire* and the second one will be referred to as the *preferences questionnaire*.

The usability questionnaire aims to collect the user's opinions on the comparative usability of the two UIs in terms of a set of usability attributes. The questionnaire shows one screenshot for each UI as a reference, which is followed by the questionnaire items in Table 6.14.

Table 6.14: Items in the Usability Questionnaire

1. Which interface makes information more visible?
2. Which interface makes interaction more intuitive?
3. Which interface makes it easier to figure out what to do next?
4. Which interface makes it clearer how to use it?
5. Which interface lends itself more to how you like to work?
6. Which interface requires less manual interaction?
7. Which interface demands less precision on your part?
8. Which interface demands less time from you?
9. Which interface makes interaction more efficient?
10. Which interface is more visually attractive?
11. Overall, which interface would you use to book a flight?

The items in the usability questionnaire are derived from the usability taxonomy, but the specific choices in attributes and wording were arrived at after several discussions with the ICT staff. Two things are particularly noteworthy in this regard. Firstly, only a small subset of subattributes from the usability taxonomy are relevant to the study, as the two UIs to be tested are identical all regards except layout. Secondly, and given the high number of participants, the length of the tests, the size of the questionnaires, and the difficulty of the comparison itself, special emphasis was put in making the questionnaire as short and self-explanatory as possible. The questionnaire literature was also consulted for this, particularly the USE Questionnaire [115], the Software Usability Measurement Inventory (SUMI) [104][102][144], the W3C's WAI Site Usability Testing Questions[7], and the Cognitive Dimensions framework [33]. Reviewing this literature served to inspire the wording of specific sentences and to confirm that no significant usability aspects were being ignored.

The correspondences between the usability questionnaire and the usability taxonomy are shown in Table 6.15. The instantiations of the usability subattributes with the particular characteristics of the system and the context of use (e.g., specific tasks) are shown in parentheses. Items number 4 and 9 are intentionally redundant questions to check for consistency in the responses, as knowability and efficiency are the two usability taxonomy attributes that are represented by more than one item in the questionnaire. This consistency check was informal, however, as the qualitative nature of the questions, along with the imprecisions and ambiguities of natural

[7] http://www.w3.org/WAI/EO/Drafts/UCD/questions.html

language, make it difficult to subject them to formal analysis. At certain points (i.e., items number 1 and 2), two usability subattributes are merged into one questionnaire item. This was done when the meaning conveyed by the distinction was considered so subtle that it did not merit two separate questionnaire items. Finally, item number 11 is a question about the overall assessment of the user. Note that the word "usability" is not actually mentioned in this questionnaire item, as the user tests are concerned with only a subset of usability attributes (i.e., the questionnaire does not assess the *absolute* usability of the UIs).

Table 6.15: Correspondences between the Usability Questionnaire and the Usability Taxonomy

Questionnaire item	**Taxonomy attribute**	**Taxonomy subattribute**
1. Which interface makes information more visible?	Knowability	Formal (visual) clarity of elements Formal (visual) clarity of structure
2. Which interface makes interaction more intuitive?	Knowability	Clarity in functioning of user tasks (data input) Clarity in functioning of system tasks (data validation)
3. Which interface makes it easier to figure out what to do next?	Knowability	Clarity in functioning of user tasks (navigation)
4. Which interface makes it clearer how to use it?	Knowability	Clarity (Redundant question)
5. Which interface lends itself more to how you like to work?	Operability	Flexibility
6. Which interface requires less manual interaction?	Efficiency	Efficiency in physical human effort (hand movements)
7. Which interface demands less precision on your part?	Efficiency	Efficiency in physical human effort (precision of touch)
8. Which interface demands less time from you?	Efficiency	Efficiency in task execution time
9. Which interface makes interaction more efficient?	Efficiency	Efficiency (Redundant question)
10. Which interface is more visually attractive?	Subjective satisfaction	(Visual) aesthetics
11. Overall, which interface would you use to book a flight?		

For each question, the user had to indicate the preferred UI on a Likert scale, with preference being "extreme", "strong", "moderate", and "equal". As all UIs are identical except in layout, users were expressly asked to only focus on the differences between UIs and ignore their similarities.

The other questionnaire, that is, the preferences questionnaire, represents the priority given by the users to certain usability attributes over the others. This serves to identify personal preferences and biases on the part of the users, which helps the usability engineer to interpret and extrapolate the results of the usability study more accurately.

The preferences questionnaire asks the users the same question for all the questionnaire items: "In my opinion, it's more important that an interface...". This is followed by pairwise comparisons of four usability criteria, which are listed in Table 6.16 along with the usability taxonomy attributes from which they are derived. The main motivation for using pairwise comparisons is to compel the user to make a choice, as all usability criteria are desirable in theory, and therefore likely to receive high ratings.

Table 6.16: Usability Criteria in the Preferences Questionnaire

Usability criterion	**Taxonomy attribute**	**Taxonomy subattribute**
Makes it clear how to use it	Knowability	Clarity
Lends itself to how you like to work	Operability	Flexibility
Makes interaction efficient	Efficiency	-
Is visually attractive	Subjective satisfaction	Visual aesthetics

The four criteria in the preferences questionnaire are basically paraphrased from items in the usability questionnaire (more specifically, the items with the maximum level of generality were chosen). Like the usability questionnaire, this aims to represent a subset of aspects of a few usability attributes, namely, knowability, operability, efficiency, and subjective satisfaction. Degree of preference is stated by using the same Likert scale as in the usability questionnaire.

Figures 6.18, 6.19, and 6.20 show examples of the usability questionnaire and the preferences questionnaire, both filled by the same user. In the usability questionnaire, one can discern a preference for T-UI over V-UI in most questionnaire items. Of these, the most salient one is the one representing efficiency in manual interaction, as this is the only questionnaire item for which the user strongly prefers one UI over the other. The reason given by the user in the comments box for this item is "No scrolling required". Similarly, on the last questionnaire item, that is, the overall assessment, the user chooses T-UI and writes in the comments box that "The structure with tabs is more comfortable".

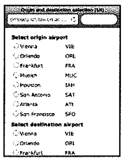
TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

USABILITY QUESTIONNAIRE

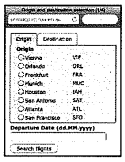
Institut für Computertechnik
Fakultät für Elektrotechnik und Informationstechnik

Instructions: For each item in the questionnaire, please check one of the boxes to indicate which interface you prefer. If you have no preference, please tick "equal". Otherwise, please indicate the degree of preference (moderate, strong or extreme). Comments are optional.

Interface #1

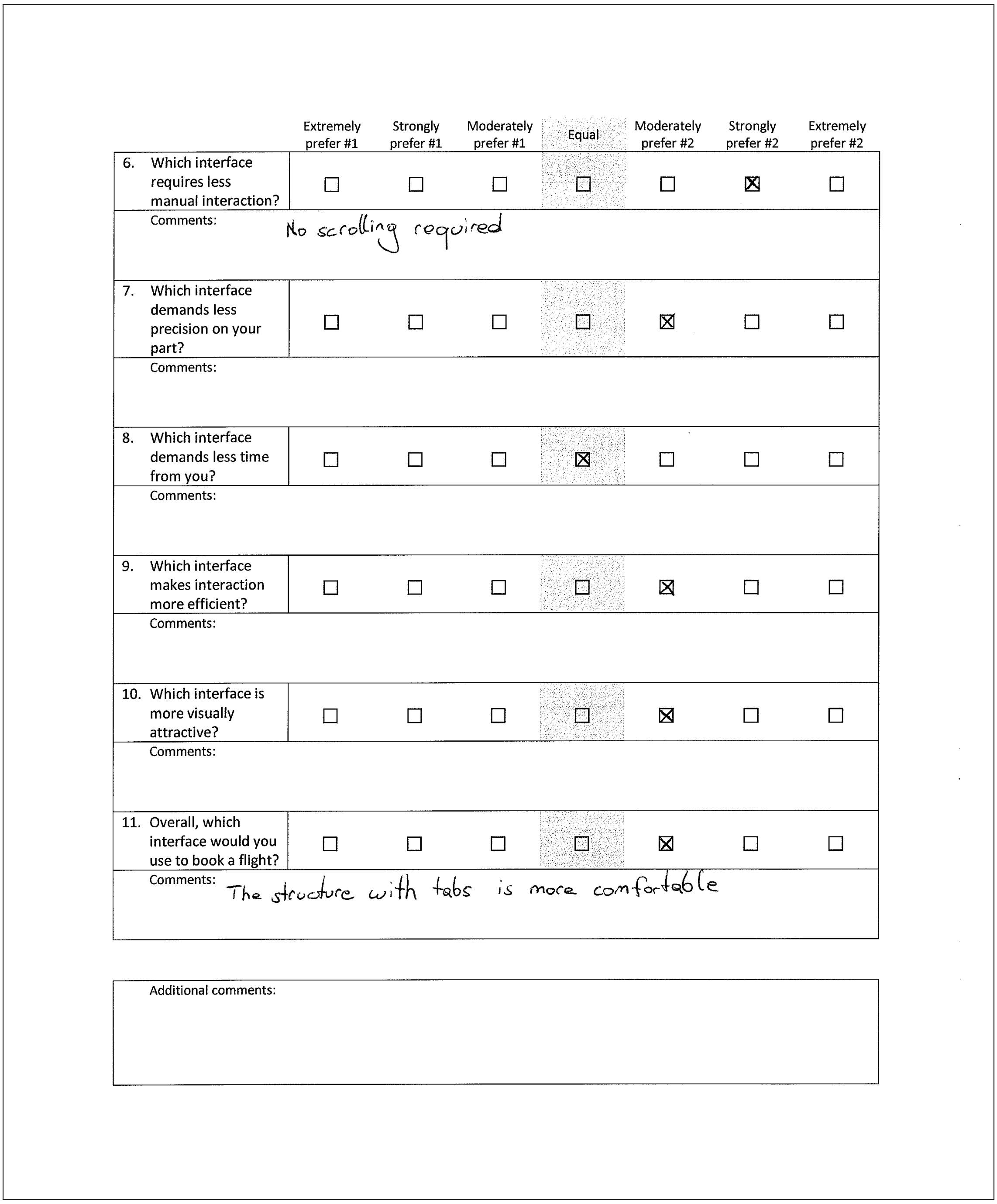


Interface #2



	Extremely prefer #1	Strongly prefer #1	Moderately prefer #1	Equal	Moderately prefer #2	Strongly prefer #2	Extremely prefer #2
1. Which interface makes information more visible?	☐	☐	☐	☐	☒	☐	☐
Comments:							
2. Which interface makes interaction more intuitive?	☐	☐	☐	☒	☐	☐	☐
Comments:							
3. Which interface makes it easier to figure out what to do next?	☐	☐	☒	☐	☐	☐	☐
Comments:							
4. Which interface makes it clearer how to use it?	☐	☐	☒	☐	☐	☐	☐
Comments:							
5. Which interface lends itself more to how you like to work?	☐	☐	☐	☒	☐	☐	☐
Comments:							

Figure 6.18: Filled usability questionnaire, page 1



	Extremely prefer #1	Strongly prefer #1	Moderately prefer #1	Equal	Moderately prefer #2	Strongly prefer #2	Extremely prefer #2
6. Which interface requires less manual interaction?	☐	☐	☐	☐	☐	☒	☐
Comments: No scrolling required							
7. Which interface demands less precision on your part?	☐	☐	☐	☐	☒	☐	☐
Comments:							
8. Which interface demands less time from you?	☐	☐	☐	☒	☐	☐	☐
Comments:							
9. Which interface makes interaction more efficient?	☐	☐	☐	☐	☒	☐	☐
Comments:							
10. Which interface is more visually attractive?	☐	☐	☐	☐	☒	☐	☐
Comments:							
11. Overall, which interface would you use to book a flight?	☐	☐	☐	☐	☒	☐	☐
Comments: The structure with tabs is more comfortable							

Additional comments:

Figure 6.19: Filled usability questionnaire, page 2

TECHNISCHE UNIVERSITÄT WIEN Vienna University of Technology

Institut für Computertechnik
Fakultät für Elektrotechnik und Informationstechnik

USER PREFERENCES REGARDING THE USABILITY CRITERIA

Instructions: Please indicate your opinions about the relative importance of each usability criterion with respect to the others. For each row, please check one of the boxes. If both criteria are equally important to you, please tick "equal". Otherwise, please choose which one is more important to you by indicating the degree of preference (moderately more important, strongly more important or extremely more important)

	Extreme	Strong	Moderate	Equal	Moderate	Strong	Extreme	
	In my opinion, it's more important that an interface…							
Makes it clear how to use it	☐	☐	☒	☐	☐	☐	☐	Lends itself to how I like to work
Makes it clear how to use it	☐	☐	☐	☐	☐	☒	☐	Makes interaction efficient
Makes it clear how to use it	☐	☐	☒	☐	☐	☐	☐	Is visually attractive
Lends itself to how I like to work	☐	☐	☐	☐	☐	☒	☐	Makes interaction efficient
Lends itself to how I like to work	☐	☐	☐	☐	☒	☐	☐	Is visually attractive
Makes interaction efficient	☐	☐	☒	☐	☐	☐	☐	Is visually attractive

Figure 6.20: Filled preferences questionnaire

Given the overall choice of interface by this user, the usability questionnaire seems to suggest that the user personally gives special importance to efficiency over, for example, aspects of clarity. This impression is confirmed by the preferences questionnaire shown in Figure 6.20, as the user shows a strong preference for an interface that makes interaction efficient. The preference for an interface that makes it clear how to use it is moderate, and no particular preference can be discerned for the other usability attributes.

6.7.4 Empirical Data

The empirical data obtained from the tests includes two types of performance measurements, namely, *completion time* and *number of errors*, which are common usability metrics.

Time to complete a task is an efficiency measurement included in standards such as ISO 9241-11 [87]. This metric has a direct counterpart in the usability taxonomy, namely, the *efficiency in task execution time* subattribute, which was also the basis for item number 8 on the usability questionnaire (i.e, "Which interface demands less time from you?").

Errors are more complex to interpret from a usability standpoint. ISO 9241-11 proposes several empirical metrics related to errors: *Percentage of errors corrected or reported by the system* and *number of user errors tolerated* are measures of *effectiveness*, whereas *number of persistent errors* and *time spent on correcting errors* are measures of *efficiency*. ISO 9241-11 also includes *rating scale for error handling* as a measure of *subjective satisfaction*, but it differs from the previous metrics in that it is subjective.

The time spent on correcting errors was not included as a separate category in the metrics of this study, however. This is because it is difficult to draw a distinction between the time spent on correcting an error and the time spent noticing or understanding that an error has occurred. This is further complicated by the difficulty of inferring the users' thoughts from the recordings, as users were asked not to make comments during the test and the camera recorded only the screens of the devices. Moreover, as all the users' interactions consist in gestures on touch devices, the absence of mouse pointers limits even more the ability to infer the users' reactions. When users are confused, they simply spend a few seconds staring at the screen, motionless. Therefore, the more realistic option was to simply measure total completion times.

Counting the number of errors was simpler, as errors are easily identified by the fact that an alert appears informing the user that some field has been left blank. As ISO 9241-11 suggests, errors have an impact on the users' perception of usability in many ways by reducing efficiency, subjective satisfaction, and so on. Therefore, the existence of errors would have some degree of influence on the responses of the users to many usability questionnaire items. The nature and degree of this influence is very subjective, which means that it can only be assessed broadly.

The empirical measurements were obtained by processing the video recordings of the tests. To achieve this, the ANVIL annotation tool [101] was used. ANVIL[8] is a free video annotation program that allows to play back a video and define events at specific points. The events have properties such as duration, and several events can take place at the same time. ANVIL automatically converts all this information to Extensible Markup Language (XML).

The total time spent using a UI was divided into the time spent on screens 1, 2, and 3 individually. The time spent on things unrelated to the purpose of the tests, that is, extraneous time that might introduce bias into the results, was also measured with the intention of being subtracted afterwards. This includes:

1. **The time spent typing at the keyboard**. One might think that this would be similar for both UIs, but it was actually what might have introduced the most noise in the results, as it was typically a very long stretch of time that varied significantly from the first UI to the second and also between users

[8] http://www.anvil-software.org/

(e.g., some users were not familiar with the different modes of the iPod Touch keyboard, while others even typed with both hands).

2. **The time wasted on loading the pages**. This is not only irrelevant to the purpose of the tests but also depends on the functioning of the WiFi connection.

3. **A separate category for other things** unrelated to the purpose of the tests.

Events for user errors were also indicated so the total number of errors could be calculated later.

Whereas it is easy to determine the occurrence of an event (e.g., user errors were easily detected from the appearance of dialog boxes), it is not always trivial to determine where it begins and where it ends. For example, finishing the task of filling and submitting a form on the iPod Touch can be achieved in several different ways: by tapping the "Done" button of the iPod keyboard or by tapping the submit button of the web page; in the latter case, some users hid the iPod keyboard first whereas others did not and moved their finger directly towards the web page. All this was taken into account in the annotations but, nevertheless, it means barely tenths of seconds, which is negligible compared to, for example, the time invested by the users while figuring out what to do next.

All the annotations were done manually by the author of this doctoral thesis. The resulting XML files were given to the ICT staff, who converted this information into Comma Separated Values (CSV) so they could perform a statistical analysis on them.

6.8 Results of the User Tests

All 60 users finished the two assigned tasks successfully, although some of them experienced difficulties with those layouts in which some information was hidden from immediate view. As users were not allowed to ask for help or hints, finding a workaround was something they had to do on their own. Nevertheless, this ended up being a matter of seconds at most.

Being quantitative measures of user performance, the empirical results were subjected to a subsequent statistical analysis on SPSS[9], which was done entirely by the ICT staff. The subjective results, on the other hand, are presented in this chapter "as is", without performing any statistical analysis on them, as they are purely qualitative.

The statistical analysis is primarily concerned with the correlations between (1) task time and layout, and (2) error rate and layout. An independent variable is the UI, with three possible values: T-UI, H-UI, and V-UI. This variable is dichotomous,

[9]http://www.ibm.com/software/analytics/spss/

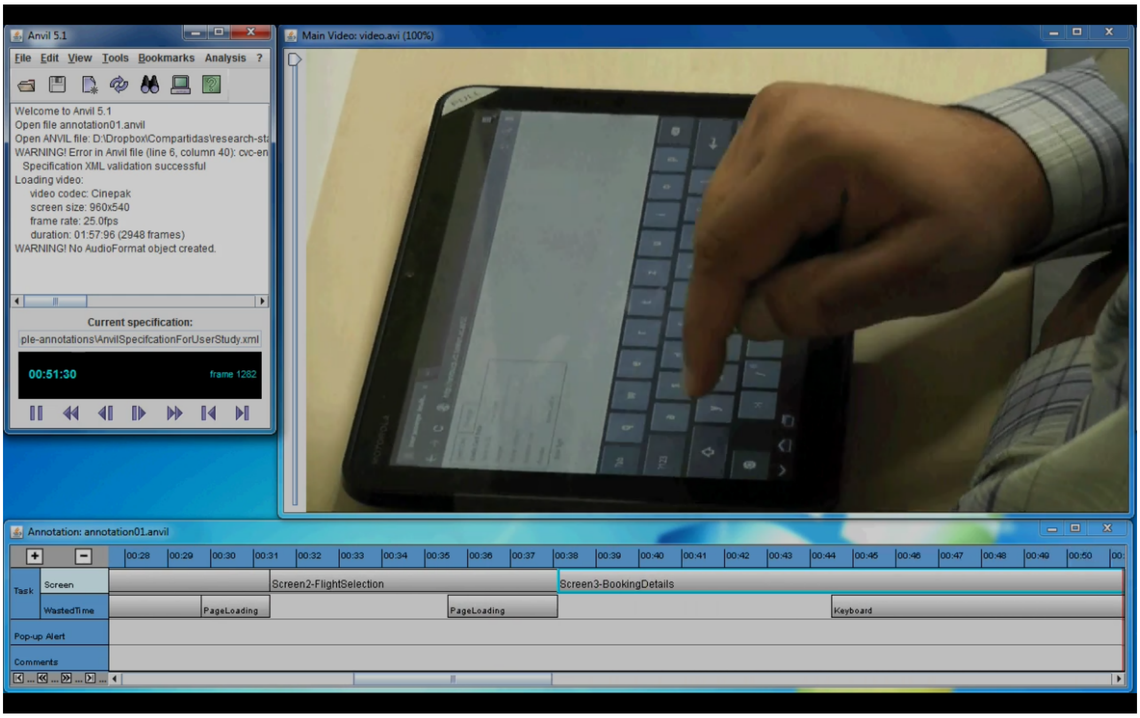


Figure 6.21: Defining interaction events on ANVIL

as only two UIs per run are compared. A second independent variable is the order in which the two layouts are tested. The dependent variables were *adjusted completion time* and *error rate*.

For the analysis of the adjusted completion time, the total time spent on each screen was measured and the time spent on extraneous interaction (e.g., typing text and loading the screen) was subtracted.

For the error rate analysis the number of errors each participant made on a given screen was measured.

As the type of UI is dichotomous and the calculated time is on an interval scale, the point-biserial Pearson correlation coefficients were calculated. A preliminary test showed that the variables are nearly normally distributed and thus satisfy the prerequisites of the point-biserial Pearson correlation. Note that the difference to the more common t-test is that the correlation covers coherence, while the t-test deals with differences of means.

The questions being investigated were formulated as null hypotheses, which are the correlation hypotheses listed in Table 6.17. A null hypothesis refers to the default position that there is no relationship between two things. The null hypothesis is then rejected or not according to the statistical significance relative to a given p-value, which is typically 0.05.

The full results of the tests are presented in the next few subsections. Each test run is discussed individually, presenting firstly the subjective results, and secondly the empirical results, which are followed by an analysis of the conclusions that can be

Table 6.17: Null Hypotheses

Test run	Name	Hypothesis
Run 1	$NH_{TH,T}$	There is no statistically significant correlation (*p*-value = 0.05) between the adjusted task time and the layout: T-UI and H-UI.
	$NH_{TH,E}$	There is no statistically significant correlation (*p*-value = 0.05) between the error rate and the layout: T-UI and H-UI.
Run 2	$NH_{VH,T}$	There is no statistically significant correlation (*p*-value = 0.05) between the adjusted task time and the layout: V-UI and H-UI.
	$NH_{VH,E}$	There is no statistically significant correlation (*p*-value = 0.05) between the error rate and the layout: V-UI and H-UI.
Run 3	$NH_{VT,T}$	There is no statistically significant correlation (*p*-value = 0.05) between the adjusted task time and the layout: V-UI and T-UI.
	$NH_{VT,E}$	There is no statistically significant correlation (*p*-value = 0.05) between the error rate and the layout: V-UI and T-UI.

obtained by putting all this information together. The description of the results of the tests concludes with a brief summary of all the previous results and the analysis of the users' responses to the preferences questionnaire.

6.8.1 Run 1: T-UI vs. H-UI

This test run compares the tabbed layout T-UI with the horizontal, non-tailored layout H-UI. As mentioned previously, the fact that H-UI is not tailored means that some information might be hidden from immediate view. The tests were performed on an iPod Touch and a Motorola Xoom tablet, with ten users each.

Table 6.18 shows the users' responses to the usability questionnaire for the iPod Touch. The participants clearly preferred T-UI on this device, based on the answers to most questions. In their overall assessment (i.e, the last item in the questionnaire) 90% of the participants preferred T-UI (moreover, all of them preferred it either "strongly" or "extremely"), whereas only 10% preferred H-UI ("moderately"). Between 80% and 90% of the participants opined that T-UI lends itself more to how they like to work, and that it is more clear and efficient to use, as well as more visually attractive. Regarding efficiency, the users did not perceive nearly as much of a difference in terms of manual effort, but they were clearly aware that operating a non-tailored and horizontally-scrolling UI was more time-consuming (which it actually was, as will be shown later).

The questionnaire results were diametrically opposed for the Motorola Xoom tablet, as can be seen in Table 6.19. For this device, 90% of the participants preferred

Table 6.18: T-UI vs. H-UI: Questionnaire Results for the iPod Touch

	T-UI preferred			equal	H-UI preferred		
	extreme	strong	moderate		moderate	strong	extreme
1. Which interface makes information more visible?	40%	50%	10%	0%	0%	0%	0%
2. Which interface makes interaction more intuitive?	20%	40%	10%	20%	10%	0%	0%
3. Which interface makes it easier to figure out what to do next?	30%	30%	30%	10%	0%	0%	0%
4. Which interface makes it clearer how to use it?	20%	40%	20%	20%	0%	0%	0%
5. Which interface lends itself more to how you like to work?	20%	20%	40%	10%	10%	0%	0%
6. Which interface requires less manual interaction?	0%	0%	40%	50%	10%	0%	0%
7. Which interface demands less precision on your part?	0%	0%	20%	60%	10%	10%	0%
8. Which interface demands less time from you?	0%	20%	60%	20%	0%	0%	0%
9. Which interface makes interaction more efficient?	0%	50%	40%	10%	0%	0%	0%
10. Which interface is more visually attractive?	10%	60%	20%	10%	0%	0%	0%
11. Overall, which interface would you use to book a flight?	40%	50%	0%	0%	10%	0%	0%

H-UI in their overall assessment. In fact, participants preferred (almost always overwhelmingly) H-UI over T-UI in all categories except visual attractiveness (for which 50% of them preferred T-UI and 40% H-UI). Interestingly, even though using tab panels leaves a considerable amount of white space on the tablet screen, some users remarked in interviews that they liked the “compact” look of the tabbed UI. This was perhaps the most noteworthy thing said by the users in the interviews for run 1, as the rest of their opinions left little room for doubt.

The results of the questionnaires are consistent with the empirical measurements obtained from the video recordings, not only in terms of performance but regarding overall preference as well. Table 6.20 shows the Pearson correlations between adjusted task time and layout for screens 1 to 3 on the iPod Touch and the corresponding statistical significance, as well as the average adjusted task time (i.e., after subtracting page loading times, and so on) for both layouts. Table 6.21 does the same for the Motorola Xoom tablet.

The correlations highlighted in bold are those with a significance lesser or equal than 0.05, which means that the results are statistically significant and the corresponding null hypothesis can be rejected. Positive correlations indicate that T-UI performed better, whereas negative correlations mean the opposite.

For the iPod Touch (Table 6.20), one can see that the participants performed better with T-UI, as the correlations for screens 1 and 3 are positive and statistically significant. Screen 2 shows almost no correlation since all information fitted on the screen and was immediately visible on both layouts. In fact, Screen 2 was identical for both UIs, and thus layout-independent, which means that a low correlation is to be expected.

Therefore, the null hypothesis $NH_{TH,T}$ can be rejected for the experiment on the iPod Touch when using screens 1 and 3, as the adjusted task time using T-UI is significantly smaller than using H-UI for those two screens. Moreover, the sum of the average adjusted task times for all screens is 31.28 seconds for T-UI and 52.88 seconds for H-UI. In other words, it is 69% longer for H-UI.

For the Motorola Xoom tablet (Table 6.21), the empirical information shows that the null hypothesis $NH_{TH,T}$ can be also rejected for that device on screens 1 and 3. The adjusted task time using T-UI is significantly higher than using H-UI for Screen 1 and 3 and the correlations were negative, which means that the participants performed better on H-UI on this device. The most plausible explanation is that all information fitted into the display on H-UI, so no scrolling was required, whereas T-UI needed additional clicks to access the information. These additional clicks should not be underestimated, as the sum of average times for all screens is 28.82 seconds for T-UI and 18.37 seconds for H-UI. That is, it took 57% longer to operate T-UI.

The other empirical measurement is error rate. Table 6.22 shows the correlations between number of errors and layout for screens 1 to 3 on the iPod Touch, whereas Table 6.23 does the same for the Motorola Xoom tablet.

As can be seen in Table 6.22, T-UI leads to significantly fewer errors on screen

Table 6.19: T-UI vs. H-UI: Questionnaire Results for the Motorola Xoom

	T-UI preferred			equal	H-UI preferred		
	extreme	strong	moderate		moderate	strong	extreme
1. Which interface makes information more visible?	0%	0%	10%	0%	30%	30%	30%
2. Which interface makes interaction more intuitive?	0%	0%	10%	10%	50%	20%	10%
3. Which interface makes it easier to figure out what to do next?	0%	10%	0%	10%	30%	30%	20%
4. Which interface makes it clearer how to use it?	0%	0%	10%	50%	40%	0%	0%
5. Which interface lends itself more to how you like to work?	0%	10%	0%	10%	30%	40%	10%
6. Which interface requires less manual interaction?	0%	0%	10%	10%	30%	20%	30%
7. Which interface demands less precision on your part?	0%	10%	0%	70%	0%	20%	0%
8. Which interface demands less time from you?	0%	0%	10%	30%	30%	20%	10%
9. Which interface makes interaction more efficient?	0%	0%	10%	20%	30%	40%	0%
10. Which interface is more visually attractive?	10%	20%	20%	10%	10%	30%	0%
11. Overall, which interface would you use to book a flight?	0%	0%	10%	0%	30%	40%	20%

Table 6.20: T-UI vs. H-UI: Correlations between Adjusted Task Time and Layout on an iPod Touch

Screen	Pearson corr.	Sig. (1-tailed)	T-UI avg. time	H-UI avg. time
Screen 1	**0.52**	0.01	14.23 s	29.57 s
Screen 2	0.12	0.31	8.96 s	9.98 s
Screen 3	**0.37**	0.05	8.09 s	13.34 s

Table 6.21: T-UI vs. H-UI: Correlations between Adjusted Task Time and Layout on a Motorola Xoom

Screen	Pearson corr.	Sig. (1-tailed)	T-UI avg. time	H-UI avg. time
Screen 1	**-0.56**	0.01	13.37 s	7.43 s
Screen 2	-0.17	0.23	7.45 s	6.38 s
Screen 3	**-0.66**	0.00	8.00 s	4.57 s

1, as the correlation is both positive and significant. In fact, users made on average almost two mistakes on screen 1, which is a high number that can be again attributed to the fact that some information was hidden from view. Users made no mistakes on screen 2, so no correlation can be calculated for this screen. On screen 3 the correlation is not significant, so no conclusions can be drawn. As a matter of fact, users made no mistakes on screen 3. Therefore, the null hypothesis $NH_{TH,E}$ can be rejected only for screen 1 on the iPod Touch.

For the Motorola Xoom tablet, on the other hand, none of the correlations in Table 6.23 are significant, which means that the null hypothesis $NH_{TH,E}$ cannot be rejected for this device. In fact, users made no mistakes on H-UI and only an average of 0.1 mistakes on T-UI for screen 1. In other words, only one out of ten users made a mistake on screen 1 of T-UI. The exact same results were obtained for screen 3 of T-UI. All the mistakes on this tabbed layout can be attributed to forgetting to switch tabs.

6.8.2 Run 2: V-UI vs. H-UI

This test run compares the vertical layout V-UI with the horizontal layout H-UI. Ten users participated in this run and all of them operated an iPod Touch. Given the small size of the iPod Touch, both UIs required (one-dimensional) scrolling on screens 1 and 3.

Table 6.24 shows the questionnaire results for this run. The participants clearly preferred V-UI over H-UI for most questions. In fact, 90% of the users preferred V-UI in their overall assessment (i.e., the last questionnaire item). As for the individual usability criteria, users preferred V-UI in all categories except for demands on precision and visual attractiveness.

In the subsequent interview, seven participants emphasized that they disliked H-UI because they found it less intuitive. They expressed this opinion in statements

Table 6.22: T-UI vs. H-UI: Correlations between Number of Errors and Layout on an iPod Touch

Screen	Pearson corr.	Sig. (1-tailed)	T-UI avg. errors	H-UI avg. errors
Screen 1	**0.65**	0.01	0.2	1.7
Screen 2	-	-	0	0
Screen 3	-0.33	0.08	0.2	0

Table 6.23: T-UI vs. H-UI: Correlations between Number of Errors and Layout on a Motorola Xoom

Screen	Pearson corr.	Sig. (1-tailed)	T-UI avg. errors	H-UI avg. errors
Screen 1	-0.229	0.17	0.1	0
Screen 2	-	-	0	0
Screen 3	-0.229	0.17	0.1	0

such as: "If I had not been informed that this was a test maybe I'd have never tried to scroll sideways", "very irritating", or "I really don't like it".

In addition, the interviews revealed further interesting issues that were not necessarily commented on by the majority of the participants, but are nevertheless worth mentioning:

1. Three participants explicitly stated that on mobile devices vertical scrolling feels "more natural" than horizontal scrolling. They said things like: "It's more intuitive to scroll downward", "for mobile devices it's natural to slide up and down, not sideways", or "vertical scrolling is more like a [desktop] computer".

2. Interestingly, three other participants stated that horizontal scrolling would have been acceptable if they had initially known that the screen continued to the right (e.g., "if you get used to [H-UI] it's easier to use"). One participant additionally stated that in this case he would even prefer horizontal scrolling, and another one opined that he would prefer horizontal browsing between different screens (i.e., sideways swiping to switch pages) rather than continuous scrolling. The latter participant stated that this preference is related to his previous use of an Android smartphone, where horizontal browsing is, for example, used to switch between different blog entries, while vertical browsing would be used to navigate within a certain blog entry.

3. Two participants remarked that seeing a submit button at the bottom left corner of screen 1 does not suggest that the screen continues horizontally. In fact, three participants suggested, unprompted, to place the submit button in the lower right corner instead. A verbatim user statement in this context was: "This button on the bottom makes me feel that I've finished."

4. One participant stated: "In general, I like to see the whole screen". This statement is interesting as it arguably depends on the display size. The screens in this application never exceed the display size more than twice (i.e., they are

Table 6.24: V-UI vs. H-UI: Questionnaire Results

	V-UI preferred			equal	H-UI preferred		
	extreme	strong	moderate		moderate	strong	extreme
1. Which interface makes information more visible?	30%	20%	30%	10%	0%	10%	0%
2. Which interface makes interaction more intuitive?	30%	40%	0%	20%	10%	0%	0%
3. Which interface makes it easier to figure out what to do next?	50%	50%	0%	0%	0%	0%	0%
4. Which interface makes it clearer how to use it?	20%	40%	10%	30%	0%	0%	0%
5. Which interface lends itself more to how you like to work?	40%	20%	30%	10%	0%	0%	0%
6. Which interface requires less manual interaction?	10%	0%	70%	20%	0%	0%	0%
7. Which interface demands less precision on your part?	10%	10%	10%	70%	0%	0%	0%
8. Which interface demands less time from you?	10%	20%	60%	10%	0%	0%	0%
9. Which interface makes interaction more efficient?	10%	20%	40%	30%	0%	0%	0%
10. Which interface is more visually attractive?	0%	10%	0%	70%	20%	0%	0%
11. Overall, which interface would you use to book a flight?	40%	10%	40%	0%	0%	10%	0%

at most twice as large as the display size of the iPod Touch). This means that the widgets would probably still be distinguishable with the screen zoomed out to the maximum, but this is not necessarily true for larger UIs.

The subjective preferences of the users are also borne out by the empirical measurements. Table 6.25 shows the correlations between the adjusted task time and the layout, whereas Table 6.26 does the same for the correlations between the number of errors and the layout.

Table 6.25: V-UI vs. H-UI: Correlations between Adjusted Task Time and Layout

Screen	Pearson corr.	Sig. (1-tailed)	V-UI avg. time	H-UI avg. time
Screen 1	**0.74**	0.000	10.20 s	36.78 s
Screen 2	-0.05	0.417	8.03 s	7.48 s
Screen 3	**0.59**	0.003	5.15 s	9.97 s

Table 6.26: V-UI vs. H-UI: Correlations between Number of Errors and Layout

Screen	Pearson corr.	Sig. (1-tailed)	V-UI avg. errors	H-UI avg. errors
Screen 1	**0.83**	0.000	0	1.8
Screen 2	-	-	0	0
Screen 3	-	-	0	0

The positive correlations in Table 6.25 show that V-UI takes less time to operate than H-UI, with highly significant results. The null hypothesis $NH_{VH,T}$ can then be rejected for this experiment, that is, the adjusted task time using V-UI is significantly smaller than using H-UI for screens 1 and 3, which are the only screens that differ depending on the layout. In fact, the sum of times for all screens is on average 23.38 seconds for V-UI and 54.23 seconds for H-UI (which is more than twice the time spent for V-UI).

As for the correlations between the number of errors and the layout (see Table 6.26), the error rate is significantly smaller for V-UI only on screen 1. Users made no errors on the other screens, so no correlations can be calculated for this experiment. Therefore, the null hypothesis $NH_{VU,E}$ can be rejected for screen 1.

6.8.3 Run 3: V-UI vs. T-UI

This test run compares the vertical layout V-UI with the tabbed layout T-UI. 30 users participated and all of them operated an iPod Touch. This run differs greatly from the previous two runs, which compared a layout tailored for small displays against what can be considered a non-tailored layout (i.e., H-UI). In this run, both UIs under study can be considered tailored for small displays, so the experiment becomes a question of investigating which optimization strategy is preferred by the users.

Table 6.27 presents the results of the subjective questionnaire. In their overall assessment (i.e., the last questionnaire item), 60% of the participants (i.e., the majority) preferred V-UI, but 30% preferred T-UI. In general, the answers to the other questionnaire items also varied much, and the users' preferences were not as extreme as they had been on the previous test runs. There was only one criterion for which T-UI was widely preferred, namely, visual attractiveness. As for V-UI, the questionnaire shows that it was considered much more intuitive to navigate and to interact with, and slightly less demanding and more efficient to use in general. However, no consensus can be determined for two things: which UI makes information more visible and which one lends itself more to how the participants like to work. The responses to these two questions are shown in Figures 6.22 and 6.23, including also the order in which the UIs were tested. On the former question, opinions were sharply divided, with little middle ground (only 6.7% of the users voted "Equal"). On the latter, the users' opinions were so dissimilar that no conclusion can be reached. It is worth mentioning that, whenever opinion was divided, analyzing the data showed no clear relationship between the answers and the background of the participants.

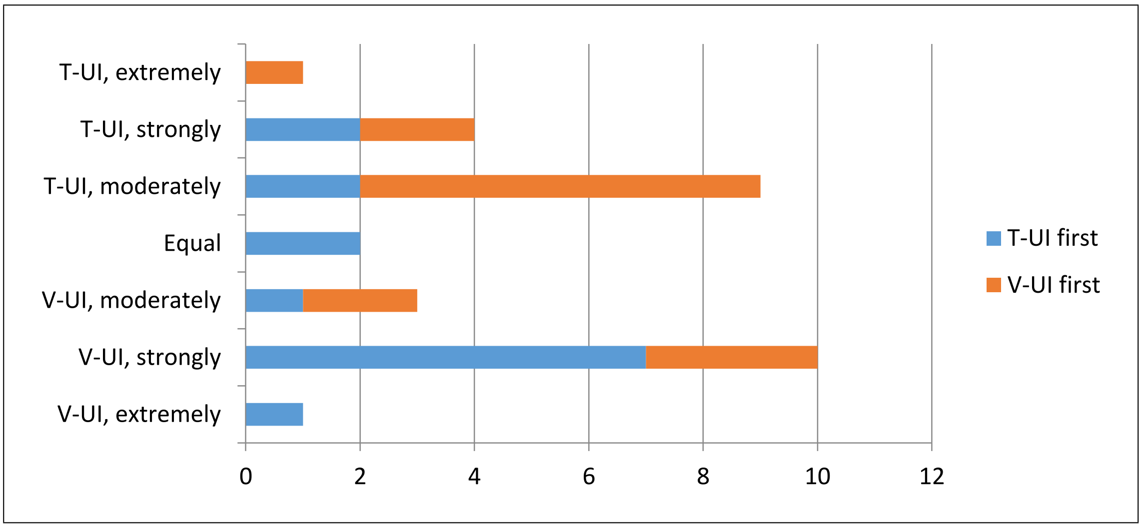


Figure 6.22: Run 3, question 1: Which interface makes information more visible?

Table 6.28 summarizes the empirical results for the adjusted task time analysis. The positive correlations show that V-UI performs better than T-UI, and the results are highly significant. The null hypothesis $NH_{VT,T}$ can then be rejected for this experiment. That is, the adjusted task time using V-UI is significantly smaller than using T-UI on screens 1 and 3. In fact, it took 54% longer to operate T-UI, as the sum of adjusted times for all screens is on average 25.81 seconds for V-UI and 36.79 seconds for T-UI. These empirical measurements are consistent with the users' subjective perceptions expressed in questionnaire item number 8 ("Which interface demands less time from you?"), which are shown in Table 6.27.

Table 6.29 shows the correlations between the error rate and the layout, together with the average number of errors for each screen. On screens 1 and 3, the error rate is significantly smaller. On Screen 2, however, there were no errors, so no correlation could be calculated. Therefore, the null hypothesis $NH_{VT,E}$ can also be rejected for

Table 6.27: V-UI vs. T-UI: Questionnaire Results

	V-UI preferred			equal	T-UI preferred		
	extreme	strong	moderate		moderate	strong	extreme
1. Which interface makes information more visible?	3.33%	33.33%	10%	6.67%	30%	13.33%	3.33%
2. Which interface makes interaction more intuitive?	13.33%	26.67%	26.67%	30%	3.33%	0%	0%
3. Which interface makes it easier to figure out what to do next?	20%	36.67%	20%	10%	10%	3.33%	0%
4. Which interface makes it clearer how to use it?	0%	13.33%	30%	53.33%	3.3%	0%	0%
5. Which interface lends itself more to how you like to work?	13.33%	23.33%	13.33%	23.33%	10%	13.33%	3.33%
6. Which interface requires less manual interaction?	6.67%	20%	26.67%	26.67%	13.33%	6.67%	0%
7. Which interface demands less precision on your part?	3.33%	13.33%	16.67%	56.67%	10%	0%	0%
8. Which interface demands less time from you?	13.33%	10%	26.67%	33.33%	10%	6.67%	0%
9. Which interface makes interaction more efficient?	10%	6.67%	26.67%	33.33%	16.67%	6.67%	0%
10. Which interface is more visually attractive?	0%	0%	6.67%	10%	40%	23.33%	20%
11. Overall, which interface would you use to book a flight?	16.67%	20%	23.33%	10%	10%	13.33%	6.67%

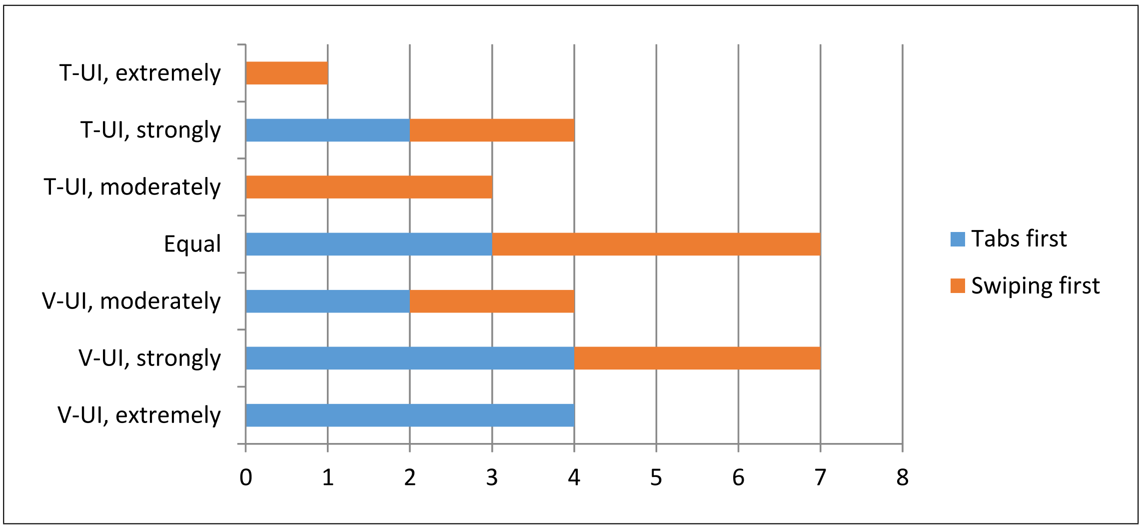


Figure 6.23: Run 3, question 5: Which interface lends itself more to how you like to work?

Table 6.28: V-UI vs. T-UI: Correlations between Adjusted Task Time and Layout

Screen	**Pearson corr.**	**Sig. (1-tailed)**	**V-UI avg. time**	**T-UI avg. time**
Screen 1	**0.35**	0.003	12.79 s	17.98 s
Screen 2	0.12	0.189	8.17 s	9.50 s
Screen 3	**0.43**	0.000	4.85 s	9.31 s

this experiment. That is, the error rate for V-UI is significantly smaller than that of T-UI on screens 1 and 3. Actually, there were no errors for V-UI, and all errors can be attributed to the participants tapping the submit button before changing tabs, so this could be another argument against using tabs. Moreover, some users stated in the interviews that it is easy to forget to switch tabs before tapping the submit button, even when they were clearly aware of the fact that there were two tabs.

Table 6.29: V-UI vs. T-UI: Correlations between Number of Errors and Layout

Screen	**Pearson corr.**	**Sig. (1-tailed)**	**V-UI avg. errors**	**T-UI avg. errors**
Screen 1	**0.363**	0.002	0	0.23
Screen 2	-	-	0	0
Screen 3	**0.285**	0.014	0	0.2

6.8.4 Summary of the Subjective and the Empirical Results

Table 6.30 shows a summary of the results of the iPod Touch tests, organized by UI. For each of the two runs in which a given UI was tested, both the subjective and the empirical results are summarized. Regarding the subjective results, Table 6.30 indicates if the UI in question was the one preferred by the users (based on the last questionnaire item). If so, the percentage of users who voted for it are indicated in parentheses. As for the empirical results, the average adjusted time and the average number of errors are indicated. These figures are the sum of all the individual results for each screen, that is, these are the results from the beginning (screen 1) to the end (finish screen) of each user task. As explained previously, adjusted time means that extraneous times such as page loading and typing time have been subtracted. Taking into account that run 3 had thrice as many users as the other runs on the iPod Touch, the average times for each UI are calculated as follows: 35.41 seconds for T-UI, 53.55 seconds for H-UI, and 25.2 seconds for V-UI.

Table 6.30: Summary of the Results for the iPod Touch UIs

UI	Run	Preferred	Avg. Adjusted Time	Avg. No. of Errors
T-UI	1	Yes (90%)	31.28 s	0.4
	3	No	36.79 s	0.43
H-UI	1	No	52.88 s	1.7
	2	No	54.23 s	1.8
V-UI	2	Yes (90%)	23.38 s	0
	3	Yes (60%)	25.81 s	0

Analogously, Table 6.31 summarizes the results for the Motorola Xoom tablet. It should be kept in mind that the Motorola Xoom and the iPod Touch are so different – as is the way the UIs are rendered on them – that it would make no sense to merge the results for both devices.

Table 6.31: Summary of the Results for the Motorola Xoom UIs

UI	Run	Preferred	Avg. Adjusted Time	Avg. No. of Errors
T-UI	1	No	28.82 s	0.2
H-UI	1	Yes (90%)	18.37 s	0

The results for the null hypotheses on the iPod Touch are summarized in Table 6.32. The specific screens for which a given null hypothesis ended up being rejected are indicated, as well as the conclusions than can be obtained from the rejection of said hypothesis. Table 6.33 does the same for the Motorola Xoom tablet.

All the information above, both subjective (i.e., questionnaires) and empirical (i.e., measurements of user performance) strongly suggests that for smartphone-sized devices V-UI is the most usable layout. The second most usable layout would be T-UI, and H-UI ranks a distant third.

For tablets, the results are quite different, as the subjective preferences of the users are overwhelmingly in favor of H-UI and against T-UI, and the empirical time measurements are also significantly smaller for H-UI.

6.8.5 User Preferences Regarding the Usability Criteria

This section summarizes the responses of the users to the additional questionnaire that asked about their preferences regarding the different usability criteria. The

Table 6.32: Summary of the Results for the Null Hypotheses on the iPod Touch

Hypothesis	**Rejected for**	**Conclusions**
$NH_{TH,T}$	Screens 1 and 3	Time on T-UI is significantly shorter than on H-UI
$NH_{TH,E}$	Screen 1	Errors on T-UI are significantly fewer than on H-UI (on screen 1)
$NH_{VH,T}$	Screens 1 and 3	Time on V-UI is significantly shorter than on H-UI
$NH_{VH,E}$	Screen 1	Errors on V-UI are significantly fewer than on H-UI (on screen 1)
$NH_{VT,T}$	Screens 1 and 3	Time on V-UI is significantly shorter than on T-UI
$NH_{VT,E}$	Screens 1 and 3	Errors on V-UI are significantly fewer than on T-UI

Table 6.33: Summary of the Results for the Null Hypotheses on the Motorola Xoom

Hypothesis	**Rejected for**	**Conclusions**
$NH_{TH,T}$	Screens 1 and 3	Time on H-UI is significantly shorter than on T-UI
$NH_{TH,E}$	None	No significant conclusions can be obtained by comparing the number of errors on T-UI and H-UI

questionnaire items consist of six pairwise comparisons of the following usability criteria:

1. **Makes it clear how to use it**.
2. **Lends itself to how you like to work**.
3. **Makes interaction efficient**.
4. **Is visually attractive**.

The statistics below are for all the 60 users that participated in the tests. For the sake of brevity, no distinction is made between individual test runs. Once more, this qualitative information is presented "as is", without performing statistical calculations on it.

Figure 6.24 shows the responses to the first questionnaire item, that is, "Makes it clear how to use it vs. Lends itself to how I like to work". 58.3% of the users preferred the former, whereas 31.7% chose the latter. Only 10% of the users answered "Equal".

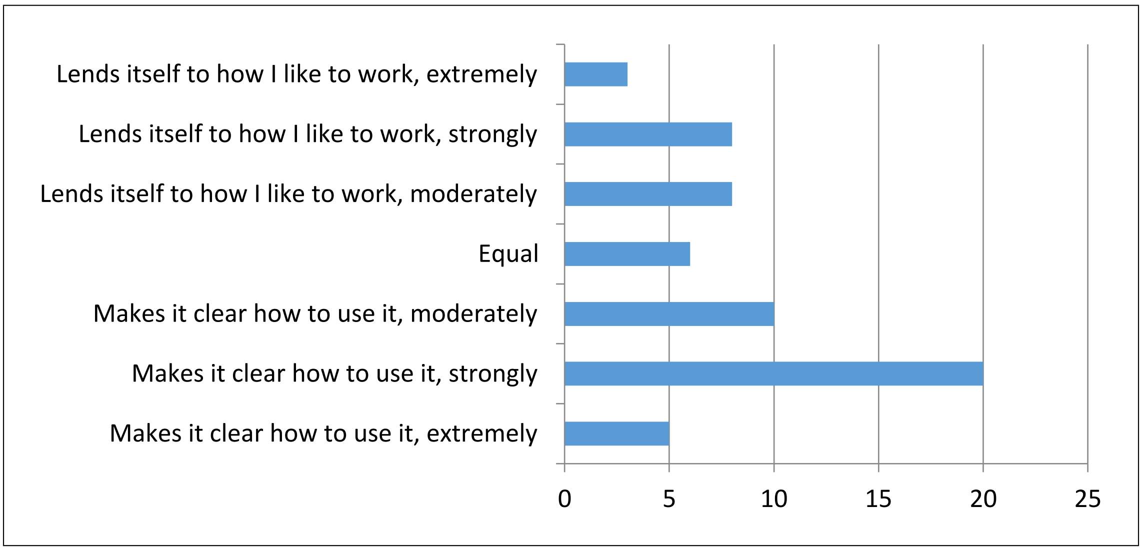


Figure 6.24: User preferences: Makes it clear how to use it vs. Lends itself to how I like to work

Figure 6.25 shows the responses to the second questionnaire item, that is, "Makes it clear how to use it vs. Makes interaction efficient". 35% of the users preferred the former, whereas 41.7% chose the latter. The rest, that is, 23.3% of the users, answered "Equal".

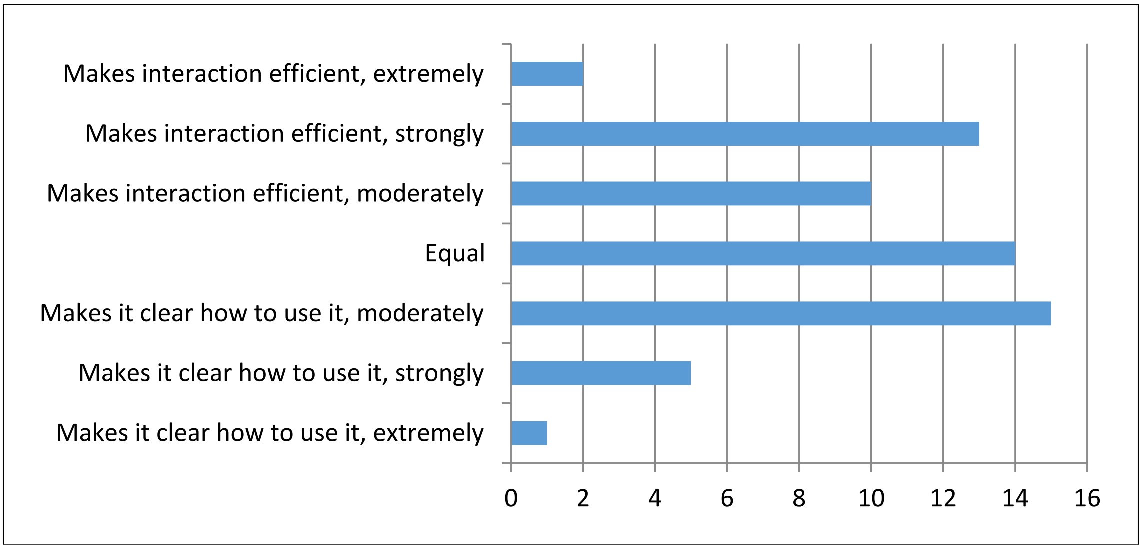


Figure 6.25: User preferences: Makes it clear how to use it vs. Makes interaction efficient

Figure 6.26 shows the responses to the third questionnaire item, that is, "Makes it clear how to use it vs. Is visually attractive". A high number of users, 61.7% of them, preferred the former, whereas only 15% chose the latter. The percentage of users who answered "Equal" is 23.3%.

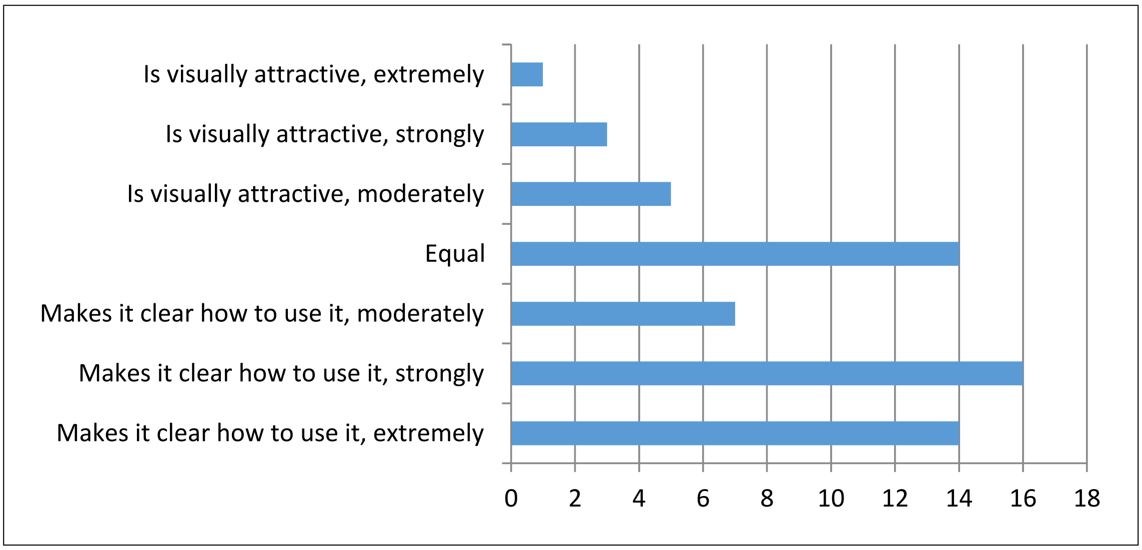


Figure 6.26: User preferences: Makes it clear how to use it vs. Is visually attractive

Figure 6.27 shows the responses to the fourth questionnaire item, that is, "Lends itself to how I like to work vs. Makes interaction efficient". 30% of the users preferred the former, whereas 46.7% chose the latter. At 23.3%, the percentage of users who answered "Equal" is not far from the percentage of users for the other two options.

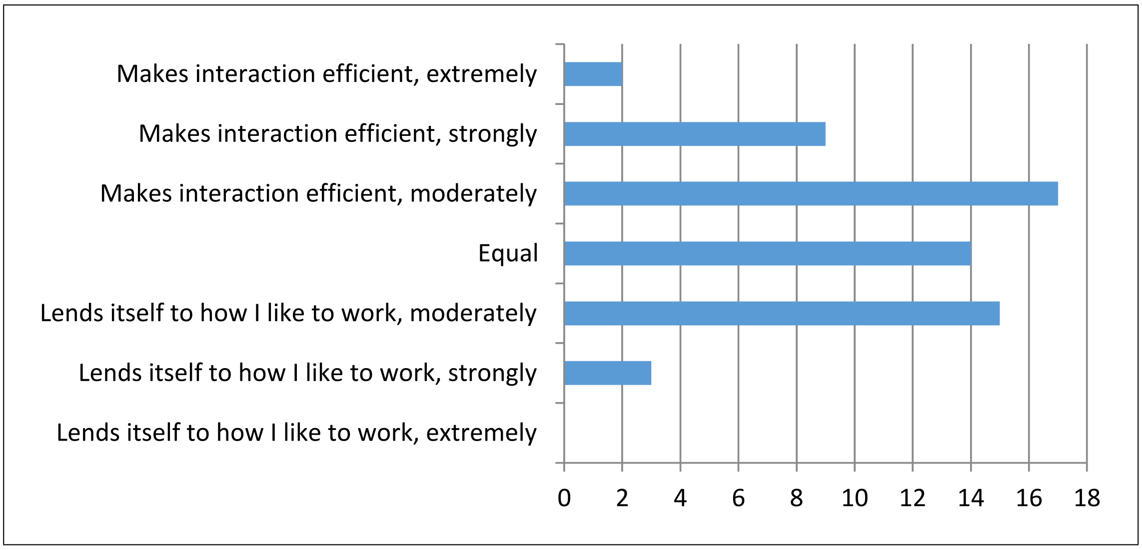


Figure 6.27: User preferences: Lends itself to how I like to work vs. Makes interaction efficient

Figure 6.28 shows the responses to the fifth questionnaire item, that is, "Lends itself to how I like to work vs. Is visually attractive". 56,7% of the users preferred the former, whereas only 13.3% chose the latter. At 30%, the percentage of users who answered "Equal" is again relatively high, although it is barely over half of the percentage for the most voted option.

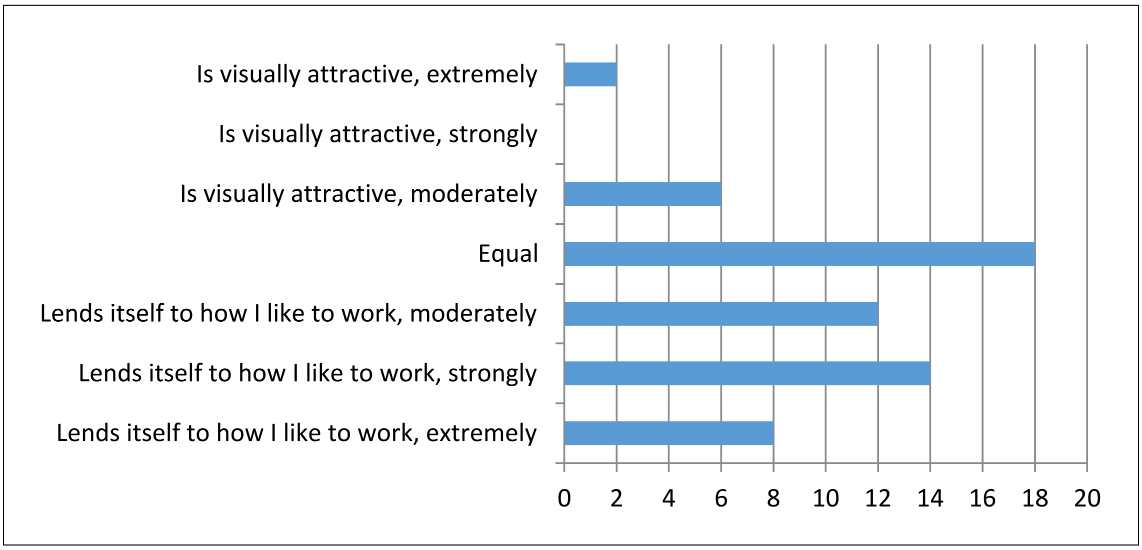


Figure 6.28: User preferences: Lends itself to how I like to work vs. Is visually attractive

Figure 6.29 shows the responses to the sixth questionnaire item, that is, "Makes interaction efficient vs. Is visually attractive". 78.3% of the users, that is, the vast majority, preferred the former, whereas only 10% chose the latter. The percentage of users who answered "Equal" is also low, at 11.7%.

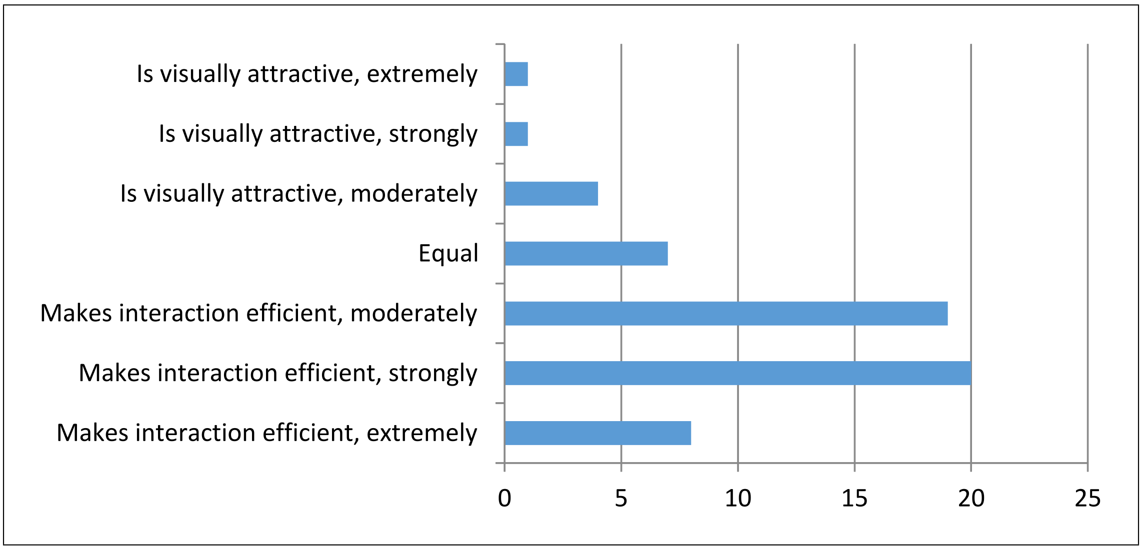


Figure 6.29: User preferences: Makes interaction efficient vs. Is visually attractive

The most noticeable thing about these results is that the users seem to give little importance to visual attractiveness. As for the importance given to the other three usability criteria, it is difficult to infer any kind of bias, as all criteria rate highly at one point or another. These preferences could be considered consistent with the responses to the usability questionnaire, as the opinions of the users regarding visual attractiveness were sometimes at odds with their overall assessment of the usability of the UIs (particularly in runs 2 and 3), whereas the other usability criteria tended to be more in agreement with their overall assessment. The fact that aesthetic judgments seem to have carried little weight in the final decisions of the users could arguably have been influenced by their background, as most of them were students of Electrical Engineering. That is to say, aesthetic opinions are the least generalizable part of these results.

During the tests, several users remarked on the difficulty of choosing between these criteria, as all of them can be considered desirable. Users were suggested to tick “Equal” in case of doubt, but, as can be seen from the results above, not many did. That is to say, regardless of the difficulty of the questionnaire, users do have definite preferences individually. Taking their opinions as a whole, however, no conclusions can be inferred with certainty beyond the aforementioned lack of interest in aesthetics.

Chapter 7

Validity of the Approach

7.1 Introduction

This chapter aims to explore and discuss the validity of the overall approach taken in this thesis. Although the question of validity can be tackled in different ways, it will be mainly investigated here in comparative terms. That is, the approach taken in this thesis will be compared with the opposite approach – more specifically, with a non-generic, context-dependent usability model constructed in an ad hoc fashion for a specific field of application. Note that the fact that the taxonomies presented in this thesis are applicable to any kind of product means that it is not necessarily guaranteed that a context-specific model will exist for any given field of application. However, a context-specific usability model happened to exist for the field of in-vehicle information systems, which was also the field of application for the usability study of the CARAT counter (see Chapter 5). This allowed to map the results of the aforementioned usability study to the context-specific usability model and compare the mappings to the results previously shown in Section 5.10. This comparison, in turn, lays the groundwork for a more formal discussion of expressiveness in general – applied to the area of usability models – and the problems caused by lack of expressiveness in this regard.

7.2 Background

The usability literature includes few attempts at investigating the validity of the existing usability models. Two examples would be the respective empirical studies by Hornbæk and Law [80] and Sauro and Lewis [164], which aimed to investigate the construct validity of the traditional ISO 9241-11 usability model [87]. The main goal of both studies was to analyze the correlations between, on the one hand, the traditional attributes of *efficiency*, *effectiveness*, and *satisfaction*, and, on the other hand, common usability metrics of real-world usability studies like *completion time* and *error rates*. The Hornbæk and Law study uses raw data from 73 usability studies, whereas the Sauro and Lewis study uses 97 raw data-sets from 90 distinct usability tests. Interestingly, they obtain diametrically opposed results, as will be discussed in Chapter 8. More empirical studies definitely need to be conducted in the usability literature, although their scarcity is understandable, as they can be very resource-demanding.

So far, this thesis has illustrated the usefulness of the taxonomies by example, by showing how they have been used in real-world usability studies to identify, model, and classify usability issues. These usability studies could be considered a way of subjecting the taxonomies to consecutive real-world tests of their usefulness. Successes in these kinds of tests do not necessarily prove the validity of the taxonomies, but failures would point to inherent limitations in them. The conditions for failure in a usability model would be that it does not represent what it aims to represent, that the concepts are difficult to understand, or that the results it yields seem arbitrary or not relevant (for example, when the usability model seems disconnected from concrete, real-world usability information). Or when the users respond to the usability questionnaires in a seemingly random manner; or they never show any particular preference; or when they leave the questions blank. Based on Sears' criteria of *validity*, *thoroughness*, and *reliability* [166] for comparing different evaluation techniques, we could conclude that usability studies should be based on usability models that focus on pertinent issues, cover all usability aspects, and obtain similar results under similar conditions.

Winter et al. [186] investigated the validity of their comprehensive usability model (see Section 1.1) not empirically but comparing it against the ISO 15005:2002 standard [86]. Their goals are, firstly, to validate their usability model by showing that it can be used to model the principles contained in the standard, and, secondly, to use their model to uncover deficiencies in the ISO standard. In other words, their model and the ISO 15005 standard are evaluated against each other. As a result, they identify problems like incompleteness, lack of explicitness, and inconsistency in the requirements and recommendations proposed by the standard. Unfortunately, their analysis is limited to a few examples, and it would have been interesting to see a full analysis, given the high specificity of the requirements and recommendations in ISO 15005.

Note that the validity of a usability model is always tested with reference to something. For Hornbæk and Law and Sauro and Lewis, it is usability metrics; for Winter et al., it is the requirements and recommendations in the ISO 15005 standard.

7.3 A Context-Specific Usability Model for In-Vehicle Information Systems

The context-specific usability model against which our approach will be compared was developed by Harvey et al. [72][73]. Their work is aimed specifically at in-vehicle information systems (IVISs) and it is representative of traditional usability studies in that it is based on an tailor-made definition of usability that was obtained by conducting an extensive research of both the literature of the specific field of application (e.g., guidelines for IVISs) and the general literature on usability (e.g., history, definitions). More specifically, their work is based on the usability models from the following authors and standards: Shackel [170][172], Stanton and Baber [176], ISO-9241-11 [87], Nielsen [129], Shneiderman [174], Norman [137], and Jordan [96].

The usability model by Harvey et al. consists of 13 usability criteria that "collectively define usability for IVISs and were developed to meet the overall needs of drivers: safety, efficiency, and enjoyment" [73, p. 514]. The criteria are listed in Table 7.1.

Table 7.1: Harvey et al.'s Usability Criteria for In-Vehicle Information Systems

Usability criterion
Effectiveness of interaction with device under varying driving conditions
Sustained effectiveness of interaction with device whilst driving
Sustained efficiency of interaction with device whilst driving
Interference between interaction with device and primary task performance
Learnability of device interactions
Efficiency on first use of the device whilst driving
Effectiveness on first use of the device whilst driving
Adaptability of device to individual users
Compatibility of device with full range of users
Satisfaction on first use of device whilst driving
Perceived usefulness of device in real driving situation
Satisfaction based on short- and long-term use of device whilst driving
Memorability of device interaction

Their approach is based on the following premise:

> "Consideration of the context of use makes a general definition of usability virtually impossible because different situations will demand different attributes from a product to optimise the interaction with particular users. Despite the desire to construct a universal definition, it appears that most people now accept that the context in which a product or system is used must be taken into account, and definitions therefore need to be constructed individually according to the product, tasks, users and environment in question." [72, p. 324]

They begin to describe their particular methodology as follows:

> "Defining the context of use for a particular product enables designers/evaluators to specify the usability factors which are important. This was done for in-vehicle devices to identify six contextual factors." [72, p. 332]

Table 7.2: Harvey et al.'s Contextual Factors for In-Vehicle Information Systems

Contextual Factor
Varying environmental conditions
Dual task environment
Training provision
Range of users
Uptake
Varying frequency of use

Their contextual factors are listed in Table 7.2.

To this they add another step:

> "Next, criteria from the general definitions of usability [...] were used as guidance to examine each context factor in more detail. This involved investigating how these general usability criteria, such as effectiveness, efficiency and satisfaction [...] applied in an in-vehicle device context. For example, efficiency of the device is important in the dual task environment and in terms of training provision. When viewed within the specific context of the dual task environment efficiency must be consistently good to ensure that interference between secondary and primary tasks is always low. In the context of training, however, the focus should be on the initial efficiency of the device, i.e. on first use. This is because initial efficiency will indicate how learnable a device is. Twelve general usability criteria were examined in this way in relation to one or more of the contextual factors. They were then translated into 13 in-vehicle device-specific criteria which cover all aspects of usability in this context of use." [72, pp. 332-333]

Table 7.3 (adapted from [72]) shows the correspondences that Harvey et al. themselves established between their IVIS usability criteria, their contextual factors, and the general criteria from the usability literature.

As Table 7.3 shows, the classification by Harvey et al. differs from the usability taxonomy by Alonso-Ríos et al. in that the former is not an actual top-down decomposition of the general usability criteria from the literature into nonredundant lower-level subattributes. Instead, most of their criteria are simultaneously related to several criteria from the literature, and vice versa. More problematically, there is also a great deal of overlap in these correspondences. All this makes their approach not very amenable to analysis, in the sense of separating a complex object into smaller parts in order to study it closely. It should also be noted that no definitions are provided for their criteria. Their contextual factors are described in one or two paragraphs each, but no definitions are provided, either.

Table 7.3: Correspondences among Harvey et al.'s Usability Criteria for IVISs and the Contextual Factors and the General Usability Criteria From Which They Are Derived

	Harvey et al.'s usability criteria	**Harvey et al.'s contextual factors**	**General usability criteria from the literature**
1.	Effectiveness of interaction with device under varying driving conditions	Varying environmental conditions	Effectiveness
2.	Sustained effectiveness of interaction with device whilst driving	Dual task environment	Effectiveness, Efficiency, Errors
3.	Sustained efficiency of interaction with device whilst driving		
4.	Interference between interaction with device and primary task performance		
5.	Learnability of device interactions	Training provision	Effectiveness, Efficiency, Learnability
6.	Efficiency on first use of the device whilst driving		
7.	Effectiveness on first use of the device whilst driving		
8.	Adaptability of device to individual users	Range of users	Task match, Task characteristics, User criteria, Flexibility
9.	Compatibility of device with full range of users		
10.	Satisfaction on first use of device whilst driving	Uptake	Perceived usefulness, Satisfaction
11.	Perceived usefulness of device in real driving situation		
12.	Satisfaction based on short- and long-term use of device whilst driving	Varying frequency of use	Attitude, Satisfaction, Memorability
13.	Memorability of device interaction		

Note. IVISs = in-vehicle information systems. Adapted from [72]

7.4 Mapping the Usability Results of the CARAT Counter

As the usability criteria by Harvey et al. belong to the same field of application as the usability study of the CARAT counter (see Chapter 5), this section attempts to establish correspondences between them as an exercise. The mapping between the results and the Alonso-Ríos usability taxonomy was previously presented in Section 5.10.

The mappings between the usability attributes proposed by Harvey et al. and the CARAT counter results are shown in Tables 7.4 and 7.5.

If one compares the different mappings, the following conclusions can be drawn:

1. The links between the criteria of Harvey et al. and specific instances of real-world usability problems are often tenuous. This does not necessarily mean that important usability problems will go unnoticed, but they would need to be detected in an ad hoc fashion and leaning very heavily on context-specific usability guidelines and the expertise of the usability specialists.

2. The classification by Harvey et al. is missing some accepted usability attributes from the literature, such as certain aspects of robustness and safety. The latter is particularly significant, as Harvey et al. explicitly regard it as very important but did not actually include it as a usability criterion. But the fact is that adding the missing attributes would not be an easy task anyway, as the entangled relationship between their usability criteria and the usability criteria in the literature (as seen in Table 7.3) makes it difficult to add new attributes without adding more redundancy.

3. The intentionally context-specific nature of the classification by Harvey et al. makes it impossible to generalize it to other areas in order to reuse knowledge.

The mapping problems outlined above (e.g., vagueness, incompleteness, redundancy) are analogous to Winter et al.'s critique [186] of ISO 15005. In fact, they are also examples of well-known problems of expressiveness that will be explained in the next section.

7.5 Expressiveness of Usability Models

This section aims to analyze the problems that can occur when trying to represent usability information with the constructs offered by usability models. From a practical standpoint, the interest of the discussion lies in:

- Showing how a model can be described in terms of another model, which will be called the *reference model*.

Table 7.4: Results of the Usability Study of the CARAT Counter (CC), Classified according to the Usability Criteria for IVISs by Harvey et al. (part I)

	Usability Results for the CC	**Harvey et al.'s Usability Criteria**
1.	Text on left side of display is small	Effectiveness of interaction with device under varying driving conditions, Sustained effectiveness of interaction, Compatibility of device with full range of users
2.	Graphics on right side of display are big	
3.	Text on left side of display has bad contrast	
4.	Graphics on right side of display have good contrast	
5.	End-of-journey button has bad contrast	
6.	Display is structurally simple	Sustained effectiveness of interaction, Learnability of device interactions, Effectiveness on first use
7.	External buttons have unclear purpose	
8.	End-of-journey button does not look like a button	
9.	Speed limit sign does not indicate units	
10.	Error messages have unclear meaning	
11.	Pushing end-of-journey button before stopping engine is not intuitive	
12.	Errors are arbitrarily represented as text messages or graphics	
13.	Messages of behavior text and smiley sometimes seem contradictory	
14.	CC includes necessary functionalities	Sustained effectiveness of interaction
15.	GPS signal is not precise enough	Effectiveness of interaction with device under varying driving conditions, Sustained effectiveness of interaction
16.	Smiley has alternatives to color for users with color blindness	Compatibility of device with full range of users
17.	Text is English only	
18.	Speed limit sign is km/h only	
19.	CC does not demand much physical effort	Sustained efficiency of interaction, Efficiency on first use
20.	CC does not demand much time	

Note. IVISs = in-vehicle information systems.

Table 7.5: Results of the Usability Study of the CARAT Counter (CC), Classified according to the Usability Criteria for IVISs by Harvey et al. (part II)

	Usability Results for the CC	**Harvey et al.'s Usability Criteria**
21.	Counting seems to stop working	-
22.	USB is vulnerable to abuse	Compatibility of device with full range of users, Sustained efficiency of interaction
23.	CC may be distracting to the point of being dangerous	Interference between interaction with device and primary task performance
24.	CC may cause emotional discomfort	-
25.	CC does not compromise legal safety	-
26.	CC does not compromise confidentiality	-
27.	External shell is unattractive	Satisfaction on first use, Satisfaction based on short- and long-term use

Note. IVISs = in-vehicle information systems.

- Providing means to compare different models in terms of their expressiveness or, in other words, their capacity for representing information.
- Offering advice on how to generate a new model from another model in the sense of identifying omissions in the former and how to fix them without diminishing expressiveness.

Edelman [56] discusses representation as a mapping that "establishes correspondences between the members of the set of distal objects (things "out there" in the world) and the members of the set of proximal entities (things "in here" inside the head)" [56, p. 20]. Thus, different kinds of distal-to-proximal mappings "are liable to result in representations of varying usefulness" [56, p. 21]. This provides a basis for comparing different models of real-world phenomena.

Analogously, but applied to the field of conceptual modeling, Wand and Weber [184] propose a framework for evaluating the ontological expressiveness of information systems analysis and design grammars. Their work is based on philosopher Mario Bunge's ontology [38][39], which consists of ontological constructs like *things*, *properties*, *systems*, *states*, *events*, and so on (i.e., this ontology is much more complex than any usability model). Wand and Weber identify two desirable characteristics of a design grammar, namely, *completeness* and *clarity*. Lack of completeness is known as *construct deficit*, whereas lack of clarity is related to undesirable problems like *construct overload* and *construct redundancy*.

This line of research is followed up by researchers like Shanks et al. [173], Recker et al. [157], and Barcelos et al. [22], and has been applied to areas like telecommunications specifications [22]. Barcelos et al. propose four different formal properties that a mapping between a reference model and a representation artifact must exhibit, namely, *lucidity*, *soundness*, *laconicity*, and *completeness*. Deficiencies in each of these properties are related to the following problems, respectively [22]:

1. **Construct overload**. A mapping from the representation to the reference model is not functional (in the set theoretical sense). That is, a construct in the representation stands for more than one domain concept.

2. **Construct excess**. A mapping from the representation to the reference model is not total (in the set theoretical sense). That is, there is a construct in the representation which lacks interpretation in terms of a domain concept.

3. **Construct redundancy**. A mapping from the reference model to the representation is not functional. That is, a domain concept is represented by more than one representation construct.

4. **Construct incompleteness**. A mapping from the reference model to the representation is not total. That is, there is a concept in the domain that is not covered by any representation construct.

These deficiencies are also depicted in Figure 7.1

The terms for describing these deficiencies will be borrowed in order to analyze in detail some of the problems that were informally identified in Sections 7.3 and 7.4. On the one hand, Harvey et al.'s criteria and the Alonso-Ríos et al. usability attributes will be used as representational artifacts, and the ultimate purpose of the analysis will be to compare the different *approaches* to constructing usability models, not to evaluate those specific models. On the other hand, the usability models in the literature will be used as reference models. As usability is not a thing in the natural world but a useful construct, the "real" meaning of the term can only be defined by consensus among practitioners.

As suggested above, the discussion that follows could also lay the groundwork for establishing principles for constructing a taxonomy. The publication of the taxonomies in this thesis predates our discovery of the conceptualization shown in Figure 7.1, so, obviously, the latter was not used in the construction process. A more systematic construction methodology that takes all these aspects into account could be a worthwhile research topic.

7.5.1 Construct Overload

Construct overload happens when a representational artifact corresponds to more than one element in the reference model. Construct overload "causes ambiguity and, hence, undermines clarity. When a construct overload exists, users have to bring additional knowledge not contained in the representation to understand the phenomenon which is being represented" [22, p. 234].

For example, let us take Harvey et al.'s usability model for IVISs as the representation and ISO 9241-11 (i.e., *effectiveness*, *efficiency*, and *satisfaction*) as the

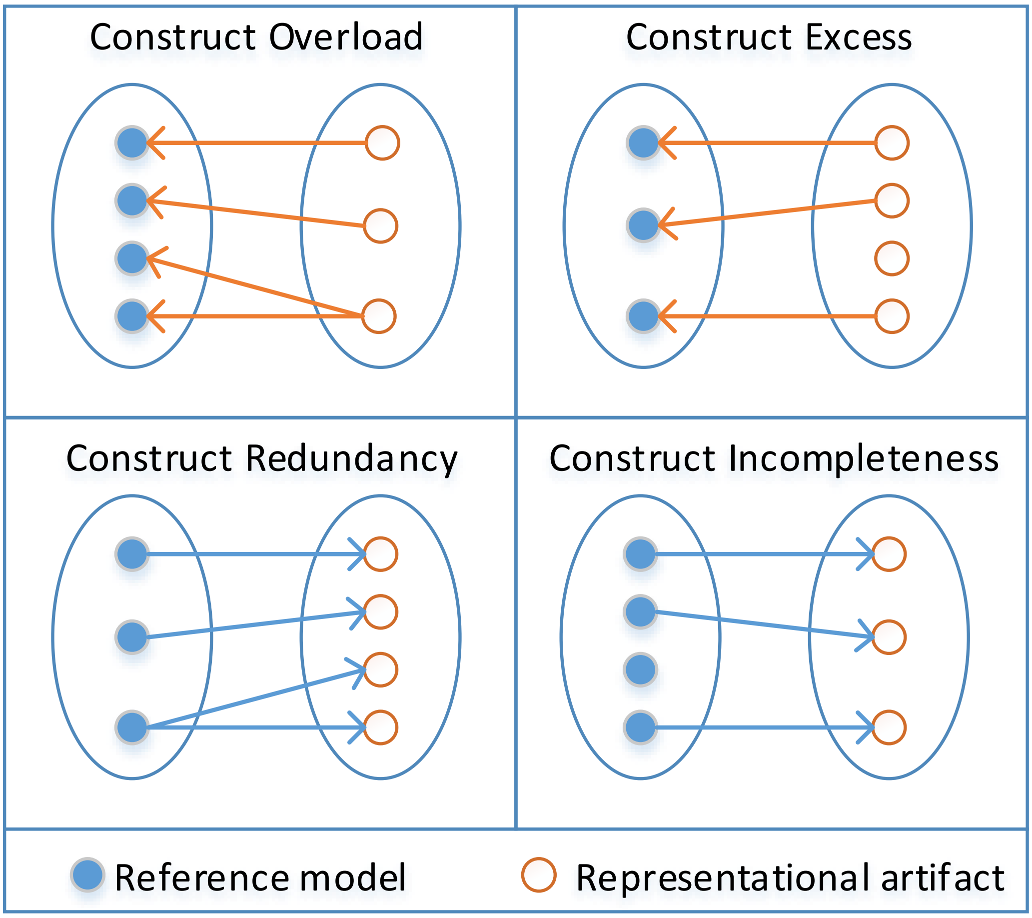


Figure 7.1: Deficiencies in the mappings between a reference model and a representation artifact (based on [22]).

reference model. As can be seen in Table 7.3, Harvey et al.'s *dual task environment* and *training provision* contextual factors map to both effectiveness and efficiency. In fact, something similar happens to almost all their contextual factors, as all but one map to several attributes from the literature at the same time, and often combining attributes from different usability models. Moreover, Harvey et al.'s usability criteria for IVISs are directly derived from their contextual factors, which means that the relationships between their criteria and the general usability from the literature are implicit rather than explicit (the relationships seem obvious in most cases, but not all). These problems only help to make the correspondences between Harvey et al.'s criteria and the usability literature more confusing.

As another example, let us take as a reference model Stanton and Baber's [176] extension of Shackel's four factors of usability [172]. Note also that Stanton and Baber's usability model is one of the models that Harvey et al.'s work is based on and that Stanton is also one of the authors of Harvey et al.'s criteria. According to Harvey et al.'s paper [72], the *range of users* contextual factor is derived from four general usability criteria from the literature: *flexibility* (from Shackel) and *task match*, *task characteristics*, and *user criteria* (from Stanton and Baber's "four additional factors to Shackel's LEAF precepts"). Harvey et al.'s *range of users* contextual factor, in turn, results in two of their usability criteria, namely, *adaptability of device to individual users* and *compatibility of device with full range of users*. As can be seen, four independent usability attributes are collapsed into a single contextual factor, which in turn yields two (seemingly non-independent) usability criteria. This simplification is in fact a conscious decision on their part, as "[t]he term 'compatibility' replaced task match, task characteristics, user criteria, and flexibility" [72, p. 333], whereas it would seem logical that these terms cannot be collapsed into one without losing information.

On the other hand, if one examines the correspondences between the Alonso-Ríos et al. usability taxonomy and the usability literature, which were previously shown on Tables 3.1 to 3.7, the only examples of a usability attribute mapping to several attributes from another usability model happen for Seffah et al.'s QUIM model [167]. Moreover, as the taxonomy by Alonso-Ríos et al. has more attributes than those shown in Tables 3.1 to 3.7, not all of these mappings are actual instances of construct overload. For example, on Table 3.3, *universality* maps to Seffah et al.'s *accessibility* and *universality*, but Alonso-Ríos et al.'s *universality* is in fact further subdivided into *accessibility* and *cultural universality* in the taxonomy. An example of actual construct overload could be *precision*, which maps to Seffah et al.'s *effectiveness* and *usefulness*, although it could also be argued that the overlap is actually in the attributes by Seffah et al., as effectiveness is defined by them as "the capability of the software product to enable users to achieve specified tasks with accuracy and completeness" [167, p. 168] and usefulness is defined as "whether a software product enables users to solve real problems in an acceptable way" [167, p. 169].

Generalizing these conclusions to the construction of new usability models, two suggestions could be made. First, the attributes need to be sufficiently specific. Second, when two attributes in the literature are merged, this must be done very carefully in order to avoid loss of expressiveness. Two attributes should be merged either when they are exact synonymous or when they are near synonymous and the semantic differences between them are covered by other attributes in the new usability model. All these are arguments against simplified models and in favor of including a sufficient number of attributes, which should also be grouped in a meaningful way. Definitions should also be included, as they help to make clear the intended meaning of the attributes. All of these were prerequisites for the usability taxonomy presented in this thesis.

7.5.2 Construct Excess

Construct excess happens when a representational artifact models something that has no correspondence in the reference model. Instances of construct excess are undesirable "because they add complexity to the representation without increasing its expressivity" [22, p. 234]. If we take any of the accepted usability models as a reference model, the existence of construct excess would mean that the representational artifact models something that is not generally accepted as a usability attribute.

Harvey et al.'s attempt to incorporate the context of use of IVISs directly into their usability criteria can be considered an example of construct excess. For example, their criteria repeatedly suggest a dichotomy between first use and sustained use in attributes like *effectiveness on first use of the device whilst driving* and *sustained effectiveness of interaction with device whilst driving*. The rationale seems to be that the "[e]valuation of *initial* effectiveness and efficiency [...] must be high in the context of in-vehicle devices because of the lack of training provision" [72, p. 331]. However, these could be considered ad hoc additions to the concept of usability as it is commonly understood. Moreover, the distinction between first use and sustained use is arguably not so clear-cut, as the effects of learning can be gradual and dependent on the characteristics of the users. These additions can be also considered superfluous on the grounds that the effects of learning are already represented by commonly-accepted context-of-use attributes like *experience*, *skills*, and *knowledge*.

As for the Alonso-Ríos usability taxonomy, Table 3.1 shows that all the first-level attributes in the taxonomy can be mapped to attributes in other usability models, although there is not a single usability model for which the mapping is total. That is, the usability taxonomy synthesizes different, but arguably partial, models of usability that can be said to collectively define the concept of usability. By far, the less commonly-accepted first-level attributes in the Alonso-Ríos et al. usability taxonomy would be *robustness* and *safety*, which only map to two usability models each. Examining the second-level usability subattributes in Tables 3.3 to 3.7, the only attributes that do not appear in any other usability model are *robustness to environment problems* and *environment safety*, which are, in turn, refinements of the concepts of *robustness* and *safety*, respectively.

Generalizing to the construction of usability models, construct excess would be an argument for basing a new usability model on commonly-accepted usability attributes from the literature. As usability is a construct collectively defined by consensus, deviating from established conceptions of the term can be problematic and needs a clear justification. Adhering to tradition, rather than reinventing it, is – at least philosophically – opposed to construct excess, and was the general approach taken in this thesis.

7.5.3 Construct Redundancy

Construct redundancy happens when an element in the reference model corresponds to more than one representational artifact. Examining Harvey et al.'s criteria, the ISO 9241-11 attribute of *effectiveness* maps to seven of their criteria, *efficiency* to six, and *satisfaction* to four. However, none of their criteria are completely redundant, as they all actually convey different meanings. Note also that Harvey et al.'s model has 13 criteria, whereas ISO 9241-11 has only three attributes, so it makes sense that several of their criteria map to the same ISO 9241-11 attribute. Likewise for the Alonso-Ríos et al. usability taxonomy.

Therefore, when these types of mappings happen, a distinction must be drawn among three situations. An element in the reference model can map to two usability attributes whose respective meanings:

1. Do not overlap.
2. Overlap partially.
3. Overlap completely.

The first situation is characteristic of refinements in which an element in the reference model is subdivided into two non-overlapping subelements (e.g., Harvey et al.'s *effectiveness on first use of the device whilst driving* and *sustained effectiveness of interaction with device whilst driving*). The second situation, on the other hand, can lead to inconsistency. For example, in Harvey et al.'s criteria, can *learnability of device interactions* be evaluated positively if *memorability of device interaction* is rated negatively? In fact, Harvey et al. acknowledge the existence of redundancy:

> "Although Nielsen and Shackel only listed one common attribute (learnability) in their definitions of usability, there is much overlap between the two descriptions. For example, memorability (Nielsen) is related to learnability (Shackel and Nielsen); efficiency (Nielsen) is a measure of effectiveness (Shackel) against some other metric such as time; errors (Nielsen) are closely linked to effectiveness and efficiency (Shackel); and satisfaction (Nielsen) is synonymous with attitude (Shackel). This is the evidence of the difficulty in defining concrete terms for usability and is perhaps one reason why a universal definition of usability has so far proved difficult." [72, p. 322]

Is this problem unavoidable? *Learnability* is a widely-used attribute in the literature, but its vagueness (and the lack of precise definitions in the literature) motivated the author of this thesis to refine the term and specify the characteristics that ensure good learnability. Hence the *knowability* attribute in the Alonso-Ríos et al. usability taxonomy, which is based on *learnability* (along with *ease of learning*, *understandability*, and *memorability*) and is further subdivided into *clarity*, *consistency*, *memorability*, and *helpfulness* (which are, in turn, further subdivided into more subattributes).

The third type of mapping, that is, mapping to two attributes whose meanings overlap completely, is the purest case of redundancy. If two usability criteria are exact synonymous, only one of them should be retained. Construct redundancy introduces ambiguity into the representational system [56] and adds complexity without increasing expressivity. Moreover, users may ascribe different interpretations to the different representations [22]. Again, resolving construct redundancy becomes difficult if no definitions are provided.

The existence of construct redundancy is an argument for constructing detailed, hierarchically structured, and precisely defined usability models that do not merely borrow attributes from different classifications in the literature. Rather, the attributes should be merged carefully, avoiding redundancy and breaking them down into non-overlapping subattributes.

7.5.4 Construct Incompleteness

Construct incompleteness happens when there is an element in the reference model that is not covered by any representational artifact. "Representation incompleteness entails lack of expressivity" [22, p. 234] as "there are phenomena [...] that cannot be represented" [22, p. 234].

For example, and as mentioned in Section 7.4, Harvey et al.'s criteria do not include some aspects of safety, which is a characteristic they deem very important (e.g., "the safety-critical nature of driving, must be accounted for in defining and evaluating the usability of in-vehicle devices" [72, p. 318]). However, in their actual criteria, safety is only related to *interference between interaction with device and primary task performance*, which leaves out other, more fine-grained aspects of safety that were already covered by authors such as Seffah et al. [167]. Note also that the lack of definitions in the Harvey et al. criteria makes it difficult to determine whether construct incompleteness actually exists or not in some cases.

The usability taxonomy by Alonso-Ríos et al., on the other hand, covers all the attributes in the seven usability models in which is based, with the exception of the *usability compliance* attribute from ISO/IEC 9126-1, which was deemed redundant. Moreover, as *usability compliance* does not appear as such in the other usability models under consideration, it could be argued that ISO/IEC 9126-1 suffers from construct excess if we take any of the other usability models as a reference. Note also that *usability compliance* does not appear in the recent ISO/IEC standard that replaces ISO/IEC 9126-1, namely, ISO/IEC 25010 [89].

Construct incompleteness limits the usefulness and the reusability of a representational model. When constructing a usability model, all of the above are arguments for following closely the usability literature, as well as including definitions in order to avoid ambiguities.

7.6 Generalizability of the Models

By their very nature, the usability criteria by Harvey et al. are not generalizable to other systems. Being tailored for a very specific context of use from the start means that they are consequently limited by it . While it is widely accepted that "it is important to understand and identify the details of [the] context, in order to guide early design decisions, and to provide a basis for evaluation" [28, p. 360], trying to construct a definitive set of usability criteria from a restricted view of the context of use has two problems: loss of information (as demonstrated in Section

7.5) and rigidity. Loss of information makes it impossible to know which context-of-use and usability attributes have been discarded as non-relevant in the process, which means that the usability classification cannot be easily applied to other areas, or even revised for the same area. As for rigidity, it should be borne in mind that the context of use can change over time, or even during the life cycle of a project. Harvey et al.'s methodology allows for revising the selection of evaluation methods [74], but their usability criteria and the context of use are fixed from the start. On the other hand, standards such as ISO 13407:1999 [85] have explicitly considered the tasks of understanding and specifying the context of use as a part of an iterative process. For these reasons, the approach in this thesis is to conceptualize usability and context of use into two separate and general-purpose taxonomies and to try and make them as comprehensive as possible. The relative importance of each usability attribute could then be assigned as a weight and the dependencies between the taxonomies could be mapped as relations, which would make the process more flexible and generalizable.

Chapter 8

Discussion

8.1 Introduction

This chapter discusses in depth several related topics that have arisen so far in this thesis. It reflects on the lessons learned while conducting this research, addresses unexplored aspects – including similar work published contemporaneously – and summarizes strong points and limitations. It also addresses explicit references and critiques made to this research by other authors. The chapter is divided into a discussion of the scope of the thesis and a discussion of the methodology.

8.2 Discussion of the Scope

One of the main premises of this work is that the concept of usability has been very inconsistently described in the literature. There is a noticeable lack of consensus between authors and sometimes there is little overlap between definitions, or the overlap is not immediately obvious. This is not helped by the fact that usability tends to be described in vague and informal ways. Although standards do exist, such as those of the ISO, these have been criticized by authors for several reasons: they "do not support all usability aspects; furthermore, they are not well integrated into current software engineering practices and they lack tool support" [12, p. 23]. This helps to explain why usability studies are often conducted in an ad hoc manner. The ad hoc nature of the process can make the study of usability very laborious, as the "developer must read a significant body of work to develop even an inkling of the global thought on usability" [49, pp. 15-16]. Moreover, it can lead to problems of inconsistency and redundancy.

This situation motivated the author of this thesis to construct a *usability* taxonomy, which was intended as a comprehensive, structured, and general-purpose synthesis of the literature. General-purpose taxonomies are, however, inherently abstract, and in order to address specific usability problems it is always necessary to negotiate the gap between the abstraction of a taxonomy and the specific context of use and the characteristics of the product being analyzed. A general-purpose taxonomy for the *context of use* has also been proposed, which is, again, a detailed synthesis of the existing literature. Moreover, the concepts of taxonomies of *components* of user interfaces (see Section 6.5) and *tasks* (see Section 5.6) have also been introduced (applied to web interfaces), although both have remained out of the scope of this thesis.

The usability taxonomy in this thesis does not limit itself to compiling attributes as they appear in the literature, but also seeks to expand and refine them. This motivation is not exactly new, and in fact constitutes some kind of small tradition in usability research. Lewis [111] coined the term "expanded models of usability" to refer to the work of authors like Bevan [29], Seffah et al. [167], Winter et al. [186], and Alonso-Ríos et al. [16].

Bevan's contributions to the usability literature have been unquestionably influential (for example, his work is included in the relevant literature for ISO 9241-11 [87]), and he participates in several international standards groups and directly contributed to the development of ISO 13407 [85][4]. Seffah et al.'s usability model, also known as the QUIM model [167], has been extensively covered in Chapter 3 and was probably the most comprehensive definition of usability that had been published when the usability taxonomy by Alonso-Ríos et al. was being developed. Concurrently, Winter et al. were developing their own "Comprehensive Model of Usability" [186], also partly based on the QUIM model. As the author of this thesis was not aware of Winter et al.'s work at the time, it was never a part of his synthesis of usability definitions. Unfortunately, Winter et al. state that their "complete model is too large to be described in total" [186, p. 113], and proceed instead to "highlight specific core parts of the model to show the main ideas" [186, p. 113]. Furthemore, they never extended this line of research [183] in later peer-reviewed publications, so a detailed comparison between their model and the usability taxonomy by Alonso-Ríos et al. would be moot.

The most interesting and unusual thing about Winter et al.'s model is the fact that it is two-dimensional. The motivation is "to ensure the well-structuredness of these model instances and foster their preciseness as well as completeness" [186, p. 111]. It also "introduces a rigorous separation of system properties and activities to be able to describe quality attributes and their impact on the usage of a software product precisely" [186, p. 111].

As with the Alonso-Ríos et al. taxonomies, the motivation for Winter et al.'s model is a result of the limitations of existing usability models. For example, they argue that since the QUIM model's "decomposition doesn't provide any means for precise structuring, the factors used in the QUIM model are not independent" [186, p. 109]. This problem was also discussed in Chapter 3, but the solution offered by the author of this thesis was to refine, merge, or eliminate the problematic factors. Moreover, the factors in the QUIM model "contain attributes that are similar, e.g. appropriateness and consistency, both of which are defined in the paper as capable of indicating whether visual metaphors are meaningful or not" [186, p. 110]. Again, the solution offered by the author of this thesis was to examine the meanings behind the attributes themselves and merge them or split them into non-redundant attributes.

Interestingly, Winter et al. also criticize the existing models for containing "a dangerous mixture of activities and actual system properties" [186, p. 111]. For example, in the QUIM model "time behavior and navigability are presented as the same type of criteria. Where navigability clearly refers to the navigation activity carried out by the user of the system, time behavior is a property of the system and not an activity" [186, p. 111]. This contrasts with the approach taken in this thesis, as the usability taxonomy limits itself to rearranging and expanding the attributes in the literature, rather than interrogating their nature or purpose. This means that if the existing definitions mix perceived qualities with "inherent" properties, this mixture is also carried out to the resulting synthesis.

In Lewis's review of the "current controversies in the field of usability" [111], he notes that all the expanded models of usability "have yet to undergo statistical testing to confirm their hypothesized structures" [111, p. 666]. This was already mentioned in Chapter 7, in which the validity of the usability taxonomy was investigated from a different angle. In contrast, Lewis argues that "there is compelling psychometric evidence for an underlying construct of usability for the traditional metrics of effectiveness, efficiency, and satisfaction" [111, p. 666].

It is important to understand the context in which Lewis makes this argument. The "compelling psychometric evidence" to which Lewis refers is a study conducted by Sauro and Lewis [164] on correlations among prototypical usability metrics. This study

> "incorporated data from usability studies conducted from 1983 to 2008, including products such as printers, accounting and human resources software, websites and portals. In total we obtained 97 raw data-sets from 90 distinct usability tests, all of which contained some combination of the prototypical usability metrics, with data from over 2000 unique users and 1000 tasks." [164, p. 1609]

Sauro and Lewis argue that their results provide evidence for an "underlying construct of usability containing two components, one objective and one subjective" [164, p. 1615]. More specifically, they are referring to the traditional attributes of effectiveness, efficiency, and satisfaction. The main implication of their study is that it would demonstrate the usefulness of these three traditional attributes of usability, as they correlate strongly with the usability metrics employed in usability tests.

The Sauro and Lewis study is intended as a refutation of a previous study by Hornbæk and Law [80], which used raw data from 73 usability studies and obtained diametrically opposed conclusions. Namely, it "reported weak correlations among efficiency, effectiveness and satisfaction" [164, p. 1609]. Weak correlations would mean that efficiency, effectiveness and satisfaction are independent, which would mean that, firstly, the three attributes need to be explicitly taken into account when making general claims about usability [63], and, secondly, the attributes cannot be meaningfully aggregated into a single usability score [80].

Hornbæk and Law argue that "attempts to reduce usability to one measure [...] are bound to lose important information, because there is no strong correlation among usability aspects. Apart from masking potentially interesting details, the use of a single usability score cannot support formative evaluation well" [80, p. 625]. Sauro and Lewis do not entirely disagree with this view, as they acknowledge that "the variety of studies used in Hornbæk and Law most likely provide a better picture of the broader area of human computer interaction (HCI), whereas the data analyzed here [i.e., in the study by Sauro and Lewis] present a more focused picture of summative usability tests" [164, p. 1616].

Hornbæk and Law's observation that single usability scores cannot support *formative* evaluation well (i.e., detecting usability problems and proposing solutions) offers evidence on something that usability practitioners have already observed in practice: the traditional attributes of effectiveness, efficiency, and satisfaction are useful for providing global, non-detailed assessments of the usability of a product, whereas the identification of more specific usability issues requires a more comprehensive view of usability. Hence the need for expanded models of usability.

The two usability studies described in this thesis have been focused on formative goals. *Summative* goals, that is, obtaining general assessments of usability, have been mostly out of the scope of this thesis. Note that the respective usability questionnaires for both usability studies contain a separate item asking users about their overall opinion, but this can only be used as an informal assessment. Overall scores like these are "constrained by how users interpret the definition of usability. For instance, they may selectively and inconsistently focus on certain aspects when assessing the usability of an object" [80, p. 618].

An aggregated usability score was informally calculated for the CARAT counter in Section 5.9. Each usability subattribute was given equal weight, as there was no basis for determining the relative importance of each usability attribute with respect to the others. An aggregated usability score could have also been calculated for the user study of the automatically-generated UIs, as aggregated scores are particularly suited for choosing among competing systems [110]. This particular course of action was not taken for a variety of reasons: statistically significant metrics had already been obtained; an overall subjective score had already been given by the users; calculating numerical aggregations of subjective data is generally problematic.

The question of determining the relative importance of a given usability attribute with respect to others is a recurring topic in the usability literature. This is fundamentally approached in two ways [111]. The first approach is based on the judgments of stakeholders (e.g., usability practitioners and developers). The second approach is based on data associated with usability problems, mainly frequency and impact (but also ease of correction, likelihood of usage of the affected part of the product, etc.). A common method of determining impact is to assign "scores according to whether the problem (a) prevents task completion, (b) causes a significant delay or frustration, (c) has a relatively minor effect on task performance, or (d) is a suggestion" [111, p. 670].

The general approach taken in this thesis has also been criticized on philosophical grounds, particularly by Riemer and Vehring [161]. As the taxonomies in this thesis are a synthesis of the usability literature, the taxonomies can be easily criticized for reinforcing conventional philosophical assumptions:

> "Often, software usability is then further broken down into more detailed lists of characteristics that a software needs to meet to be deemed usable [...] For example, Alonso-Ríos et al. (2010), in a recent article, have criticized existing usability definitions as being too brief. Instead, the authors have developed an "exhaustive and thorough taxonomy of the attributes underlying the concept of usability". In essence, by drawing on the extensive body of usability literature they have synthesized a comprehensive catalogue of artifact properties to describe all facets of usability in meticulous detail. Consequently, usability, while not an atomic characteristic, is seen as describing an artifact, as comprising of a set of quality attributes or properties of that artifact." [161, p. 3]

Similar criticisms are leveled at the approach to the context of use, which is

> "treated in very abstract ways. For example, Alonso-Ríos et al. [...] see the necessity to derive formalized attributes for describing context aspects. Hence, while context is sometimes acknowledged as an influence factor, the call typically is not for a contextualized understanding of usability, but for a more complex set of variables to capture a generalized understanding of usability properties." [161, p. 3]

Describing the context of use through formalized attributes is, of course, a common approach in the usability literature. Interestingly, this demand for "a contextualized understanding of usability" brings to mind Harvey et al.'s rejection of universal definitions of usability:

> "Consideration of the context of use makes a general definition of usability virtually impossible because different situations will demand different attributes from a product to optimise the interaction with particular users. Despite the desire to construct a universal definition, it appears that most people now accept that the context in which a product or system is used must be taken into account, and definitions therefore need to be constructed individually according to the product, tasks, users and environment in question." [72, p. 324]

However, Riemer and Vehring's critique is more philosophical in nature. Riemer and Vehring use the failings of one of their own usability studies as an example of the limitations of the traditional conception of usability, with the aim of "illustrating that software cannot be usefully conceptualized as a bundle of features with certain objectively agreed upon properties, one of which is usability" [161, p. 11]. Their usability study "clearly has to be regarded a failure, as [they] were not able to establish any coherent understanding of the usability of the client software" [161, p. 7].

Just for the record, the author of this thesis does not believe that usability is an objective property or a set of objective properties. Even though the definitions in the taxonomies include sentences like “knowability is defined as the property by means of which the user can understand, learn, and remember how to use the system” [16, p. 56], perhaps the word “property” was not the best way to describe them.

Riemer and Johnson [160] expand on Riemer and Vehring’s line of research to attack what they consider a Cartesian subject/object dualism in IT. Using Heidegger’s *ways of Being* [75] and his conception of *equipment*, they aim to “demonstrate that, under this worldview, IT in the users’ world is *ontologically* different to what it is in the designers’ world” [160, p. 6]. They

> “argue that phenomena of IT in use cannot be grasped appropriately with a worldview that takes the distinction between user subject and IT artefact to be a fundamental principle. At the same time however, unlike users, designers do indeed encounter the IT artefacts of their creation as objects, since these IT artefacts are at the focus of their concern. This is why design is commonly conceptualised as the act of endowing an IT object with specific features (properties).” [160, p. 6]

Therefore, this would point to a split in how the usability of a system is understood by users and designers (or, analogously, usability experts). What sets Riemer and Johnson apart from other authors is their belief “that this problem is better thought of as ontological than epistemological” [160, p. 14]. This is not surprising, as Heidegger prioritizes ontology over consciousness and subjectivity [6], but the *practical* usefulness of the ontological angle is not obvious to the author of this thesis, and neither does it seem clear how this can help Riemer et al. overcome their impasse.

Riemer and Johnson describe several consequences for their analysis, but this is probably the most relevant for this thesis:

> “It is a well-known phenomenon that during absorbed use we do not experience the objects we are using [...] When driving a car we can engage in conversation or thought while our body does the driving to the extent that we find ourselves at our destination without quite remembering how we got there. Even if our attention is with the street ahead, the car itself remains withdrawn, as we move effortlessly in traffic. The same happens to the word processing software and the computer keyboard when our attention is with the text we are writing.” [160, p. 11]

This also points to the different purposes of, for example, heuristic evaluations and user questionnaires. Heuristic evaluation is more analytic and works at a very detailed level, whereas questionnaires are more general, and it would be unproductive to ask the users to remember every element in the display. Instead, it makes more sense to ask them about their general impressions. Note, however, that usability techniques like retrospective testing, in which the tester and the user review a video recording of the test, allow the users to regain perspective by distancing themselves from the test. Note also that, for example, the users in the usability study of automatically-generated UIs were not familiar with the UIs, and therefore also felt too distanced from them to fully achieve the state of absorbed use that Riemer and Johnson describe, especially when they were confused by the horizontal layout. Consequently, there are situations in which the users do not perceive the systems so differently from the way they are perceived by the usability experts.

In fact, the phenomenon of absorbed use has long been investigated in the field of Human-Machine Interaction from a psychological point of view [155][61][181]. For example, in Rasmussen's *skills, rules, and knowledge* taxonomy [155]:

> "The skill-based behavior represents sensory-motor performance during acts or activities which, following a statement of an intention, take place without conscious control as smooth, automated, and highly integrated patterns of behavior." [155, p. 258]

The author of this thesis would argue that approaching these phenomena from a psychological angle appears more promising from a pragmatic standpoint, which is what the discipline of engineering ultimately demands. Moreover, a critical stance towards conventional philosophical wisdom is not necessarily incompatible with pragmatism. Even though Heidegger's interests were in interrogating the philosophical foundations of science, he himself acknowledged that "in suggesting that anthropology, psychology, and biology all fail to give an unequivocal and ontologically adequate answer to the question about the kind of Being which belongs to those entities which we ourselves are, we are not passing judgment on the positive work of these disciplines" [75]. Usability techniques, like science, deal with simplifications of reality in order to produce useful results and predictions, so, from a pragmatic point of view, it makes sense to continue using them for the time being.

The usability taxonomy presented in this thesis has also been used by some researchers in order to construct their own usability models, as described below.

Dubey et al.'s integrated model for software usability [54] consists of the following attributes (which are in turn further subdivided into several levels of subattributes):

1. **Effectiveness**, which is subdivided into *task accomplishment*, *operability*, *universality*, *flexibility*, and *errors*.

2. **Efficiency**, which is subdivided into *user effort*, *finance*, *resource utilization*, and *performance*.

3. **Satisfaction**, which is subdivided into *likeability*, *trustfulness*, *comfort*, and *attractiveness*.

4. **Comprehensibility**, which is subdivided into *clarity*, *learnability*, *memorability*, and *helpfulness*.

5. **Safety**, which is subdivided into *user safety*, *third party safety*, and *environmental safety*.

Gupta et al.'s usability model [68] consists of:

1. **Efficiency**, which is further subdivided according to *resource*, *time*, *economic cost*, *documentation*, and *user effort*.
2. **Effectiveness**, which is subdivided into *task accomplishment*, *operability*, *extensibility*, *reusability*, and *scalability*.
3. **Satisfaction**, which is subdivided into *likeability/attractiveness/interest*, *convenience*, and *aesthetics*.
4. **Memorability**, which is subdivided into *learnability*, *memorability*, *comprehensibility*, and *consistency*.
5. **Security**, which is subdivided into *safety*, and *error tolerance*.
6. **Universality**, which is subdivided into *approachability*, *utility*, *faithfulness*, and *cultural universality*.
7. **Productivity**.

Selamat et al.'s taxonomy of user interface acceptance [169] consists of the following goals:

1. **Knowledge ability**.
2. **Motivation**.
3. **Learning style**.
4. **Knowability**.
5. **Operability**.
6. **Efficiency**.
7. **Robustness**.
8. **Safety**.
9. **Subjective satisfaction**.
10. **Media element**.
11. **Communicativeness**.
12. **User expectation**.

Buccella et al.'s usability framework [36] consists of:

1. **Knowability**, which is subdivided into *clarity of elements and structure*, *consistency of elements and structure*, and *helpfulness*.

2. **Operability**, which is subdivided into *completeness and configurability.*

3. **Efficiency**, which is subdivided into *efficiency in human efforts* and *efficiency in task execution time.*

Mullins's taxonomy of evaluation methods to address cartographic symbol usability [127] consists of:

1. **Comprehension**, which is subdivided into *clarity*, *consistency*, and *memorability.*

2. **Operability**, which is subdivided into *accuracy*, *precision*, *completeness*, and *flexibility.*

3. **Efficiency**, which is subdivided into *human effort* and *interpretation time.*

4. **Robustness**, which is subdivided into *context* and *improper use.*

Finally, Moorthy [125] has merged the taxonomies by Alonso-Ríos et al. [16] and Dubey et al. [54] in order to construct a taxonomy of software risks classified according to the affected usability attributes.

8.3 Discussion of the Methodology

The general methodology followed in this thesis could be described as analytic in the sense of breaking down the things being investigated into their constituent parts. These things include not only the concepts of usability and the context of use, but also the systems whose usability is being studied. The topic of decomposing the system itself into its constituent parts has also been investigated by researchers like Han et al. [69], but, as far as the author of this thesis knows, this line of research has remained basically underexplored.

This analytic approach contrasts strongly with most usability studies, which tend to rely solely on simple usability models and are of a more intuitive nature.

The author of this thesis would argue that a taxonomy of attributes (for usability, the context of use, etc.) would ideally feature the following characteristics:

- Meaningful (as opposed to vague) concepts that can be understood by the participants in a usability study.

- Explicit and precise definitions that help to clarify the concepts.

- A hierarchical structure for these concepts, which would facilitate scalability and the estimation of the complexity of the usability aspects under study.

Based on this point of view, the goals for constructing the usability and the context-of-use taxonomies were as follows:

- To make the taxonomies comprehensive, covering all the commonly-accepted attributes from the literature but avoiding contradictions and redundancy.
- To follow the usability and context-of-use models from the literature as closely as possible, prioritizing a synthesis and refinement from within, instead of a radical departure from the way they are commonly understood.
- To structure the taxonomies hierarchically into several levels of detail, from the general to the particular. Typically, the usability classifications in the literature only have one level.
- To make the taxonomies general-purpose and applicable to any type of product. This contrasts strongly with traditional usability models, which are intentionally restricted to IT systems.
- To provide definitions for all the attributes and subattributes, which is something often neglected in the literature.

These taxonomies were used in two usability studies, which are described in detail in Chapters 5 (CARAT counter) and 6 (automatically-generated user interfaces). From a methodological point of view, the most novel aspect of these usability studies is in how deeply the taxonomies are integrated with them. While usability studies typically start from some agreed-on model of usability, the usability models tend to be simple, which is one of the reasons why studies are heavily based on guidelines and ad hoc judgments.

Even though the level of detail of the usability taxonomy (compared to traditional models of usability) allows to get closer to concrete usability problems, the taxonomy cannot be expected to address the whole range of usability problems of formative usability evaluation. Additional context-specific guidelines are needed, as these tend to be more detailed than any usability model. But the fact that guidelines were used in this thesis to complement the limitations of a general-purpose usability model does not mean that guidelines are not needed for tailor-made, context-dependent usability models. As Harvey et al. [72, p. 328] explain:

> "Detailed guidance on a specific product may be issued in the form of guidelines. These are based on the original usability factors, however, they specify a far greater level of detail, usually including performance thresholds, and will be targeted at a particular product range, e.g. in-vehicle devices." [72, p. 328]

When discussing the guidelines and authors they had consulted, Harvey et al. ask us to "note that there is no link from guidelines to specific authors [...] This is because guidelines are context-specific and so will need to be selected by designers on an individual basis according to the type of product they are designing" [72, p. 328].

While the author of this thesis agrees with the idea that guidelines are necessary, he disagrees with the a priori rejection of the possibility of establishing meaningful connections between a general-purpose usability model and low-level guidelines. For example, Green et al. suggest the following guideline for driver information systems: “Provide flexibility in terms of the sequence of actions” [66, p. 14]. While this is a critical functionality in these particular types of systems, it is also a desirable usability characteristic in general, and is closely related to the Alonso-Ríos et al. subattribute *freedom in tasks*, which is in turn a subattribute of *flexbility*.

In other respects, the usability studies in this thesis are rather typical in that they make use of a fairly standard combination of usability techniques, which is also very similar for both studies. Broadly, a heuristic evaluation is followed by user tests and questionnaires. There were, however, some small but significant differences in the two studies, which allowed to explore different angles of usability engineering. More specifically:

- The system’s characterization was fairly different for both studies. Firstly, the CARAT counter was structurally simple (at the user interface level), whereas the automatically-generated UIs had more types of elements. Secondly, the CARAT counter consisted of a very specific and purpose-built combination of hardware and software, whereas the automatically-generated UIs were multi-platform HTML/CSS web pages displayed on mass-produced smartphones. The standardized nature of HTML and CSS gave rise to a more formalized and generalizable characterization than the one obtained for the CARAT counter, as the latter was very tied to the particular characteristics of the device itself. In fact, the characterization of the automatically-generated UIs constitutes a first step towards constructing a general-purpose – and, therefore, more reusable – taxonomy of system components, which is something that could not have been achieved with the CARAT counter.

- There were 60 participants in the user tests of the automatically-generated UIs, whereas four persons participated in the user tests of the CARAT counter.

- The user tests of the CARAT counter only yielded qualitative data, whereas the user tests of the automatically-generated UIs also yielded a great amount of quantitative data, including statistically significant usability measurements, thus covering all three categories (i.e., expert, subjective, and empirical) of usability evaluation in the Adelman and Riedel conceptual framework [13].

- Whereas the user tests of the CARAT counter were focused on one system, the user tests of the automatically-generated UIs aimed to investigate the *comparative* usability of different layouts for the same application. The differences were exclusively in the layout, as the UIs were otherwise identical. For this reason, the usability study of automatically-generated UIs was intentionally restricted to a subset usability attributes, whereas the study of the CARAT counter examined its usability in full.

- Information on the user preferences (or biases) regarding the usability criteria was collected for the usability study of automatically-generated UIs.

The *comparative* nature of the user tests of the automatically-generated UIs poses significant challenges from a methodological standpoint, as this means that the tests are concerned with only a subset of usability attributes. While it is certainly easy to prune specific branches of the taxonomy and retain only a few selected attributes and subattributes, difficulties arise if one tries to "merge back" an arbitrary selection of attributes with the aim of, for instance, simplifying a user questionnaire. For example, the subattributes of *operability* are structured in a specific way in the usability taxonomy and if one wants to focus exclusively on, for instance, *universality* and *flexibility* (and therefore ignore *completeness* and *precision*), it becomes difficult to find a name for a new usability attribute that would synthesize these two subattributes and only these two subattributes.

The problem is further complicated by the fact that the items in a usability questionnaire need to be self-explanatory for the respondents and not too numerous. For example, as mentioned in Section 6.7.3, it was decided to merge the attributes *visual clarity of elements* and *visual clarity of structure* into one questionnaire item. After much deliberation, it was decided to phrase this as "which interface makes information more visible?". This means that *elements* and *structure* were merged into "information", and *visual clarity* (as opposed to *conceptual clarity*) was reworded as "visibility". Once more, this needs to be done very carefully, as two risks exists when merging several attributes into a new one. On the one hand, there is the risk of failing to convey all the intended meanings. On the other hand, there is the risk of suggesting unintended connotations. The second questionnaire in the usability study of the automatically-generated UIs, that is, the one about the users' preferences regarding the usability criteria (see Section 6.8.5) posed the same kinds of problems, as it consisted of four questions taken from the usability questionnaire.

As mentioned in Section 6.7.3, the usability questionnaire included redundant questions with the aim of allowing informal checks for consistency in the responses, as *knowability* and *efficiency* were represented by three questionnaire items each. This, perhaps inevitably, led to the above-mentioned problems of "merging back" an arbitrary selection of attributes and phrasing the result as self-explanatory questions for the participants. From an informal check of the results, the responses for the redundant items seem consistent with the responses for the items they are meant to summarize, with the notable exception of the Motorola Xoom test (see Table 6.19). Item number 4 is meant to summarize items number 1 to 3, but the users' preference for the chosen UI is stronger on items number 1 to 3 than it is on item number 4. Looking closely at the *individual*[1] responses of the participants, it is strinking how two users responded "Swiping, moderately" to item number 4 but responded "Swiping, strongly" or "Swiping, extremely" for all the first three items. Similarly, one user responded "Tabs, moderately" to item number 4 but responded "Swiping, moderately", "Swiping, moderately", and "Swiping, strongly" to items number 1, 2, and 3, respectively. This might perhaps be a blip, but it calls into question the accuracy with which item number 4 summarizes the three previous items.

[1]Please note that this thesis has so far only shown the aggregated data, not individual responses. This was done for conciseness and to respect the confidentiality of the information.

The construction of the questionnaires for the automatically-generated UIs has been, by far, the most ad hoc activity carried out during the two usability studies. Even though the aims of this thesis are clearly opposed to ad hoc activities, sometimes compromises are hard to avoid. In this case, it is because usability studies that are focused on a subset of usability aspects inevitably lead to custom-made questionnaires. Although it is justified in this case, ad hoc questionnaires should be clearly avoided if other options exist. According to Hornbæk [78]:

> "While specific studies assessing specific aspects of usability may justify the need for custom-made questions, a large group of studies do no more than measure [...] ease-of-use. Those studies add in their reinvention [...] little but lack of clarity and difficulties in comparing to other studies." [78, p. 91]

For "absolute" usability studies like the one for the CARAT counter, the question also arises as to whether it is preferable to use a questionnaire directly based on a usability model or if it is preferable to use a standardized usability questionnaire from the literature. While standardized questionnaires were consulted for this thesis, they were not directly used for a variety of reasons: they were too short, or they were too long, they contained redundant questions, and so on. Therefore, they were used mainly to inspire the wording of specific sentences and to confirm that no significant usability aspects were being ignored, as stated in Chapter 6. Hornbæk responds to criticisms of standardized questionnaires in this manner:

> "Some readers may be skeptical about the use of standardized questionnaires because they feel that such questionnaires are not applicable in the context-of-use under consideration or are unnecessarily limiting the scope of studies of usability. While we acknowledge that questionnaires for all constructs of interest to HCI do not exist, skeptical readers may appreciate that comparing studies using standardized questionnaires would be easier than comparing the studies reviewed." [78, p. 94]

Along similar lines, Sauro and Lewis [164] conducted a statistical study on the internal reliability of post-test questionnaires and found that standardized questionnaires were more reliable than ad hoc questionnaires. It should be noted, however, that the ad hoc questionnaires they chose are not as representative as they could have been, as all of them

> "asked questions about ease of use and at least one additional construct. For example, one questionnaire asked whether the product met the user's business needs and another asked about the perceived attractiveness of the interface. The inclusion of these items reduced the internal reliability, suggesting that they were getting at a construct other than usability." [164, p. 1615]

If one decides to use standardized questionnaires, it is generally advisable to avoid modifications. Lewis describes two points of view regarding this issue. On the one hand,

> "a standardized questionnaire has a specific set of questions presented in a specified order (which could be random order) using a specified format, with specific rules for producing metrics. Any deviation, no matter how slight, from these specifications makes it possible that the resulting measures would be invalid. These deviations might include changing the wording of an item, changing the number of scale steps or step labels, or using the questionnaire in a setting different from its development setting." [111, p. 673]

But, on the other hand:

> "Robust psychometric instruments should be able to tolerate some deviation from specification. When there is deviation from specification, the question is whether the practitioner or researcher has merely bent the rules or has broken the instrument." [111, p. 673]

Therefore, the questions about the advisability of using standardized usability questionnaires remain open.

Chapter 9

Conclusions and Future Work

The main goals of this thesis have been as follows:

- To study the different ways in which usability and the context of use have been defined in the literature, and to investigate the problems with this multiplicity of definitions.
- To propose comprehensive and general-purpose taxonomies of attributes for the concepts of usability and context of use, with the aim of synthesizing and refining the most prominent definitions for both terms.
- To present and discuss a taxonomy-based approach to the study of usability as an alternative to context-specific (or ad hoc) approaches.
- To illustrate how this taxonomy-based approach can be integrated with selected techniques for the study of usability in specific moments of the life cycle of a product.
- To document how this approach has been used in real-world international research projects.

The field of application has been Intelligent Systems, focusing specifically on the area of user interfaces. Nevertheless, the taxonomies presented in this thesis and the general methodology followed in it are meant to be applicable to any kind of system.

For its conceptualization of usability, this thesis has prioritized established interpretations of the concept, such as the ones proposed in standards like ISO 9241-11 [87] and ISO/IEC 9126-1 [90] and by prominent authors (such as Nielsen [129] or Seffah et al. [167]) over unconventional reinventions (of which there are many). Usability is ultimately a convenient construct that is defined by consensus among researchers and practitioners, and the reason for its existence is its practical usefulness.

However, researching the literature reveals that, so far, no real consensus has been reached among researchers or practitioners on the exact attributes that make up usability. Even though the different usability models that can be found in the literature tend to overlap significantly, there is also much divergence. Moreover, they tend to overlap partially and unevenly, with different terms used to designate the same usability attribute or with the same term used to describe different concepts. This ambiguity is also exacerbated by other facts: the usability models in the literature are often described in broad terms; they tend to include few attributes; and precise definitions are rarely provided for the attributes.

Similar criticisms could be made at the way the context of use is defined in the literature: there is also a lack of consensus among researchers and practitioners, and the existing models are typically ambiguous.

For the author of this thesis, this situation provided the motivation for constructing detailed taxonomies of attributes for the concepts of usability [16] and the context of use [17]. The goals for constructing the taxonomies were as follows:

- To make the taxonomies comprehensive, covering all the commonly-accepted attributes from the literature but avoiding contradictions and redundancy.
- To follow the usability and context-of-use models from the literature as closely as possible, prioritizing a synthesis and refinement from within, instead of a radical departure from the way they are commonly understood.
- To structure the taxonomies hierarchically into several levels of detail, from the general to the particular (typically, the usability models in the literature have only one level).
- To make the taxonomies general-purpose and applicable to any type of product (this contrasts strongly with traditional usability models, which are intentionally restricted to IT systems).
- To provide definitions for all the attributes and subattributes, which is something often neglected in the literature.

The usability taxonomy includes every usability aspect covered by the existing usability models except for the ISO/IEC 9126-1 [90] *usability compliance* attribute, whose inclusion was considered redundant. Similarly, all the first-level attributes in the usability taxonomy are, at least to some degree, included in two or more of the aforementioned usability models, and all the second-level attributes are included in at least one classification, except for *robustness to environment problems* and *environment safety*.

The *context-of-use* taxonomy is similar in that it expands on the attributes of existing context-of-use models while attempting to avoid redundancies. Regarding this, it should be noted that, compared to the models for usability, the context-of-use models in the literature are more complex but also more similar to each other at a superficial level, while the most significant differences are located at the lower levels of the respective hierarchies.

Shortly after the two taxonomies described in this thesis had been published [16][126][17], the International Organization for Standardization replaced its famous ISO 9241-11 and ISO/IEC 9126-1 standards with ISO FDIS 9241-210 [88] and ISO/IEC 25010 [89], respectively. While the old standards provided two definitions of usability that were significantly different from each other, the two new standards basically share the same definition, which can only be good news to those of us who had tried to raise awareness about those inconsistencies.

The taxonomies proposed in this thesis were integrated into the usability studies of two types of systems. Firstly, an Intelligent Speed Adaptation device [14]; secondly, the user interfaces generated by an automated multi-device UI generator [15][153][154]. This thesis has described in detail not only the methodology of these usability studies, but also the results. In fact, describing the results in detail is not very common in the literature, as the work conducted in usability engineering is typically kept confidential and companies are reluctant to expose the usability defects of their products and services.

The system under study in the first usability study is called CARAT counter, and both its development and its usability study are part of an international research project called Galileo Speed Warning [2], which proposes a reward-based alternative to traditional speed enforcement systems that are typically punitive. The CARAT counter monitors good driving and accumulates points if the driver keeps within specified speed limit thresholds. These points can be then traded for real-world rewards.

The system under study in the second usability study is a collection of automatically-generated UIs that are composed of HTML/CSS/Javascript web pages. The web pages are automatically produced by a multi-device UI generator [152] developed at the Institut für Computertechnik of Vienna's University of Technology. The generation process can be customized according to certain parameters and the resulting pages can be also modified manually.

The first step in both usability studies was to characterize the system. This involves breaking down the system into its constituent parts and identifying their usability-relevant attributes. As a result, this characterization provides a checklist of the aspects to be assessed through the usability taxonomy proposed in this thesis.

Even tough each system has different characteristics, some elements of the user interfaces and some tasks are commonplace and can be generalized to different systems. In fact, this thesis has made some steps towards constructing a general taxonomy of components for user interfaces, although this topic still needs more research.

The usability study of any system also needs to take into account the particular characteristics of its context of use, as usability is always relative to it. In the two usability studies mentioned above, the context of use was analyzed by following the context-of-use taxonomy presented in this thesis.

From a methodological point of view, the most novel aspect of both usability studies lies in how deeply the taxonomies are integrated with them. While usability studies typically start from some agreed-on model of usability, the usability models are typically simple, which is one of the reasons why studies tend to lean heavily on guidelines and ad hoc judgments in order to compensate for these limitations. The level of detail of the taxonomies presented in this thesis aims to facilitate the study of usability in this regard.

The taxonomies helped to identify the following types of usability problems in the CARAT counter: readability and consistency of the displayed elements, intuitiveness of the error messages and the user tasks, universality of the speed measurement units and the interface language, GPS precision, user safety, robustness to abuse and internal error, and aesthetic appeal.

Similarly, the heuristic analysis of the automatically-generated UIs helped to identify usability problems in areas like: clarity (of meaning, of functioning, of structure), consistency (of terminology, of look and feel), memorability, completeness, accessibility, cultural universality (of language, of date formats), flexibility, efficiency, robustness to improper use, confidentiality, and aesthetics.

Both usability studies are similarly structured in that, broadly, a heuristic evaluation is followed by user tests and questionnaires. There are some important differences, however. Firstly, whereas the CARAT counter consists of a very specific and purpose-built combination of hardware and software, the automatically-generated UIs were multi-platform web pages displayed on mass-produced smartphones or tablets. Secondly, the usability study of automatically-generated UIs was not focused on one interface, but it was instead a comparative study of different layouts for the same application. Thirdly, the usability study of automatically-generated UIs was also focused on a subset of usability attributes, as the UIs were identical in everything except layout. Fourthly, the latter usability study yielded not only qualitative data (e.g., the users' opinions) but also quantitative data, including statistically significant performance measurements obtained from annotating video recordings of 60 user tests.

This thesis has also investigated the validity of the approach taken. Note, however, that an empirical investigation of the construct validity of the taxonomies has been out of the scope of this thesis due to material limitations[1]. Instead, the usefulness of the taxonomies has been discussed in two ways. Firstly, by example, by showing their practical application to the above-mentioned real-world usability studies. Secondly, in theoretical terms, by investigating their validity in a way that is comparable – but more conceptually rigorous – to Winter et al.'s critique [186] of the ISO 15005 standard [86]. Whereas Winter et al. used their own comprehensive usability model to show the limitations of ISO 15005, this thesis has used the results of the usability study of the CARAT counter (which are derived from the taxonomies presented in this thesis) to analyze the limitations of the opposite approach to the one taken in this thesis, that is, an ad hoc and non-comprehensive usability model tailored for an specific field of application. Obviously, such a model does not necessarily exist for any field of application, but one [72] happened to exist for the field

[1] As an example, researchers like Hornbæk and Law [80] and Sauro and Lewis [164] have used data collected in over 70 usability studies in order to investigate the construct validity of the ISO 9241-11 usability model [87].

of in-vehicle information systems. The limitations found in this particular ad hoc model are in fact symptomatic of a certain lack of clarity and completeness that is typical of the ad hoc approach. Extrapolating from this, the problems of the ad hoc approach have been discussed in terms of *construct overload*, *construct excess*, *construct redundancy*, and *construct incompleteness* [184][173][157][22]. Furthermore, ad hoc usability models that are tailored for a specific field of application lack generalizability to other areas, which is usually a desirable quality in engineering.

The usability studies in this thesis are focused on *formative* usability goals, that is, detecting usability problems and proposing solutions. While the comprehensiveness of the taxonomies have facilitated the completion of these formative tasks, the suitability of the usability taxonomy for the other type of usability goals, that is, *summative* goals, has been mostly out of the scope of this thesis. The purpose of summative goals is to obtain general (as opposed to specific) assessments of usability. For example, a single usability score calculated from the rating given to the individual usability attributes. This is a complex topic that needs more research, as there is not even a consensus among researchers on whether the three ISO 9241-11 attributes of *effectiveness*, *efficiency*, and *satisfaction* can be meaningfully aggregated into a single usability (summative) score. Moreover, there is also the problem of which weights to assign to the individual attributes, which is also an important topic among researchers.

The comprehensiveness of the taxonomies has been particularly useful for systematic, thorough usability techniques like heuristic evaluation (which was done for the two usability studies) and user questionnaires that address the whole spectrum of usability (which was done for the usability study of the CARAT counter). However, the usability taxonomy is perhaps less suited to user questionnaires that are restricted to a subset of usability aspects (like the one for the usability study of the automatically-generated UIs). Even tough the taxonomy is structured into several levels of detail, and this allows to focus on specific subattributes, some particular combinations of subattributes can easily lead to custom-made questionnaires, and these kinds of ad hoc activities are precisely what this thesis endeavored to avoid. This problem is sometimes inevitable, and an alternative option in these situations is to simply use a standardized usability questionnaire from the literature, although those are not beyond criticism, either.

As can be seen, this thesis has touched upon complex questions such as: Is it possible to aggregate the attributes in a usability model into a single usability score or should they be measured independently? How should weights be assigned to the different attributes? When is it preferable to use standardized usability questionnaires instead of custom-made questionnaires? These are fundamental questions of the usability literature that remain open for future research. Another topic worth investigating is the integration of the taxonomies with metrics and guidelines, in order to make the process more comprehensive and systematic.

This thesis has also provided the foundations for the construction of a third taxonomy that could be integrated into the general methodology proposed in this thesis – namely, a taxonomy of components for user interfaces. As the usability study of the automatically-generated UIs has been concerned with HTML interfaces, and

HTML provides a standardized and comprehensive set of interface components, the characterization of the system done for that usability study can be easily generalized to any type of user interface.

As this thesis belongs to the discipline of engineering, the formalization and systematization of all these things – usability, context of use, UI components – have not been simply done for their own sake. Rather, they are prerequisites for a specific goal that is proposed as future work: the construction of an decision support system for usability studies.

Appendix A

Relevant Publications by the Author of this Work

A.1 IJHCI-09

Reference

- **Usability: A Critical Analysis and a Taxonomy**. D. Alonso-Ríos, A. Vázquez-García, E. Mosqueira-Rey, V. Moret-Bonillo. International *Journal of Human-Computer Interaction* Vol. 26(1), 2009

Abstract

A major obstacle to the implantation of User-Centered Design in the real world is the fact that no precise definition of the concept of usability exists that is widely accepted and applied in practice. Generally speaking, the literature tends to define usability in overly brief and ambiguous terms and to describe its application in informal terms. This is one of the main reasons why ad hoc techniques predominate in usability study methodologies. The aims of this article are to investigate the concept of usability and to describe it by means of a detailed taxonomy that is organized hierarchically and that contains exhaustive descriptions of usability attributes. This taxonomy can be used to support different stages in the development of usable systems.

A.2 SMC-09

Reference

- **Usability Taxonomy and Context-of-Use Taxonomy for Usability Analysis**. E. Mosqueira-Rey, D. Alonso-Ríos, V. Moret-Bonillo. International *Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 812-817, 2009

Abstract

The interest in developing usable products has led to the inclusion of usability aspects in product development processes. Generally speaking, the literature tends to define usability in overly brief and ambiguous terms and to describe its application in informal terms. Also there is a tendency to overlook characteristics of the context in which a product is to be used, and this usually means that the usability of a product in its operational environment is often diminished. For these reasons we propose in this work a detailed taxonomy which contains exhaustive descriptions of usability attributes, and a comprehensive taxonomy that describes context of use and its attributes by means of precise definitions.

A.3 IJHCI-10

Reference

- **A Context-of-Use Taxonomy for Usability Studies**. D. Alonso-Ríos, A. Vázquez-García, E. Mosqueira-Rey, V. Moret-Bonillo. *International Journal of Human-Computer Interaction* Vol. 26(10), 2010

Abstract

The interest in developing usable products has led to the inclusion of usability aspects in product development processes. Nonetheless, the fact that there is a tendency to overlook characteristics of the context in which a product is to be used means that the usability of a product in its operational environment is often diminished. One of the main reasons why this is the case is because there is no clear and sufficiently detailed model available for defining the concept of context of use. A comprehensive taxonomy that describes context of use and its attributes by means of precise definitions is proposed. This taxonomy will serve as a basis for improving the validity of usability activities by enabling an analysis of the conditions of use of a product in usability studies in a structured way.

A.4 IJHCI-14

Reference

- **A Taxonomy-based Usability Study of an Intelligent Speed Adaptation Device**. D. Alonso-Ríos, E. Mosqueira-Rey, V. Moret-Bonillo. *International Journal of Human-Computer Interaction* Vol. 30(7), pp. 585–603, 2014

Abstract

Usability studies are often based on ad hoc definitions of usability. These studies can be difficult to generalize, they might have a steep learning curve, and there is always the danger of being inconsistent with the concept of usability as defined in standards and the literature. This alternative approach involves comprehensive, general-purpose, and hierarchically structured taxonomies that follow closely the main usability literature. These taxonomies are then instantiated for a specific product. To illustrate this approach, a usability study for a prototype of an Intelligent Speed Adaptation device is described. The usability study consists of usability requirements analysis, heuristic evaluation, and subjective analysis, which helped identify problems of clarity, operability, robustness, safety, and aesthetics. As a context-specific usability taxonomy for this particular field of application happened to exist, the way that real-world usability results can be mapped to that taxonomy compared to the taxonomy in this article is examined, with the argument that this study's taxonomy is more complete and generalizable.

A.5 SMC-13

Reference

- **A User Study with GUIs Tailored for Smartphones and Tablet PCs**. D. Raneburger, R. Popp, D. Alonso-Ríos, H. Kaindl, J. Falb. *Proceedings of the 2013 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 3727-3732, 2013

Abstract

Usually, Web-based graphical user interfaces (GUIs) are not specifically tailored for different devices with touchscreens, such as smartphones and tablet PCs, where interaction is affected mainly by screen size. There is little scientific evidence on the conditions under which additional taps for navigation are better than scrolling or vice versa. Therefore, we conducted a user study in which we experimentally evaluated GUIs tailored for a smartphone and a tablet PC, respectively. Each participant performed the same task with two different layouts of the same GUI, either on a given smartphone or tablet PC. We collected quantitative data through measuring task completion time and error rates, as well as qualitative data through subjective questionnaires. The main result is that tailoring a GUI specifically for a smartphone or tablet PC, respectively, is important, since screen size matters. Users performed significantly better when they could use the tailored version on the given device. This preference was also reflected in their subjective opinions.

A.6 SAC-14

Reference

- **A User Study on Tailoring GUIs for Smartphones**. D. Alonso-Ríos, D. Raneburger, R. Popp, H. Kaindl, J. Falb. *Proceedings of the 29th ACM Symposium on Applied Computing (SAC)*, pp. 186-192, 2014

Abstract

Usually, Web-based graphical user interfaces (GUIs) are not specifically tailored for different devices with touchscreens, such as smartphones, where interaction is affected mainly by screen size. There is little scientific evidence on the benefits of tailoring in general, and in particular on the conditions where scrolling is good or bad. Therefore, we conducted a user study in which we experimentally evaluated a GUI tailored for a smartphone and another non-tailored one. The tailoring in this case only rearranges widgets in a way that the width of the device screen is sufficient but vertical scrolling may be necessary. Each participant performed the same task with these two different layouts. We collected quantitative data through measuring task completion time and error rates, as well as qualitative data through subjective questionnaires and interviews. The main result is that tailoring a GUI for a smartphone is important, since task performance time was significantly shorter when using a tailored GUI requiring only vertical scrolling as compared to a non-tailored one. This preference was also reflected in the subjective opinions of the users.

A.7 INTERACT-13

Reference

- **A User Study with GUIs Tailored for Smartphones**. D. Raneburger, D. Alonso-Ríos, R. Popp, H. Kaindl, J. Falb. *Human-Computer Interaction – INTERACT 2013. Lecture Notes in Computer Science* Vol. 8118, pp. 505-512, 2013

Abstract

Web-based graphical user interfaces (GUIs) are mostly not tailored for small devices with touchscreens, such as smartphones. There is little scientific evidence on the conditions where additional taps for navigation are better or scrolling. Therefore, we conducted a user study in which we evaluated different ways of tailoring a GUI for a smartphone. Each participant performed the same task with two different layouts of the same GUI. We collected quantitative data through measuring task completion time and error rates, as well as qualitative data through subjective questionnaires. The main result is that minimizing the number of taps is important on a smartphone. Users performed significantly better when they could scroll (vertically), instead of tapping on widget elements (tabs). This preference was also reflected in their subjective opinions.

Bibliography

[1] CARATs for safe drivers. Pinpoint: The Newsletter of the Location and Timing Knowledge Transfer Network. Summer 2009.

[2] Galileo Speed Warning (GSW). european GNSS agency. http://www.gsa.europa.eu/galileo-speed-warning-gsw-0. Accessed: 2014-06-01.

[3] iOS human interface guidelines. https://developer.apple.com/library/ios/documentation/userexperience/conceptual/MobileHIG/MobileHIG.pdf. Accessed: 2014-06-01.

[4] Nigel Bevan - Biography. http://www.nigelbevan.com/about.htm. Accessed: 2014-06-01.

[5] OS X human interface guidelines. https://developer.apple.com/library/mac/documentation/UserExperience/Conceptual/AppleHIGuidelines/OSXHIGuidelines.pdf. Accessed: 2014-06-01.

[6] Phenomenology. Stanford encyclopedia of philosophy. http://plato.stanford.edu/entries/phenomenology/. Accessed: 2014-06-01.

[7] Windows user experience interaction guidelines. http://www.microsoft.com/en-us/download/details.aspx?id=2695. Accessed: 2014-06-01.

[8] Council directive 90/270/EEC of 29 May 1990 on the minimum safety and health requirements for work with display screen equipment. http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:31990L0270, May 1990. Accessed: 2014-06-01.

[9] Speed management. Technical report, Organisation for Economic Co-operation and Development (OECD), 1996.

[10] Section 508 standards for electronic and information technology. http://www.access-board.gov/guidelines-and-standards/communications-and-it/about-the-section-508-standards/section-508-standards, December 2000. Accessed: 2014-06-01.

[11] A. Abran, A. Khelifi, and W. Suryn. Usability meanings and interpretations in ISO standards. *Software Quality Journal*, 11:325–338, 2003.

[12] A. Abran, A. Khelifi, W. Suryn, and A. Seffah. Consolidating the ISO usability models. In *Proceedings of 11th international software quality management conference*, pages 23–25. Citeseer, 2003.

[13] L. Adelman and S. L. Riedel. *Handbook for evaluating knowledge-based systems: Conceptual framework and compendium of methods.* Kluwer Academic Publishers, 1997.

[14] D. Alonso-Ríos, E. Mosqueira-Rey, and V. Moret-Bonillo. A taxonomy-based usability study of an Intelligent Speed Adaptation device. *International Journal of Human-Computer Interaction*, 30(7):585–603, 2014.

[15] D. Alonso-Ríos, D. Raneburger, R. Popp, H. Kaindl, and J. Falb. A user study on tailoring GUIs for smartphones. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, pages 186–192. ACM, 2014.

[16] D. Alonso-Ríos, A. Vázquez-García, E. Mosqueira-Rey, and V. Moret-Bonillo. Usability: A critical analysis and a taxonomy. *International Journal of Human-Computer Interaction*, 26(1):53–74, 2009.

[17] D. Alonso-Ríos, A. Vázquez-García, E. Mosqueira-Rey, and V. Moret-Bonillo. A context-of-use taxonomy for usability studies. *International Journal of Human-Computer Interaction*, 26(10):941–970, 2010.

[18] R. Anderson, A. McLean, M. Farmer, B. Lee, and C. Brooks. Vehicle travel speeds and the incidence of fatal pedestrian crashes. In *International IRCOBI Conference on the Biomechanics of Impacts, 1995, Brunnen, Switzerland*, 1995.

[19] J. A. M. Association et al. Guideline for in-vehicle display systems, version 3.0. *Retrieved November*, 14:2008, 2004.

[20] J. R. Baker. The impact of paging vs. scrolling on reading online text passages. *Usability News*, 5(1), 2003.

[21] F. Balagtas-Fernandez, J. Forrai, and H. Hussmann. Evaluation of user interface design and input methods for applications on mobile touch screen devices. In *Human-Computer Interaction–INTERACT 2009*, pages 243–246. Springer, 2009.

[22] P. Barcelos, G. Guizzardi, A. Garcia, and M. E. Monteiro. Ontological evaluation of the ITU-T recommendation G. 805. In *Telecommunications (ICT), 2011 18th International Conference on*, pages 232–237. IEEE, 2011.

[23] J. C. Bastien and D. L. Scapin. Evaluating a user interface with ergonomic criteria. *International Journal of Human-Computer Interaction*, 7(2):105–121, 1995.

[24] N. Bevan. The MUSIC methodology for usability measurement. In *Posters and short talks of the 1992 SIGCHI conference on Human factors in computing systems*, pages 123–124. ACM, 1992.

[25] N. Bevan. Measuring usability as quality of use. *Software Quality Journal*, 4(2):115–130, 1995.

[26] N. Bevan. Usability is quality of use. *Advances in Human Factors Ergonomics*, 20:349–349, 1995.

[27] N. Bevan. Quality in use: Meeting user needs for quality. *Journal of Systems and Software*, 49(1):89–96, 1999.

[28] N. Bevan. Quality in use for all. *User Interfaces for All, Concepts, Methods and Tools. London: Lawrence Erlbaum Publications*, pages 352–68, 2001.

[29] N. Bevan. *Extending quality in use to provide a framework for usability measurement*. Springer, 2009.

[30] N. Bevan, J. Kirakowski, and J. Maissel. What is usability? In *Proceedings of the 4th International Conference on HCI*, pages 651—655, 1991.

[31] N. Bevan and M. Macleod. Usability measurement in context. *Behaviour & Information Technology*, 13(1-2):132–145, 1994.

[32] T. Biding and G. Lind. Intelligent Speed Adaptation (ISA). results of large-scale trials in Borlaenge, Linkoeping, Lund and Umeaa during the period 1999-2002. *VAEGVERKET. PUBLIKATION*, (2002: 89E), 2002.

[33] A. F. Blackwell, C. Britton, A. Cox, T. R. Green, C. Gurr, G. Kadoda, M. Kutar, M. Loomes, C. L. Nehaniv, M. Petre, et al. Cognitive dimensions of notations: Design tools for cognitive technology. In *Cognitive Technology: Instruments of Mind*, pages 325–341. Springer, 2001.

[34] S. Borsci, R. D. Macredie, J. Barnett, J. Martin, J. Kuljis, and T. Young. Reviewing and extending the five-user assumption: A grounded procedure for interaction evaluation. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 20(5):29, 2013.

[35] B. Bos, T. Çelik, I. Hickson, and H. W. Lie. Cascading Style Sheets level 2 revision 1 (CSS 2.1) specification. *W3C working draft, W3C, June*, 2005.

[36] A. Buccella, A. Cechich, M. Pol, M. Arias, M. del Socorro Doldan, E. Morsan, et al. Marine ecology service reuse through taxonomy-oriented SPL development. *Computers & Geosciences*, 73:108–121, 2014.

[37] G. Buchanan, S. Farrant, M. Jones, H. Thimbleby, G. Marsden, and M. Pazzani. Improving mobile Internet usability. In *Proceedings of the 10th international conference on World Wide Web*, pages 673–680. ACM, 2001.

[38] M. Bunge. *Treatise on Basic Philosophy: Ontology I: The Furniture of the World*. Springer, 1977.

[39] M. Bunge. *Treatise on Basic Philosophy: Ontology II: A World of Systems*. Reidel Publishing Company, 1979.

[40] G. Burnett. Usable vehicle navigation systems: Are we there yet? In *European Conference on Vehicle Electronic Systems (2000: Stratford-upon-Avon, England). Vehicle electronic systems*, 2000.

[41] M. D. Byrne, B. E. John, N. S. Wehrle, and D. C. Crow. The tangled web we wove: a taskonomy of WWW use. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 544–551. ACM, 1999.

[42] G. Calvary, J. Coutaz, L. Bouillon, M. Florins, Q. Limbourg, L. Marucci, F. Paternò, C. Santoro, N. Souchon, D. Thevenin, et al. The CAMELEON reference framework. deliverable 1.1, CAMELEON project, 2002.

[43] G. Calvary, J. Coutaz, D. Thevenin, Q. Limbourg, L. Bouillon, and J. Vanderdonckt. A unifying reference framework for multi-target user interfaces. *Interacting with Computers*, 15(3):289–308, 2003.

[44] J. M. Cantera Fonseca, J. M. González Calleros, G. Meixner, F. Paternò, J. Pullmann, D. Raggett, D. Schwabe, and J. Vanderdonckt. Model-based UI XG final report. `http://www.w3.org/2005/Incubator/model-based-ui/XGR-mbui-20100504/`, May 2010. Accessed: 2015-01-07.

[45] O. Carsten and S. Comte. UK work on automatic speed control. In *Proceedings of the ICTCT 97 conference*, pages 5–7, 1997.

[46] O. Carsten, M. Fowkes, F. Lai, K. Chorlton, S. Jamson, F. Tate, and B. Simpkin. ISA UK-Intelligent Speed Adaptation-final report. 2008.

[47] A. Chandor, J. Graham, and R. Williamson. *The Penguin dictionary of computers*. Penguin Books, 1985.

[48] D. Churchill and J. Hedberg. Learning object design considerations for small-screen handheld devices. *Computers & Education*, 50(3):881–893, 2008.

[49] S. Conger. Software development life cycles and methodologies: Fixing the old and adopting the new. 2010.

[50] C. K. Coursaris and D. J. Kim. A research agenda for mobile usability. In *Proceedings of the 2007 Conference on Human Factors in Computing Systems (CHI 2007)*, pages 2345—-2350, 2007.

[51] A. K. Dey. Understanding and using context. *Personal and ubiquitous computing*, 5(1):4–7, 2001.

[52] A. Dillon and M. Song. An empirical comparison of the usability for novice and expert searchers of a textual and a graphic interface to an art-resource database. 1997.

[53] T. Dingus, M. Hulse, S. Jahns, J. Alves-Foss, S. Confer, A. Rice, I. Roberts, R. Hanowski, and D. Sorenson. Development of human factors guidelines for advanced traveler information systems and commercial vehicle operations: Literature review supplement. Technical report, 1996.

[54] S. K. Dubey, A. Gulati, and A. Rana. Integrated model for software usability. *International Journal on Computer Science & Engineering*, 4(3), 2012.

[55] J. Dumas. The great leap forward: The birth of the usability profession (1988-1993). *Journal of Usability Studies*, 2(2):54–60, 2007.

[56] S. Edelman. *Representation and recognition in vision.* MIT press, 1999.

[57] P. Estrella, A. Popescu-Belis, and N. Underwood. Finding the system that suits you best: towards the normalization of MT evaluation. In *27th International Conference on Translating and the Computer (ASLIB)*, page 23–34, 2005.

[58] X. Fang, S. Xu, J. Brzezinski, and S. S. Chan. A study of the feasibility and effectiveness of dual-modal information presentations. *International journal of human-computer interaction*, 20(1):3–17, 2006.

[59] B. Fernandez-Saavedra, R. Alonso-Moreno, J. Uriarte-Antonio, and R. Sanchez-Reillo. Evaluation methodology for analyzing usability factors in biometrics. *Aerospace and Electronic Systems Magazine, IEEE*, 25(8):20–31, 2010.

[60] D. Finch, P. Kompfner, C. Lockwood, and G. Maycock. Speed, speed limits and accidents. *TRL Project Report*, (PR 58), 1994.

[61] J. M. Flach, P. A. Hancock, J. E. Caird, and K. J. Vicente. *Global perspectives on the ecology of human–machine systems, Vol. 1.* Lawrence Erlbaum Associates, Inc, 1995.

[62] E. Folmer and J. Bosch. Architecting for usability: a survey. *Journal of systems and software*, 70(1):61–78, 2004.

[63] E. Frøkjær, M. Hertzum, and K. Hornbæk. Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 345–352. ACM, 2000.

[64] J. D. Gould and C. Lewis. Designing for usability: key principles and what designers think. *Communications of the ACM*, 28(3):300–311, 1985.

[65] W. D. Gray and M. C. Salzman. Damaged merchandise? a review of experiments that compare usability evaluation methods. *Human–Computer Interaction*, 13(3):203–261, 1998.

[66] P. Green, W. Levison, G. Paelke, and C. Serafin. Suggested human factors design guidelines for driver information systems. *University of Michigan Transportation Research Institute Technical Report UMTRI-93-21*, 1993.

[67] J. Grudin et al. Adapting a psychophysical method to measure performance and preference tradeoffs in human-computer interaction. In *Human-Computer Interaction-INTERACT'84*. Citeseer, 1985.

[68] D. Gupta, A. Ahlawat, and K. Sagar. A critical analysis of a hierarchy based usability model. In *Contemporary Computing and Informatics (IC3I), 2014 International Conference on*, pages 255–260. IEEE, 2014.

[69] S. H. Han, M. Hwan Yun, K.-J. Kim, and J. Kwahk. Evaluation of product usability: development and validation of usability dimensions and design elements based on empirical models. *International Journal of Industrial Ergonomics*, 26(4):477–488, 2000.

[70] I. Harms and W. Schweibenz. Usability engineering methods for the web: results from a usability study. 2007.

[71] V. H. Hartkopf, V. E. Loftness, and P. A. D. Mill. The concept of total building performance and building diagnostics. In G. Davis, editor, *Building performance: Function, preservation, and rehabilitation*, pages 5–22. American Society for Testing and Materials, 1986.

[72] C. Harvey, N. A. Stanton, C. A. Pickering, M. McDonald, and P. Zheng. Context of use as a factor in determining the usability of in-vehicle devices. *Theoretical issues in ergonomics science*, 12(4):318–338, 2011.

[73] C. Harvey, N. A. Stanton, C. A. Pickering, M. McDonald, and P. Zheng. In-vehicle information systems to meet the needs of drivers. *Intl. Journal of Human–Computer Interaction*, 27(6):505–522, 2011.

[74] C. Harvey, N. A. Stanton, C. A. Pickering, M. McDonald, and P. Zheng. A usability evaluation toolkit for in-vehicle information systems (IVISs). *Applied ergonomics*, 42(4):563–574, 2011.

[75] M. Heidegger. Being and time. 1927. *Trans. John Macquarrie and Edward Robinson. New York: Harper*, 1962.

[76] T. Hollingsed and D. G. Novick. Usability inspection methods after 15 years of research and practice. In *Proceedings of the 25th annual ACM international conference on Design of communication*, pages 249–255. ACM, 2007.

[77] A. Holzinger. Usability engineering methods for software developers. *Communications of the ACM*, 48(1):71–74, 2005.

[78] K. Hornbæk. Current practice in measuring usability: Challenges to usability studies and research. *International journal of human-computer studies*, 64(2):79–102, 2006.

[79] K. Hornbæk and E. Frøkjær. Comparing usability problems and redesign proposals as input to practical systems development. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 391–400. ACM, 2005.

[80] K. Hornbæk and E. L.-C. Law. Meta-analysis of correlations among usability measures. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 617–626. ACM, 2007.

[81] G. Hu and K. H. Chang. A methodology for structured use-centered quantitative full-life-cycle usability requirements specification and usability evaluation of web sites. In *International Conference on Internet Computing*, pages 133–140, 2006.

[82] E. Hull, K. Jackson, and J. Dick. *Requirements engineering*. Springer, 2005.

[83] L. Hull. Accessibility: it's not just for disabilities any more. *interactions*, 11(2):36–41, 2004.

[84] E. T. Hvannberg, E. L.-C. Law, and M. K. Lérusdóttir. Heuristic evaluation: Comparing ways of finding and reporting usability problems. *Interacting with computers*, 19(2):225–240, 2007.

[85] ISO 13407:1999 - Human-centred design processes for interactive systems, 1999.

[86] ISO 15005:2002 - Road vehicles – Ergonomic aspects of transport information and control systems – Dialogue management principles and compliance procedures, 2002.

[87] ISO 9241-11:1998 - Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability, 1998.

[88] ISO 9241-210:2010 - Ergonomics of human-system interaction – Part 210: Human-centred design for interactive systems. (ISO 9241-210), 2010.

[89] ISO/IEC 25010:2011 - Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models, 2011.

[90] ISO/IEC 9126-1:2001 - Software engineering – Product quality – Part 1: Quality model, 2001.

[91] M. Y. Ivory. Web TANGO: towards automated comparison of information-centric web site designs. In *CHI'00 Extended Abstracts on Human Factors in Computing Systems*, pages 329–330. ACM, 2000.

[92] M. Y. Ivory and M. A. Hearst. The state of the art in automating usability evaluation of user interfaces. *ACM Computing Surveys (CSUR)*, 33(4):470–516, 2001.

[93] M. Jones, G. Marsden, N. Mohd-Nasir, K. Boone, and G. Buchanan. Improving web interaction on small displays. *Computer Networks*, 31(11):1129–1137, 1999.

[94] S. Jones, M. Jones, G. Marsden, D. Patel, and A. Cockburn. An evaluation of integrated zooming and scrolling on small screens. *International Journal of Human-Computer Studies*, 63(3):271–303, 2005.

[95] S. Joo and J. Y. Lee. Measuring the usability of academic digital libraries: Instrument development and validation. *Electronic Library, The*, 29(4):523–537, 2011.

[96] P. W. Jordan. *An introduction to usability.* CRC Press, 1998.

[97] A. Kaikkonen and V. Roto. Navigating in a mobile XHTML application. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 329–336. ACM, 2003.

[98] C.-M. Karat. Usability engineering in dollars and cents. *Software, IEEE*, 10(3):88–89, 1993.

[99] B. Keevil. Measuring the usability index of your web site. In *Proceedings of the 16th annual international conference on Computer documentation*, pages 271–277. ACM, 1998.

[100] L. Kim and M. J. Albers. Web design issues when searching for information in a small screen display. In *Proceedings of the 19th annual international conference on Computer documentation*, pages 193–200. ACM, 2001.

[101] M. Kipp. Multimedia annotation, querying and analysis in ANVIL. *Multimedia information extraction*, 19, 2010.

[102] J. Kirakowski. The software usability measurement inventory: background and usage. *Usability evaluation in industry*, pages 169–178, 1996.

[103] J. Kirakowski and B. Cierlik. Context of use: introductory notes. `http://www.ucc.ie/hfrg/baseline/filearchive.html`, 1999.

[104] J. Kirakowski and M. Corbett. SUMI: The software usability measurement inventory. *British journal of educational technology*, 24(3):210–212, 1993.

[105] G. E. Krasner, S. T. Pope, et al. A description of the Model-View-Controller user interface paradigm in the Smalltalk-80 system. *Journal of object oriented programming*, 1(3):26–49, 1988.

[106] A. Kullgren, H. Stigson, F. Achterberg, E. Townsend, and J. Crettaz. In-car enforcement technologies today. *European Transport Safety Council*, 2005.

[107] J. Kwahk and S. H. Han. A methodology for evaluating the usability of audiovisual consumer electronic products. *Applied Ergonomics*, 33:419–431, 2002.

[108] H. Lahrmann, N. Agerholm, N. Tradisauskas, J. Juhl, and L. Harms. Intelligent Speed Adaption based on pay as you drive principles. In *ITS for a Better Life-14 th World Congress on Intelligent Transport Systems*, 2007.

[109] A. Lasch and T. Kujala. Designing browsing for in-car music player: effects of touch screen scrolling techniques, items per page and screen orientation on driver distraction. In *Proceedings of the 4th International Conference on*

Automotive User Interfaces and Interactive Vehicular Applications, pages 41–48. ACM, 2012.

[110] J. R. Lewis. A rank-based method for the usability comparison of competing products. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 35, pages 1312–1316. SAGE Publications, 1991.

[111] J. R. Lewis. Usability: Lessons learned… and yet to be learned. *International Journal of Human-Computer Interaction*, 30(9):663–684, 2014.

[112] G. Lindgaard. Notions of thoroughness, efficiency, and validity: Are they valid in HCI practice? *International journal of industrial ergonomics*, 36(12):1069–1074, 2006.

[113] R. Liu and J. Tate. Network effects of Intelligent Speed Adaptation systems. *Transportation*, 31(3):297–325, 2004.

[114] Y.-C. Liu, C. S. Schreiner, and T. A. Dingus. Development of human factors guidelines for advanced traveler information systems (ATIS) and commercial vehicle operations (CVO): Human factors evaluation of the effectiveness of multi-modality displays in advanced traveler information systems. Technical report, 1999.

[115] A. M. Lund. Measuring usability with the USE questionnaire. *Usability interface*, 8(2):3–6, 2001.

[116] K. R. Lynch, D. J. Schwerha, and G. A. Johanson. Development of a weighted heuristic for website evaluation for older adults. *International Journal of Human-Computer Interaction*, 29(6):404–418, 2013.

[117] L. W. MacDonald. Using color effectively in computer graphics. *Computer Graphics and Applications, IEEE*, 19(4):20–35, 1999.

[118] R. L. Mack, C. H. Lewis, and J. M. Carroll. Learning to use word processors: problems and prospects. *ACM Transactions on Information Systems (TOIS)*, 1(3):254–271, 1983.

[119] A. MacLean, P. Barnard, and M. Wilson. Evaluating the human interface of a data entry system: User choice and performance measures yield different tradeoff functions. *People and Computers: Designing the Interface. Cambridge: Cambridge University*, 1985.

[120] M. Macleod. Usability: practical methods for testing and improvement. In *Proceedings of the Norwegian Computer Society Software Conference, Sandvika, Norway. Retrieved July*, volume 3, page 2005, 1994.

[121] M. Maguire. Context of use within usability activities. *International Journal of Human-Computer Studies*, 55(4):453–483, 2001.

[122] J. Maissel, A. Dillon, M. Maguire, R. Rengger, and M. Sweeney. *Context Guidelines Handbook*. National Physical Laboratory, Teddington, UK, 1991.

[123] M. M. Mantei and T. J. Teorey. Cost/benefit analysis for incorporating human factors in the software lifecycle. *Communications of the ACM*, 31(4):428–439, 1988.

[124] J. Martin. After a 4 year trial: what the swedes think of ISA. *Traffic engineering & control*, 43(10):376–379, 2002.

[125] J. T. S. Moorthy. Identification of usability risk as a risk in software development projects. *International Journal of Software Engineering and Technology*, 1(2), 2014.

[126] E. Mosqueira-Rey, D. Alonso-Ríos, and V. Moret-Bonillo. Usability taxonomy and context-of-use taxonomy for usability analysis. In *Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on*, pages 812–817. IEEE, 2009.

[127] R. S. Mullins. *Interpretive uncertainty and the evaluation of symbols and a taxonomy of symbol evaluation methods and mobile evaluation tool.* PhD thesis, The Pennsylvania State University, 2014.

[128] J. Nielsen. Finding usability problems through heuristic evaluation. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 373–380. ACM, 1992.

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

[130] J. Nielsen. Usability inspection methods. In *Conference companion on Human factors in computing systems*, pages 413–414. ACM, 1994.

[131] J. Nielsen. *Designing for the Web.* New Riders Publishing, 2000.

[132] J. Nielsen. How many test users in a usability study? `http://www.nngroup.com/articles/how-many-test-users/`, June 2012. Accessed: 2014-06-01.

[133] J. Nielsen and T. K. Landauer. A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pages 206–213. ACM, 1993.

[134] J. Nielsen and J. Levy. Measuring usability: preference vs. performance. *Communications of the ACM*, 37(4):66–75, 1994.

[135] J. Nielsen and H. Loranger. *Prioritizing web usability.* Pearson Education, 2006.

[136] J. Nielsen and R. Molich. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 249–256. ACM, 1990.

[137] D. A. Norman. Design principles for human-computer interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 1–10. ACM, 1983.

[138] C. Nowakowski, P. Green, and O. Tsimhoni. Common automotive navigation system usability problems and a standard test protocol to identify them. In *ITS-America 2003 Annual Meeting*, 2003.

[139] W. Oyomno, P. Jäppinen, E. Kerttula, and K. Heikkinen. Usability study of ME2.0. *Personal and Ubiquitous Computing*, 17(2):305–319, 2013.

[140] M. Paine, D. Paine, M. Griffiths, and G. Germanos. In-vehicle Intelligent Speed advisory systems. *Proceedings of 20th ESV, http://tinyurl. com/367gpf*, 2007.

[141] J. Pascoe. Adding generic contextual capabilities to wearable computers. In *Wearable Computers, 1998. Digest of Papers. Second International Symposium on*, pages 92–99. IEEE, 1998.

[142] A. Peytchev, M. P. Couper, S. E. McCabe, and S. D. Crawford. Web survey design paging versus scrolling. *Public Opinion Quarterly*, 70(4):596–607, 2006.

[143] R. Popp, D. Raneburger, and H. Kaindl. Tool support for automated multi-device GUI generation from discourse-based communication models. In *Proceedings of the 5th ACM SIGCHI symposium on Engineering interactive computing systems*, pages 145–150. ACM, 2013.

[144] M. Porteous, J. Kirakowski, and M. Corbett. SUMI user handbook. *Human Factors Research Group, University College Cork, Ireland*, 1993.

[145] J. Preece, D. Benyon, et al. *A guide to usability: Human factors in computing.* Addison-Wesley Longman Publishing Co., Inc., 1993.

[146] J. Preece, Y. Rogers, H. Sharp, D. Benyon, S. Holland, and T. Carey. *Human-computer interaction.* Addison-Wesley, 1994.

[147] W. Quesenbery. What does usability mean: Looking beyond “ease of use”. In *Proceedings of the 18th Annual Conference Society for Technical Communications*, 2001.

[148] W. Quesenbery. Dimensions of usabilility: Opening the conversation, driving the process. In *Proceedings of the UPA 2003 Conference*, volume 2003, 2003.

[149] W. Quesenbery. Balancing the 5Es: Usability. *Cutter IT Journal*, 17(2):4–11, 2004.

[150] J. Radatz, A. Geraci, and F. Katki. IEEE standard glossary of software engineering terminology. *IEEE Std*, 1990.

[151] D. Raggett, A. Le Hors, I. Jacobs, et al. HTML 4.01 specification. *W3C recommendation*, 24, 1999.

[152] D. Raneburger. *Interactive Model-Driven Generation of Graphical User Interfaces for Multiple Devices.* PhD thesis, Institute of Computer Technology, Vienna University of Technology, 2014.

[153] D. Raneburger, D. Alonso-Ríos, R. Popp, H. Kaindl, and J. Falb. A user study with GUIs tailored for smartphones. In *Human-Computer Interaction–INTERACT 2013*, pages 505–512. Springer, 2013.

[154] D. Raneburger, R. Popp, D. Alonso-Ríos, H. Kaind, and J. Falb. A user study with GUIs tailored for smartphones and tablet PCs. In *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*, pages 3727–3732. IEEE, 2013.

[155] J. Rasmussen. Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *Systems, Man and Cybernetics, IEEE Transactions on*, (3):257–266, 1983.

[156] A. Raza, L. F. Capretz, and F. Ahmed. Users' perception of open source usability: an empirical study. *Engineering with Computers*, 28(2):109–121, 2012.

[157] J. Recker, M. Rosemann, P. F. Green, and M. Indulska. Do ontological deficiencies in modeling grammars matter? *MIS Quarterly*, 35(1):57–79, 2011.

[158] R. M. Reffat and E. L. Harkness. Environmental comfort criteria: weighting and integration. *Journal of Performance of Constructed Facilities*, 15:104—-108, 2001.

[159] M. Regan, K. Young, D. Healy, P. Tierney, and K. Connelly. Evaluating in-vehicle intelligent transport systems: a case study. In *Proceedings of Road Safety Research, Policing and Education Conference, Adelaide*, 2002.

[160] K. Riemer and R. B. Johnston. What is IT in use and why does it matter for IS design? *Systems, Signs & Actions*, 7(1):5–21, 2013.

[161] K. Riemer and N. Vehring. It's not a property! exploring the sociomateriality of software usability. 2010.

[162] C. Rohrer. When to use which user–experience research methods. `http://www.nngroup.com/articles/which-ux-research-methods`, 2014. Accessed: 2015-01-07.

[163] T. L. Saaty. *What is the analytic hierarchy process?* Springer, 1988.

[164] J. Sauro and J. R. Lewis. Correlations among prototypical usability metrics: evidence for the construct of usability. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1609–1618. ACM, 2009.

[165] B. Schilit, N. Adams, and R. Want. Context-aware computing applications. In *Mobile Computing Systems and Applications, 1994. WMCSA 1994. First Workshop on*, pages 85–90. IEEE, 1994.

[166] A. Sears. Heuristic walkthroughs: Finding the problems without the noise. *International Journal of Human-Computer Interaction*, 9(3):213–234, 1997.

[167] A. Seffah, M. Donyaee, R. B. Kline, and H. K. Padda. Usability measurement and metrics: A consolidated model. *Software Quality Journal*, 14:159–178, 2006.

[168] A. Seffah and E. Metzker. The obstacles and myths of usability and software engineering. *Communications of the ACM*, 47(12):71–76, 2004.

[169] M. H. Selamat, N. C. Pa, R. Abdullah, et al. E-learning user interface acceptance based on analysis of user's style, usability and user benefits. *Jurnal Sistem Informasi*, 9(1):6–12, 2013.

[170] B. Shackel. Ergonomics in design for usability. In *Proceedings of the Second Conference of the British Computer Society, human computer interaction specialist group on People and computers: designing for usability*, pages 44–64. Cambridge University Press, 1986.

[171] B. Shackel. Usability-context, framework, definition, design and evaluation. *Human factors for informatics usability*, pages 21–37, 1991.

[172] B. Shackel. Human-computer interaction—whence and whither? *Journal of the American Society for Information Science*, 48(11):970–986, 1997.

[173] G. Shanks, E. Tansley, J. Nuredini, D. Tobin, and R. Weber. Representing part-whole relations in conceptual modeling: an empirical evaluation. *MIS Quarterly*, pages 553–573, 2008.

[174] B. Shneiderman. *Designing the user interface: strategies for effective human-computer interaction*, volume 2. Addison-Wesley Reading, MA, 1992.

[175] S. L. Smith and J. N. Mosier. *Guidelines for designing user interface software.* Citeseer, 1986.

[176] N. Stanton and C. Baber. Usability and EC Directive 90270. *Displays*, 13(3):151–160, 1992.

[177] A. Stevens, A. Quimby, A. Board, T. Kersloot, and P. Burns. *Design guidelines for safety of in-vehicle information systems.* TRL Limited, 2002.

[178] S. Suzuki, V. Bellotti, N. Yee, B. E. John, Y. Nakao, T. Asahi, and S. Fukuzumi. Variation in importance of time-on-task with familiarity with mobile phone models. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2551–2554. ACM, 2011.

[179] C. Thomas and N. Bevan. *Usability Context Analysis: A Practical Guide, V4.04.* National Physical Laboratory, Teddington, UK, 1996.

[180] M. Träskbäck and M. Haller. Mixed reality training application for an oil refinery: user requirements. In *Proceedings of the 2004 ACM SIGGRAPH International Conference on Virtual Reality Continuum and its Applications in Industry*, page 324–327, 2004.

[181] K. J. Vicente. Ecological interface design: Progress and challenges. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 44(1):62–78, 2002.

[182] R. A. Virzi. Refining the test phase of usability evaluation: how many subjects is enough? *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 34(4):457–468, 1992.

[183] S. Wagner. Re: Questions about "a comprehensive model of usability". personal communication.

[184] Y. Wand and R. Weber. On the ontological expressiveness of information systems analysis and design grammars. *Information Systems Journal*, 3(4):217–237, 1993.

[185] A. Whitefield, F. Wilson, and J. Dowell. A framework for human factors evaluation. *Behaviour & Information Technology*, 10(1):65–79, 1991.

[186] S. Winter, S. Wagner, and F. Deissenboeck. A comprehensive model of usability. In *Engineering interactive systems*, pages 106–122. Springer, 2008.