UNIVERSIDAD DE OVIEDO

Doctorado en Informática

"AUTOMATIC USER MODELLING THROUGH DEMOGRAPHIC ANALYSIS OF USER INTERACTION"

"MODELADO DE USUARIO AUTOMÁTICO A TRAVÉS DEL ANÁLISIS DEMOGRÁFICO DE LA INTERACCIÓN DEL USUARIO"

AUTOR: Beatriz Pariente Martínez

Acknowledgements

To my tutors Martin González Rodriguez, Daniel Fernández Lanvin and Javier de Andrés Suarez, without your help, I would not have gone this far. To my family and friends for their support and for cheering me up to don't give up.
I'll always be grateful to all of you.

Abstract

Knowing the user is crucial to improve websites in many of their dimensions. An understanding of user's profile and demographic data can help designers, for example, to adapt the application interface for elderly or handicapped users (increasing the size of the components), filter the contents for children, or, in the scenario of an e-commerce site, customize both interface and catalog in order to improve customer perception of the site quality and therefore, increasing the sales.

This thesis analyzes whether it would be possible to build an automatic system that categorizes visitors of a website depending on their age or gender through the analysis of the execution time required to complete *Point & Click*, *Drag & Drop*, *Selection Text*, *Text Edit* and *Item Selection* tasks. The performance of 592 volunteers who executed these five tasks through different tests designed using GOMS (Goals, Operators, Methods, and Selection rules) was analyzed using several different statistical methods. Consistencies in the execution times of individuals across the different tasks were found in this study, revealing whether age and gender are sufficiently determining factors to support an automatic profiling system. According to that, Machine learning algorithms like Bagging (Bagging, Random forest), Boosting (AdaBoost.M1, LogitBoost), SVM(SMO), C4.5(J48), Logistic and Stacking were tested in order to build this automatic profiling system.

Key words

Usability, Human behavior, HCI, GOMS, Performance, Age, Gender, Machine learning, Automatic profiling system, User classification

Resumen

Conocer al usuario es crucial para mejorar los sitios web en muchas de sus dimensiones. La comprensión del perfil del usuario y los datos demográficos pueden ayudar a los diseñadores, por ejemplo, para adaptar las interfaces para personas mayores o usuarios discapacitados (aumentando el tamaño de sus componentes), filtrar los contenidos para niños o, en el escenario de un sitio web de comercio electrónico, personalizar tanto la interfaz como el catálogo de de productos para mejorar la percepción que tiene el cliente de la calidad del sitio y, por lo tanto, aumentar las ventas.

Esta tesis analiza si es posible construir un sistema que categorize a los visitantes de un sitio web de forma automática en base a su edad o género a través del análisis de la ejecución del tiempo requerido para completar tareas de "Point & Click", "Drag & Drop", selección de texto, edición de texto y selección de elementos. El rendimiento de 592 voluntarios en la ejecución de esas cinco tareas en diferentes test diseñados empleando GOMS (Goals Operators, Methods and Selection Rules) fué analizado empleando diferentes métodos estadísticos. Consistencias en los tiempos de ejecución de los voluntarios sobre diferenes tareas revelaron si la edad y el género son factores determinantes para la categorización automática de usuarios. En base a estos primeros estudios, se evaluaron diferentes algoritmos de Aprendizaje Automático (Machine Learning) para la construción de aplicaciones de categorización automática. Entre los algoritmos estudiados se incluyen los tipo Bagging (Bagging, Random Forest), los tipo Boosting (AdaBoost,.M1, LogitBoost), SVM(SMO), C4.5(J48), Logistic y Stacking.

Palabras clave

Usabilidad, Comportamiento humano, HCI, GOMS, Rendimiento, Edad, Género, Machine learning, Sistema de perfilado automático, Clasificación de usuerios

Contents

List of Figures

Table index

1 Introduction

The success of online information systems is highly determined by the way they manage the information they contain. Selecting the right information to show to each specific user is a critical factor for the success of the site [98]. Moreover, its appropriate representation and layout according to the user's specific needs can also determine the user's satisfaction [150]. These two decisions are strongly determined by several factors that are usually referred to as user profile or user model [101]. Therefore, the quick identification of common interaction patterns among users is critical to web designers in order to adapt their information systems to meet the users' diverse needs [82].

This user profiling can be determined by several factors that have a great relevance in the success of the online information systems of companies [134, 144] since differences in age, gender, and income indicate different interaction intentions [158]. There is strong empirical evidence showing that important behavior differences caused by gender, age, social grouping, and household income have a great influence on how the user accesses online services [65, 92]. In fact, there is a wide literature studying patterns of child–computer interaction [2, 46, 47, 77], older people's performance rates [28, 42, 161, 135], motion-impaired users [76, 83, 75], and similar factors.

However, given that most of the potential customers of online sites are anonymous users, performing user profiling at the design stage is a difficult task to carry out without falling into misconceptions, since both age and gender parameters are usually unknown. Even when the user provides this data as part of the registration process, its reliability can be determined by social or cultural factors. Underage users, for example, usually lie to avoid controls in adult-oriented sites or social networks [137].

The measurement and analysis of the performance in certain tasks could help to draft a user's profile. The work of Freudenthal et al.[55] and Weiser et al. [151] evidenced than both age and gender influenced in user behaviour and could be a key to obtaining a profile of an individual. However, these specific experiments do not demonstrate if this influence is significant enough to support user classification in the context of web interaction.

This work intends to fill this GAP, by exploring the feasibility of classifying anonymous users in age and gender groups through the observation of the way they interact with the UI. For that, we gathered interaction data of a sample of users (which for the first time includes right- and left-handed people, individuals of both genders and of all ages) performing – also for first time – five basic tasks required to interact with

information systems: (i) *Point & Click*, (ii) *Drag & Drop* , (iii) *Text Selection*, (iv) *Text Edit* and (v) *Menu Item Selection*.

The automatic classification of users according to their ages or gender can be done by analyzing the performance of the users when they perform these tasks, comparing it with the performance of individuals whose demographic characteristics are known. This work required a previous statistical study that allowed us to verify whether these demographic characteristics influenced on the performance of the users while executing these interaction tasks.

This statistical study showed that age and gender influence the performance in some of the tasks. Once these discoveries were found out, Machine learning algorithms like Bagging (Bagging, Random forest), Boosting (AdaBoost.M1, LogitBoost), SVM(SMO), C4.5(J48), Logistic and Stacking were tested in order to build this automatic profiling system. These results could allow the design of systems able to infer the gender and the age of the user, employing adaptive web design techniques to enhance the user experience of a web information system, adapting its contents to the user's requirements.

2 Background/Related Work

2.1 GOMS

GOMS [24] is a reduction of a user's interaction in terms of **G**oals, **O**perators, **M**ethods, and **S**election rules allowing for different aspects of an interface to be accurately studied and predicted. Goals are what the user intends to accomplish. An operator is an action performed in service of a goal. A method is a sequence of operators that accomplish a goal and if more than one method exists, then one of them is chosen by some selection rule. GOMS analysis has been successfully used to estimate the user's performance in different interaction scenarios including web sites [133, 112], touch screens [1] and motor vehicles [156].

GOMS analysis assumes that users are computer literate and know perfectly how to do the task under observation, without generating any errors in the process. Due to their degree of expertise, users are able to interact as fast as possible. Therefore, GOMS focuses on estimating the user's effectiveness (performance) instead of the user's efficiency.

GOMS simplifies the user's interaction process by considering complex tasks as combinations of a limited set of basic actions called operators. Complex task estimations are simplified by estimating the performance in the basic operators required by the high-level task. Some of the main basic operators used by GOMS are "pointing" (denoted by P), "dragging" (D), "Key pressing" (K) and "Mental" (M). The latter is used whenever the user has to make a simple decision (e.g. selecting an item in a menu).

The execution of each operator requires a specific amount of time (denoted respectively as T_P, T_D, T_K, T_M, etc.), so GOMS estimates the execution time of complex interaction tasks as the sum of the execution times of the different operators required to complete the tasks. So, for example, the estimated execution time for a *Drag & Drop* interaction task would be $T_P + T_K + T_D + T_K$; that is, the time needed to move the mouse pointer over the movable object (T_P) plus the time required to press the mouse's button once the pointer is over the target (T_K) plus the time used to drag the object to a new position (T_D) plus the time required to release the mouse's button (T_K) [36, 77]. Text Selection, Text Edit and Menu Item Selection tasks use similar approaches.

The execution time for each operator (T_P, T_D, T_K, T_M, etc.) is estimated using well-known psychological laws and regularities. Fitts's Law [162] estimates the time required by a user to point/drag to/from a target (T_P and T_D) (e.g. moving the mouse

pointer over an object on the screen) as a function of the distance to the target and its size. Salthouse's regularities predict the time required by different kinds of users (ranging from novices to experts) to type texts of a given size (T_K) [127]. The Hicks-Hymans law estimates the time required to make a decision (T_M) (e.g. the selection of a menu item) as a linear increase related to the logarithm of the number of alternatives the user has to evaluate (a + b Log_2(n+1)) where n represents the number of available options and a and b again are user dependent correction factors [126, 131].

It was noted that there are external variables that may increase or decrease the estimations predicted by these psychology laws. So, for instance, aiming (P) at a square target is faster (T_P) than aiming over a rounded target when both objects have the same width [115, 36, 77]. Pointing (P) requires several interactions when aiming for the target, reducing its speed in each one to recalculate the trajectory in order to increase the accuracy [115]. Fitts's Law estimates the runtime needed to perform all these recalculations as a + b Log_2 (D/S). Where 'D' is the distance to the target and 'S' represents its size, being the parameters 'a' and 'b' which are user dependent correction factors [162, 63]. Since the surface of the squared object is slightly greater than that of the rounded object, the estimated time to point at the squared object is smaller.

Although these laws help to estimate the execution time required by an average user, they have to be adapted to the specific needs of individual users. That is the case of the correction factors used by the Fitts' law and the Hicks-Hymans law which have to be obtained through the analysis of performance records previously obtained for specific users. The values for these correction factors rely on the external variables under analysis in this research, as it is the case of age and gender.

2.2 Age

Ageing negatively impacts the ability to use computers [51, 70] and is typically defined in Human Computer-Interaction through an emphasis on decline in abilities and associated reductions in performance when using technology [146]. It produces poorer motor control and sensory deficits [42]. Related studies show that older people have slower reaction times [54], delayed movements, a decline in motor skills [149]. Reduced mobility, caused by a loss in muscle strength [138], produces difficulties in the execution of movements [138]. This process of losing muscle strength begins in people aged over 40 [104].

Other studies revealed that ageing negatively influences the learning strategies used to operate online systems, as perception and cognition declines [106, 157]. Senior users have been found to be slower than young adults when retrieving information [55, 108], performing 3D navigation on desktop systems [129] or browsing the web [110].

Studies analyzing information search behavior [143, 70] enforce the importance of ageing. The study of the behavior of expert older adults daily using the web, compared

to their younger colleagues, concluded that age is a determining factor [70]. This work is particularly interesting because it is specifically focused on web interaction. However, the analyzed activities (search behavior and related) require different operators than those involved in mouse motion.

On the other hand, very young users reveal a poor execution time in the development of certain tasks because their speed on their coordinate movements is still in development until the age of 12 years [89]. According to that, the young users also have difficulties on basic interaction tasks like *Drag & Drop*. In this case, it is especially difficult for them to keep the finger pressed down while controlling its trajectory requires a high demand of motor skills [102], perception and cognitive skills [27, 37]. Otherwise, the execution time slows down if it is possible to replace *Drag & Drop* by *Point & Click*, as the operation can be easily resumed from the last pointing task in case of failure [81]. Attaching and lifting objects in the real world causes some difficulties for children under 8 years old as these tasks require subtle hand-eye coordination [89]. At that age, the coordinate movements are further determined by cognitive factors rather than motor skills [2]. In order to avoid the effect on our study of these difficulties that young users experience, data from users with less than 12 years old was discarded.

Some authors reported how different interfaces influence the interaction of specific groups of users regarding their age [25], but no study was found about whether there are significant differences between the time required to execute different -alternative- interaction tasks conducting to the same result (e.g. *Point & Click* as an alternative to *Drag & Drop* to obtain the same result).

If these differences do not exist and the execution time keeps coherence in each basic interaction task, that is, if the time required by each group of users is similar in each task (*Point & Click*, *Drag & Drop* and *Item Selection* it would not be necessary to analyze the three interaction tasks in the same user interface to detect the users' age. It would be enough to analyze the users' performance in only one of them. However, if those differences exist, it would be necessary to measure and to analyze the users' performance in all the three different proposed interaction tasks to categorize users according to their age.

2.3 Gender

Women process information in different ways than men [9]. Gender-associated differences in decision making, learning, and problem solving can be a determining factor in user's effectiveness [7, 8]. Even more, it has been observed that the self-perceptions concerning computer competence as well as the level of ICT-related social interactions is different for boys and girls [33].

It was observed that men's performance in navigating through virtual environments is better than women's when small displays are used. The use of larger displays reduces

the gender performance gap since the women's performance improves while the men's performance is not negatively affected [38, 139].

Inkpen [77] compared *Drag & Drop* tasks as opposed to *Point & Click* in children. Although there was no significant gender difference in the overall movement time and/or general error rates, there were relevant differences in pickup and drop errors. Girls performed poorly when executing *Drag & Drop* tasks, as opposed to *Point & Click*. There were also performance correlation differences between gender and target size.

Rohr [125] evidenced that gender-specific movement biases emphasize speed for men and accuracy for women. Wahlstrom et al. [148] observed that when operating the mouse, women worked with greater extension and had a greater range of motion in the wrist when compared to men. This observation could explain Rohr's results regarding speed versus accuracy. They also found gender differences for musculoskeletal load. For most of the measured variables, women worked with higher loads than men. These differences are not limited to the low-level interaction. Collazos et al. [34] found significant differences in the way women and men face collaborative work in computer-mediated communication.

The rest of the document is structured as follows. In Chapter 3 the interaction experiment that was carried out is detailed. Chapter 4 contains the Part 1 of the study that validates the influence of age in the performance of computer interaction tasks. Chapter 5 explains the Part 2 that identifies the demographic differences based on how users interact with web applications. Chapter 6 contains the Part 3 of the study that covers the assessment whether gender and age are sufficiently significant determining factors to support an automatic profiling system and studies whether the individuals perform consistently across these basic interaction tasks. In Chapter 7, the Part 4 of the study can be found describing the process followed on building an automatic classification system by age or gender using machine learning algorithms. Chapter 8 summarizes the conclusions of this study. Finally, Chapter 10 contains the Limitations and Future Work.

3 Design of the empirical study

The use of GOMS in this context has two main advantages. First, it helps to structure the study of the different interaction tasks using a common research framework to other similar studies. Second, the experimental measurement of the users' runtime for each specific task, to a high degree of accuracy, facilitates a quick and accurate estimation of the global execution time for web sites whose user interfaces combine several of these interaction tasks.

A website was designed to include five tests based on the combined use of GOMS operators. As the GOMS analysis requires participants who know how to perform the tasks under observation, it was considered that users who are familiar with the relatively complex user interfaces of discussion web forums matched this profile.

Each test was designed to replicate the behavior of a real web application but hiding the features that might allow the user to identify it, thus avoiding the effect that the familiarity with the real product might have on the measurements to be obtained. Hence, the lexical and semantic levels (related to mouse movement and object recognition and perception) of the user interface of the corresponded web application were recreated in the most realistic way, while the semantic (iconic representation) and conceptual (final goal of the application) levels of the interface were ignored or hidden to avoid the mention familiarity.

Previous works suggest that user laterality could affect the performance. Several studies have reported differences between left- and right-handed individuals regarding movement execution ([93]; [105]), movement preparation ([10]; [109]), stimulus velocity effect ([123]) or interactions between hand preference and hand performance ([114]). These precedents suggest that user laterality can have a sensible influence on user behaviour and performance.

Also, the number of hours per week the user spends interacting with computers seems to be another determining factor. These two variables, laterality and hours of use were incorporated into the study as control variables.

Prior to executing the tests proposed for this experiment, the users provided specific information about themselves using an online questionnaire: age, gender, laterality and number of hours per week spent using computers.

During the experiment, the users had to complete five tests, requiring the use of different GOMS operators to execute each one. The average execution time for each

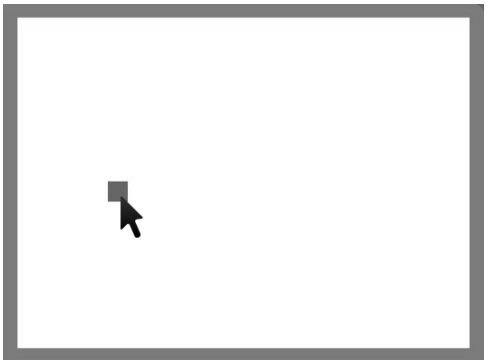


Figure 3.1. *Point & Click* task. GOMS operators are required to complete a *Point & Click* task. Step 1: users move the mouse pointing (P) to the target. Step 2: the user clicks on the target Key pressing (K) the mouse button.

operator (P, D, M or K), needed to complete each high-level task, was recorded in milliseconds and was denoted by T_P, T_D, T_M and T_K.

The first test (task 1) was designed to analyze the behavior of users executing the *Point & Click* tasks required to select objects in web documents by moving the mouse pointer across the display to click on links, buttons, scrolling boxes, etc.

The test showed a sequential series of rectangles in different locations across the screen. Participants in the test had to click inside each rectangle to make it disappear before a new one appeared in a different location. The test used fourteen different rectangles distributed in positions that followed a Z pattern layout to keep a fair balance between left-handed and right-handed users. Along the test, the location of the different targets was changed using the horizontal (left to right, right to left) and vertical dimensions (top to bottom, bottom to top).

At the same time, Fitts' law was used to increase the difficulty of each interaction, increasing the distance to the target (D) and reducing its size (S), thus increasing the time required to click on the target by a factor of Log_2(D/S).

To click on the target, users had to use two GOMS operators: P and K (see Fig. 3.1). First, the users moved the mouse over the display to place the pointer over the square using T_P units of time (step 1 in Fig. 3.1). Next, users needed to click the mouse button using a K operator (step 2 in Fig. 3.1). The time estimated by GOMS to complete each *Point & Click* action is therefore $T_P + T_K$. The time required to complete each point and click action ($T_P + T_K$) was recorded (in milliseconds) for each click interaction. The sum of the execution times required to complete the full test was recorded for later statistical analysis.

The second test (task 2) was designed to measure the time required to complete *Drag & Drop* tasks. In this second test, users were asked to drag a red rectangle over a second one, which had a size two thirds bigger than the red one. Every time the user completed the task, both rectangles disappeared, and two new rectangles appeared in separate locations of the display. The process was repeated along fourteen

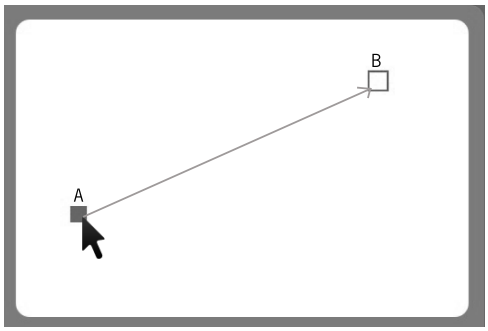


Figure 3.2. *Drag & Drop* task. GOMS operators are required to complete a *Drag & Drop* task in the test application. Steps 1 and 2 are the same as in the *Point & Click* task described in Fig. 3.1. In step 3 users had to drag (D) the small rectangle over the big one.

interactions. Each time, the rectangles were distributed using a Z shaped layout to keep a fair balance between left-handed and right-handed users. The distance between objects was incremented and its size was reduced in each interaction, using the Fitts' law to increase the time required by the users to complete each interaction.

To drag the first rectangle over the second, the users had to select it first. Therefore, they needed to use the GOMS operators required in a *Point & Click* task. The P operator is required to point to the rectangle (step 1 in Fig. 3.2) and the K operator is needed to select it (step 2 in Fig. 3.2) clicking the mouse button. Next, users had to drag the rectangle using the dragging operator (D) until the first rectangle was over the second one (step 3 in Fig. 3.2) releasing it with a mouse button action (T_K). The time estimated by GOMS to complete each *Drag & Drop* action is therefore $T_P + T_D + 2T_K$. The time required to drag the object (T_D) was recorded (in milliseconds) to be used in the statistical analysis.

The third test was designed to measure the user performance in the execution of *Text Selection* tasks. In each interaction, users were asked to select a word in an appropriate text context using a pointer (see Figure 3.3). This process was repeated eight times using a different word each time. This task required a combination of different GOMS operators. The M operator was needed to select the word. The P operator was required to aim for the beginning (or the end) of the text. Next, a K operator was needed to activate the text selection. The D operator was required to drag over the desired word. Last, the K operator was needed to deactivate the text selection. The recorded performance time was obtained as $T_M + T_P + T_D + 2T_K$.

The fourth test recorded the user performance in *Text Edit* tasks. Each iteration required users to write a provided sentence (see Figure 3.4). The test included five iterations (sentences). In order to write the sentences, the users needed to execute the K operator as many times as letters are included in the sentences. Therefore, T_K is the variable measured in this test.

Finally, the test 5 measured the user performance in the execution of *Menu Selection* tasks, which are used to select items in a user interface (e.g. menus, combo boxes, radio

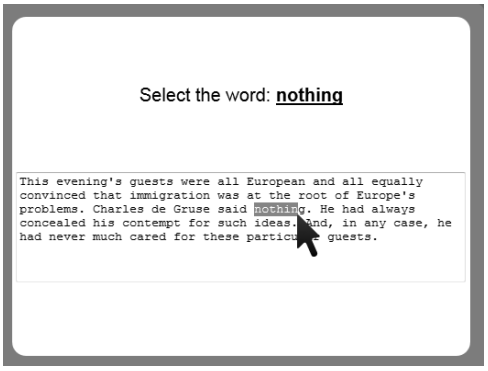


Figure 3.3. *Text Selection* task. During the *Text Selection* test users were asked to select a given piece of text included in a paragraph.

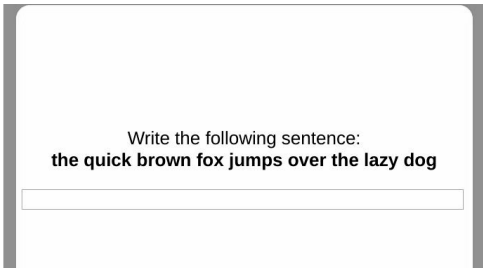


Figure 3.4. *Text Edit* task. In the *Text Edit* test, users had to write known pieces of text as fast as possible.

button groups, etc.). In this test, users were asked to select a given color in a popup menu. To achieve this operation participants needed to execute a *Point & Click* task to display the menu items available clicking on the menus title. Then, users were asked to select a specific menu item whose name was displayed on the screen. Then, participants executed a second *Point & Click* task to click on the menu item corresponding to the asked color. The process was repeated ten times. In each interaction, the menu was placed in a different position using the Z shaped layout described before. Each menu contained five items. Each volunteer had to select each menu item two times across the interactions.

The time required to achieve the first *Point & Click* task (see Fig. 3.5) was denoted by $T_{P1} + T_{K1}$. It represents the Point (P_1) and Key pressing (K_1) operators required to activate the menu. The second runtime was denoted by $T_{P2} + T_{K2}$. Finally, the time needed by the mental operator M to take the decision (selecting which menu item satisfies the search constraints) was denoted by T_M. The resulting execution time predicted by GOMS for the entire test process, denoted by $T_{P1} + T_{K1} + T_M + T_{P2} + T_{K2}$ was recorded for later statistical analysis.

GOMS assumes that the volunteers know how to use the web system under evaluation (either because they got some previous training or because they have used the system previously). GOMS also assumes that users will not commit any error during the process. Due to this high degree of expertise, users are supposed to interact as fast as possible.

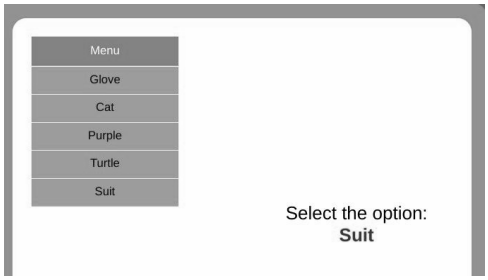


Figure 3.5. *Menu Item Selection* task. GOMS operators are required to complete an *Menu Item Selection* task. Steps 1 and 2 define the *Point & Click* task required to activate the menu clicking on its title. Once the menu items are displayed, a Mental (M) operator is executed (step 3) to select the required color (displayed in the bottom-right corner of the screen). Steps 4 and 5 represent the Pointing (P) and Key pressing (K) GOMS's operators required to complete the *Point & Click* task needed to select the menu item.

Based on these precepts, GOMS is a reliable tool to estimate the user's effectiveness (execution time) instead of estimating the user's efficiency (success/failure rate).

To meet these strong requirements, there were two rounds of individuals participating. In the first round we gather 630 individuals. With the data gathered in this first round, we carried out the first part of our research, in which we proved that the results were significant enough to be able to follow with our research. This first analysis is included in the following section Analysis 1. After these results, we conducted a second round of participants that allowed us to accomplish the rest of this research. In this second round, 592 individuals participated in it and the data collected from it were used in the rest of the investigation.

The case study was recruited through "Facebook", "Twitter" and "Foro Coches"[4], the most popular general purpose online community in Spain, thereby ensuring that participants were familiar with the basic interaction tasks frequently found in online systems. Therefore, participants could execute *Point & Click*, *Drag & Drop*, *Text Edit*, *Text Selection* and *Menu Item Selection* tasks in a natural way that they did not need to think about the steps needed to complete them.

This approach not only complied with the GOMS requirements but also allowed the participation of a high number of users. The sample used in this study includes 592 individuals. It is large when compared with the samples used in the studies described in the Related Work section, which were mostly based on samples whose size ranges between 10 and 20 individuals.

This high number allowed the use of multivariate regression analysis to obtain more accurate results when compared with those of prior studies. In addition, it allowed the inclusion of some variables in the model that may bias the results if they are not adequately controlled for (handedness and prior experience with computers).

4 Part 1: Analysis of user's age as significant factor for user classification

4.1 Introduction

A sample made up of 630 individuals was used. As a first step, individuals who were unable to complete any of the tasks were removed, so the final analyzed sample was reduced to 557 individuals. The age breakdown is detailed in Table 4.1.

Age group	Number of cases	Percentage	Cumulative percentage
1	18	03.23	03.23
2	302	54.21	57.45
3	147	26.39	83.84
4	52	09.33	93.17
5	24	04.30	97.48
6	14	02.51	100.00
TOTAL	557	100.00	

Table 4.1. Sample breakdown by age group

It can be seen that the sample used is a convenience sample where the proportions that represent each group do not match those that exist in the real population. This led the authors to use an experimental design, which is discussed later, that includes data transformation and robustness checks.

First of all, some descriptive statistics about both the dependent and independent variables were computed. Examination of such data gives us a first idea of the features of the individuals in the sample, their behavior in the experiment, and the statistical properties of the results.

Second, as an exploratory study of the data suggests the presence of remarkable degrees of skewness and kurtosis, data winsorizing was used. Winsorizing consists of replacing extreme observations (usually those above the 99% or below the 1% percentile) by a close value not considered as an outlier. This procedure has the advantage that it helps to mitigate the influence of outliers without reducing the sample size. Although winsorizing has not been previously used in usability research, it is a common

practice in other fields where non-normal observations are analyzed (in this regard see [120] and [155] as examples).

In this research, winsorizing was carried out by replacing the top and bottom 1% of the observations of each task by the closest value not considered an extreme value. Data were subsequently transformed using natural logarithms, and a series of Lilliefors tests was applied to test for the normality of the distributions. The results provide evidence that in all cases normality is not rejected at the 5% significance level [60]. Once the normality of the data was ensured, a set of inferential tests was conducted for each of the tasks included in the experiment. Specifically, three types of tests were conducted:

1. Performance comparison between age groups. Each age group was compared with each one of the others with regard to performance on the three tasks. The two-independent-samples t-test for the difference of means was used.

2. Performance comparison between subsamples. For each of the tasks, the sample was divided into two subsamples, using as cut-off points the age that separated the two adjacent age groups. The performance of one of the subsamples of the binary partition was also compared with that of the other by using the two-independent-samples t-test for the difference of means.

3. Equality of proportions tests. For each of the tasks, individuals were classified into five groups depending on their performance. The first subgroup comprised 20% of the individuals who achieved the best performance, the second subgroup the following 20%, and so on. The cumulative percentages were analyzed and a series of binomial tests were conducted to assess whether the observed age proportions in each of the performance groups differed significantly from those which would be obtained if individuals of each age group were equally distributed across the performance groups.

Finally, as a robustness check, Stages 1 and 2 of the inferential tests were repeated, though in this case the t-tests were replaced by the non-parametric Mann-Whitney test. Tests were conducted using raw data (data prior to winsorizing and logarithm transformation). The results, which are not included in this thesis, do not differ significantly from those explained in the following section.

4.2 Results

4.2.1 Descriptive analysis

Table 4.2 contains some descriptive statistics (mean, standard deviation, and quartiles) relative to the performance of each age group on each of the tasks. These statistics are referred to the raw data, that is, prior to the winsorization and log transformation.

Panel A: Task 1 : Pointing					
Age group	Mean	Std Dev.	Q25	Median	Q75
1	14488.00	2301.73	12835.00	14099.50	15993.75
2	16479.53	4263.67	13796.75	15392.50	18644.00
3	17429.87	4047.95	14702.00	16319.00	19032.00
4	18056.48	3144.48	16115.25	17345.00	19144.00
5	21001.25	5279.07	17679.50	19136.00	24392.25
6	23084.64	5257.08	19635.50	22150.00	26533.50
TOTAL	17174.05	4371.48	14232.50	16064.00	19090.50
Panel B: Task 2 : Drag and Drop					
Age group	Mean	Std Dev.	Q25	Median	Q75
1	31083.28	7307.67	25177.50	28659.50	36762.25
2	32608.76	9051.14	26267.75	30392.00	37006.75
3	36605.12	11290.19	28933.00	33965.00	41330.00
4	36954.67	8131.59	30393.00	36443.50	42719.75
5	43797.25	11485.15	32931.75	42101.00	53780.50
6	54937.07	17154.84	42020.75	50924.00	65630.00
TOTAL	35063.18	10763.09	27649.00	32273.00	39855.00
Panel C: Task 3 : Text Selection					
Age group	Mean	Std Dev.	Q25	Median	Q75
1	74133.28	24530.68	53921.00	67826.50	93035.25
2	114344.95	352440.39	61293.75	77536.00	99655.00
3	179229.82	995348.76	70725.00	84596.00	102347.00
4	100064.92	40787.62	68141.75	93952.00	119024.50
5	142488.88	59073.93	102351.25	139893.50	171880.75
6	126847.50	46977.84	84355.50	120512.00	172525.00
TOTAL	130363.25	573350.58	64861.50	82093.00	106830.00

Table 4.2. Performance of the individuals on each task by age group

The data suggest that individuals from older groups take more time to complete the three tasks. However, dispersion levels are very high in general, extreme values are somewhat common, and the distributions are always positively skewed (the mean is always higher than the median). This confirms the need for the winsorization/log transformation procedure that was outlined in the previous section.

Table 4.3 displays the same raw data descriptive statistics but relative to the different subsamples obtained by making binary partitions as explained in the previous section.

The results obtained are very similar to those of Table 4.2. Older individuals take more time to complete the tasks, but the observed distributions are leptokurtic and positively skewed. So, prior processing of the data is also evidenced.

4.2.2 Inferential Test

Table 4.4 contains the results of the pairwise comparison of the different age groups with regard to the performance on Task 1. Each cell contains the t statistics of the t

Panel A: Task 1 : Pointing					
Under the cut-off point					
Cut-off point	Mean	Std dev.	Q25	Median	Q75
20 yrs	14488.00	2301.739	12835.00	14099.50	15993.75
30 yrs	16367.50	4200.794	13737.00	15320.00	18372.25
40 yrs	16701.91	4178.250	13993.00	15504.00	18671.00
50 yrs	16837.63	4104.205	14136.00	15864.00	18724.00
60 yrs	17021.66	4244.401	14200.00	15930.00	18939.00
Above the cut-off point					
Cut-off point	Mean	Std dev.	Q25	Median	Q75
20 yrs	17263.75	4396.858	14313.00	16115.00	19161.00
30 yrs	18263.05	4369.853	15203.00	17146.00	19749.50
40 yrs	19623.91	4553.667	16453.25	18689.50	21527.25
50 yrs	21768.82	5298.237	17893.25	20367.00	24687.75
60 yrs	23084.64	5257.082	19635.50	22150.00	26533.50
Panel B: Task 2 : Drag and Drop					
Under the cut-off point					
Cut-off point	Mean	Std dev.	Q25	Median	Q75
20 yrs	31083.28	7307.676	25177.50	28659.50	36762.25
30 yrs	32522.96	8959.375	26231.75	30381.50	36972.75
40 yrs	33807.92	9924.089	26801.00	31832.00	38413.00
50 yrs	34123.20	9798.232	27385.00	31921.00	38850.00
60 yrs	34550.78	10065.422	27591.00	32159.00	39295.00
Above the cut-off point					
Cut-off point	Mean	Std dev.	Q25	Median	Q75
20 yrs	35196.09	10839.050	27744.00	32290.00	39998.00
30 yrs	38493.03	11994.038	30541.00	35395.00	43943.50
40 yrs	41576.62	12524.194	31873.00	39379.00	47240.50
50 yrs	47901.39	14664.648	34437.75	46655.00	56194.75
60 yrs	54937.07	17154.840	42020.75	50924.00	65630.00
Panel C: Task 3 : Text Selection					
Under the cut-off point					
Cut-off point	Mean	Std dev.	Q25	Median	Q75
20 yrs	74133.28	24530.687	53921.00	67826.50	93035.25
30 yrs	112083.04	342525.116	61121.00	76033.50	98331.50
40 yrs	133219.18	625847.634	63660.00	79677.00	100301.00
50 yrs	129897.37	593825.385	64198.00	80566.00	102232.00
60 yrs	130453.90	580662.412	64389.00	81713.00	105735.00
Above the cut-off point					
Cut-off point	Mean	Std dev.	Q25	Median	Q75
20 yrs	132241.06	582752.943	65338.00	82487.00	107575.00
30 yrs	155045.39	784102.854	71703.00	89030.00	117845.00
40 yrs	115544.16	50332.776	77759.25	107922.50	148157.00
50 yrs	136726.26	54801.106	95536.50	136265.50	169424.75
60 yrs	126847.50	46977.842	84355.50	120512.00	172525.00

Table 4.3. Descriptive statistics, splitting the sample according to the different age cut-off points

test for the difference of means (the test was performed by computing the difference between the column group and the row group). The corresponding p-value is also shown below (in parentheses).

	1	2	3	4	5	6
1	–					
2	2.253 (0.025)	-				
3	3.439 (0.001)	2.884 (0.004)	-			
4	5.042 (<0.001)	3.519 (<0.001)	1.480 (0.141)	-		
5	5.588 (<0.001)	5.432 (<0.001)	3.939 (<0.001)	2.945 (0.004)	-	
6	6.773 (<0.001)	6.024 (<0.001)	4.893 (<0.001)	4.480 (<0.001)	1.316 (0.197)	-

Table 4.4. Age groups comparisons for Task 1: *Point & Click*

It is evidenced that for Task 1, at the usual significance levels (1 and 5%) younger groups perform better than older groups. The only exceptions are the comparisons between Groups 3 and 4 and between Groups 5 and 6, where no significant differences arise.

Table 4.5 follows the same structure as Table 4.4 but contains information related to Task 2.

	1	2	3	4	5	6
1	–					
2	0.691 (0.490)	-				
3	2.224 (0.028)	4.238 (< 0.001)	-			
4	2.997 (0.004)	3.839 (< 0.001)	0.725 (0.469)	-		
5	4.401 (<0.001)	5.793 (<0.001)	3.245 (0.001)	2.877 (0.005)	-	
6	6.277 (<0.001)	7.341 (<0.001)	5.058 (<0.001)	5.298 (<0.001)	2.194 (0.035)	-

Table 4.5. Age groups comparisons for Task 2: *Drag & Drop*

The results indicate that for Task 2 the performance of younger groups is also better than that of older groups in general. Exceptions are the comparisons between Groups 1 and 2 and between Groups 3 and 4.

Table 4.6 follows the same structure as Tables 4.4 and 4.5 but contains information related to Task 3.

	1	2	3	4	5	6
1	–					
2	1.302 (0.194)	-				
3	2.347 (0.020)	2.139 (0.033)	-			
4	2.549 (0.013)	2.535 (0.012)	1.158 (0.248)	-		
5	5.026 (<0.001)	5.959 (<0.001)	5.210 (<0.001)	3.367 (0.001)	-	
6	4.152 (<0.001)	3.760 (<0.001)	3.240 (0.001)	2.062 (0.043)	0.682 (0.499)	-

Table 4.6. Age group comparisons for Task 3: *Text Selection*

The data provide evidence that the pattern observed for Tasks 1 and 2 is once again present; that is, younger groups perform better. This time the exceptions are the comparisons between Groups 1 and 2, 3 and 4, and 5 and 6.

Table 4.7 contains the results of the t-tests conducted once the sample was split using the different cut-off points determined by the extremes of the age groups. In order to ease the interpretation of the results, summarized descriptive statistics (mean and standard deviation) are also provided, which in this case were computed after the winsorization of data but prior to the log transformation.

The results confirm the findings evidenced in Tables 4.4, 4.5, and 4.6. That is, younger individuals perform better. The only exception is Task 2 (drag and drop), considering 20 years as the cut-off point. Individuals under this age do not perform significantly better than those above.

Finally, the results of the analysis of the composition of each performance band for each of the three tasks are displayed in Table 4.8.

It is evidenced that in the high performance bands (100–80%, that is, the top 20% of performers), the proportion of younger individuals is significantly higher than the proportion observed in the total sample, while older individuals are underrepresented. For the low performance bands (the bottom 20% of performers) the opposite phenomenon is observed; that is, the proportion of younger people is lower than that of the sample, while that of older people is higher. For the intermediate band (60–40%), the observed percentages do not differ significantly from those of the total sample.

4.3 Conclusions

The results show several interesting facts that must be taken into consideration in the following stages of this research. Aging, as expected and pointed out by most of the

Panel A: Task 1 : Pointing						
Cut-off	Under		Above			
point	Mean	Std Dev.	Mean	Std Dev.	t-statistic	p-value
20 years	11506.89	2274.49	17204.09	4048.39	3.069	0.002
30 years	16290.20	3642.62	18233.17	4261.31	6.198	< 0.001
40 years	16637.89	3756.36	19602.56	4489.20	6.860	< 0.001
50 years	16780.03	3721.43	21718.24	5185.60	7.318	< 0.001
60 years	16963.06	3880.47	23084.64	5257.08	5.407	< 0.001
Panel B: Task 2 : Drag and Drop						
Cut-off	Under		Above			
point	Mean	Std Dev.	Mean	Std Dev.	t-statistic	p-value
20 years	31083.28	7307.67	35022.22	10013.08	1.706	0.089
30 years	32436.08	8370.27	38214.90	10935.35	7.213	< 0.001
40 years	33683.90	9229.58	41178.81	11209.89	6.856	< 0.001
50 years	34011.61	9171.16	46959.21	12317.95	7.614	< 0.001
60 years	34444.12	9488.74	52379.71	12180.96	6.176	< 0.001
Panel C: Task 3 : Text Selection						
Cut-off	Under		Above			
point	Mean	Std Dev.	Mean	Std Dev.	t-statistic	p-value
20 years	74133.28	24530.68	92908.75	43422.78	2.004	0.046
30 years	85792.78	40246.96	101090.68	45245.94	4.890	< 0.001
40 years	87822.78	40071.42	133683.90	49229.58	5.832	< 0.001
50 years	89049.35	40272.30	136726.36	54801.10	6.620	< 0.001
60 years	91411.32	42649.69	146847.50	46977.84	6.232	< 0.001

Table 4.7. Tests of difference of means, splitting the sample according to the different cut-off points determined by age

works discussed in the background section, has a strong influence on the performance of computer users.

After a more detailed analysis based on the GOMS task division, it can be argued that, as pointed out in [32], this influence differs between specific activities. In fact, results show that the aging effect is not the same for the three tasks that were analyzed. Although these facts were already observed in [32] and [127], the more fine-grained grouping used in this work allowed the profiling of the six age groups for each of the three basic tasks that were analyzed. That provides the pillars for an automatic classification system that would help classify an anonymous user based on his or her performance, and thus to dynamically customize a user interface for his or her specific needs and constraints.

The results obtained are encouraging enough to be considered as the starting point for introducing more complex measurements (error rates, biometrics, behavior patterns) and more variables under study, such as the user's gender and laterality.

Panel A: Task 1 : Pointing						
Performance	Cumulative percentage and binomial tests					
Bands	1	1, 2	1, 2, 3	1, 2, 3, 4	1, 2, 3, 4, 5	All
100–80%	07.2%*	82.0%**	98.2%**	100.0%**	100.0%	100.0%
80–60%	04.5%	60.4%	92.8%**	97.3%*	99.1%	100.0%
60–40%	01.8%	58.0%	83.9%	97.3%	100.0%	100.0%
40–20%	02.7%	38.7%**	71.2%**	89.2%	97.3%	100.0%
20–0%	00.0%*	48.2%*	73.2%**	82.1%**	91.1%**	100.0%
Sample cum. percentage	03.23%	57.45%	83.84%	93.17%	97.48%	100.0%

Panel B: Task 2 : Drag and Drop						
Performance	Cumulative percentage and binomial tests					
Bands	1	1, 2	1, 2, 3	1, 2, 3, 4	1, 2, 3, 4, 5	All
100–80%	05.4%	78.4%**	98.2%**	100.0%**	100.0%	100.0%
80–60%	03.6%	68.5%*	89.2%	99.1%**	100.0%	100.0%
60–40%	01.8%	51.8%	82.1%	92.0%	99.1%	100.0%
40–20%	04.5%	53.2%	82.9%	95.5%	98.2%	100.0%
20–0%	00.9%*	35.7%**	67.0%**	79.5%**	90.2%**	100.0%
Sample cum. percentage	03.23%	57.45%	83.84%	93.17%	97.48%	100.0%

Panel C: Task 3 : Text Selection						
Performance	Cumulative percentage and binomial tests					
Bands	1	1, 2	1, 2, 3	1, 2, 3, 4	1, 2, 3, 4, 5	All
100–80%	06.3%	72.1%**	91.1%*	97.3%*	99.1%	100.0%
80–60%	03.6%	67.6%*	91.0%*	99.1%**	100.0%	100.0%
60–40%	01.8%	53.6%	87.5%	95.5%	97.3%	100.0%
40–20%	03.6%	51.4%	83.8%	93.7%	97.4%	100.0%
20–0%	00.9%*	42.9%**	66.1%**	80.4%**	93.8%*	100.0%
Sample cum. percentage	03.23%	57.45%	83.84%	93.17%	97.48%	100.0%

Table 4.8. Cumulative percentages of age groups in each performance band and binomial tests (* Significant at the 5% level; ** Significant at the 1% level)

This analysis was published in Universal Access in the Information Society under the name of "*Measuring the role of age in user performance during interaction with computers.*" [113] attached in the Appendix B.

5 Part 2: Analysis of user's gender as significant factor for user classification

5.1 Introduction

This section aims to identify demographic differences based on how users interact with web applications. The research is needed to develop future systems able to adapt the representation of online information to the user's specific needs and preferences, improving its usability. Results from our first analysis had led us to some gaps found in the related work. In this section we will be focused on them.

Age related studies suggest that we can expect a negative influence of ageing in most facets of computer use. However, most of the experiments based on ageing are too specific and/or high level and do not evidence if these differences are significant enough to be observed in more general and fine-grained operations that could be required to interact with common web applications. Furthermore, some of them cover only specific age groups and cannot be extrapolated to the whole Internet user scope. These gaps lead us to formulate the first hypothesis to be verified in the experiment:

(h_1) There is a direct relationship between age and performance time

In gender related studies, there was evidence of differences in the way women and men use the mouse for common actions like drag-and-drop or point-and-click. However, it is not clear enough whether these differences are significant enough to support user classification while executing fine-grained basic operations, so we formulate the following hypotheses:

(h_2) Women's performance in point-and-click mouse operations is better than men's

(h_3) Men's performance in drag-and-drop operations is better than women's

5.2 Variables in the study

The experiment was completed by 592 participants. Table 5.1 indicates the variables used in the study.

<table>
<tr><td>Name</td><td colspan="3">Definition</td></tr>
<tr><td>Point & Click</td><td colspan="3">Time (T_P) required to pointing (P) each object during the test (measured in milliseconds)</td></tr>
<tr><td>Drag & Drop</td><td colspan="3">Time (T_D) required to drag (D) each object during the test (measured in milliseconds)</td></tr>
<tr><td>Text Selection</td><td colspan="3">Time required to select each word during the test. It was calculated as $T_M + T_P + T_D + T_K$ (measured in milliseconds)</td></tr>
<tr><td>Text Edition</td><td colspan="3">Time required to write each word during the test. It was calculatedasssssss as T_K (measured in milliseconds)</td></tr>
<tr><td>Menu Selection</td><td colspan="3">Time required selecting each menu item during the test. It was calculated as $T_{P1} + T_{K1} + T_M + T_{P2} + T_{K2}$ (measured in milliseconds)</td></tr>
<tr><td>Total time</td><td colspan="3">Sum of $Task_1$ to $Task_5$</td></tr>
<tr><td rowspan="14">Age</td><td>Age Group</td><td>Minimum Age</td><td>Maximum Age</td></tr>
<tr><td>0</td><td>0</td><td>15</td></tr>
<tr><td>1</td><td>16</td><td>20</td></tr>
<tr><td>2</td><td>21</td><td>25</td></tr>
<tr><td>3</td><td>26</td><td>30</td></tr>
<tr><td>4</td><td>31</td><td>35</td></tr>
<tr><td>5</td><td>36</td><td>40</td></tr>
<tr><td>6</td><td>41</td><td>45</td></tr>
<tr><td>7</td><td>46</td><td>50</td></tr>
<tr><td>8</td><td>51</td><td>55</td></tr>
<tr><td>9</td><td>56</td><td>60</td></tr>
<tr><td>10</td><td>61</td><td>65</td></tr>
<tr><td>11</td><td colspan="2">>= 66</td></tr>
<tr><td>HoursUse</td><td colspan="3">Weekly number of hours interacting with computers.</td></tr>
<tr><td>Gender</td><td colspan="3">1 Female, 0 Male</td></tr>
<tr><td>Handedness</td><td colspan="3">1 Left handed, 0 Right handed</td></tr>
</table>

Table 5.1. Independent variables in the study

5.3 Methodology

First, we produced some descriptive statistics about both the dependent and independent variables. The examination of such data gives us a first impression of the features of the individuals in the sample and their behaviour in the experiment. Second, as we had a sample size which is much bigger than those of prior works in this area, we used multivariate statistical analysis instead of the univariate techniques employed by other authors. As explained above, this allowed the inclusion of control variables in the models, as well as the joint analysis of the two studied factors (age and gender).

So, for each of the tasks and the total time, and in order to test hypotheses h_1, h_2 and h_3, we estimated an ordinary least squares (OLS) linear regression model. The regression equations have the following form:

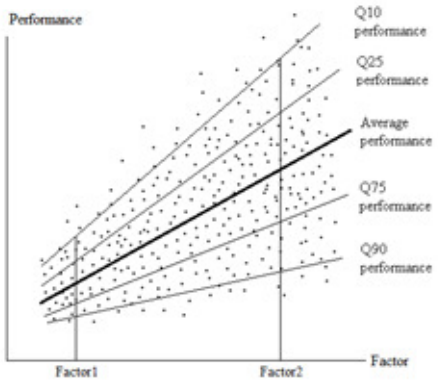


Figure 5.1. Quantile regression approach

$$Task_i = a_0 + a_1 xAge + a_2 xHoursUse + a_3 xGender + a_4 xLeftHanded + \varepsilon$$

where $Task_i$ is the dependent variable in each one of the models, a0 is the intercept term, a_0 to a_5 are the coefficients of the independent variables in the models and ε is the error term.

With regard to these equations, and as prior robustness checks, we tested for multicollinearity and heteroscedasticity. Multicollinearity tests were conducted through the calculation of the condition indices (CI) and the variance inflation factors (VIF). For all the models, the CIs are below 15 and the VIFs below 10, which are common thresholds to discard the presence of significant multi-collinearity among the variables of a linear regression model [90], However, the heteroscedasticity test [22, 35] suggests the presence of a significant level of heteroscedasticity. So, we used robust standard errors for the computation of the p-values which are reported in the results section.

The aforementioned heteroscedasticity in the regression results also suggests that the influence of each one of the hypothesized factors on the performance of the considered tasks could vary depending on the relative level of performance achieved by the individual. That is, the influence of each one of the independent variables could differ depending on whether the individual performs better (or worse) than expected according to his/her characteristics. Figure 5.1 exemplifies this for the univariate relationship between a generic factor and the performance.

It can be seen that the dots representing the behaviour of each one of the individuals follow a heteroscedastic pattern, that is, the dispersion is higher as the level of the factor grows. The central line represents the average relationship between the factor and the measured performance. This line is obtained through standard approaches such as, for example, OLS regression explained above. The slope of this line (coefficient of the regression) is a measure of the mean effect of variations in the factor on the performance. However, Figure 5.1 also shows that as the dispersion of performance increases with the factor, when we trace conditional distributions we obtain lines with different slopes. In the figure, we have represented the lines which pass by the 10, 25,

75 and 90 conditional quantiles. For example, the Q75 performance line indicates that, for each level of the factor, 75% of the individuals having that level perform worse than the threshold stated by the Q75 line. Please note that, as in our case, performance is measured by the time used for the completion of a task; lines for the best performers are below those for bad performers. It is evidenced that, because of the heteroscedasticity, when the level of the factor is low (i.e. Factor1 in the figure) the difference between good and bad performers is not very high. However, for higher levels of the factor (i.e. Factor2 in the figure), the differences are higher.

So, the estimation of conditional regression lines can give us further information with regard to the behaviour of the individuals in the different tests carried out during the experiment. As an estimation method we used linear quantile regression (LQR) [86]. LQR has been applied to a number of research tasks in the field of information systems, e.g., the artificial intelligence system for the design of a consumer credit scoring system[153], the artificial intelligence system for the analysis of firm solvency [40], analysis of downloads from an electronic commercial web site [166]. However, to our knowledge, no prior research papers on usability have applied such a methodology.

We computed 19 quantile regression estimates (5%, 10%, 15%...95%), using the regression equations indicated above. For each of the equations, 500 bootstrapping replications were used for the estimation of the robust standard errors.

Finally, and in order to ascertain whether the performance of individuals with regard to one task is related to the performance in the other tasks we conducted a correlation analysis. We computed both parametric (Pearson) and non-parametric (Spearman) correlation coefficients. For the calculations of these statistics, as well as for all the other tests and equations indicated above, we used the statistical package STATA-11.

5.4 Results

5.4.1 Descriptive statistics

Table 5.2, shows the descriptive statistics for the variables in the study.

Table 5.3, Table 5.4, Table 5.5 and Table 5.6 contain the frequency distributions for the independent variables considered in our model.

Table 5.6 shows that around 11% of the sample are left-handed. That fits with the global proportion of left-handed people (estimated between 8% and 13%).

As we might expect, the average execution times for the first four tests increase depending on their complexity. As we have seen in previous sections, some authors [102, 27, 37] reported a higher level of complexity in the execution of *Drag & Drop* tasks compared with *Point & Click*. Thereby, Table 5.2 shows that *Drag & Drop* tasks required a higher amount of time than *Point & Click* tasks.

	Mean	Std. Dev.	Minimum	Maximum
Point & Click	16864.77	4294.92	9319	45792
Drag & Drop	32832.77	10615.61	19595	159867
Text Selection	59389.97	29777.94	16780	534397
Text Selection	84064.47	53974.94	13689	635472
Menu Selection	61139.34	14069.62	38351	147630
Total Time	254291.30	82849.83	133378	902551

Table 5.2. Descriptive statistics for the dependent variables in the study

Age	Number of observations
0	2
1	85
2	182
3	145
4	77
5	38
6	25
7	13
8	12
9	9
10	2
11	2
TOTAL	592

Table 5.3. Frequency distribution for Age

HoursUse	Number of observations
0	1
1	16
2	58
3	67
4	63
5	33
6	354
TOTAL	592

Table 5.4. Frequency distribution for Hours of Use

Sex	Number of observations
0 (Male)	462
1 (Female)	130
TOTAL	592

Table 5.5. Frequency distribution for Gender

Laterality	Number of observations
0 (Right Handed)	462
1 (Left Handed)	130
TOTAL	592

Table 5.6. Frequency distribution for Laterality

In turn, GOMS predicts that *Text Selection* tasks are more complex than both *Drag & Drop* and *Point & Click* tasks, since they require a higher number of operators ($T_M + T_P + T_D + 2T_K$) that must be applied for more time in order to achieve the goal. Table 5.2 reports a higher average time to complete *Text Selection* tasks than that taken in the first two tests. Similarly, the mean time measured for the *Text Edit* test is consistent with its higher complexity when compared with the previous three tests. Lastly, the relatively low complexity required to achieve menu selection tasks is reflected in its low runtime.

5.4.2 Regression results

Table 5.7 indicates the main results of the six OLS regression models. In each column we show the statistics for each of the dependent variables. The first five cells contain the estimates for each of the independent variables and the intercept. In each cell, the upper figure is the coefficient estimate, that in the middle is the t statistic and the figure shown in the lower part of the cell is the p-value. For each model, the F statistic and its p-value are also shown, as well as the R^2 and the results of the Breusch-Pagan/Cook-Weisberg test on heteroscedasticity (the upper figure is the statistic and the lower figure is the p-value).

First, it is noticeable that the Breusch-Pagan/Cook-Weisberg tests evidence of significant heteroscedasticity in the data. This reinforces the need for the calculation of robust standard errors and robust p-values (including White's correction). This also justifies the estimation of quantile regression equations in order to gain further understanding of the behaviour of the data; second, it is remarkable that although the R^2s are not very high, the F tests we conducted to determine if the coefficients of the variables are jointly equal to zero indicate that all the models are significant, that is, the set of variables considered as a whole has an influence on performance in all tests.

With regard to the parameter estimates, coefficients for age are always positive and significant. This data confirms the first hypothesis (h_1*: there is a direct relationship between age and performance time*) as it means that older users perform worse in all the tasks (needing more time to complete). For the case of *Hours of Use* the result is the opposite, as more hours of computer use always imply a better performance. These results confirm that performance declines with age. These results are similar to those obtained in the research cited in the background section, which focused on observations based on the execution of high-level interaction tasks, mostly related to cognition and perception. Our findings confirm that the same effect is observable at the low level of interaction required by the GOMS analysis, which is mostly based on the human motoric system.

Gender coefficient is significantly positive for *Point & Click*, *Drag & Drop*, *Menu Selection* and the total time, but is not significantly different from zero for the case of *Text Selection* and *Text Edit*. So, women perform worse on *Point & Click*, *Drag & Drop* and *Menu Selection*. These results suggest that hypothesis h_3 *(Men's performance in*

	Point & Click	Drag & Drop	Text Selection	Text Edition	Menu Selection	Total Time
Age Group	500.74	1721.47	1098.10	10118.24	3077.63	16516.21
	4.40	6.86	2.01	6.18	8.82	7.33
	0.000	0.000	0.045	0.000	0.000	0.000
HoursUse	-394.89	-895.07	-1460.54	-8202.50	-1483.98	-12435.01
	-3.02	-2.46	-2.35	-5.81	-4.73	6.33
	0.003	0.014	0.019	0.000	0.000	0.000
Sex	1742.03	5077.81	6656.58	1377.96	6293.93	21148.33
	4.29	3.91	1.42	0.30	4.32	2.52
	0.000	0.000	0.156	0.761	0.000	0.012
LeftHanded	707.98	2986.57	-1598.63	6281.345	107.61	8484.88
	1.18	1.53	-0.64	0.63	0.07	0.69
	0.238	0.127	0.520	0.529	0.946	0.489
Intercept	16759.52	30363.66	61785.65	91378.54	57363.51	257651.30
	19.21	16.99	16.43	10.49	27.88	20.65
	0.000	0.000	0.000	0.000	0.000	0.000
F	17.48	13.46	4.46	17.82	33.27	27.07
(p-value)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)
R^2	11.41%	17.59%	2.19%	20.74%	25.49%	23.78%
Het test	7.07	388.80	107.87	109.79	44.60	53.01
(p-value)	(0.007)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)

Table 5.7. OLS regression results

drag-and-drop operations is better than women's) holds, corroborating the observations made by [77] regarding girls having difficulty using a *Drag & Drop*. However, our data do not support hypothesis h_2 *(Women's performance in point-and-click mouse operations is better than men's)*. This hypothesis was formulated based on the results of [77] that were obtained from children. That leads us to conclude that Inkpen's results are more related to children's different learning styles than directly to gender.

Moreover, left-handed users perform neither significantly better nor significantly worse than right-handed.

In addition to model estimations, we conducted some additional robustness tests with the aim to further check the soundness of our results. First, we computed Cook's D statistics in order to detect the presence of influential cases in the regressions. Ds are always lower than 1 for all the individuals in all the regression models. Second, we tested the existence of non-linear effects for the age variable (that is, whether middle-aged users perform better than both younger and older users). This was done by adding a quadratic term (Age^2) to the equations. None of these terms was found to be significant.

Finally, we re-estimated the models for the different subsamples defined considering the browser used for the test (three subsamples: Chrome, Firefox, IExplorer, as the number of persons using other navigators was not enough to allow regression equation estimation) and the operating system (Windows, Linux and Mac). The results are qualitatively the same as those displayed in Table 5.7. For the sake of brevity, we did not include the results of all these robustness checks. However, they are available from the authors on request.

5.4.3 Quantile regression results

As indicated above, heteroscedasticity in the data suggests that quantile regression could provide further understanding of the behaviour of data. We estimated conditional quantiles 5%, 10%, ... 95% for all the six equations defined above. Due to space limitations, we report only the results which add new information to the discussion in the comments on the OLS regressions results (complete results are also available from the authors on request). These results refer to the behaviour of the gender variable. Table 5.8 indicates the p-value of gender for each of the considered quantiles, and each of the regression models.

With regard to *Point & Click* tasks, it is remarkable that, for the case of the upper quantiles, the effect of gender is less significant (above Q70 gender is not significant at the 1% level). In other words, for low performers gender is not as influential as for intermediate or high performers. For the case of *Drag & Drop*, the results indicate the opposite, that is, for the best performers (quantiles under Q25) the effect of gender is less marked. In the case of *Text Selection* and *Text Edit* tasks the coefficient of gender is not significant for the majority of the quantiles. This is in accordance with

Percentile	Point & Click	Drag & Drop	Text Selection	Text Edition	Menu Selection	Total Time
5	0.001	0.031	0.226	0.003	0.021	0.000
10	0.007	0.019	0.357	0.005	0.001	0.000
15	0.000	0.043	0.195	0.025	0.001	0.000
20	0.000	0.017	0.196	0.110	0.000	0.000
25	0.000	0.003	0.224	0.009	0.000	0.000
30	0.000	0.005	0.525	0.035	0.000	0.000
35	0.002	0.011	0.259	0.074	0.000	0.000
40	0.000	0.002	0.218	0.188	0.000	0.001
45	0.000	0.000	0.152	0.485	0.000	0.004
50	0.000	0.001	0.224	0.369	0.000	0.008
55	0.000	0.000	0.429	0.368	0.000	0.007
60	0.000	0.001	0.593	0.455	0.002	0.009
65	0.000	0.000	0.867	0.277	0.007	0.009
70	0.001	0.000	0.891	0.070	0.001	0.070
75	0.022	0.000	0.732	0.087	0.004	0.145
80	0.011	0.000	0.785	0.157	0.024	0.370
85	0.027	0.014	0.597	0.191	0.005	0.012
90	0.053	0.003	0.790	0.685	0.007	0.026
95	0.047	0.003	0.830	0.466	0.063	0.020

Table 5.8. P-values for the quantile regression for Gender

the results of the OLS models. The results for *Menu Selection* indicate that gender is significant at the 1% level in all cases except some of the extreme quantiles (Q5, Q80, Q95). Finally, the results for the total time suggest that the influence of gender is more significant for intermediate and best performers (only quantiles below Q70 have a p-value which is under 1%).

5.5 Conclusions and Practical Implications

The main goal of this work was to assess whether gender and age are sufficiently significant determining factors in mouse motion behaviour to support an automatic profiling system, as well as to evaluate the role that laterality and user experience have on the overall performance.

Regarding age, the results obtained are consistent with previous research, especially with the work of Hill et al. [70]. The results obtained in tests 1 to 5 show a negative impact on the performance of the GOMS operators P, K, and D when age is increased. The increment in the overall execution time for each task proved to be much bigger in people aged over 40. This result makes us believe that it would be relatively easy to classify people above and below this age using data gathering agents.

With respect to gender, the results obtained are consistent with the findings of Inkpen [77] who found that girls perform worse with regard to drag-and-drop tasks. In our study males obtained better results when executing interaction operators based on pointing and dragging, which are required by the *Point & Click*, *Drag & Drop* and *Menu Selection* tasks. Nevertheless, there were no significant differences in the typing operators required by the *Text Selection* and *Text Edit* tasks. The differences were also so relevant that they can be used for the design of algorithms for automatic classification.

Surprisingly, no significant performance differences were detected between left- and right-handed users and we did not find any evidence regarding the influence of this factor on the overall performance.

The performance differences in gender and age are relevant enough to be gathered by a data gathering agent hidden in the user interface of the online information system. Since the main differences in performance are based on mouse pointing tasks (which are used more frequently than those based on text editing) the gender and age of the user could be estimated automatically shortly after user arrival on the web site.

Once the feasibility of using age and gender in the automatic classification of users on certain parameters of web interaction performance has been demonstrated, the next step is determining the best strategy to implement a classification device. Several systems using different machine learning strategies (neural networks, regression trees, etc.) are suitable for such applications. A profitable avenue of research could be the comparison of the accuracy of different machine learning models. Although this

research found significant differences in the studied variables regarding their influence on performance, there are still other important factors that could contribute to the development of an accurate automatic profiling system.

This analysis was published in Online Information Review under the name of "*Towards an automatic user profiling system for online information sites.*" [41] attached in the Appendix C.

6 Part 3: Analysis of the correlation of Interaction tasks' significance in age and gender based users classification

6.1 Introduction

The related previous studies evidence that there are significant differences between the times required by children and adults to execute different basic interactions tasks. However, to date, we found no studies evidencing these differences in adults, something that led us to conjecture that the performance of one specific adult in these tasks could be correlated. If so, it would mean that the analysis of performance in one of them would be enough to identify adults, simplifying users' age classification. On the other hand, even though there is no evidence of differences between genders in adults for these basic interaction tasks, some studies identified some differences between men and women in other activities that could determine the correlation between the performances in these basic interaction tasks. That leads us to wonder whether these correlations could be determined by the user's gender. Therefore, we formulate the following hypotheses to be verified/refuted by the empirical study:

- *(h_1) The execution time of the different tasks increases with the age of the subject under study*
- *(h_2) Women's execution time for the different tasks is longer than men's*
- *(h_2) The execution times of basic interaction tasks (Point & Click, Drag & Drop and Item Selection) are significantly correlated*

To assess whether the hypothesis formulated in the prior section holds, the performance of 592 individuals was analyzed in the execution of three basic interaction tasks.

6.2 Variables of the study

Apart from the variables used to test our first two hypotheses (age and gender) and the execution times of the analyzed tasks we considered some additional variables for the testing of h_1 and h_2.

Specifically, we included handedness and previous user experience with computers. Several studies reported differences regarding movement between left and right handed individuals [93, 105, 145], movement preparation [68, 14, 10, 109], stimulus velocity effect [123] and interactions between hand preference and hand performance [114]. Besides this, other studies suggest that skill performance and the amount of practice are correlated [73] following an exponential law [67].

These precedents suggest that users' handedness and the users' experience may have a sensible influence on the user behavior and therefore in their execution time. As a result, these two variables (handedness and amount of practice) were incorporated as control variables in the regression models that are explained in the next section.

As was noted above, the experiment was completed by 592 participants. As a summary, we indicate in Tables 6.1 and 6.2 the variables used in the study. Before participating in the tests, users were asked to fulfill a questionnaire to provide information about their age, gender, handedness (tendency to use either the right or the left hand) and experience in the use of computers. This last parameter was provided in terms of the number of weekly hours spent by the users interacting with computers. Some of the users were reluctant to provide their actual age (especially older users). As a result, we were forced to discretize the age value in ranges of 5 years. This way we sacrifice some of the statistical analysis to obtain this parameter from all the users participating in the tests.

Name	Definition
Point & Click	Time (T_P) required to pointing (P) each object during the test (measured in milliseconds)
Drag & Drop	Time (T_D) required to drag (D) each object during the test (measured in milliseconds)
Item Selection	Time required selecting each menu item during the test. It was calculated as $T_{P1} + T_{K1} + T_M + T_{P2} + T_{K2}$ (measured in milliseconds)

Table 6.1. Dependent variables in the study

Age	Age Group	Minimum Age	Maximum Age
	0	0	15
	1	16	20
	2	21	25
	3	26	30
	4	31	35
	5	36	40
	6	41	45
	7	46	50
	8	51	55
	9	56	60
	10	61	65
	11	>= 66	
HoursUse	Weekly number of hours interacting with computers.		
Gender	1 Female, 0 Male		
Handedness	1 Left handed, 0 Right handed		

Table 6.2. Independent variables in the study

6.3 Statistical Methods

First, we computed some descriptive statistics about both the dependent and independent variables. The exam of such data gives us a first idea of the features of the individuals in the sample and their behavior in the experiment.

Second, to test hypotheses h_1 and h_2, we estimated a Linear Regression model for each of the tasks. The regression equations have the following form:

$$Task_i = a_0 + a_1\ x\ Age + a_2\ x\ Gender + a_3\ x\ HoursUse + a_4\ x\ LeftHanded + \epsilon_i$$

Where $Task_i$ is the dependent variable in each one of the models, a_0 is the intercept term, a_1 to a_4 are the coefficients of the independent variables in the models and ϵ_i is the error term.

Regarding these equations, and as prior robustness checks, we tested for multicollinearity and heteroskedasticity. Multicollinearity tests were conducted through the calculation of the Condition Indices (CI) and the Variance Inflation Factors (VIF). To assess whether heteroskedasticity represents a problem we used the Cook and Weisberg test [35].

Furthermore, we also conducted some post-estimation additional tests which allow shedding light on specific concerns about whether a) there are extreme values which have an abnormal influence on the results, b) the model is not correctly specified and c) results are sensitive about the browser/operating system used. First, to detect the presence of influential cases we computed Cook's D statistics for each data point in the regressions. Second, and regarding model specification, we tested for the existence

of non-linear effects for the age variable (that is, whether middle-age users perform better than both younger and older users). This was done by adding a quadratic term (Age^2) to the equations and reestimating the models. Finally, we also re-estimated the models for different subsamples defined considering the browser used for the test (three subsamples: Chrome, Firefox, IExplorer, as the number of persons using other navigators was not enough to allow regression equation estimation) and the operating system (Windows, Linux and Mac).

Finally, and to know whether the execution time of individuals about one task is related to the performance in the other tasks (h_3) we conducted a correlation analysis. We computed Nonparametric correlation coefficients (Spearman's Rho) to avoid the problems caused by nonnormality of data. To test normality of data we used the Lilliefors test, and in all cases data distributions departed significantly from normality (results not reported due to space limitations). For the calculations of these statistics, as well as for all the other tests and equations indicated above, we used the statistical package STATA 11.

6.4 Results

6.4.1 Descriptive statistics

	Mean	**Std. Dev.**	**Min.**	**Max.**
Point & Click	16864.77	4294.92	9319	45792
Drag & Drop	32832.77	10615.61	19595	159867
Item Selection	61139.34	14069.62	38351	147630

Table 6.3. Descriptive statistics for the dependent variables in the study (execution times are measured in milliseconds).

Age	Number of observations
0	2
1	85
2	182
3	145
4	77
5	38
6	25
7	13
8	12
9	9
10	2
11	2
TOTAL	592

Table 6.4. Frequency distribution for Age.

Gender	Number of observations
0 (Male)	462
1 (Female)	130
TOTAL	592

Table 6.5. Frequency distribution for Gender.

HoursUse	Number of observations
0	1
1	16
2	58
3	67
4	63
5	33
6	354
TOTAL	592

Table 6.6. Frequency distribution for HoursUse.

As we might expect the average execution time depends on the complexity of the test. As mentioned previously, some authors [102, 27, 37] reported a higher level of complexity in the execution of *Drag & Drop* tasks when compared with *Point & Click*. Thereby, Table 6.3 shows that *Drag & Drop* tasks require a higher amount of time than *Point & Click* tasks. Furthermore, the runtime of menu selection tasks is higher than that of the other two. This result is consistent with predictions provided by GOMS analysis studied before. Notice that while the *Point & Click* and the *Drag & Drop* tasks required the execution of single P or D operator, the *Item Selection* tasks requires the execution of two P operators (one for menu activation and another one for item selection). Besides that, item selection requires the execution of a complex M operator to take the decision of what item to select.

With regard to the sample descriptive indicators (tables 6.4 to 6.7) it is noticeable that the sample is mainly composed by individuals which are young, male and have intensive experience in the use of computers. However, the number of observations that correspond to the other types of web applications users (women, elder and low experienced users) is sufficient to conduct a valid statistical study. Furthermore, and regarding handedness, around 11% of the individuals in the sample are left-handed. This value is consistent with the global rate of left-handed people, that is estimated between 10% and 13% [121].

Handedness	Number of observations
0 (Right Handed)	524
1 (Left Handed)	68
TOTAL	592

Table 6.7. Frequency distribution for handedness.

6.4.2 Regression analysis results (hypotheses h_1 and h_2)

Table 6.8 indicates the main results of the three regression models and the related tests. Prior to the comment of the results we must highlight that all CIs of the different variables in the three regression models are below 15. In accordance with this, all VIFs are below 10. These values are common thresholds to discard the presence of significant multicollinearity among the variables of a linear regression model [90]. For the sake of clarity in the presentation of the results we do not include CI and VIF values in table 6.8, but data are available from the authors upon request.

Results of the Cook-Weisberg test for heteroscedasticity are detailed in the last row of the table. We indicate the values of the chi-squared test statistic and the corresponding p value. As the null hypothesis for this test is that variance is constant we can conclude that such hypothesis is rejected in the three cases and heteroscedasticity is significant. So, we repeated the estimation of the regression equations using a robust estimation procedure, which consisted in the calculation of robust standard errors for the coefficients in the different regression equations and robust p-values, including White's correction [152]. Such results are those displayed in table 6.8. The layout of the rest of the table is as follows: in each column, we show the statistics for each one of the regression equations (where, the dependent variables are, respectively, time for completion of point & click, drag & drop, and item selection tasks). The first five cells of each column contain the estimates for each one of the independent variables and the intercept of each model. In each cell, the upper figure is the coefficient estimate, that in the middle is the robust t statistic (computed using the standard error that includes White's correction) and the figure shown in the lower part of the cell is the robust p-value. In addition, the table displays for each model, the F statistics for the test of the joint significance of the coefficients and its p-value, as well as the adjusted R^2 and the results of the aforementioned Cook-Weisberg test for heteroscedasticity.

With regard to the results, we must first underline that although R^2s are not very high, conducted F tests evidence the joint significance of the coefficients of the variables, that is, the set of variables, considered as a unit, influence the performance in all the tests.

Regarding the parameter estimates, coefficients for age are always positive and significant. These results give support to the first hypothesis (h_1) as it is evidenced that older users perform worse for all the tasks (needing more time to complete. So, the performance decline regarding the age is confirmed.

Furthermore, the gender coefficient is significantly positive in all cases. So, women perform worse on *Point & Click*, *Drag & Drop* and *Item Selection* tasks. These results suggest that hypothesis h_2 also holds corroborating the observations made by Inkpen [77] regarding girls having difficulty with *Drag & Drop* tasks. However, our data does not support Inkpen's other observations related to the absence of any significant gender difference in the overall movement time. That leads us to conclude that Inkpen's results are more related to children's different learning styles than directly to the gender.

	Point & Click	Drag & Drop	Item Selection
Age Group			
Parameter estimate	500.74	1721.47	3077.63
t statistic	4.40	6.86	8.82
p-value	<0.001	<0.001	<0.001
Gender			
Parameter estimate	1742.03	5077.81	6293.93
t statistic	4.29	3.91	4.32
p-value	<0.001	<0.001	<0.001
HoursUse			
Parameter estimate	-394.89	-895.07	-1481.98
t statistic	-3.02	-2.46	-4.73
p-value	0.003	0.014	<0.001
Handedness			
Parameter estimate	707.98	2986.57	107.61
t statistic	1.18	1.53	0.07
p-value	0.238	0.127	0.946
Intercept			
Parameter estimate	16759.52	30363.66	57363.51
t statistic	19.21	16.99	27.88
p-value	<0.001	<0.001	<0.001
F test			
F-statistic	17.48	13.46	33.27
p-value	<0.001	<0.001	<0.001
Adjusted R^2	11.41%	17.59%	25.49%
Cook-Weisberg test for heteroscedasticity			
Chi-squared	7.07	388.80	44.60
p-value	0.007	<0.001	<0.001

Table 6.8. Regressions results and related tests.

Regarding the control variables in the model, it is first noticeable that prior experience with computers is significant in all cases. Coefficients for *HoursUse* are significantly negative in all cases, meaning that more hours of computer use always imply a better performance. These results are similar to those obtained by a prior study [41] which was focused on observations based on the execution of top level interaction tasks, mostly related with cognition and perception. Our findings confirm that the same effect is observable at the low level of interaction required by the GOMS analysis, which is mostly based on the human motoric system. With regard to the other control variable, handedness does not seem to have an influence, as left-handed users perform neither significantly best nor significantly worse than right-handed.

With respect to the additional post-estimation tests, we must first underline that Cook's D values are always lower than 1 for all the individuals in all the regressions so there are no influential cases in the models. Second, none of the quadratic terms (Age^2) that we included in alternative versions of the equations was found to be significant. So, we can reject the existence of non-linear effects for the age variable. Finally, the

re-estimation of the models for different subsamples defined considering the browser used for the test (Chrome, Firefox, IExplorer,) and the operating system (Windows, Linux, Mac) produced results which are qualitatively the same as those displayed in Table 6.8. For the sake of brevity, we did not include in the paper the results. However, they are available from the authors upon request.

6.4.3 Correlation analysis results (hypothesis h_3)

The results of the correlation analysis we conducted to assess whether individuals that perform well in a certain task also perform well in the others (h_3) are shown in table 6.9. In each of the cells we display the non-parametric Spearman correlation coefficient (upper figure) and the corresponding p-value (lower figure). Cells below the main diagonal contain the results of tests.

	Point & Click	Drag & Drop	Item Selection
Point & Click			
Drag & Drop	0.731 0.000		
Item Selection	0.660 0.000	0.674 0.000	

Table 6.9. Results of the correlation analysis.

Data in table 6.9 evidence that correlations are significant among all the tasks. This finding supports the hypothesis h_3, suggesting that the execution time (performance) of an individual in a specific task, keeps its coherence in the other tasks as well. So, for example, if a person has superior performance in the execution *Point & Click* tasks, she/he is expected to also have superior performance in the execution of *Drag & Drop* and *Item Selection* tasks.

This finding may have a relevant impact in the future design of automatic user modeling algorithms. As the three proposed interaction tasks have the same usefulness in terms of user categorization, any of them can be used separately to achieve this goal. Moreover, the amount of data required to automatically infer the type of user may be notably reduced (as only one task is analyzed), which is crucial for the execution of real time algorithms.

6.5 Conclusions

This work had two interrelated goals. First, we wanted to assess whether the gender and age are sufficiently significant determining factors to support an automatic profiling system based on the analysis of mouse motion behavior when executing *Point & Click*, *Drag & Drop* and *Item Selection* tasks. Second, to figure out whether the individuals

perform consistently across these basic interaction tasks, that is, if their performance in one of them are extrapolable (or not) to the others.

Regarding the first, the results of the empirical study reveal that both age and gender factors are significantly determinant. While older users performed worse than younger in each of the interaction tasks, men obtained better results than women. On the other hand, in relation with the analysis of correlations between the execution times of the target basic interaction tasks, data gathered in the tests revealed consistency in the execution times of individuals across them. User's performance measured in any of these tasks is coherent to their execution time in the other tasks.

These results open the door to implement a system that automatically classifies users in age and gender groups by observing the way they interact and perform in these basic interaction tasks with any web interface. However, this evidence must be taken carefully, given that the data was gathered through artificial and isolated ad-hoc tests, and not in a real web interface where the behavior of the user can differ from the one evidenced during the tests. On the other hand, the existing correlation between the way individuals perform across the different interaction tasks makes it more flexible not only to integrate the data gathering processes into the final system (since developers do not need to force the use of all of them), but also it expands the data gathering possibilities to a number of observations whose results could be combined in a hybrid voting algorithm or a machine learning based system.

This analysis was published in Computer Standards & Interfaces under the name of "*The dimension of age and gender as user model demographic factors for automatic personalization in e-commerce sites.*" [50] attached in the Appendix D.

7 Part 4: Implementation of a ML-based operative classification system

7.1 Introduction

The main objective of this study is to evaluate different Machine Learning (ML) algorithms and compare each other in order to classify users by age or gender, based on the user's interactions on a website by using a machine learning approach.

The quick identification of common interaction patterns among users is critical to web designers in order to adapt their information systems to meet the users' diverse needs [82]. This user profiling can be determined by several factors that have a great relevance in the success of the online information systems of companies [134, 144] since differences in age, gender, and income indicate different interaction intentions [158]. There is strong empirical evidence showing that important behaviour differences caused by gender, age, social grouping, and household income have a great influence on how the user accesses online services [65, 92].

The measurement and analysis of the performance in certain tasks could help to draft a user's profile. Age and gender are two user features which could be a key to obtaining a profile of an individual and whose influence in user behaviour has been reported in some scenarios [151, 55]. However, these specific experiments do not demonstrate if this influence is significant enough to support user classification in the context of web interaction. The lack of control for these factors may bias the results as they may have an influence on the performance level.

This work intends to fill this gap by measuring the performance of an algorithm classifying users based on their age or gender (which for the first time includes individuals of both genders and of all ages) performing – also for first time – five basic tasks required to interact with information systems: (i) *Point & Click*, (ii) *Drag & Drop*, (iii) *Text Selection*, (iv) *Text Edit* and (v) *Menu Item Selection*. The automatic classification of users according to their ages or gender can be done by analyzing the performance of the users when they perform these tasks, comparing it with the performance of individuals whose demographic characteristics are known.

Data were gathered with a test that collected the performance metrics while the user executed five different tasks corresponding to the basic interaction tasks while navigating a website. These tasks were: *Point & Click*, *Drag & Drop*, *Select text*, *Edit Text* or *Select Items* on a menu. Machine learning classification requires input features or dimensions. In our scenario, these features are represented by the time spent by the user in each task of our test. In previous research, we concluded that age and gender influenced users' performance while executing these web interactions[113, 41]. Thus, even if we gather more characteristics about the users like the browsers used, their laterality(right or left-handed), etc., we chose age and gender as the two demographic characteristics to do our classification study [50]. In order to create simpler models, we splitted this classification analysis in two, one based on the gender and the other one on the age.

Regarding age, it has been proved in a previous research that there is a evident threshold at the age of 40 [41] where the performance decreased noticeably at that point. Thus, to simplify our analysis using a binary classifier and gain accuracy on the predictions, we regrouped the individuals in two groups based on their age, under and above 40 years old.

Along this study, different algorithms of Machine learning were used: the algorithms used were Bagging, Random forest, AdaBoost.M1, LogitBoost, SVM (SMO), C4.5(J48), Logistic and Stacking. They were executed 10 times using 10-fold cross validation. The performance of the results were evaluated using the AUC-PR metric among other metrics.

The rest of the paper is organized as follows. Following the introduction, Section 2 gives a background introduction on age and gender differences. Section 3 explains the experiment that was conducted to gather the data needed for our research, the tested machine learning algorithms and the validation methodology used to measure their performance. In Section 4, the obtained results are shown. Finally, Section 5 includes a summary with the conclusions and the future work.

7.1.1 Tested Machine Learning Algorithms

After gathering all the data from the interaction experiment explained in Chapter 3 and carrying out all the studies detailed in Chapters 4, 5, and 6, we found enough evidence that age and gender interfere in the user's performance while interacting with the computers. Because of that, we continued our research exploring whether with Machine Learning algorithms, we will be able to classify the user's based on their age or gender.

The Machine Learning approach aims to find a relationship between an input X = $\{x_1, x_2, \ldots, x_N\}$ and an output Y. In our case, we inferred the relationship between the user's performance and the participant's age and gender. In other words, we determine whether a participant is a male or female in the case of the gender, or if he is under

or above 40 years old based on the performance times obtained in our tests. More precisely, we aimed at determining the best combination(s) of features (times gathered during the *Point & Click*, *Drag & Drop*, *Select text*, *Text Edit* and *Menu Selection* tasks) showing the most predictive power in these binary classification tasks. In this chapter, theoretical explanations of several machine learning concepts are given. In order to fit the models, Weka was used [64].

Nowadays, the most used algorithms are the ensemble learning algorithms. Their goal is to construct a collection (an ensemble) of individual classifiers that are diverse and yet accurate. If this can be achieved, then highly accurate classification decisions can be obtained by voting the decisions of the individual classifiers in the ensemble. Many authors have demonstrated significant performance improvements through ensemble methods [17, 87, 5, 103]. Two of the most popular techniques for constructing ensembles are bootstrap aggregation ("Bagging" [16]) and the Adaboost family of algorithms ("Boosting" [57]). Both of these methods operate by taking a base learning algorithm and invoking it many times with different training sets. Breiman [15] explores the causes of instability in learning algorithms and discusses ways of reducing or eliminating it. Bagging (and to a lesser extent, Boosting) can be viewed as ways of exploiting this instability to improve classification accuracy.

We evaluated the following set of classifier families: Bagging (Bagging, Random forest), Boosting (AdaBoost.M1, LogitBoost), SVM(SMO), C4.5(J48), Logistic and Stacking. These algorithms were chosen due to their good results obtained in other studies like the one from Goh et al [62], Breiman [16], Freund [57] and others. In addition, other reasons to chose them was that the use of boosting based algorithms and cost-sensitive learning has been proved work better with unbalanced data as it is our case [30, 45].

Bagging Algorithms

In bagging, each training set is constructed by forming a bootstrap replica of the original training set. In other words, given a training set A with N number of examples, a new training set A' is constructed by drawing N examples uniformly (with replacement) from A. Bagging generates diverse classifiers only if the base learning algorithm is unstable or in other words, small changes to the training set can cause large changes in the learned classifier.

Bagging Bagging utilizes the Breiman's bagging procedure[16]. In statistics, bootstrap is a technique that allows any test or metric rely on random sampling with replacement. Bootstrapping allows assigning measures of accuracy (defined in terms of bias, variance, confidence intervals, prediction error or some other such measure) to sample estimates[49]. The word Bagging comes from bootstrap aggregation as it uses this technique and generates bootstrap samples of the training data. It builds the distinctive training set consisting of frequent and plentiful data sets. Classification

results are based upon the highest number of votes. In this study, Bagging was used using the RepTree algorithm as classifier.

Random Forest Random Forest, which was proposed by Leo Breimans in 1996[19, 20], is considered to be a special type of ensembles using bagging[16] and random splitting methods[71] for growing multiple trees[19, 20]. It is a combination of tree predictors such that each tree depends on the values of a bootstrap sampled independently and with the same distribution for all trees in the forest to fit a classification tree. At each node, it selects several features at random from all possible features, which are independent at each node. According to some objective function, the feature which can provide the best split is used to do a binary split on that node. Once the best split point is found on the selected variables, the tree grows bigger. Then, at the next node, another feature is chosen with the same number at random and the same process is repeated. In order to calculate the random number, Breiman[20] suggested the possible values as square root of the number of features, half of square-rooted number of features or twice of the square-rooted number. Because of the randomness, Random Forest can avoid overfitting in most of the cases and is relatively robust against outliers and noise.

It has shown robust and improved results of classifications on standard data sets. It is throwing very good competition to neural networks, other ensemble techniques and support vector machines on various classification problems[165]. In addition, It is highly recommended the use of random forest classifiers when learning from imbalanced data[85].

Boosting Algorithms

Boosting is a method of converting weak learners into strong learners. In boosting, each new tree is a fit on a modified version of the original data set. The previous two boosting algorithms [130, 56] were designed based on the assumption that a uniform upper bound, strictly smaller than $1/2$, exists on the weighted error of all weak hypotheses. In practice, the common behavior of learning algorithms is that their error gradually increases with the number of boosting iterations and as a result, the number of boosting iterations required for AdaBoost and their followers is far smaller than the number of iterations required for the previous boosting algorithms.

AdaBoost In the AdaBoost Algorithm[18] the weak learner is defined as a classifier that is slightly correlated to true classification, but slightly better than random probability. According to Friedman[59], each step consists of a weighted minimization and recomputation of both classifier and weight.

In other words, the Adaboost algorithm maintains a set of weights over the original training set and adjusts these weights after each classifier is learned by the base learning algorithm. The adjustments increase the weight of examples that are misclassified by

the base learning algorithm and decrease the weight of examples that are correctly classified. There are two ways that Adaboost can use these weights to construct a new training set to use in the base learning algorithm. One method is boosting by sampling, in which examples are drawn with replacement from the training set with probability proportional to their weights. The second method, boosting by weighting, can be used with base learning algorithms that can accept a weighted training set directly. With such algorithms, the entire training set (with associated weights) is given to the base learning algorithm.

Adaboost requires less instability than bagging, because Adaboost can make much larger changes in the training set (e.g., by placing large weights on only a few subset of the examples). The AdaBoost boosting algorithm has become over the last few years a very popular algorithm to use in practice[74, 147, 3]. The two main reasons for this popularity are simplicity and adaptivity[44]. We say that AdaBoost is adaptive because the amount of update is chosen as a function of the weighted error of the hypotheses generated by the weak learner. While the success of AdaBoost is indisputable, there is increasing evidence that the algorithm is quite susceptible to noise into the training data (i.e., training and test examples with incorrect class labels)[43]. During this study, the noise didn't affect as the used data was well labeled by each individual during the realization of our interaction test.

AdaBoost.M1[57] is a version of AdaBoost which is based on boosting by weighting. Like AdaBoost, it requires a weak classifier algorithm. In this research, AdaBoost.M1 we used with Random Forest as the weak classifier[100]. The Adaboost.M1-RF algorithm combined the merit of the Adaboost and the Random Forest. It also has been proved that the AdaBoost.MI-RF algorithm is better than the other weak learners in Adaboost, such as DecisionStump and J48[165]. This integration manifests improved metrics of performance in binary classification in different scenarios like tumor classes[124], traffic predictions[94, 95], and others [13, 12].

LogitBoost It was observed [21] that AdaBoost had very good generalization (the ability to classify new data), but the exponential loss function of AdaBoost algorithm changes exponentially with the classification error, rendering it vulnerable while handling noisy data. To overcome such a problem, Friedman et al. [58] proposed to use LogitBoost, which can reduce training errors linearly and hence yield better generalization. LogitBoost classifier is based on AdaBoost procedure using an adaptive Newton algorithm to fit an additive multiple logistic regression model[58] and changing the exponential loss of the Adaboost algorithm to a conditional Bernoulli possibility loss[111].

Decision tree is a quite suitable model to build the weak classifiers for LogitBoost[163, 167]. In this study, Random Forest was used as the base procedure for boosting (AdaBoost and LogitBoost).

Support Vector Machines

A support vector machine(SVM)[132] is a linear classifier which attempts to learn the maximum margin hyperplane separating two classes in the data. This maximum margin hyperplane is usually determined by a small subset of instances called support vectors. In order to learn nonlinear decision boundaries, the data can be transformed by a kernel function. The linear maximum margin hyperplane then found for the transformed data can represent a nonlinear partitioning of the original feature space. SMO (Sequential Minimal Optimization) is the implementation used of SVM[116]. This implementation solves the SVM QP (support vector machine quadratic programming) problem by decomposing it into SVM QP Sub-problems and solving the smallest possible optimization problem, involving the two Lagrange multipliers, at each step. The advantage of SMO respect SVM is that it avoids QP optimization that was made in SVM calling a QP library routine; instead, SMO can express the inner loop algorithm in a short amount of C code.

C4.5

The C4.5 algorithm, called J48 in Weka, is a benchmark decision tree algorithm proposed by Quinlan[128] and is based on the univariate decision trees technique. In this technique, the decision tree is built using an entropy-based splitting criterion performed by using one attribute at internal nodes. J48 algorithm is an extension of the ID3 algorithm[118] that uses a divide and conquer approach to growing decision trees[119].

The steps needed to construct a tree are[11]: First, check whether all cases belong to the same class, then the tree is a leaf and is labeled with t indicator, which measures the amount of disorder of the data. Finally, find the best splitting attribute[88].

Pruning is a very important technique to be used in tree creation because of outliers. It also addresses overfitting. Datasets may contain little subsets of instances that are not well defined. To classify them correctly, pruning can be used. Separate and Conquer rule learning algorithms are the basis to prune any tree.

To conclude, J48 is quite a robust algorithm that improves Quinlan's older ID3 decision tree algorithm by adding support for tree pruning and dealing with missing values and numeric attributes in spite of its run-time complexity depending on the size of the tree. It also remains as a popular classifier for research on the class imbalance problem [29].

Logistic

Logistic Regression is a classification technique for building models with a ridge estimator. The ridge estimator improves the parameter calculation and diminishes the error made by further predictions[91]. It works on the association between categor-

ical dependent variables, that has only two possible values, and a set of independent variables. Logistic Regression is used to predict this binary outcome by estimating the parameters of a logistic model. In the logistic model, the log-odds for the value labeled "1" is a linear combination of one or more independent variables. Although original Logistic Regression[91] does not deal with instance weights, in this study the version used is a version that is a little bit modified to handle the instance weights.

Stacking

Stacked generalization[154], or stacking, is an approach for constructing classifier ensembles. A classifier ensemble is a set of classifiers whose individual decisions are combined in some way to classify new instances[142]. "Stacking" is a technique in which the predictions of a collection of models are given as inputs to a second-level learning algorithm. This second-level algorithm is trained to combine the model predictions optimally to form a final set of predictions. Ground-level classifiers often make different classification errors.

AdaBoost.M1 and Random Forest were used as the classifiers and LogitBoost was the meta classifier, that is, the second-level algorithm. These classifiers were used due to their good performance in other studies [31].

7.1.2 Performance Metrics

Classification accuracy is the most commonly used metric representing the percentage of correctly predicted examples. However, it is cost insensitive and attaches the same loss to different types of errors. In cases of extremely imbalanced datasets or cost-sensitive learning, accuracy fails and precision/recall statistics might be used. They are defined as a percentage of positive examples in population and a percentage of true positives in the population of positively classified examples, respectively, and they are usually used together. We can either use precision/recall for a given threshold or we can generate a function of threshold which returns the ratios. The function that will give us those thresholds, can be simplified by a single number as the so-called F-measure developed by Van Rijsbergen [122]. The F-score provides a measure for how well a binary classifier can classify positive cases (given a threshold value). The F-score is calculated from the harmonic mean of the precision and recall. An F-score of 1 means both precision and recall are perfect and the model correctly identified all the positive cases and didn't mark a negative case as a positive case. If either precision or recall are very low it will be reflected with a F-score closer to 0.

Nowadays, many researchers found[96, 69] that a method of evaluating performance is finding the optimal threshold if the model predictions are in the form of real numbers or probability, called ROC (Receiver Operator Characteristic) analysis. This method is based on dividing the prediction using all possible thresholds and calculating the specific measures for each threshold and plot the graph for each measure. The ROC

plot shows the dependency between specificity and sensitivity. Resulting curve starts from (0,0) and ends at (1,1) position. The better a model is, the closer the curve reaches to (0,1). The result of ROC analysis is a performance measure called Area Under Curve, simply as AUC[97]. When AUC is 1, then it is a perfect prediction. Ideally, higher AUC shows the model performs better, while a random classification would score 0.5. This is the recommended metric score for the classifiers by Provost[117]

ROC curves are commonly used to present results for binary decision problems in machine learning. However, when dealing with highly skewed datasets, Precision-Recall (PR) curves give a more informative picture of an algorithm's performance[39]. The metric in charge of measuring the performance of this curve is called Area Under Precision-Recall Curve (AUC-PR)[39, 99].

The literature has suggested that the "information loss" assumption may be more suitable for some highly imbalanced data sets, particularly with F-measure and the use of the AUC-PR metrics as an evaluation parameters for prediction models which have imbalanced data as a defect[140, 84]. The Root Mean Square Error (RMSE) is another well known used metric [6, 23]

Working with unbalanced datasets is one of the biggest challenges in data mining[80, 29]. The class imbalance problem arises when the class of interest is relatively rare as compared with other class(es). In this case, we will assume that the positive class (or class of interest) is the minority class, and the negative class is the majority class. The dataset used in this study contains unbalanced data as we will show in the following sections. The minority classes are represented by individuals above 40 years old (in the case of age) or females (in case of gender). Both classes will be considered as positive classes, and for that, labeled as 1. On the contrary, males and individuals under 40 years old are more common and were considered as class label 0 or negative classes.

One commonly used strategy for handling the class imbalance problem involves re-sampling techniques[78, 79], which aim to balance the class distributions of a data-set before feeding the output into a classification algorithm. There are two main re-sampling techniques: over-sampling techniques, which amplify positive instances (class of interest), and undersampling techniques, which suppress negative instances. However, over-sampling techniques are negatively impacted by the overfitting problem[141], while under-sampling techniques typically suppress important parts of a dataset. Other strategies deal with this problem differently, for example a boosting based algorithm[30] and cost-sensitive learning [45].

To assess the performance of this study, we report a list of selected metrics, mostly based on current practices in the field (AUC, AUC-PR, F-score, RMSE, TPR(or recall) and TNR(or specificity)). Despite reporting 6 performance measures, the final ordering was generated by AUC-PR, as it is the most widely accepted regarding our unbalanced case[140, 84]. In addition, boosting based algorithms and cost-sensitive learning were used to avoid the unbalanced data problem[30, 45].

7.1.3 Validation Strategy

The validation strategy is commonly used to prevent overfitting and to have a good assessment of model validity[26]. There are some procedures that are frequently used, here we exposed the evolution of some of them. Cross-validation (CV)[160, 136, 61, 136, 159] is a procedure for estimating the generalization performance in data mining. Cross-validation is the most commonly used method for predictive performance evaluation of a model[48]. Data is split usually into two parts and based on this splitting, on one part, training is done while the predictive performance is tested on the other part. Thus, cross-validation is widely accepted in the data mining and machine learning community, and serves as a standard procedure for the sake of model selection or modeling procedure selection[66, 164].

K-fold Cross-validation is an improvement of the cross validation procedure[160]. With this technique, the data is equally partitioned into k equal or nearly equal segments or folds. On these partitioned folds, training and testing is done in k iterations such that in each iteration, we leave one fold for testing and train the model on the remaining k-1 folds. The accuracy obtained in each iteration is then averaged to get the model accuracy. An important thing to note is that data is commonly stratified before being split into k segments. Stratification is the process of rearranging data in such a way that each fold is a good representative of the whole. In 10-fold cross validation, the model is trained and tested 10 different times and then, mean accuracy is considered as the accuracy of the model. In other words, 10-fold cross validation produces 10 equal sized sets. Each set is divided into two groups: 90 labeled data, that will be used for training, and 10 labeled data, that will be used for testing. Now, each set will use the training labeled data to train the model and will get its performance using its testing data group, producing as the result the prediction performance. After repeating it among the 10 folds, ten prediction performances will be obtained. The procedure will average those to obtain the final prediction performance of the model.

To obtain reliable performance estimation or comparison, large numbers of estimates are always preferred. In k-fold cross-validation, the number of estimates obtained is equal to k. In order to achieve a further increase in the number of estimates, we run k-fold cross-validation multiple times. The data is reshuffled and re-stratified before each round. This is called Repeated k-fold Cross-validation. In other words, with 10-fold cross validation, as we mentioned before, we will obtain 10 prediction performances for each 10-fold cross validation. If we repeat this process 10 times, reshuffling and re-stratifying the data before each round, we will obtain 100 prediction performances (10 times x 10 performances = 100). To get the final prediction performance for the model the average is calculated.

Within this experiment, in order to obtain reliable performances, we used a 10 repeated 10-fold cross validation procedure in order to get the models performances. As mentioned before, it will give us the mean of 100 performances for each model.

Furthermore, after obtaining those metrics, we need to compare the performance of each model against the others to be able to know which one fits better with our dataset. In order to do that, we used the paired two-sample t-test comparing the means of the AUC-PR of the different models considered two by two. The t-test, is used to determine whether the means of two groups are equal to each other. The null hypothesis(H0) is that the two means of the AUC-PR of each pair of models are equal, and the alternative is that they are not. This test is carried out at the 0.05 level of significance. In addition, we calculate a confidence interval of 95% for their mean value.

7.2 Results

7.2.1 Descriptive Analysis

We examined 592 participants who collaborated in our study. Based on previous research[113], it was found that age and gender influenced the performance while executing interaction tasks. Because of that, we are going to use the time taken by those users while performing each of the tests explained in Chapter 3, measured in milliseconds, as features for our machine learning models. In addition, it has been proved that older people perform worse than young people [41] and there is a threshold in the age of 40 where the performance was notably decreasing[41]. Therefore, to allow us to classify the users based on their age, we split the data based on their age into two groups: under and over 40 years old. Regarding gender classification, the data is split based on the user's gender (male, female). Their frequency distributions are shown in table 7.1.

	male	female	Total
under 40	411	117	528
over 40	51	13	64
Total	462	130	592

Table 7.1. Frequency distribution

In order to verify that the proportions across the two groups are homogeneous, thus, the distribution of gender is the same for individuals over and under 40 years old, a Chi squared test for homogeneity was used. The Null hypothesis (H_0) considered was that the distributions are the same for individuals over and under 40 years old. On the contrary the Alternative Hypothesis(H_1) is that the distributions are not the same. The results of this test are shown in Table 7.2.

X-squared	df	p-value
0.031381	1	0.8594

Table 7.2. Chi squared test results for homogeneity

At a 5% significance level, the data provided insufficient evidence (P-value = 0.8594) to conclude that the distribution of gender is different for individuals over and under 40. In other words, the Null hypothesis is accepted and the distribution of gender is homogeneous for individuals of both age groups.

As we might expect the average execution time depends on the complexity of the test. As mentioned previously, some authors [102, 27, 37] reported a higher level of complexity in the execution of *Drag & Drop* tasks when compared with *Point & Click*. Thereby, Table 7.3 shows that *Drag & Drop* tasks require a higher amount of time than *Point & Click* tasks. Furthermore, the runtime of menu selection tasks is higher than that of the other two. This result is consistent with predictions provided by GOMS analysis studied before. Notice that while the *Point & Click* and the *Drag & Drop* tasks required the execution of single P or D operator, the *Item Selection* tasks requires the execution of two P operators (one for menu activation and another one for item selection). Besides that, *Text Selection*, *Text Edit* and *Menu Selection* require the execution of a complex M operator to take the decision of what item to select or key to press in case of the *Text Edit* task.

	Mean	**Std. Dev.**	**Min.**	**Max.**
Point & Click	16870.3970	4292.5680	9319	45792
Drag & Drop	32858.7872	10620.8291	19595	159867
Text Selection	59382.8260	29775.9585	16780	534397
Text Edit	83985.5980	53980.0728	13689	635472
Menu Selection	61155.9392	14075.6995	38351	147630

Table 7.3. Descriptive statistics for the dependent variables in the study (execution times are measured in milliseconds).

7.2.2 Age classification analysis

After executing a 10 repeated 10-fold cross validation with each of the models, the algorithms classification produced a set of classification devices with varying performance results for each bootstrap replicate. The means of the most important metrics are shown below in table 7.4. The results are ordered by AUC-PR as it is the metric that we selected to evaluate the performance of each model. We can see that LogitBoost is the algorithm that gave us more AUC-PR.

The interval of confidence was calculated at 95% for the mean value of AUC-PR for each model, it is shown in table 7.5. In this table, we can see that the LogitBoost interval almost doesn't intersect with the Random Forest and AdaBoost intervals. In other words, those intervals are disjointed and for that, LogitBoost has the best AUC-PR metric. On the other hand, Random Forest and Adaboost are quite similar. On the contrary, the SMO interval doesn't intersect with other intervals, is disjointed, and for that the rest of models have better performance than the SMO model.

In order to determine the best performing model, the sets of overall classification results are compared two by two using the paired two-sample t-test against the Null

	AUC PR	AUC	F-score	RMSE	TPR	TNR
LogitBoost	0.9785	0.8705	0.9546	0.2689	0.9776	0.4190
RandomForest	0.9743	0.8620	0.9580	0.2518	0.9841	0.4186
AdaBoost	0.9728	0.8566	0.9577	0.2547	0.9848	0.4076
Bagging	0.9699	0.8203	0.9457	0.2783	0.9811	0.2290
Stacking	0.9680	0.8403	0.9544	0.2595	0.9780	0.4117
Logistic	0.9631	0.7891	0.9377	0.2969	0.9801	0.0910
J48	0.9183	0.6218	0.9420	0.3028	0.9771	0.1962
SMO	0.8919	0.5000	0.9429	0.3285	1	0

Table 7.4. Models metrics for Age classification.

	AUC PR confidence interval	
LogitBoost	0.9755	0.9815
RandomForest	0.9706	0.9780
AdaBoost	0.9688	0.9767
Bagging	0.9664	0.9735
Stacking	0.9646	0.9714
Logistic	0.9586	0.9675
J48	0.9130	0.9236
SMO	0.8904	0.8935

Table 7.5. Confidence interval at 95% of the mean value of AUC-PR metrics for Age classification

hypothesis (H_0) of equal means, with a significance probability of 5%. The results for this test can be seen in table 7.6. Looking at the table, we can observe that in most of the cases the Null hypothesis of equal means is rejected. Those cells are highlighted with a grey background in the table. According to the results of this table and the previous ones, the LogitBoost model mean is considered significantly similar to the Random Forest model mean, thus, these models as the ones with better performance. This coincides with the use of Random Forest as the weak learning for the LogitBoost algorithm. In addition, the AdaBoost model mean is different from the LogitBoost model mean, in other words, the LogitBoost model is better than the AdaBoost one, that agrees with some researches that say that LogitBoost reduces the training errors linearly and hence yields better generalization [58].

On the other hand and according to the results shown in table 7.5, SMO model mean is not equal to any other mean leaving it as the model with the worst performance. The SMO bad performance agrees with the literature regarding this algorithm is highly affected by the unbalanced data and data transformation techniques need to be done in order to improve the unbalanced effect[140].

7.2.3 Gender classification analysis

After executing a 10 repeated 10-fold cross validation with each of the models, the algorithms produced a set of classification performance results for each bootstrap rep-

AUC-PR	LogitB.	AdaB.	RF	J48	Stack.	SMO	Logistic
LogitB.	-						
statistic							
p-value							
AdaB.	1 > 2	-					
statistic	-2.29						
p-value	0.023						
RF	1 = 3	2 = 3	-				
statistic	1.72	-0.57					
p-value	0.086	0.567					
J48	1 > 4	2 > 4	3 > 4	-			
statistic	19.58	16.38	-17.17				
p-value	<0.001	<0.001	<0.001				
Stacking	1 > 5	2 = 5	3 > 5	4 < 5	-		
statistic	4.55	1.81	-2.48	15.63			
p-value	<0.001	0.072	0.014	<0.001			
SMO	1 > 6	2 > 6	3 > 6	4 > 6	5 > 6	-	
statistic	-50.61	-37.98	-40.61	-9.48	-40.12		
p-value	<0.001	<0.001	<0.001	<0.001	<0.001		
Logistic	1 > 7	2 > 7	3 > 7	4 < 7	5 = 7	6 < 7	-
statistic	-5.7	-3.24	-3.86	12.84	-1.75	30	
p-value	<0.001	<0.001	<0.001	<0.001	0.082	<0.001	
Bagging	1 > 8	2 = 8	3 = 8	4 < 8	5 = 8	6 < 8	7 < 8
statistic	-3.61	1.05	-1.68	16	0.77	-39.55	-2.39
p-value	<0.001	0.294	0.094	<0.001	0.443	<0.001	0.018

Table 7.6. Results of paired two-sample t-test with equal means as Null Hypothesis for Age classification. LogitBoost(1), AdaBoost(2), Random Forest(3), J48(4), Stacking(5), SMO(6), Logistic(7), Bagging(8)

licate. The mean of each of their most important metrics are shown below in table 7.7. The results are ordered by AUC-PR as it is the metric that we selected to evaluate the performance of each model. We can see that Bagging is the algorithm that gave us more AUC-PR. In addition, if we compare these results with the results obtained in the age classification, we can appreciate that the age's results are more accurate than these. As we saw in the literature review there are a lot of studies that analyze the effects of age and prove that ageing negatively impacts the ability to use computers [51, 70] reducing the performance when using technology [146]. Those effects are more evident in these results than the differences in gender.

The interval of confidence was calculated at 95% for the mean value of AUC-PR for each model, it is shown in table 7.8. In this table, we can see that the Bagging, AdaBoost and Random Forest model intervals are quite similar, although in the Bagging interval, the upper boundary is higher than the other two models. On the other hand, SMO and J48 interval models are quite similar and disjointed with the rest of algorithms having worse values.

	AUC PR	AUC	F-score	RMSE	TPR	TNR
Bagging	0.8666	0.6403	0.8672	0.4104	0.9638	0.0800
AdaBoost	0.8592	0.6312	0.8600	0.4192	0.9416	0.1200
RandomForest	0.8584	0.6307	0.8638	0.4153	0.9463	0.1308
Logistic	0.8522	0.6405	0.8747	0.4040	0.9805	0.0715
LogitBoost	0.8515	0.6144	0.8554	0.4648	0.9258	0.1531
Stacking	0.8236	0.5773	0.8719	0.4160	0.9832	0.0346
J48	0.7811	0.5012	0.8711	0.4179	0.9870	0.0100
SMO	0.7804	0.5000	0.8767	0.4686	1	0

Table 7.7. Models metrics for Gender classification

	AUC PR confidence interval	
Bagging	0.8584	0.8747
AdaBoost	0.8510	0.8674
RandomForest	0.8503	0.8666
Logistic	0.8425	0.8618
LogitBoost	0.8425	0.8605
Stacking	0.8167	0.8306
J48	0.7793	0.7828
SMO	0.7801	0.7807

Table 7.8. Confidence interval at 95% of the mean value of AUC-PR metrics for Gender classification

In order to determine the best performing model, the sets of overall classification results are compared two by two using the paired two-sample t-test against the Null hypothesis (H_0) of equal means, with a significance probability of 5%. The results for this test can be seen in table 7.9. Looking at the table, we can observe that in most of the cases the Null hypothesis of equal means is rejected. Those cells are highlighted with a grey background in the table. According to the results of this table and the previous ones, the Bagging model mean is considered significantly similar to the Random Forest and AdaBoost model means, thus, these models are the ones with better performance.

On the contrary, SMO and J48 model means are equal. In addition, according to the results shown in table 7.7 and 7.8, these models have the worst performance comparing them with the other models. The SMO and J48 bad performance agrees with the literature regarding these algorithms are considered to be sensitive to the unbalanced data and data transformation techniques need to be done in order to improve the unbalanced effect[140, 52].

7.3 Conclusions

The main goal of this research is to find an automatic way to predict the demographic of an user while interacting with a website. For this purpose, an interaction test was

AUC-PR	LogitB.	AdaB.	RF	J48	Stack.	SMO	Logistic
LogitB.	-						
statistic							
p-value							
AdaB.	1 = 2	-					
statistic	1.25						
p-value	0.212						
RF	1 = 3	2 = 3	-				
statistic	-1.12	0.14					
p-value	0.262	0.891					
J48	1 > 4	2 > 4	3 > 4	-			
statistic	15.27	18.45	-18.39				
p-value	<0.001	<0.001	<0.001				
Stacking	1 > 5	2 > 5	3 > 5	4 < 5	-		
statistic	4.87	6.55	-6.43	11.74			
p-value	<0.001	<0.001	<0.001	<0.001			
SMO	1 > 6	2 > 6	3 > 6	4 = 6	5 > 6	-	
statistic	-15.71	-19.03	-18.97	-0.72	-12.29		
p-value	<0.001	<0.001	<0.001	0.475	<0.001		
Logistic	1 = 7	2 = 7	3 = 7	4 < 7	5 < 7	6 < 7	-
statistic	0.09	-1.1	-0.98	14.35	4.75	14.72	
p-value	0.925	0.272	0.329	<0.001	<0.001	<0.001	
Bagging	1 < 8	2 = 8	3 = 8	4 < 8	5 < 8	6 < 8	7 < 8
statistic	2.46	-1.26	1.4	20.28	7.93	-20.91	-2.26
p-value	0.015	0.208	0.162	<0.001	<0.001	<0.001	0.025

Table 7.9. Results of paired two-sample t-test with equal means as Null Hypothesis for Gender classification. LogitBoost(1), AdaBoost(2), Random Forest(3), J48(4), Stacking(5), SMO(6), Logistic(7), Bagging(8)

conducted with the participation of 592 individuals. In this experiment, the individuals had to execute 5 interaction tasks: *Point & Click*, *Drag & Drop*, *Select text*, *Type Text* or *Select Items* on a menu. During the execution of these tasks, the time spent on milliseconds by each individual on each of the tasks were collected and used in this research. Before completing these tasks a questionnaire was filled by each user with their demographics: age, gender, laterality, etc. This allowed us to apply machine learning algorithms in order to classify the individuals by age(under 40 and over 40 years old) or gender(male, female).

After gathering the data, it was analysed and we observed that the classes were unbalanced. Been females, in the gender case and individuals over 40 the minority classes. In order to avoid this problem, boosting based algorithms and cost-sensitive learning were used. In addition, we used the AUC-PR and F-score metrics to measure models performance, the literature suggests that the "information los" assumption is more suitable for some highly imbalanced data sets and these two metrics were recommended [140, 84]. AUC, RMSE, TPR and TNR were reported too.

During this study, the machine learning algorithms tested using Weka were: Bagging (Bagging, Random forest), boosting (AdaBoost.M1, LogitBoost), SVM (SMO), C4.5(J48), Logistic and Stacking. Each of the algorithms was executed 10 times using a 10-fold cross validation technique. The metrics reported for each algorithm were the mean of all the iterations. In order to validate which algorithms were better than the others a pair two-sample t-test was executed comparing the algorithms two by two. This test validated as the Null hypothesis(H_0), that the means of the two algorithms evaluated in each time were the same with a significance level of 95%. This process was executed for the case of age classification and the case of gender.

Regarding age, the results of the t-test showed us that few algorithms had the same means of AUC-PR metric like LogitBoost and Random Forest, and for that, they performed the same. This coincides with the use of Random Forest as the weak learner of LogitBoost. These algorithms gave the best performance for the AUC-metric with a 97%, F-score of 95% and with an RMSE of 25-26%. On the other hand, SMO gave us the worst result caused by its weakness with unbalanced data[107, 140].

In the case of gender, we got the best results from Bagging, Random Forest and AdaBoost that behave in a significant same way as shown in the results. The AUC-PR mean confidence interval was between 85 and 87% and with an RMSE of 41%. On the other hand, SMO and J48 that are considered with equivalent performances, also gave us the worst result caused by their weakness with unbalanced data and their need of data processing for better results [140, 52]. This also agrees with the criticism raised by the community that there is too much reliance on the class-imbalance research on C4.5 when it is not the best classifier for dealing with class imbalances and that the community should focus less on it [72].

In addition, if we compare the results obtained in the age classification with the results obtained in gender classification, we can appreciate that the age's results are more accurate than the gender ones. This agrees with a lot of studies that analyzed the effects of age and proved that ageing negatively impacts the ability to use computers [51, 70] reducing the performance when using technology [146]. In addition, as this study was carried out by the participants in their computers using larger displays(monitors), the gender performance gap is reduced[38, 139] as these results agreed on.

To conclude, these results proved that machine learning algorithms can be used in order to predict the age or the gender of an user while browsing a website and do common human computer interaction tasks like *Point & Click*, *Drag & Drop*, *Select text*, *Type Text* or *Select Items* on a menu.

7.4 Limitations and Future Work

On the other hand, one of the limitations of this experiment is that the interaction test was carried out by participants who were collected in Facebook, Twitter and Foro coches. As they were contacted on the Internet, they could belong to any country and

the different cultures might have influenced the study. According to that, considering more variables might be good to study in future research like their country, language, errors produced in the execution of the tasks.

We have seen that in both classifications, Random Forest and AdaBoost are the two algorithms that in both cases have one of the best performances with our models. In future research, we could combine these two classifications in one and see if the results are significant enough. In addition, once having a trained model the next step would be productivize it and use it in real website scenario. This could help to improve the usability of the websites adapting their interface depending on the user browsing on it, i.e. an old user would need bigger fonts and bigger buttons in order to compensate for the deficiencies caused by the age. In addition, it can be used to help marketing in order to give them a specific target regarding age or gender.

8 Conclusions

The main goal of this research is to find an automatic way to predict the demographic of an user while interacting with a website. For this purpose, an interaction test was conducted with the participation of 592 individuals. In this experiment, the individuals had to execute 5 interaction tasks: *Point & Click*, *Drag & Drop*, *Select text*, *Type Text* or *Select Items* on a menu. During the execution of these tasks, the time spent on milliseconds by each individual on each of the tasks were collected and used in this research. Before completing these tasks a questionnaire was filled by each user with their demographics: age, gender, laterality, hours of computer use, etc.

The study was splitted in four parts:

8.1 Part 1: Analysis of user's age as significant factor for user classification

The main goal of this part was to validate if the age has an influence in the performance of computer interaction tasks. For this study the task of *Point & Click*, *Drag & Drop*, *Select text* were taken into account. The results showed several interesting facts, aging, as expected and pointed out by most of the works discussed in the background section, has a strong influence on the performance of computer users.

After a more detailed analysis based on the GOMS task division, it can be argued that, as pointed out in [32], this influence differs between specific activities. In fact, results show that the aging effect is not the same for the three tasks that were analyzed. Although these facts were already observed in [32] and [127], the more fine-grained grouping used in this work allowed the profiling of six age groups for each of the three basic tasks that were analyzed.

8.2 Part 2: Analysis of user's gender as significant factor for user classification

This section aims to identify demographic differences based on how users interact with web applications. Results from our first analysis and some gaps found in the related work, lead us to verified if:

- There is a direct relationship between age and performance time,
- Women's performance in point-and-click mouse operations is better than men's
- Men's performance in drag-and-drop operations is better than women's
- Evaluate the role of laterality in computer interaction tasks

Regarding age, the results obtained were consistent with previous research, especially with the work of Hill et al. [70]. The results obtained in tests 1 to 5 show a negative impact on the performance of the GOMS operators P, K, and D when age is increased. The increment in the overall execution time for each task proved to be much bigger in people aged over 40. This result makes us believe that it would be relatively easy to classify people above and below this age using data gathering agents.

With respect to gender, the results obtained are consistent with the findings of Inkpen [77] who found that girls perform worse with regard to *Drag & Drop* tasks. In our study males obtained better results when executing interaction operators based on pointing and dragging, which are required by the *Point & Click*, *Drag & Drop* and *Menu Selection* tasks. Nevertheless, there were no significant differences in the *Text Selection* and *Text Edit* tasks. The differences were also so relevant that they can be used for the design of algorithms for automatic classification.

Surprisingly, no significant performance differences were detected between left- and right-handed users and we did not find any evidence regarding the influence of this factor on the overall performance.

8.3 Part 3: Analysis of the correlation of Interaction tasks' significance in age and gender based users classification

This work had two interrelated goals. First, we wanted to assess whether gender and age are sufficiently significant determining factors to support an automatic profiling system based on the analysis of mouse motion behavior when executing *Point & Click*, *Drag & Drop* and *Item Selection* tasks. Second, to figure out whether the individuals perform consistently across these basic interaction tasks, that is, if their performance in one of them are extrapolable (or not) to the others.

Regarding the first, the results of the empirical study reveal that both age and gender factors are significantly determinant. While older users performed worse than younger in each interaction task, men obtained better results than women. On the other hand, in relation with the analysis of correlations between the execution times of the target basic interaction tasks, data gathered in the tests revealed consistency in the execution times of individuals across them. User's performance measured in any of these tasks is coherent to their execution time in the other tasks.

8.4 Part 4: Implementation of a ML-based operative classification system

The main goal of this research is to find an automatic way to predict the demographic of an user while interacting with a website. In order to do that, we applied machine learning algorithms in order to classify the individuals by age(under 40 and over 40 years old) or gender(male, female).

AUC-PR and F-score metrics were used to measure the models performance. The literature suggests that the "information loss" assumption is more suitable for some highly imbalanced data sets and these two metrics were recommended [140, 84]. AUC, RMSE, TPR and TNR were reported too.

During this study, the machine learning algorithms tested using Weka were: Bagging (Bagging, Random forest), boosting (AdaBoost.M1, LogitBoost), SVM (SMO), C4.5(J48), Logistic and Stacking. Each of the algorithms was executed 10 times using a 10-fold cross validation technique. The metrics reported for each algorithm were the mean of all the iterations. In order to validate which algorithms were better than the others a pair two-sample t-test was executed comparing the algorithms two by two. This test validated as the Null hypothesis(H0), that the means of the two algorithms evaluated in each time were the same with a significance level of 95%. This process was executed for the case of age classification and the case of gender.

Regarding age, the results of the t-test showed us that few algorithms had the same means of AUC-PR metric like LogitBoost and Random Forest, and for that, they performed the same. This coincides with the use of Random Forest as the weak learner of LogitBoost. These algorithms gave the best performance for the AUC-metric with a 97%, F-score of 95% and with an RMSE of 25-26%. On the other hand, SMO gave us the worst result caused by its weakness with unbalanced data [107, 140].

In the case of gender, we got the best results from Bagging, Random Forest and AdaBoost that behave in a significant same way as shown in the results. The AUC-PR mean confidence interval was between 85 and 87% and with an RMSE of 41%. On the other hand, SMO and J48 that are considered with equivalent performances, also gave us the worst result caused by their weakness with unbalanced data and their need of data processing for better results [140, 52].

In addition, if we compare the results obtained in the age classification with the results obtained in gender classification, we can appreciate that the age's results are more accurate than the gender ones. This agrees with a lot of studies that analyzed the effects of age and proved that ageing negatively impacts the ability to use computers [51, 70] reducing the performance when using technology [146]. In addition, as this study was carried out by the participants in their computers using larger displays(monitors), the gender performance gap is reduced [38, 139] as these results agreed on.

To conclude, these results proved that machine learning algorithms can be used in order to predict the age or the gender of an user while browsing a website and do common human computer interaction tasks like *Point & Click*, *Drag & Drop*, *Select text*, *Type Text* or *Select Items* on a menu.

This automatic profiling can be used to provide different contents to different kinds of users and also to adapt the appearance and behaviour of different elements of the user interface. For example, the size of the clickable elements (buttons, links, menu items, etc.) could be dynamically increased when they are being used by elderly users. Since Fitts's function is logarithmic, a small increase in the size of the object would represent a drastic reduction in the time required by the user to click on it. This approach could not only increase the user's overall performance but could also significantly contribute to enhancing the user's experience on the site.

9 Conclusiones

El objetivo principal de esta investigación es encontrar una forma automática de predecir valores demográficos de un usuario mientras interactúa con un sitio web. Para este propósito, se realizó una prueba de interacción con la participación de 592 personas. En este experimento, los individuos tuvieron que ejecutar 5 tareas de interacción: *Point & Click*, *Drag & Drop*, *selección de texto*, *edición de texto* y *selección de elementos de un menú*. Durante la ejecución de estas tareas, el tiempo dedicado en milisegundos por cada individuo en cada una de las tareas fue recopilado y empleado en esta investigación. Antes de completar estas tareas, cada usuario llenó un cuestionario con edad, género, lateralidad, horas de uso semanal de su ordenador, etc.

El estudio se dividió en cuatro partes:

9.1 Parte 1: Análisis de la edad del usuario como factor significativo para la clasificación del usuario

El objetivo principal de esta parte era validar si la edad tiene una influencia en el rendimiento de tareas de interacción. En este estudio se analizaron tareas de *Point & Click*, *Drag & Drop* y *selección de texto*. Los resultados mostraron varios hechos interesantes. El envejecimiento, como se esperaba y así ha sido señalado en la mayoría de los trabajos discutidos en la sección de antecedentes, tiene una fuerte influencia en el rendimiento de los usuarios.

Después de un análisis más detallado basado en la división de tareas GOMS, se puede argumentar que, como se señaló en [32], esto influye en las diferencias entre actividades específicas. De hecho, los resultados muestran que el efecto de envejecimiento no es el mismo para las tres tareas que se analizaron. Aunque estos hechos ya se observaron en [32] y [127], en análisis de más bajo nivel utilizado en este trabajo permitió la categorización de seis grupos de edad para cada una de los tres tareas básicas analizadas.

9.2 Parte 2: Análisis del género del usuario como factor significativo para la clasificación del usuario

Esta sección tiene como objetivo identificar las diferencias demográficas en función de cómo interactúan los usuarios con aplicaciones web. Los resultados de nuestro primer análisis y algunas lagunas encontradas en el el estado del arte, nos llevaron a verificar si:

- Existe una relación directa entre la edad y el tiempo de rendimiento,
- El desempeño de las mujeres en las operaciones de *Point & Click* con el ratón es mejor que el de los hombres
- El rendimiento de los hombres en las operaciones de *Drag & Drop* r es mejor que el de las mujeres.
- Evaluar el papel de la lateralidad en las tareas de interacción.

En cuanto a la edad, los resultados obtenidos fueron consistentes con investigaciones previas, especialmente con el trabajo de Hill et al. [70] Los resultados obtenidos en las pruebas 1 a 5 muestran un resultado negativo. impacto en el rendimiento de los operadores GOMS P, K y D cuando aumenta la edad. El incremento en el tiempo de ejecución general para cada tarea resultó ser mucho mayor en personas mayores de 40 años. Este resultado nos hace creer que sería relativamente fácil clasificar a las personas mayores y menores de esta edad utilizando agentes de recopilación de datos.

Con respecto al género, los resultados obtenidos son consistentes con los resultados de Inkpen [77], quien descubrió que las niñas tienen un peor desempeño con respecto a las tareas de *Drag & Drop*, en nuestro estudio, los hombres obtuvieron mejores resultados al ejecutar operadores de interacción basados en señalar y arrastrar, que son requeridos por *Point & Click*, *Drag & Drop* y tareas de selección de elementos en menús. Sin embargo, no hubo diferencias significativas en las tareas Selección de texto y Edición de texto. Las diferencias también fueron tan relevantes que demostraron pueden ser utilizadas para el diseño de algoritmos de clasificación automática.

Sorprendentemente, no se detectaron diferencias de rendimiento significativas entre zurdos y deistros y no encontramos ninguna evidencia con respecto a la influencia de este factor en el rendimiento general.

9.3 Parte 3: Análisis de la correlación de la importancia de las tareas de interacción en la clasificación de usuarios basados en edad y género

Este trabajo tenía dos objetivos interrelacionados. Primero, queríamos evaluar si el género y la edad son factores determinantes suficientemente importantes para respaldar la construcción de sistemas de clasificación automática basados en el análisis del comportamiento del movimiento del ratón al al ejecutar *Point & Click*, *Drag & Drop* y tareas de selección de elementos.En segundo lugar, averiguar si los individuos realizan consistentemente estas tareas básicas de interacción, es decir, si su rendimiento en una de ellas es extrapolable (o no) a las demás.

Con respecto al primer objetivo, los resultados del estudio empírico revelan que tanto la edad como el género son factores significativamente determinantes. Mientras que los usuarios mayores se desempeñaron peor que los más jóvenes en cada tarea de interacción, los hombres obtuvieron mejores resultados que las mujeres. Por otra parte, en relación con el análisis de correlaciones entre los tiempos de ejecución, los datos recopilados en las pruebas revelaron consistencia en la ejecución. El rendimiento del usuario medido en cualquiera de estas tareas es coherente con su tiempo de ejecución en las otras tareas.

9.4 Parte 4: Implementación de un sistema clasificador basado en Machine Learning

El objetivo principal de esta investigación es encontrar una forma automática de predecir la demografía de un usuario mientras interactúa con un sitio web. Para hacer eso, aplicamos algoritmos de Machine Learning para clasificar a los individuos por edad (menores de 40 y mayores de 40 años) o género (hombre, mujer).

Las métricas AUC-PR y F-score se utilizaron para medir el rendimiento de los modelos. La literatura sugiere que el supuesto de "pérdida de información" es más adecuado para algunos conjuntos de datos altamente desbalanceados y sugieren el uso de estas dos métricas para estos casos [140, 84]. AUC RMSE, TPR y TNR se muestran también.

Durante este estudio, los algoritmos de Machine Learning probados con Weka fueron: Bagging (Bagging, Random forest), boosting (AdaBoost.M1, LogitBoost), SVM (SMO), C4.5(J48), Logistic y Stacking. Cada algoritmo fue ejecutado 10 veces usando la técnica de "10-fold cross valitation" 10 veces. La métrica recogidas para cada algoritmo fue la media de las métricas de todas las iteraciones. Para validar qué algoritmos eran mejores que los otros se ejecutó la prueba t-test para dos muestras comparando los algoritmos de dos en dos. Esta prueba valida como la hipótesis nula (H_0), que las

medias de los dos algoritmos evaluados en cada tiempo son los mismos con un nivel de significación del 95%. Este proceso fue ejecutado para el caso de clasificación por edad y el de género.

En cuanto a la edad, los resultados de la prueba t-test nos mostraron que para el caso de LogitBoost y Random Forest, las medias de la métrica AUC-PR eran iguales y por ello, se considera que ambos obtuvieron el mismo rendimiento. Esto coincide con el uso de Random Forest como "weak learner" de LogitBoost. El mejor resultado fue: para la métrica AUC un 97%, un F-score de 95% y un RMSE de 25-26%. Por otro lado, SMO dió el peor resultado probablemente causado por su debilidad con datos desbalanceados [107, 140].

En el caso del género, obtuvimos los mejores resultados de Bagging, Random Forest y AdaBoost, los cuales se comportan significativamente de la misma manera tal y como se muestra en los resultados. El intervalo de confianza medio para AUC-PR fue entre 85 y 87% y con un RMSE de 41%. Por otro lado, SMO y J48, los cuales tienen rendimientos significativamente equivalentes, también dieron el peor resultado probablemente causado por su debilidad ante datos desbalanceados y su necesidad de usar datos procesados previamente para obtener mejores resultados [140, 52].

Además, si comparamos los resultados obtenidos en la clasificación por edad con los obtenidos en la clasificación por género, se puede apreciar unos resultados más precisos para el caso de la edad. Esto concuerda con muchos estudios que analizan los efectos de la edad provando que el envejecimiento impacta negativamente en la habilidad de usar ordenadores [51, 70] reduciendo el rendimiento mientras se usa tecnología [146]. Adicionalmente, debido a que este estudio se llevó a cabo por los participantes en sus ordenadores usando una pantalla grande, un monitor, la brecha en el rendimiento obtenido por ambos géneros se reduce [[38, 139] como podemos observar en nuestros resultados.

Para finalizar, estos resultados demuestran que se pueden utilizar algoritmos de Machine Learning para predecir la edad o el género de un usuario cuando navega en un sitio web y realiza comunes tareas de interacción persona-ordenador tales como *Apuntar y señalar*, *Arrastrar y soltar*, *Seleccionar texto* o *Seleccionar elementos en un menú*.

Este perfilado automático de usuarios puede usarse para proveer diferente contenido para diferentes usuarios y también para adaptar la apariencia y comportamiento de diferentes elementos de la interfaz un sitio web. Por ejemplo, el tamaño de los elementos en los que se puede hacer click (botones, enlances, menús, etc.) podría dinámicamente aumentar cuando una persona mayor está navegando. Como la función de Fitt es logarítmica, un pequeño incremento en el tamaño del objeto podr'ia representar una drástica reducción en el tiempo de ejecución requerido por el usuario para hacer click en ello. Este enfoque podría no sólo aumentar el rendimiento general del usuario sino que también podría contribuir significativamente a mejorar la experiencia del usuario en el sitio web.

10 Limitations and Future Directions

One of the limitations of this experiment is that the interaction test was carried out by participants who were collected in Facebook, Twitter and Foro coches. As they were contacted on the Internet, they could belong to any country and the different cultures might have influenced the study. Some studies suggest that the culture of an individual could determine his/her performance. Ford et al. [53] designed an experiment to evaluate if any of Hofstede's cultural dimensions can affect human performance while interacting with computers. Even though their study did not provide sufficient evidence to reach any determining conclusion, we consider it would be interesting to extend this work to a multicultural sample of individuals to study such influence in these specific types of interaction.

In addition, we consider there are other factors that could somehow determine a user's performance in basic interaction tasks and that should be considered to extend this study in the future. For example, user accuracy is another important factor that can be measured in online information systems that could bring a different approach to classifying users. Therefore, our next step will be the design of a similar case study in order to analyse the role that accuracy plays in classifying users. In this way, it would be possible to design classifying systems based on a trade-off between speed and accuracy.

The possible benefits of such a classification system are straightforwardly applicable in e-commerce sites, the main target of this work, since the information architecture of the site (and the list of products or sales offered) could be adapted accordingly to the preferences of this target user. However, there are other possible applications like preventing some users from claiming the identities of other users or from pretending to be a different age and/or have a different gender. In addition, detecting old users would support the automatic adaptation of the interface to the specific features of this group, using for example bigger fonts and simpler interfaces.

A Impact Factor Report

Q3/T3. Pariente-Martinez, Beatriz; Gonzalez-Rodriguez, Martin; Fernandez-Lanvin; Daniel; De Andres-Suarez, Javier (2016) Measuring the Role of the Age in the Performance when interacting with computers. Universal Access in the Information Society (UAIS) Vol 15, Issue 2, Pages 237-247. ISSN: 1615-5289 (print) ISSN: 1615-5297 (online). Springer, Berlín, Heidelberg. DOI: 10.1007/s10209-014-0388-6 [JCR 2016; Area: COMPUTER SCIENCE, CYBERNETICS; Journals in Area: 22; Citations: 520; Rank: 15; Impact Factor: **1.219**].

Q3/T2. De Andres-Suarez, Javier; Pariente-Martinez, Beatriz; Gonzalez-Rodriguez, Martin; Fernandez-Lanvin; Daniel; (2015) Towards an automatic user profiling system for online information sites: Identifying demographic determining factors. Online Information Review. Vol 39, Issue 1 2015, Pages 61-80. ISSN: 1468-4527. Emerald Group Publishing Limited. Bingley, West Yorkshire, UK. DOI 0.1108/OIR-106-2014-0134 [JCR 2015; Area: COMPUTER SCIENCE, INFORMATION SYSTEMS; Journals in Area: 144; Rank: 73 Impact Factor: **1.152**].

Q2/T1. Fernandez-Lanvin; Daniel; De Andres-Suarez, Javier; Gonzalez-Rodriguez, Martin; Pariente-Martinez, Beatriz (2018) The dimension of age and gender as user model demographic factors for automatic personalization in e-commerce sites. Computer Standards & Interfaces: The International Journal on the Development and Application of Standards for Computers, Software Quality, Data Communications, Interfaces and Measurement. Vol 59, (August), Pages 1-9. ISSN: 0920-5489. Elsevier, Amsterdam, The Netherlands. DOI: 10.1016/j.csi.2018.02.001 [JCR 2018; Area: COMPUTER SCIENCE, HARDWARE AND ARCHITECTURE; Journals in Area: 52; Citations: 1,499; Rank: 17; Impact Factor: 1.465; Area: COMPUTER SCIENCE, SOFTWARE ENGINEERING; Journals in Area: 107; Citations: 1,499; Rank: 29; Impact Factor: **2.441**].

B Article 1: Measuring the role of age in user performance during interaction with computers

Univ Access Inf Soc
DOI 10.1007/s10209-014-0388-6

LONG PAPER

Measuring the role of age in user performance during interaction with computers

Beatriz Pariente-Martinez · Martin Gonzalez-Rodriguez · Daniel Fernandez-Lanvin · Javier De Andres-Suarez

Abstract The influence of aging on computer interaction has been widely analyzed in human–computer interaction research literature. Despite this, there are no age-based user maps that could support the user-interface customization. Studying the specific needs and constraints of these groups is crucial in order to adapt a user interface to the user's interaction requirements. This work studies the performance of a sample of participants on three different basic tasks (pointing, dragging and dropping, and text selection) and the influence of age for each of them. It is concluded that this influence differs between specific activities. A group profile map that can support automatic classification in the future has been obtained.

Keywords Personalization · User categorization · User performance · HCI · GOMS · Fitts's law · Hick's law

1 Introduction

HCI research and usability studies focus on the measurement of several factors that determine the user's experience [1]. A popular strategy in HCI research literature is to study a group of users determined by one or several of these factors, such as age, social status, or specific impairments, among others. In fact, there is a wide literature studying patterns of child–computer interaction [2–5], older people's performance rates [6–9], motion-impaired users [10–12], and similar factors. However, most user interfaces are designed for younger users [8], ignoring other groups' specific requirements. Studying the specific needs and constraints of these groups is crucial in order to adapt a user interface to the user's interaction requirements. But this categorization is not always particularly evident or known. In the specific case of the age, this parameter is unknown unless the user has previously logged onto the system (having provided that information in his or her profile). Even when this happens, the reliability of this information can be determined by social or cultural factors. Underage users, for example, often tend to lie in order to avoid controls on adult-oriented sites or social networks [13].

Several authors consider age as a determining factor in users' performance [15]. Furthermore, the automatic detection of users' ages would gather extra and useful information to drive the user-interface design of applications according to their target users, given that there are differences among age groups in what they do online [16]. For instance, young adults lead the age groups in their use of communication tools, such as instant messaging and chatting, and are also more likely to pursue hobby or entertainment activities, such as downloading music or surfing for fun, while those aged between 29 and 59 years tend to use the Internet more frequently to perform job research and/or to use government sites [17].

The automatic classification of users according to their ages can be done by analyzing the performance of the users when they perform certain tasks, comparing it with the

B. Pariente-Martinez (✉) · M. Gonzalez-Rodriguez · D. Fernandez-Lanvin
Department of Computer Science, University of Oviedo, Oviedo, Spain
e-mail: bparientem@gmail.com

M. Gonzalez-Rodriguez
e-mail: martin@uniovi.es

D. Fernandez-Lanvin
e-mail: dflanvin@uniovi.es

J. De Andres-Suarez
Department of Accounting, University of Oviedo, Oviedo, Spain
e-mail: jdandres@uniovi.es

Published online: 25 September 2014

performance of individuals whose age is known. This process requires a previous study and classification into groups according to the age and performance of the users.

This work covers part of the previous analysis necessary to build this automatic classification system, focusing on the impact of the influence of age on the user's interaction. Considering that it was pointed out that this influence could be heterogeneous or relevant to specific activities [18], the analysis was divided into three common actions required to interact with computers on modern interfaces: (1) pointing, (2) drag and drop, and (3) text selection (a combination of pointing and dragging).

For this purpose, a small Web application including three types of tests and gathered results from 557 people was designed. The analysis of the gathered data presented some interesting results. First, the results show that, as expected, the impact of age depends on the activity considered. Second, although some studies state that age is not always the determining factor [19], its influence on the performance of users is strong. Third, it was possible to identify the thresholds that limit the groups according to their performance related to each of the activities.

2 Background

CMN-GOMS (Card, Moran, and Newell's Goals, Operators, Methods, and Selection rule) and its variations (KLM-GOMS, NGOMSL, CPM-GOMS, etc.) are a well-known user model employed to estimate the user performance time in interactive systems [20]. GOMS analysis was successfully used to determine the usability of Web sites for disabled users [21], the performance of automobile human–machine interfaces [22], or the navigational structure of Web sites [23]. Although it was designed to predict task execution time on mouse-and-keyboard systems, it is also an accurate tool to measure performance time on touch screens [24].

GOMS reduces the user interaction to a basic set of operators, each of which represents an elementary action required to reach a goal. The main operators that a user needs to interact with in GOMS are 'pointing' (P) to something on the display, 'dragging' (D) objects around, 'key pressing' (K) on the keyboard, mouse, or touch screen, and 'mental' (M) decisions.

GOMS analysis is based on the assumption that users know how to perform the tasks and will not generate any errors. It is also assumed that the users may interact as quickly as possible due to their previous experience. These assumptions allow its application in the analysis of systems where the users are computer literate or veteran, enabling GOMS to focus on the users' effectiveness (performance) instead of their efficiency.

As GOMS proved that it is possible to reduce the user's interaction with a computer to these elementary actions, it is feasible to obtain the global user performance if the performance on each operator is analyzed separately. The use of this analysis commonly results in a measure of the global performance in the development of simple tasks (e.g., logging into an email account), which is obtained from the performance recorded in the execution of the individual operators required to achieve the task. Therefore, if a high-level task requires the use of mental (M), pointing (P), and dragging (D) operators, performance measurement would be the sum of the time required for each one ($T_M + T_P + T_D$).

Pointing and clicking tasks require two GOMS operators: pointing (P) to an object (with the mouse or with the finger) and key pressing (K) (on the mouse or touch screen). The first operator (P) requires the user to move the pointer to the target, reducing its speed and aiming accurately [25]. The time required by an 'average' user to execute any GOMS operator may be estimated by mean psychological laws such as Fitts's law for T_P and T_D, Salthouse's regularities for T_K, and Hick's law for T_M.

Fitts's law [26] is used to obtain accurate estimations of the time required by a user to point to a target (e.g., clicking on a button with a mouse) depending upon the distance to the target and its size [26]. Salthouse observed regularities in typing in a classic research on skilled typing performance [27]. These regularities are robust enough to estimate the efficiency and speed of overlapping successive keystrokes. Hick's law predicts a linear increase in choice reaction time with the logarithm of the number of alternatives the user has to evaluate before making a decision (e.g., the number of items in a combo box or radio button group) [28].

There are external variables that may increase or decrease the speed of the pointing action, which affects the error rate of the subjects. It was noted that aiming at a square target can be done faster than aiming at a rounded one when the target objects share the same width [5, 25, 29–32]. This effect is caused by the fact that the surface of the squared object is slightly greater than that of the rounded object. Fitts's law estimates the runtime needed to point to an object as $a + b \log_2(D/S)$. Parameters 'a' and 'b' are device-dependent constants, where 'D' is the distance to the target and 'S' represents its size. Since the surface of a rounded shape is slightly smaller than that of a squared one, the estimated time required to click on the first one is longer than in the second case.

Drag and drop tasks allow the user to move objects from one place to another. The basic GOMS sequence involves placing the pointer over a target (P), holding down the finger over the display or on the mouse button (K) to select the object, dragging (D) the object to the desired

location without releasing the button, and finally releasing it to drop the object (K) [5, 29].

Unlike *pointing* tasks, *drag and drop* may entail a different effect of aging, given that it requires holding the mouse button down, which causes a motor demand [33]. This effect is especially relevant to children and older adults. For these kinds of users, it is more comfortable to perform *point–move–point* operations rather than *drag and drop* because if a failure is produced during a *drag and drop* task it must be repeated from scratch, while if this kind of failure is produced during a *point–move–point* operation, the operation could be resumed from the last pointing task [34].

The relevance of aging to the way users interact with computers has been widely studied. Aging, as evidenced by cognitive change, poorer motor control, and sensory deficits, negatively impacts on the ability to use computers [15]. It is associated with changes in characteristics such as visual perception, fine motor control, and some aspects of memory and cognition [7]. All of them have an impact on the overall performance of the pointing (P), key pressing (K), dragging (D), and mental (M) operators.

In the particular case of the mouse (considered by some researchers as the most efficient input device [17, 35]), the studies mostly agree on the effect of aging on its use. With the exception of very young children, who interact at a slower rate than adults due to the fact that their skills are further determined by cognitive factors and psychometric abilities [2], the ability of users with this device decreases with aging.

During the developmental period between childhood and youth, a subject's abilities evidence important changes. Children between four and 5 years old have less direct and less accurate trajectories with the mouse than adults. Children up to 8 years old evidence major difficulties in attaching and lifting objects, and the speed of movement execution is constantly evolving up to 12 years [36]. Thus, differences in development can cause a variety of difficulties when making subtle and delicate movements that require hand–eye coordination.

On the other hand, the effects of aging on the motor and the musculoskeletal systems, as well as loss of muscle strength lead to reduced mobility [37], causing difficulties in the execution of movements [38]. In [39], Metter et al. stated that muscle strength begins to decline in people aged over 40, probably due to changes in the number and size of muscle fibers. Considering also the slowing of the conduction velocity of nerve signals [40], changes could be related to studies showing that older people have slower reaction times [41], delayed movements, a decline in motor skills [38], and a decrease in coordination capacity, agility, and balance [42, 43]. However, in older adults, age could be not such a determining factor [14], given that although fitness decreases markedly with age from young adulthood to old age, individual differences increase [44].

The deterioration of cognitive, perceptual, and motor skills may have even worse effects when several tasks are performed simultaneously, a common practice in graphical user interfaces where pointing (P), dragging (D), or key pressing (K) operators are activated at the same time [45, 46].

Chaparro et al. concluded that there is a significant correlation between movement speed and age. People over 65 years are slower than people aged 40 years irrespective of the task carried out [6]. However, some authors consider that these effects can be mitigated by other factors. Elderly users consider that the use of the Internet not only helps them to improve their independence, but also prevents their cognitive decline [9], something that, if true, would minimize the aging effect. Other studies state that even assuming that aging decreases motor control [38, 47], it does not have direct effects on pointing (P) operators because older adults compensate for the decrease in perceptual efficiency by adjusting the velocity and number of sub-movements required to hit the target [38].

Salthouse et al. [27] reported that while the reaction time for mental operators (M) declined with age, key pressing (K), which was thought to be highly correlated with age, was rather consistent across age groups [18]. Since something similar could happen with the use of the pointer, the different tasks that can be performed (*pointing* and *drag and drop*) should be analyzed separately.

Loos [19] also questions the role of aging in performance when interacting with computers. In [19], he analyzes the results of an eye-tracking case study focused on information search behavior. He concluded that the greatest factor affecting information search behavior is not only always age but also gender, educational background, and frequency of use of the Internet. This interesting result must be carefully taken into consideration, since the main activity of the experiment (information search on a Webpage) requires abilities that are essentially different from the motion-based ones, which are the focus of this work (in that respect, the impact of other factors may be quite different in both activities). Furthermore, other studies analyzing the same activity [48, 14] evidence the importance of aging. The latter, based on a study of the behavior of expert older adults using the Web daily, compared to their younger colleagues, concluded that age is a determining factor.

In summary, studies mainly conclude that aging brings general changes (declines) in perception, cognition, and movement control, which have an influence on learning and use of digitized systems [49, 50]. Senior users have been found to be slower than young adults when performing information retrieval tasks [51, 52], 3D navigation on desktop systems [53], and Web navigation [54].

3 Goals

The results described in the previous section can easily be taken into consideration for user-interface design whenever the age of the target audience is known (for example, online services targeted toward senior citizens or child-oriented didactic games), although do not provide enough information to support the automatic classification of users. In the studies that identify a threshold for specific activities, there is not enough information to classify users into age-based groups for each of the mouse usage motion-based related activities. Only [18] divided users into groups, but limited them to (1) young, (2) middle aged, and (3) older users, a classification that is too raw for the purposes of the current study. Also, according to [19], the influence of age should also be checked.

In order to find out more information about how aging affects interaction, an experiment was designed to observe the behavior of different users in *pointing, drag and drop*, and *text selection* activities. The data gathered in this case study will be later used to:

- compare the performance of different age groups;
- identify (if they exist) the different groups based on their performance, and the age thresholds that should be considered for the performance-based classification;
- determine the discriminating power of age as classification criteria by means of the statistical analysis that will be developed in the subsequent sections.

4 Case study

Volunteers invited to participate in the case studied were classified depending upon their age into groups of 10 years each. Each volunteer had to specify the group to which he or she belonged without giving his or her real age. Although this abstraction can reduce the quality of the results, the idea of asking a direct question about the user's age was discarded due to the resistance of users to revealing their specific ages, which was observed in previous attempts during similar experiences.

Data gathering was performed by means of a Web application created ad hoc for the tests. The application implements three different tests to measure user performance on the three target interaction tasks (pointing, drag and drop, and text selection) using the GOMS principles. The user's performance was measured in milliseconds for the pointing (P), dragging (D), and key pressing (K) GOMS operators. These values were denoted as T_P (average time required by user to execute a pointing operator), T_D (average time spent executing dragging operators), and T_K (average key pressing time).

At the beginning of each test, the mouse pointer was located in the bottom right corner. This is the position in which the button the user needs to click in order to start each test is located.

Prior to the execution of these tests, participants completed a questionnaire providing information about their ages and the kinds of input devices they would use during the experiment (mouse, touchpad, or touch screen). The questionnaire was used to categorize participants into six age groups: A (ages 0–20), B (ages 20–30), C (ages 30–40), D (ages 40–50), E (ages 50–60), and F (ages 60–80).

One of the fundamentals of GOMS is to design the test for the users who will most likely use the system which is being analyzed. Therefore, participants in this case study were recruited through social networks such as Facebook, Twitter, and Google Plus. The initial participants invited relatives and friends to participate in the research through their own social networks. This recruiting approach was designed to restrict the population sample to people who were already computer literate, avoiding volunteers with a low motoric response due to their low profile in the use of computers.

The *pointing* task (Task 1) was used to evaluate the speed at which users were able to complete the GOMS pointing (P) operator. During the test, users had to click on a red square that appeared in different positions on the screen (Fig. 1). Whenever they clicked on the square, it disappeared from its current position and moved to a new one. In each interaction, the size of the square changed too. This process was carried out 14 times, and the sequence of movements and changes in the square's size were exactly the same for all participants. During the process, hidden agents observed the users' behavior and recorded the time

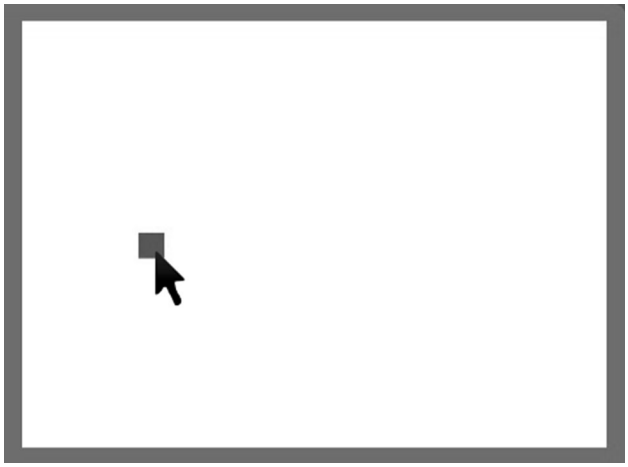


Fig. 1 During the pointing test, the user had to click anywhere on each *square* as fast as possible. Once the user clicked on it, the *square* moved to a different position on the screen

required by the users to click on each square. The average time was estimated according to the precepts of Fitts's law (distance to the target and target size). The T_P value obtained as a result of this test is a proxy of the global time required to execute the GOMS pointing (P) operators.

The *drag and drop* task (Task 2) was used to evaluate *drag and drop* speed. In this test, for each iteration, participants had to drag red squares and drop them onto empty ones. Once the user managed to place the red square inside the target, a visual and auditory cue was displayed. Then, a new red square and target appeared in a different position on the display. The size of the red square was two-thirds of the surface of the target.

The user had to repeat this process twelve times, and the size of the squares varied in each of the attempts. A screenshot of the test is shown in Fig. 2. Hidden agents recorded the time required for each drag and drop action as well as the full amount of time required to complete the test. The time taken by each user in the execution of the GOMS dragging (D) operators for this test was recorded and denoted by T_D.

The location and size of the square targets used in Tasks 1 and 2 were carefully designed to force users to make use of the two parameters on which Fitts's law is based: distance to the target and target size. At the beginning of each of the mentioned tasks, the size of the target was relatively big, though it decreased in each iteration. In the same way, the distance to the targets also increased in each iteration, making the task more difficult to complete. To increase the difficulty, the positions of the targets in each iteration varied from left to right and from the top to the bottom of the display using variations of a *Z*-shaped pattern favoring left-handed users in some iterations and right-handed ones in others. As the size of the target decreased and the distance increased in each iteration, the time required to complete the pointing (P) and dragging (D) GOMS operators increased in each iteration.

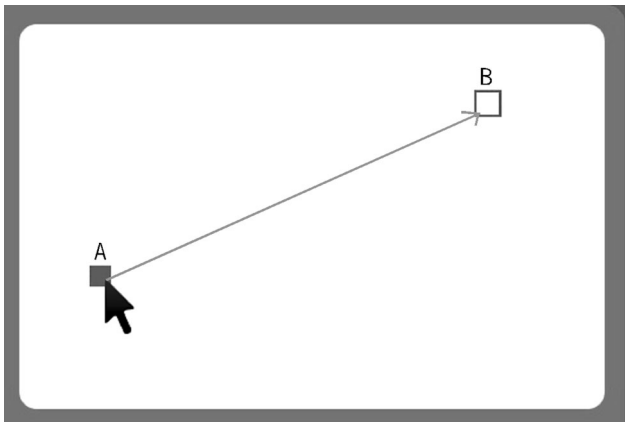


Fig. 2 During the *drag and drop* test, users had to drag a *dark rectangle* over an empty one. At each iteration, the position and size of both *rectangles* changed

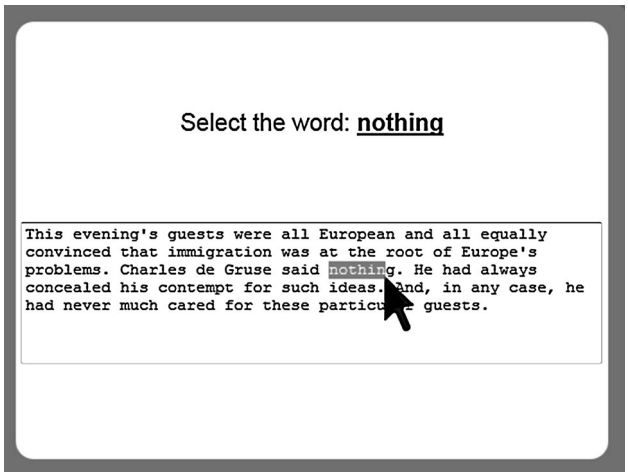


Fig. 3 In the *text selection* task, users need to point to one end of the text to be highlighted (P), click on it (K), drag the pointer over the text (D), and finally release the pointer (K)

The *text selection* task (Task 3) was used to measure the selection speed, requiring a complex combination of GOMS operators to reach the goal of selecting a text using the mouse. In this test, eight different texts and eight different words appeared successively. Each participant had to select the word in its appropriate context. The process required a pointing (P) operator to aim at the beginning (or end) of the text, a mouse/touch screen click operation (K) to activate the text selection, a drag (D) operation to move the pointer to the end of the text to be selected, and finally a mouse/touch screen click operation (K) to stop the selection process. It also required the mental (M) operator needed by the user to understand the text, which depends on his or her speed of reading. However, the mental operator was not considered or measured, since the test was constrained to measure the user's motoric speed.

Whenever a piece of text had been selected, a new text to be selected was shown to the user. A screenshot of this test is shown in Fig. 3. This time, the hidden agents recorded the time spent selecting each piece of text. The global amount of time required by each user to perform every single text selection operation was recorded as the result of this test.

5 Empirical methods

A sample made up of 630 individuals was used. As a first step, individuals who were unable to complete any of the tasks were removed, so the final analyzed sample was reduced to 557 individuals. The age breakdown is detailed in Table 1.

Table 1 Sample breakdown by age group

Age group	Number of cases	Percentage	Cumulative percentage
1	18	03.23	03.23
2	302	54.21	57.45
3	147	26.39	83.84
4	52	09.33	93.17
5	24	04.30	97.48
6	14	02.51	100.0
Total	557	100.0	

It can be seen that the sample used is a convenience sample where the proportions that represent each group do not match those that exist in the real population. This led the authors to use an experimental design, which is discussed later, that includes data transformation and robustness checks.

First of all, some descriptive statistics about both the dependent and independent variables were computed. Examination of such data gives us a first idea of the features of the individuals in the sample, their behavior in the experiment, and the statistical properties of the results.

Second, as an exploratory study of the data suggests the presence of remarkable degrees of skewness and kurtosis, data winsorizing was used. Winsorizing consists of replacing extreme observations (usually those above the 99 % or below the 1 % percentile) by a close value not considered as an outlier. This procedure has the advantage that it helps to mitigate the influence of outliers without reducing the sample size. Although winsorizing has not been previously used in usability research, it is a common practice in other fields where non-normal observations are analyzed (in this regard see [55, 56] as examples).

In this research, winsorizing was carried out by replacing the top and bottom 1 % of the observations of each task by the closest value not considered an extreme value. Data were subsequently transformed using natural logarithms, and a series of Lilliefors tests was applied to test for the normality of the distributions. The results provide evidence that in all cases, normality is not rejected at the 5 % significance level [1].

Once the normality of the data was ensured, a set of inferential tests was conducted for each of the tasks included in the experiment. Specifically, three types of tests were conducted:

1. Performance comparison between age groups. Each age group was compared with each one of the others with regard to performance on the three tasks. The two independent samples *t* test for the difference of means was used.
2. Performance comparison between subsamples. For each of the tasks, the sample was divided into two subsamples, using as cutoff points the age that separated the two adjacent age groups. The performance of one of the subsamples of the binary partition was also compared with that of the other by using the two independent samples *t* test for the difference of means.
3. Equality of proportions tests. For each of the tasks, individuals were classified into five groups depending on their performance. The first subgroup comprised 20 % of the individuals who achieved the best performance, the second subgroup the following 20 %, and so on. The cumulative percentages were analyzed, and a series of binomial tests were conducted to assess whether the observed age proportions in each of the performance groups differed significantly from those which would be obtained if individuals of each age group were equally distributed across the performance groups.

Finally, as a robustness check, Stages 1 and 2 of the inferential tests were repeated, though in this case the *t* tests were replaced by the nonparametrical Mann–Whitney test. Tests were conducted using raw data (data prior to winsorizing and logarithm transformation). The results, which are not included in this paper due to space restrictions, do not differ significantly from those explained in the following section.

6 Results

6.1 Descriptive statistics

Table 2 contains some descriptive statistics (mean, standard deviation, and quartiles) relative to the performance of each age group on each of the tasks. These statistics are referred to the raw data, that is, prior to the winsorization and log transformation.

The data suggest that individuals from older groups take more time to complete the three tasks. However, dispersion levels are very high in general, extreme values are somewhat common, and the distributions are always positively skewed (the mean is always higher than the median). This confirms the need for the winsorization/log transformation procedure that was outlined in the previous section.

Table 3 displays the same raw data descriptive statistics but relative to the different subsamples obtained by making binary partitions as explained in the previous section.

The results obtained are very similar to those of Table 2. Older individuals take more time to complete the tasks, but the observed distributions are leptokurtic and positively skewed. So, prior processing of the data is also evidenced.

Table 2 Performance of the individuals on each task by age group

Age group	Mean	Std Dev.	Q25	Median	Q75
Panel A: Task 1: pointing					
1	14,488.00	2,301.73	12,835.00	14,099.50	15,993.75
2	16,479.53	4,263.67	13,796.75	15,392.50	18,644.00
3	17,429.87	4,047.95	14,702.00	16,319.00	19,032.00
4	18,056.48	3,144.48	16,115.25	17,345.00	19,144.00
5	21,001.25	5,279.07	17,679.50	19,136.00	24,392.25
6	23,084.64	5,257.08	19,635.50	22,150.00	26,533.50
Total	17,174.05	4,371.48	14,232.50	16,064.00	19,090.50
Panel B: Task 2: drag and drop					
1	31,083.28	7,307.67	25,177.50	28,659.50	36,762.25
2	32,608.76	9,051.14	26,267.75	30,392.00	37,006.75
3	36,605.12	11,290.19	28,933.00	33,965.00	41,330.00
4	36,954.67	8,131.59	30,393.00	36,443.50	42,719.75
5	43,797.25	11,485.15	32,931.75	42,101.00	53,780.50
6	54,937.07	17,154.84	42,020.75	50,924.00	65,630.00
Total	35,063.18	10,763.09	27,649.00	32,273.00	39,855.00
Panel C: Task 3: text selection					
1	74,133.28	24,530.68	53,921.00	67,826.50	93,035.25
2	114,344.95	352,440.39	61,293.75	77,536.00	99,655.00
3	179,229.82	995,348.76	70,725.00	84,596.00	102,347.00
4	100,064.92	40,787.62	68,141.75	93,952.00	119,024.50
5	142,488.88	59,073.93	102,351.25	139,893.50	171,880.75
6	126,847.50	46,977.84	84,355.50	120,512.00	172,525.00
Total	130,363.25	573,350.58	64,861.50	82,093.00	106,830.00

Table 3 Descriptive statistics (splitting the sample according to the different age cutoff points)

Cutoff point (years)	Under the cutoff point					Above the cutoff point				
	Mean	Std dev.	Q25	Median	Q75	Mean	Std dev.	Q25	Median	Q75
Panel A: Task 1: Pointing										
20	14,488.00	2,301.739	12,835.00	14,099.50	15,993.75	17,263.75	4,396.858	14,313.00	16,115.00	19,161.00
30	16,367.50	4,200.794	13,737.00	15,320.00	18,372.25	18,263.05	4,369.853	15,203.00	17,146.00	19,749.50
40	16,701.91	4,178.250	13,993.00	15,504.00	18,671.00	19,623.91	4,553.667	16,453.25	18,689.50	21,527.25
50	16,837.63	4,104.205	14,136.00	15,864.00	18,724.00	21,768.82	5,298.237	17,893.25	20,367.00	24,687.75
60	17,021.66	4,244.401	14,200.00	15,930.00	18,939.00	23,084.64	5,257.082	19,635.50	22,150.00	26,533.50
Panel B: Task 2: drag and drop										
20	31,083.28	7,307.676	25,177.50	28,659.50	36,762.25	35,196.09	10,839.050	27,744.00	32,290.00	39,998.00
30	32,522.96	8,959.375	26,231.75	30,381.50	36,972.75	38,493.03	11,994.038	30,541.00	35,395.00	43,943.50
40	33,807.92	9,924.089	26,801.00	31,832.00	38,413.00	41,576.62	12,524.194	31,873.00	39,379.00	47,240.50
50	34,123.20	9,798.232	27,385.00	31,921.00	38,850.00	47,901.39	14,664.648	34,437.75	46,655.00	56,194.75
60	34,550.78	10,065.422	27,591.00	32,159.00	39,295.00	54,937.07	17,154.840	42,020.75	50,924.00	65,630.00
Panel C: Task 3: text selection										
20	74,133.28	24,530.687	53,921.00	67,826.50	93,035.25	132,241.06	582,752.943	65,338.00	82,487.00	107,575.00
30	112,083.04	342,525.116	61,121.00	76,033.50	98,331.50	155,045.39	784,102.854	71,703.00	89,030.00	117,845.00
40	133,219.18	625,847.634	63,660.00	79,677.00	100,301.00	115,544.16	50,332.776	77,759.25	107,922.50	148,157.00
50	129,897.37	593,825.385	64,198.00	80,566.00	102,232.00	136,726.26	54,801.106	95,536.50	136,265.50	169,424.75
60	130,453.90	580,662.412	64,389.00	81,713.00	105,735.00	126,847.50	46,977.842	84,355.50	120,512.00	172,525.00

Table 4 Age groups comparisons for Task 1: pointing

	1	2	3	4	5	6
1	–					
2	2.253 (0.025)	–				
3	3.439 (0.001)	2.884 (0.004)	–			
4	5.042 (<0.001)	3.519 (<0.001)	1.480 (0.141)	–		
5	5.588 (<0.001)	5.432 (<0.001)	3.939 (<0.001)	2.945 (0.004)	–	
6	6.773 (<0.001)	6.024 (<0.001)	4.893 (<0.001)	4.480 (<0.001)	1.316 (0.197)	–

Table 5 Age groups comparisons for Task 2: drag and drop

	1	2	3	4	5	6
1	–					
2	0.691 (0.490)	–				
3	2.224 (0.028)	4.238 (<0.001)	–			
4	2.997 (0.004)	3.839 (<0.001)	0.725 (0.469)	–		
5	4.401 (<0.001)	5.793 (<0.001)	3.245 (0.001)	2.877 (0.005)	–	
6	6.277 (<0.001)	7.341 (<0.001)	5.058 (<0.001)	5.298 (<0.001)	2.194 (0.035)	–

Table 6 Age group comparisons for Task 3: text selection

	1	2	3	4	5	6
1	–					
2	1.302 (0.194)	–				
3	2.347 (0.020)	2.139 (0.033)	–			
4	2.549 (0.013)	2.535 (0.012)	1.158 (0.248)	–		
5	5.026 (<0.001)	5.959 (<0.001)	5.210 (<0.001)	3.367 (0.001)	–	
6	4.152 (<0.001)	3.760 (<0.001)	3.240 (0.001)	2.062 (0.043)	–0.682 (0.499)	–

6.2 Inferential tests

Table 4 contains the results of the pairwise comparison of the different age groups with regard to the performance on Task 1. Each cell contains the t statistic of the *t* test for the difference of means (the test was performed by computing the difference between the column group and the row group). The corresponding *p* value is also shown below (in parentheses).

It is evidenced that for Task 1, at the usual significance levels (1 and 5 %), younger groups perform better than older groups. The only exceptions are the comparisons between Groups 3 and 4 and between Groups 5 and 6, where no significant differences arise.

Table 5 follows the same structure as Table 4 but contains information related to Task 2.

The results indicate that for Task 2, the performance of younger groups is also better than that of older groups in general. Exceptions are the comparisons between Groups 1 and 2 and between Groups 3 and 4.

Table 6 follows the same structure as Tables 4, 5 but contains information related to Task 3.

The data provide evidence that the pattern observed for Tasks 1 and 2 is once again present; that is, younger groups perform better. This time, the exceptions are the comparisons between Groups 1 and 2, 3 and 4, and 5 and 6.

Table 7 contains the results of the *t* tests conducted once the sample was split using the different cutoff points determined by the extremes of the age groups. In order to ease the interpretation of the results, summarized descriptive statistics (mean and standard deviation) are also provided, which in this case were computed after the winsorization of data but prior to the log transformation.

The results confirm the findings evidenced in Tables 4, 5 and 6. That is, younger individuals perform better. The only exception is Task 2 (drag and drop), considering 20 years as the cutoff point. Individuals under this age do not perform significantly better than those above.

Finally, the results of the analysis of the composition of each performance band for each of the three tasks are displayed in Table 8.

It is evidenced that in the high-performance bands (100–80 %, that is, the top 20 % of performers), the proportion of younger individuals is significantly higher than the proportion observed in the total sample, while older individuals are under-represented. For the low-performance bands (the bottom 20 % of performers), the opposite phenomenon is observed; that is, the proportion of younger people is lower than that of the sample, while that of older people is higher. For the intermediate band (60–40 %), the observed percentages do not differ significantly from those of the total sample.

Table 7 Tests of difference of means, splitting the sample according to the different cutoff points determined by age

Cutoff point (years)	Under		Above		*t* statistic	*p* value
	Mean	Std Dev.	Mean	Std Dev.		
Panel A: Task 1: pointing						
20	11,506.89	2,274.49	17,204.09	4,048.39	3.069	0.002
30	16,290.20	3,642.62	18,233.17	4,261.31	6.198	<0.001
40	16,637.89	3,756.36	19,602.56	4,489.20	6.860	<0.001
50	16,780.03	3,721.43	21,718.24	5,185.60	7.318	<0.001
60	16,963.06	3,880.47	23,084.64	5,257.08	5.407	<0.001
Panel B: Task 2: drag and drop						
20	31,083.28	7,307.67	35,022.22	10,013.08	1.706	0.089
30	32,436.08	8,370.27	38,214.90	10,935.35	7.213	<0.001
40	33,683.90	9,229.58	41,178.81	11,209.89	6.856	<0.001
50	34,011.61	9,171.16	46,959.21	12,317.95	7.614	<0.001
60	34,444.12	9,488.74	52,379.71	12,180.96	6.176	<0.001
Panel C: Task 3: text selection						
20	74,133.28	24,530.68	92,908.75	43,422.78	2.004	0.046
30	85,792.78	40,246.96	101,090.68	45,245.94	4.890	<0.001
40	87,822.78	40,071.42	133,683.90	49,229.58	5.832	<0.001
50	89,049.35	40,272.30	136,726.36	54,801.10	6.620	<0.001
60	91,411.32	42,649.69	146,847.50	46,977.84	6.232	<0.001

Table 8 Cumulative percentages of age groups in each performance band and binomial tests

Performance bands (%)	Cumulative percentage and binomial tests					
	1 (%)	1, 2 (%)	1, 2, 3 (%)	1, 2, 3, 4 (%)	1, 2, 3 4, 5 (%)	All (%)
Panel A: Task 1: pointing						
100–80	07.2*	82.0**	98.2**	100.0**	100.0	100.0
80–60	04.5	60.4	92.8**	97.3*	99.1	100.0
60–40	01.8	58.0	83.9	97.3	100.0	100.0
40–20	02.7	38.7**	71.2**	89.2	97.3	100.0
20–0	00.0*	48.2*	73.2**	82.1**	91.1**	100.0
Sample cum. percentage	03.23	57.45	83.84	93.17	97.48	100.0
Panel B: Task 2: drag and drop						
100–80	05.4	78.4**	98.2**	100.0**	100.0	100.0
80–60	03.6	68.5*	89.2	99.1**	100.0	100.0
60–40	01.8	51.8	82.1	92.0	99.1	100.0
40–20	04.5	53.2	82.9	95.5	98.2	100.0
20–0	00.9*	35.7**	67.0**	79.5**	90.2**	100.0
Sample cum. percentage	03.23	57.45	83.84	93.17	97.48	100.0
Panel C: Task 3: text selection						
100–80	06.3	72.1**	91.1*	97.3*	99.1	100.0
80–60	03.6	67.6*	91.0*	99.1**	100.0	100.0
60–40	01.8	53.6	87.5	95.5	97.3	100.0
40–20	03.6	51.4	83.8	93.7	97.4	100.0
20–0	00.9*	42.9**	66.1**	80.4**	93.8*	100.0
Sample cum. percentage	03.23	57.45	83.84	93.17	97.48	100.0

* Significant at the 5 % level; ** Significant at the 1 % level

7 Conclusions

The results show several interesting facts that must be taken into consideration in the following stages of this research. Aging, as expected and pointed out by most of the works discussed in the background section, has a strong influence on the performance of computer users.

After a more detailed analysis based on the GOMS task division, it can be argued that, as pointed out in [18], this influence differs between specific activities. In fact, results show that the aging effect is not the same for the three tasks that were analyzed. Although these facts were already observed in [18, 27], the more fine-grained grouping used in this work allowed the profiling of the six age groups for each of the three basic tasks that were analyzed. That provides the pillars for an automatic classification system that would help classify an anonymous user based on his or her performance, and thus to dynamically customize a user interface for his or her specific needs and constraints.

The results obtained are encouraging enough to be considered as the starting point for introducing more complex measurements (error rates, biometrics, and behavior patterns) and more variables under study, such as the user's gender and laterality.

Acknowledgments This work has been funded by the Department of Science and Technology (Spain) under the National Program for Research, Development and Innovation: project TIN2011-25978, entitled Obtaining Adaptable, Robust and Efficient Software by Including Structural Reflection to Statically Typed Programming Languages, and project TIN2009-12132, entitled SHUBAI: Augmented Accessibility for Handicapped Users in Ambient Intelligence and in Urban Computing Environments.

References

1. Frøkjær, E., Hertzum, M., Hornbæk, K.: Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? In: ACM CHI 2000 conference on human factors in computing systems, pp. 345–352 (2000)
2. Agudo, J.E., Sánchez, H., Rico, M.: Playing games on the screen: adapting mouse interaction at early ages. In: 2010 10th IEEE international conference on advance learning technology, pp. 493–497 (2010)
3. Donker, A., Reitsma, P.: Aiming and clicking in young children's use of the computer mouse. Comput. Hum. Behav. **23**, 2863–2874 (2007)
4. Donker, A., Reitsma, P.: Young children's ability to use a computer mouse. Comput. Educ. **48**, 602–617 (2007)
5. Inkpen, K.M.: Drag-and-drop versus point-and-click mouse interaction styles for children. ACM Trans. Comput. Interact. **8**, 1–33 (2001)
6. Chaparro, A., Bohan, M., Fernandez, J., Choi, S.D., Kattel, B.: The impact of age on computer input device use: psychophysical and physiological measures. Int. J. Ind. Ergon. **24**, 503–513 (1999)
7. Dickinson, A., Arnott, J., Prior, S.: Methods for human–computer interaction research with older people. Behav. Inf. Technol. **26**, 343–352 (2007)
8. Gregor, P., Newell, A.F., Zajicek, M.: Designing for dynamic diversity—interfaces for older people. In: Proceedings of the fifth international ACM conference on assistive technologies, pp. 151–156. ACM, New York, NY, USA (2000)
9. Slegers, K., van Boxtel, M.P.J., Jolles, J.: Computer use in older adults: determinants and the relationship with cognitive change over a 6 year episode. Comput. Hum. Behav. **28**, 1–10 (2012)
10. Hwang, F., Keates, S., Langdon, P., Clarkson, J.: Mouse movements of motion-impaired users: a submovement analysis. In: Proceedings of ACM SIGACCESS conference on Computer Access. ASSETS'04, p. 102 (2004)
11. Keates, S., Hwang, F., Langdon, P., Clarkson, P.J., Robinson, P.: Cursor measures for motion-impaired computer users. In: Proceedings of fifth international ACM conference assistive technology—Assets'02, p. 135 (2002)
12. Hurst, A., Mankoff, J., Hudson, S.E.: Understanding pointing problems in real world computing environments. In: Proceedings of 10th international ACM SIGACCESS conference computer access. Assets'08, p. 43 (2008)
13. Strom, P.S., Strom, R.D., Wingate, J.J., Kraska, M.F., Beckert, T.E.: Cyberbullying: assessment of student experience for continuous improvement planning. NASSP Bull. **96**, 137–153 (2012)
14. Hill, R., Dickinson, A., Arnott, J., Gregor, P., Mciver, L.: Older web users' eye movements: experience counts. In: CHI'11 proceedings of the SIGCHI conference on human factors in computing systems. pp. 1151–1160. ACM, New York, NY, USA (2011)
15. Fisk, A.D., Rogers, W.A., Charness, N., Czaja, S.J., Sharit, J.: Designing for older adults: principles and creative human factors approaches. CRC Press, Boca Raton (2009)
16. Hargittai, E., Hinnant, A.: Differences in young adults' use of the internet. Commun. Res. **35**, 602–621 (2008)
17. Fox, S., Madden, M.: Generations Online. http://www.pewinternet.org/2006/01/22/generations-online/
18. Cheong, Y., Shehab, R.L., Ling, C.: Effects of age and psychomotor ability on kinematics of mouse-mediated aiming movement. Ergonomics **56**, 1006–1020 (2013)
19. Loos, E.: In search of information on websites: a question of age? University access human–computer interact. Users Divers. Lect. Notes Comput. Sci. **6766**, 196–204 (2011)
20. Card, S.K., Moran, T.P., Newel, A.: The keystroke-level model for user performance time with interactive systems. Commun. ACM **23**, 396–410 (1980)
21. Schrepp, M.: GOMS analysis as a tool to investigate the usability of web units for disabled users. Univ. Access Inf. Soc. **9**, 77–86 (2010)
22. Xiang, L.X.L., Xiaoli, C.X.C.: The research on performance of automobile human-machine interface based on BHR-GOMS behavior model. In: Intelligent computer intelligent systems (ICIS), 2010 IEEE international Conference 2 (2010)
23. Oyewole, S.A., Haight, J.M.: Determination of optimal paths to task goals using expert system based on GOMS model. Comput. Hum. Behav. **27**, 823–833 (2011)
24. Abdulin, E.: Using the keystroke-level model for designing user interface on middle-sized touch screens. In: Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems, pp. 673–686. ACM (2011)
25. Phillips, J.G., Triggs, T.J.: Characteristics of cursor trajectories controlled by the computer mouse. Ergonomics **44**, 527–536 (2001)
26. Guiard, Y., Olafsdottir, H.B., Perrault, S.T.: Fitt's law as an explicit time/error trade-off. In: Proceedings of the 2011 annual conference on human factors in computing systems—CHI'11, p. 1619. ACM Press, New York, New York, USA (2011)
27. Salthouse, T.A.: Effects of age and skill in typing. J. Exp. Psychol. Gen. **113**, 345–371 (1984)

28. Schneider, D.W., Anderson, J.R.: A memory-based model of Hick's law. Cogn. Psychol. **62**, 193–222 (2011)
29. Crook, C.: Young children's skill in using a mouse to control a graphical computer interface. Comput. Educ. **19**, 199–207 (1992)
30. Phillips, J.G., Triggs, T.J., Meehan, J.W.: Cursor orientation and computer screen positioning movements. Hum. Factors J. Hum. Factors Ergon. Soc. **43**, 435–441 (2001)
31. Whisenand, T.G., Emurian, H.H.: Analysis of cursor movements with a mouse. Comput. Hum. Behav. **15**, 85–103 (1999)
32. Tränkle, U., Deutschmann, D.: Factors influencing speed and precision of cursor positioning using a mouse. Ergonomics **34**, 161–174 (1991)
33. Mackenzie, I.S., Buxton, W.: A comparison of input devices in element pointing and dragging tasks. In: CHI'91 Proceedings of the SIGCHI conference on human factors in computing systems, pp. 161–166. ACM, New York, NY, USA (1991)
34. Joiner, R., Messer, D., Light, P., Littleton, K.: It is best to point for young children: a comparison of children's pointing and dragging. Comput. Hum. Behav. **14**, 513–529 (1998)
35. Wood, E., Willoughby, T., Schmidt, A., Porter, L., Specht, J., Gilbert, J.: Assessing the use of input devices for teachers and children in early childhood education programs. Inf. Technol. Child. Educ. Annu. **2004**, 261–280 (2004)
36. Kuhtz-Buschbeck, J.P., Stolze, H., Jöhnk, K., Boczek-Funcke, A., Illert, M.: Development of prehension movements in children: a kinematic study. Exp. Brain Res. **122**, 424–432 (1998)
37. Stubbs, N.B., Fernandez, J.E., Glenn, W.M.: Normative data on joint ranges of motion of 25- to 54-year-old males. Int. J. Ind. Ergon. **12**, 265–272 (1993)
38. Walker, N., Philbin, D.A., Fisk, A.D.: Age-related differences in movement control: adjusting submovement structure to optimize performance. J. Gerontol. B Psychol. Sci. Soc. Sci. **52**, P40–P52 (1997)
39. Metter, E.J., Conwit, R., Tobin, J., Fozard, J.L.: Age-associated loss of power and strength in the upper extremities in women and men. J. Gerontol. Ser. A Biol. Sci. Med. Sci. **52A**, B267–B276 (1997)
40. Wagman, I.H., Lesse, H.: Maximum conduction velocities of motor fibers of ulnar nerve in human subjects of various ages and sizes. J. Neurophysiol. **15**, 235–244 (1952)
41. Fozard, J.L., Vercryssen, M., Reynolds, S.L., Hancock, P.A., Quilter, R.E.: Age differences and changes in reaction time: the Baltimore Longitudinal Study of Aging. J. Gerontol. **49**, P179–P189 (1994)
42. Spirduso, W.W.: Physical fitness, aging, and psychomotor speed: a review. J. Gerontol. **35**, 850–865 (1980)
43. Rikli, R., Busch, S.: Motor performance of women as a function of age and physical activity level. J. Gerontol. **41**, 645–649 (1986)
44. Kallman, D.A., Plato, C.C., Tobin, J.D.: The role of muscle loss in the age-related decline of grip strength: cross-sectional and longitudinal perspectives. J. Gerontol. **45**, M82–M88 (1990)
45. Chadwick-Dias, A., McNulty, M., Tullis, T.: Web usability and age. ACM SIGCAPH Comput. Phys. Handicap. **73–74**, 30–37 (2002)
46. Czaja, S.J., Lee, C.C.: The impact of aging on access to technology. Univ. Access Inf. Soc. **5**, 341–349 (2006)
47. Walker, N., Millians, J., Worden, A.: Mouse accelerations and performance of older computer users. In: Proceedings of human factors Ergon Society Annual Meeting. **40**, 151–154 (1996)
48. Tullis, T.S.: Older adults and the web: lessons learned from eye-tracking. In: UAHCI'07 proceedings of the 4th international conference on Universal access in human computer interaction: coping with diversity, pp. 1030–1039. Springer (2007)
49. Xie, B.: Older adults, computers, and the internet: future directions. Gerontechnology **2**, 289–305 (2003)
50. Morrell, R.W.: Older Adults, Health Information, and the World Wide Web. L. Erlbaum Associates Inc., Mahwah (2001)
51. Nap, H.H., De Greef, H.P., Bouwhuis, D.G.: Access for all by cognitive engineering. Gerontechnology **3**, 258 (2005)
52. Freudenthal, D.: Age differences in the performance of information retrieval tasks. Behav. Inf. Technol. **20**, 9–22 (2001)
53. Sayers, H.: Desktop virtual environments: a study of navigation and age. Interact. Comput. **16**, 939–956 (2004)
54. Neerincx, M.A., Lindenberg, J., Rypkema, J.A., van Besouw, N.J.P.: A practical cognitive theory of web-navigation: explaining age-related performance differences. In: Position paper. Workshop basic research symposium CHI2000. ACM (2000)
55. Ramprasath, L., Singh, K.: Statistical options: crash resistant financial contracts based on robust estimation. Stat. Probab. Lett. **77**, 196–203 (2007)
56. Wu, M., Zuo, Y.: Trimmed and Winsorized means based on a scaled deviation. J. Stat. Plan. Inference. **139**, 350–365 (2009)

C Article 2: Towards an automatic user profiling system for online information sites.

Towards an automatic user profiling system for online information sites

Identifying demographic determining factors

Javier De Andrés
Department of Accounting, University of Oviedo, Oviedo, Spain, and
Beatriz Pariente, Martin Gonzalez-Rodriguez and Daniel Fernandez Lanvin
Department of Computing, University of Oviedo, Oviedo, Spain

Received 3 July 2014
First revision approved 21 October 2014

Abstract

Purpose – The purpose of this paper is to identify demographic differences based on how users interact with web applications. The research is needed to develop future systems able to adapt the representation of online information to the user's specific needs and preferences improving its usability. The following question guides this quest: is there a direct relationship between age and/or gender and interaction?

Design/methodology/approach – GOMS (goals, operators, methods, and selection rules) analysis was used to reduce complex interaction tasks into basic operators like pointing, dragging, typing, etc. An experiment was designed to analyse the user performance in the use of these operators through five complex tasks: point-and-click, drag-and-drop, text selection, text edition and menu selection. The sample comprises 592 individuals which took part in the experiment. The performance was analysed using multivariate regression analysis. User laterality and the the user experience were used as control variables.

Findings – The factors studied are significant enough to support user classification. The analysis evidenced that men performed significantly better than women when executing interaction pointing and dragging GOMS's operators, but no significant differences arose with regard to the performance in the typing operators. Older users performed worse in all the interaction tasks. No significant performance differences were detected between left and right-handed users.

Research limitations/implications – The study pretends to lay the ground for developing artificial intelligence-based classification systems (e.g. neural networks, decision trees, etc.) able to detect significant differences in user performance, classifying users according to their age, gender and laterality.

Practical implications – This user profiling would drive the organisation, selection and representation of the online information according to the specific preferences and needs of each user. This would allow the design of new personalisation algorithms able to perform dynamic adaptation of user interfaces in order to improve the usability of online information systems.

Originality/value – This work extends previous research on user performance under a new approach and improved accuracy. First, it relies on the combined and simultaneous analysis of ageing and gender and the use of user laterality and experience as control variables. Second, the use of the GOMS analysis allowed the design of tests that closely resemble the user interaction in online

This work has been funded by the Department of Science and Technology (Spain) under the National Programme for Research, Development and Innovation: project TIN2011-25978 entitled Obtaining Adaptable, Robust and Efficient Software by Including Structural Reflection to Statically Typed Programming Languages and project TIN2009-12132 entitled SHUBAI: Augmented Accessibility for Handicapped Users in Ambient Intelligence and in Urban Computing Environments.

Online Information Review
Vol. 39 No. 1, 2015
pp. 61-80

1468-4527
DOI 10.1108/OIR-06-2014-0134

information systems. Third, the size of the sample used in this analysis is much bigger than those used in previous works, allowing a more thorough data analysis which includes the estimation of an advanced model which is quantile regression.

Keywords Quantile regression, Demographic factors, GOMS, User interaction, User profiling
Paper type Research paper

Introduction

The success of online information systems is highly dependent on the way they manage the information they contain. Selecting the right information to show to each specific user is a critical factor for the success of a site (Liu and Arnett, 2000). Moreover, appropriate representation and layout according to the user's specific needs can also influence the user's satisfaction (Yan and Guo, 2010). These two decisions are strongly affected by several factors that are usually referred to as "user profile" or "user model" (Lynch and Beck, 2001). Therefore the quick identification of common interaction patterns among users is critical for web designers in order to adapt their information systems to meet the users' diverse needs (Kambil *et al.*, 2000).

This "user profiling" can be determined by several factors that are of great relevance to the success of the online information systems of companies (Sebora *et al.*, 2008; Turban *et al.*, 2000) since differences in age, gender and income indicate different interaction intentions (Xu, 2006). There is strong empirical evidence showing that important behaviour differences caused by gender, age, social grouping and household income have a great influence on how the user accesses online services (Haque *et al.*, 2006; Lee, 2009).

However, given that most of the potential customers of online sites are anonymous users, performing user profiling at the design stage is a difficult task to carry out without falling into misconceptions, since both age and gender parameters are usually unknown. Even when the user provides this data as part of the registration process, its reliability can be reduced by social or cultural factors. Underage users, for example, usually lie to avoid controls in adult-oriented sites or social networks (Strom *et al.*, 2012).

The measurement and analysis of performance in certain tasks could help to draft a user's profile. Age and gender are two user features which could be a key to obtaining a profile of an individual and whose influence on user behaviour has been reported in some scenarios (Weiser, 2000; Freudenthal, 2001). However, these specific experiments do not demonstrate whether this influence is significant enough to support user classification in the context of web interaction. Also, those authors did not conduct a joint evaluation of the effect of such factors in user interaction. Moreover, neither did they take into account other features such as laterality (left handed or right handed) and experience in the use of computers. The lack of control for these factors may bias the results as they may have an influence on the performance level.

This work intends to fill this gap by measuring the performance of a representative sample of users (which for the first time includes right- and left-handed people, individuals of both genders and of all ages) performing – also for the first time – five basic tasks required to interact with information systems: point-and-click, drag-and-drop, text selection, text editing and menu item selection.

The analysis shows that both age and gender influence performance in some of the tasks. These promising results could allow the design of systems able to infer the gender and the age of the user, employing adaptive web design techniques to enhance

the user experience of a web information system, adapting its contents to the user's requirements.

The rest of the document is structured as follows. In the next section we introduce GOMS (goals, operators, methods and selection rules) as the performance analysis method. We also analyse previous research based on age and gender differences when interacting with computers. The subsequent section describes the five different tests executed by the volunteers, the recruitment method and experiment considerations. The section after that describes the methodology used for the statistical analysis of the gathered data. The final section discusses the findings of the tests and their implications for the design of adaptive user interfaces for information systems.

Background

GOMS analysis has been successfully used to estimate the user's performance in different interaction scenarios including web sites (Schrepp, 2010; Oyewole and Haight, 2011), touch screens (Abdulin, 2011) and motor vehicles (Xiang and Xiaoli, 2010).

GOMS analysis assumes that users are computer literate and know how to do the task under observation, without generating any errors in the process. Due to their degree of expertise, users are able to interact as fast as possible. Therefore GOMS focuses on estimating the user's effectiveness (performance) instead of the user's efficiency.

GOMS simplifies the user's interaction process by considering complex tasks as combinations of a limited set of basic actions called operators. Complex task estimations are simplified by estimating the performance in the basic operators required by the high-level task.

Some of the main basic operators used by GOMS are "pointing" (denoted by P), "dragging" (D), "key pressing" (K) and "mental" (M). The latter is used whenever the user has to make a simple decision (e.g. selecting an item in a menu).

So, for example, the sequence of GOMS operators required to execute a drag-and-drop task would be: placing the pointer over the movable object (P), holding it down (K) to keep the object selected, dragging the object (D) to the desired location and finally releasing it (K) (Crook, 1992; Inkpen, 2001). Text selection, text editing and menu item selection tasks use similar approaches.

The execution of each operator requires a specific runtime which is mostly based on the individual performance of each user. Therefore the runtime needed to achieve a complex task would be estimated as the sum of the runtimes required by each operator. In the drag-and-drop example the estimated execution time would be calculated as the pointing time (T_P), dragging time (T_D) and the time consumed by the two key-press operators (T_K); that is $T = T_P + T_D + 2T_K$.

The execution time required for any of the GOMS operators can be roughly estimated by means of psychology laws.

Hick's law estimates the time required to make a decision (T_M) (e.g. selecting an item from a radio button group, from a menu, etc.) as a linear increase related to the logarithm of the number of alternatives the user has to evaluate (e.g. the number of items in a radio button group) (Schneider and Anderson, 2011). Fitts' Law (Zhai, 2004) estimates the time required by a user to point/drag to/from a target (T_P and T_D) (e.g. moving the mouse pointer over an object on the screen) as a function of the distance to the target and its size (Guiard *et al.*, 2011). The Salthouse (1984) regularities predict the time required by a user to type text of a given size (T_K).

It was noted that there are external variables that may increase or decrease the estimations predicted by these psychology laws. So for instance aiming (P) at a square target is faster (T_P) that n aiming at a round target when both objects have the same width (Phillips and Triggs, 2001; Crook, 1992; Inkpen, 2001). Pointing (P) requires several interactions when aiming for the target, reducing its speed in each one to recalculate the trajectory in order to increase the accuracy (Phillips and Triggs, 2001). Fitts' Law estimates the runtime needed to perform all these recalculations as $a+b\ \mathrm{Log}_2\ (D/S)$ where D is the distance to the target and S represents its size, being the parameters a and b which are user dependent correction factors. Since the surface of the square object is slightly greater than that of the round object, the estimated time taken to point at the square object is less.

Although these psychology laws provide predictions for the expected reaction time of the average user, they must be tuned using correction factors to provide accurate estimations for individual users. That is the case for the a and b parameters commented on in the example of Fitts' Law. These correction factors rely on the external variables we study in this research, such as the user's age and gender.

Age

Prior research demonstrates that ageing is associated with changes in characteristics such as visual perception, poorer motor control, sensory deficits and some aspects of memory and cognition (Dickinson *et al.*, 2007). These factors could negatively affect the ability to use computers (Hill *et al.*, 2011).

Muscle strength begins to decline in people aged over 40 (Metter *et al.*, 1997), leading to reduced mobility and causing difficulties in the execution of movements (Walker *et al.*, 1997). Considering the slowing of the conduction velocity of nerve signals, studies have shown that the changes could be related to older people having slower reaction times (Fozard *et al.*, 1994), delayed movements, a decline in motor skills, along with decreases in coordination capacity, agility and balance, all closely related to the coordination capacity (Rikli and Busch, 1986). However, some studies claim that even assuming that age decreases motor control (Walker *et al.*, 1997) there is no strong negative effect on the performance of the P operator because older users may compensate for the decrease in perceptual efficiency by adjusting the velocity and number of sub-movements required by Fitts' Law to hit the target (Walker *et al.*, 1997).

Only very young children perform worse in P and D operators. Children up to eight years old have major difficulties in attaching and lifting objects, causing a variety of difficulties when making subtle and delicate movements that require hand/eye coordination (Kuhtz-Buschbeck *et al.*, 1998). This is due to the fact that their skills are further determined by cognitive factors and psychometric abilities rather than motor skills (Agudo *et al.*, 2010). As a result the speed of movement execution constantly evolves up to the age of 12 years (Kuhtz-Buschbeck *et al.*, 1998).

Complex high-level tasks such as drag-and-drop are especially difficult for such users due to the fact that the K operator, used to hold down the mouse button/finger, is highly demanding with respect to motor skills (MacKenzie *et al.*, 1991) and requires the combined use of several operators (D and K) at the same time (Chadwick-Dias *et al.*, 2002; Czaja and Lee, 2006). These users may feel more comfortable if the drag-and-drop tasks are replaced by pointing-move-pointing tasks, as they could be easily resumed from the last pointing task in the event of failure (Joiner *et al.*, 1998).

Other studies revealed that ageing also affects perception and cognition, which have an influence on learning and using digitised systems (Xie, 2003). Senior users have been found to be slower than young adults when performing information retrieval tasks (Nap *et al.*, 2005), 3D navigation on desktop systems (Sayers, 2004) and web navigation (Neerincx *et al.*, 2000).

Nevertheless, although the reaction time of the M operator declines with age (Salthouse, 1984), the performance of the K operator, thought to be highly correlated with age, is in fact consistent across age groups (Cheong *et al.*, 2013).

All these previous results suggest that we can expect a negative influence of ageing on most facets of computer use. However, most of the experiments based on ageing are too specific and/or high level and do not determine whether these differences are significant enough to be observed in more general and fine-grained operations that could be required when interacting with common web applications. Furthermore, some of them cover only specific age groups and cannot be extrapolated to the whole internet user population. These gaps lead us to formulate the first hypothesis to be verified in the experiment:

H1. There is a direct relationship between age and performance time.

Gender

Gender differences have several effects related to computer interaction. It was observed, for example, that males' performance in navigating through virtual environments was better with smaller displays than females', while with larger displays, females' performance improved and males' performance was not negatively affected (Tan *et al.*, 2003). Regarding control devices, Crook (1992) concluded that there are no gender-correlated differences in children using a mouse, but other researchers observed different results. Inkpen (2001) states that girls have more difficulties than boys using a drag-and-drop interaction style as opposed to a point-and-click interaction style. In her experiment she compared both interactions in children, observing that while no significant effect of gender was found in overall movement time and general errors, there was in pickup and drop errors, and also between gender and target size. The experiment conducted by Rohr (2006) found that gender-specific movement biases emphasise speed for men and accuracy for women.

These related studies demonstrate differences in the way women and men use the mouse for common actions such as drag-and-drop or point-and-click. However, it is not sufficiently clear whether these differences are significant enough to support user classification while executing fine-grained basic operations, so we formulate the following hypotheses:

H2. Women's performance in point-and-click mouse operations is better than men's.

H3. Men's performance in drag-and-drop operations is better than women's.

Design of the experiment

A web site was designed to include five tests based on the combined use of GOMS operators. As the GOMS analysis requires participants who know how to perform the tasks under observation, it was considered that users who are familiar with the relatively complex user interfaces of discussion web forums matched this profile.

Therefore an invitation was sent to one of the most popular Spanish discussion forums (www.forocoches.com) asking for volunteers to undertake the online tests.

After the completion of the test, the users were asked to invite their colleagues and friends through Twitter and Facebook and most of them did so. The data gathering process was stopped after three days.

This online recruiting process not only agrees with the GOMS principles but also allows the gathering of large samples. Unlike most of the studies described in the previous section, which were based on samples which contained fewer than 15 individuals, this experiment involved 592 participants. The large size of the sample allowed us to use multivariate regression analysis and to obtain more accurate results in comparison with those of prior studies.

Previous works suggest that user laterality could affect performance. Several studies have reported differences between left- and right-handed individuals regarding movement execution (Lenhard and Hoffmann, 2007; Mieschke *et al.*, 2001), movement preparation (Bestelmeyer and Carey, 2004; Neely *et al.*, 2005), stimulus velocity effect (Rodrigues *et al.*, 2012) or interactions between hand preference and hand performance (Peters and Ivanoff, 1999). These precedents suggest that user laterality can have a reasonable influence on user behaviour and performance.

Also, the number of hours per week the user spends interacting with computers seems to be another determining factor. These two variables – laterality and hours of use – were incorporated into the study as control variables.

Prior to executing the tests proposed for this experiment, the users provided specific information about themselves using an online questionnaire: age, gender, laterality and number of hours per week spent using computers.

During the experiment the users had to complete five tests, requiring the use of different GOMS operators to execute each one. The average execution time for each operator (P, D, M or K) needed to complete each high-level task was recorded in milliseconds and was denoted by $T_{P,}$ T_{D}, T_{M} and $T_{K.}$

Test 1 was designed to measure the performance of point-and-click tasks. Users had to click on several red squares which appeared in different screen locations, one at a time (see Figure 1). Every time the user clicked on a square, it disappeared and a new smaller one appeared in a different location. A total of 14 different squares were used in the test and their locations and size were the same for every person tested.

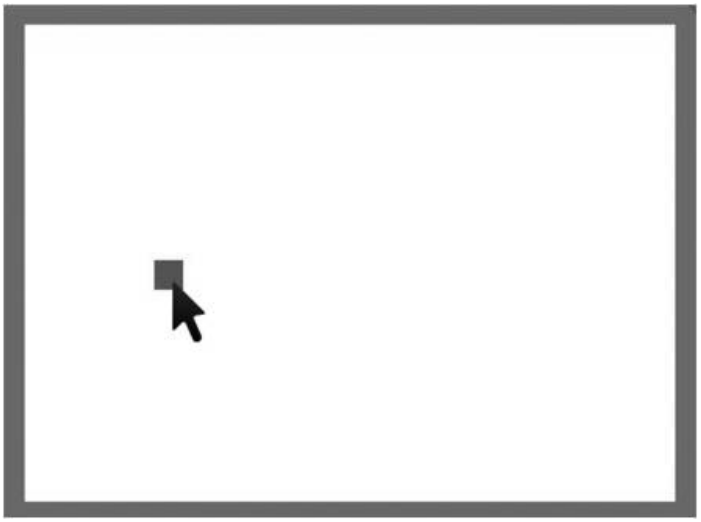


Figure 1.
During the point-and-click task of test 1, the user had to click anywhere inside each square as quickly as possible

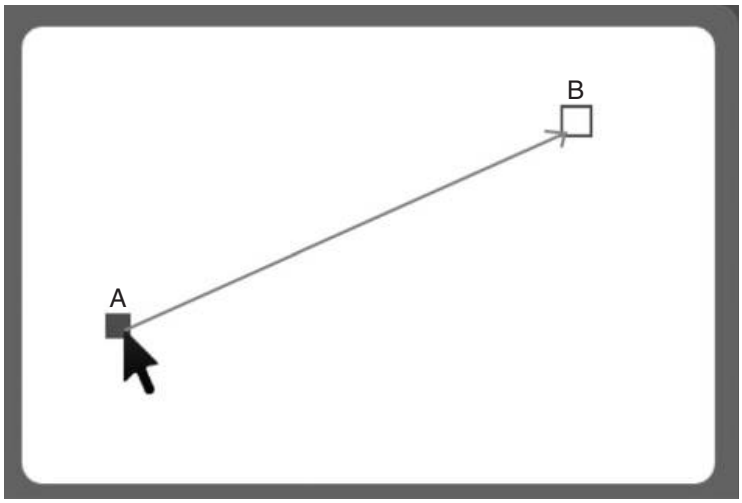
Although the execution of this task requires two GOMS operators (P to point to the target and K to click on it), only T_P was recorded.

Test 2 analysed the user's performance in the execution of drag-and-drop tasks. During the test users had to drag a red square (A), dropping it over an empty one (B) (see Figure 2). Afterwards a new pair with a smaller red square appeared in a different screen location. The size of the dragging square was two thirds of the surface of the target. This drag-and-drop process was repeated 12 times during the test. Each task involved three different GOMS operators: P, K and D, but only the T_D time was recorded.

The location and size of every object used in tests 1 and 2 were selected, modifying the parameters on which Fitts' Law is based, increasing the level of difficulty in iteration. At the beginning of each task the size of the squares was relatively big, but it decreased in each interaction. In the same way the initial distance to the target was relatively short but it was increased in each interaction. As a result the T_P and T_D increased during the test. On the other hand, the target location was changed from top to bottom and from left to right using variations of a Z shaped pattern. The objective was to favour left-handed individuals in some iterations and right-handed in others, keeping a balanced level of difficulty for each kind of user.

Test 3 measured the user's performance in the execution of text selection tasks. In each interaction users were asked to select a word in an appropriate text context using a pointer (see Figure 3). This process was repeated eight times using a different word each time. This task required a combination of different GOMS operators. The M operator was needed to select the word. The P operator was required to aim for the beginning (or the end) of the text. Next, a K operator was needed to activate the text selection. The D operator was required to drag over the desired word. Last, the K operator was needed to deactivate the text selection. The recorded performance time was obtained as $T_M+T_P+T_D+T_K$.

Test 4 recorded the user's performance in text editing tasks. Each iteration required users to type a provided sentence (see Figure 4). The test included five iterations (sentences). In order to write the sentences the users need to execute the K operator as many times as letters are included in the sentences. Therefore T_K is the variable measured in this test.



Note: Each time, the position and size of both squares changed

Figure 2.
During the drag-and-drop test users had to drag a red square over an empty one

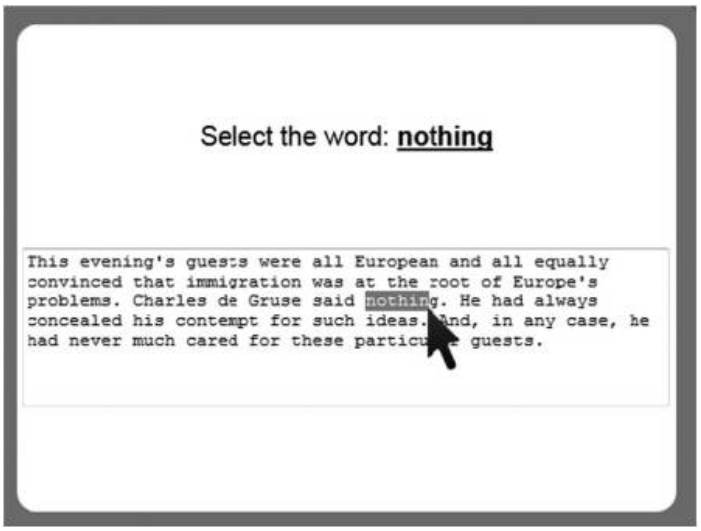


Figure 3.
During the text selection test users were asked to select a given piece of text included in a paragraph

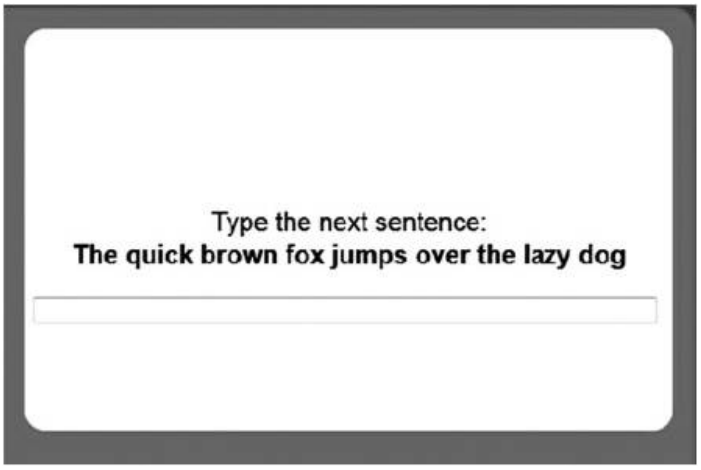


Figure 4.
In the text editing test users had to write known pieces of text as quickly as possible

Finally, test 5 measured the user's performance in the execution of menu item selection tasks. The user was asked to select a specific item from a menu in each interaction. In order to achieve this task the execution of P_1 and K_1 operators was required to activate the menu (clicking on its title), followed by the execution of an M operator needed to make the decision about which menu item should be selected. The task was completed executing a P_2 operator to move the pointer over the item followed by a K_2 operator to select it. This test measured the average amount of time required to select all the items proposed in the test and was calculated as $T_{P1}+T_{K1}+T_M+T_{P2}+T_{K2}$. The test included ten iterations and the number of options in all the menus remained the same for all of them (Figure 5).

Empirical methods

Variables in the study

The experiment was completed by 592 participants. Table I lists the variables used in the study.

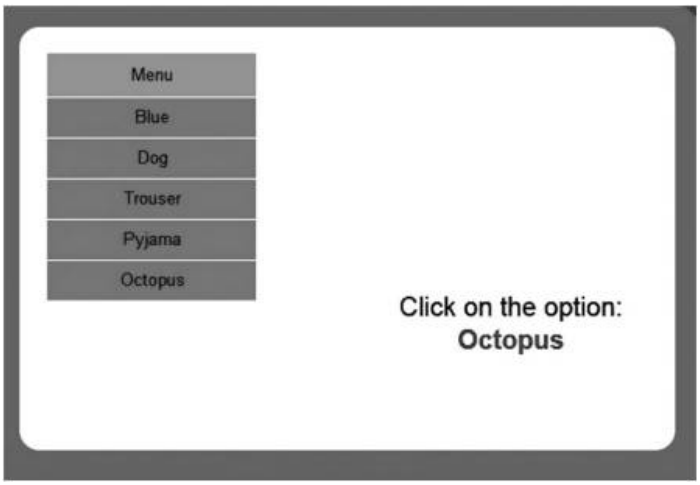


Figure 5.
In the menu selection test users had to select a given item in a menu

Name	Definition	
Point-and-click	Time (T_P) required to pointing (P) each object during the test (measured in milliseconds)	
Drag-and-drop	Time (T_D) required to drag (D) each object during the test (measured in milliseconds)	
Text selection	Time required to select each word during the test. It was calculated as $T_M+T_P+T_D+T_K$ (measured in milliseconds)	
Text edition	Time required to select write each word during the test. It was calculated as T_K (measured in milliseconds)	
Menu selection	Time required to select each menu item during the test. It was calculated as $T_{P1}+T_{K1}+T_M+T_{P2}+T_{K2}$ (measured in milliseconds)	
Timetotal	Sum of $Task_1$ to $Task_5$	
Age	Group	years
	0	0-15
	1	16-20
	2	21-25
	3	26-30
	4	31-35
	5	36-40
	6	41-45
	7	46-50
	8	51-55
	9	56-60
	10	61-65
	11	66+
HoursUse	Number of hours of computer use	
Gender	1 = female, 0 = male	
LeftHanded	1 = Left handed, 0 = right handed	

Table I.
Variables in the study

Methodology

First, we produced some descriptive statistics about both the dependent and independent variables. The examination of such data gives us a first impression of the features of the individuals in the sample and their behaviour in the experiment.

Second, as we had a sample size which is much bigger than those of prior works in this area, we used multivariate statistical analysis instead of the univariate techniques

employed by other authors. As explained above this allowed the inclusion of control variables in the models, as well as the joint analysis of the two studied factors (age and gender).

So, for each of the tasks and the total time, and in order to test hypotheses *H1*, *H2* and *H3*, we estimated an ordinary least squares (OLS) linear regression model. The regression equations have the following form:

$$\text{Task}_i = a_0 + a_1 \times \text{Age} + a_2 \times \text{HoursUse} + a_3 \times \text{Gender} + a_4 \times \text{LeftHanded} + \varepsilon$$

where Task_i is the dependent variable in each one of the models, a_0 is the intercept term, a_1 to a_5 are the coefficients of the independent variables in the models and ε is the error term.

With regard to these equations, and as prior robustness checks, we tested for multi-collinearity and heteroscedasticity. Multi-collinearity tests were conducted through the calculation of the condition indices and the variance inflation factors. For all the models the condition indices are below 15 and the variance inflation factors below 10, which are common thresholds to discard the presence of significant multi-collinearity among the variables of a linear regression model (Kutner *et al.*, 2004), However, the heteroscedasticity test (Breusch and Pagan, 1979; Cook and Weisberg, 1983) suggests the presence of a significant level of heteroscedasticity. Therefore we used robust standard errors for the computation of the p-values which are reported in the results section.

The aforementioned heteroscedasticity in the regression results also suggests that the influence of each one of the hypothesised factors on the performance of the considered tasks could vary depending on the relative level of performance achieved by the individual. That is, the influence of each one of the independent variables could differ depending on whether the individual performs better (or worse) than expected according to his/her characteristics. Figure 6 exemplifies this for the univariate relationship between a generic factor and the performance.

It can be seen that the dots representing the behaviour of each one of the individuals follow a heteroscedastic pattern, that is, the dispersion is higher as the level of the

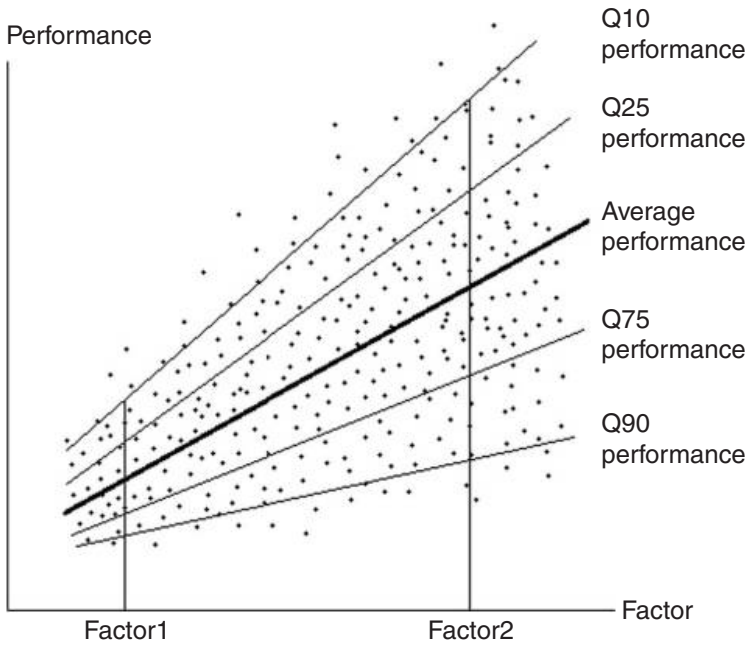


Figure 6.
The quantile regression approach

factor grows. The central line represents the average relationship between the factor and the measured performance. This line is obtained through standard approaches such as, for example, OLS regression explained above. The slope of this line (coefficient of the regression) is a measure of the mean effect of variations in the factor on the performance. However, Figure 6 also shows that as the dispersion of performance increases with the factor, when we trace conditional distributions we obtain lines with different slopes. In the Figure we have drawn the lines which pass by the 10, 25, 75 and 90 conditional quantiles. For example, the Q75 performance line indicates that, for each level of the factor, 75 per cent of the individuals having that level perform worse than the threshold stated by the Q75 line. Please note that, as in our case, performance is measured by the time used for the completion of a task; lines for the best performers are below those for bad performers. It is evident that, because of the heteroscedasticity, when the level of the factor is low (i.e. Factor1 in the figure) the difference between good and bad performers is not very high. However, for higher levels of the factor (i.e. Factor2 in the figure), the differences are greater.

Thus the estimation of conditional regression lines can give us further information with regard to the behaviour of the individuals in the different tests carried out during the experiment. As an estimation method we used linear quantile regression (Koenker *et al.*, 1978), which has been applied to a number of research tasks in the field of information systems, e.g. the artificial intelligence system for the design of a consumer credit scoring system (Whittaker *et al.*, 2005), the artificial intelligence system for the analysis of firm solvency (De Andrés *et al.*, 2012) and analysis of downloads from an electronic commercial web site (Zhou and Duan, 2012). However, to our knowledge, no prior research papers on usability have applied such a methodology.

We computed 19 quantile regression estimates (5, 10, 15 … 95 per cent), using the regression equations indicated above. For each of the equations, 500 bootstrapping replications were used for the estimation of the robust standard errors.

Finally, and in order to ascertain whether the performance of individuals with regard to one task is related to the performance in the other tasks we conducted a correlation analysis. We computed both parametric (Pearson) and non-parametric (Spearman) correlation coefficients. For the calculations of these statistics, as well as for all the other tests and equations indicated above, we used the statistical package STATA-11.

Results

Descriptive statistics

Table II shows the descriptive statistics for the variables in the study.

Tables III-VI contain the frequency distributions for the independent variables considered in our model.

	Mean	SD	Minimum	Maximum
Point-and-click	16,864.77	4,294.92	9,319	45,792
Drag-and-drop	32,832.77	10,615.61	19,595	159,867
Text selection	59,389.97	29,777.94	16,780	534,397
Text edition	84,064.47	53,974.94	13,689	635,472
Menu selection	61,139.34	14,069.62	38,351	147,630
Time total	254,291.30	82,849.83	133,378	902,551

Table II. Descriptive statistics for the dependent variables in the study

Table III. Frequency distribution for age

Age	Number of observations
0	2
1	85
2	182
3	145
4	77
5	38
6	25
7	13
8	12
9	9
10	2
11	2
Total	592

Table IV. Frequency distribution for HoursUse

HoursUse	Number of observations
0	1
1	16
2	58
3	67
4	63
5	33
6	354
Total	592

Table V. Frequency distribution for gender

Gender	Number of observations
0 (Male)	462
1 (Female)	130
Total	592

Table VI. Frequency distribution for lefthanded

LeftHanded	Number of observations
0 (Right handed)	524
1 (Left Handed)	68
Total	592

Table VI shows that around 11 per cent of the sample are left-handed. That fits with the global proportion of left-handed people (estimated between 8 and 13 per cent).

As we might expect the average execution times for the first four tests increase depending on their complexity. As we have seen in previous sections some authors (MacKenzie *et al.*, 1991; Chadwick-Dias *et al.*, 2002; Czaja and Lee, 2006) reported a higher level of complexity in the execution of drag-and-drop tasks compared with point-and-click. Thus Table II shows that drag-and-drop tasks required a higher amount of time than point-and-click tasks.

In turn GOMS predicts that text selection tasks are more complex than both drag-and-drop and point-and-click tasks, since they require a higher number of operators ($T_M+T_P+T_D+T_K$) that must be applied for more time in order to achieve the goal. Table II reports a higher average time to complete text selection tasks than that taken in the first two tests. Similarly the mean time measured for the text editing test is consistent with its higher complexity when compared with the previous three tests. Last, the relatively low complexity required to achieve menu selection tasks is reflected in its low runtime.

Regression results

Table VII provides the main results of the six OLS regression models. In each column we show the statistics for each of the dependent variables. The first five cells contain the estimates for each of the independent variables and the intercept. In each cell the upper figure is the coefficient estimate, that in the middle is the t statistic and the figure shown in the lower part of the cell is the *p*-value. For each model the F statistic and its *p*-value are also shown, as well as the R^2 and the results of the Breusch-Pagan/Cook-Weisberg test on heteroscedasticity (the upper figure is the statistic and the lower figure is the *p*-value).

First, it is noticeable that the Breusch-Pagan/Cook-Weisberg tests show significant heteroscedasticity in the data. This reinforces the need for the calculation of robust standard errors and robust *p*-values (including White's correction). This also justifies the estimation of quantile regression equations in order to gain further understanding of the behaviour of the data; second, it is remarkable that although the R^2s are not very high, the F-tests we conducted to determine whether the coefficients of the variables are jointly equal to zero indicate that all the models are significant, that is, the set of variables considered as a whole has an influence on performance in all tests.

	Point-and-click	Drag-and-drop	Text selection	Text edition	Menu selection	$\text{Time}_{\text{total}}$
Age Group	500.74	1,721.47	1,098.10	10,118.24	3,077.63	16,516.21
	4.40	6.86	2.01	6.18	8.82	7.33
	0.000	0.000	0.045	0.000	0.000	0.000
HoursUse	−394.89	−895.07	−1,460.54	−8,202.50	−1,481.98	−12,435.01
	−3.02	−2.46	−2.35	−5.81	−4.73	−6.33
	0.003	0.014	0.019	0.000	0.000	0.000
Gender	1,742.03	5,077.81	6,656.58	1,377.96	6,293.93	21,148.33
	4.29	3.91	1.42	0.30	4.32	2.52
	0.000	0.000	0.156	0.761	0.000	0.012
LeftHanded	707.98	2,986.57	−1,598.63	6,281.345	107.61	8,484.88
	1.18	1.53	−0.64	0.63	0.07	0.69
	0.238	0.127	0.520	0.529	0.946	0.489
Intercept	16,759.52	30,363.66	61,785.65	91,378.54	57,363.51	257,651.30
	19.21	16.99	16.43	10.49	27.88	20.65
	0.000	0.000	0.000	0.000	0.000	0.000
F	17.48	13.46	4.46	17.82	33.27	27.07
(*p*-value)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)
R^2	11.41%	17.59%	2.19%	20.74%	25.49%	23.78%
Het test	7.07	388.80	107.87	109.79	44.60	53.01
(*p*-value)	(0.007)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)

Table VII. OLS regressions results

With regard to the parameter estimates, coefficients for age are always positive and significant. This data confirms the first hypothesis (*H1*: there is a direct relationship between age and performance time) as it means that older users perform worse in all the tasks (needing more time to complete). For the case of HoursUse the result is the opposite, as more hours of computer use always imply a better performance. These results confirm that performance declines with age. These results are similar to those obtained in the research cited in the background section, which focused on observations based on the execution of high-level interaction tasks, mostly related to cognition and perception. Our findings confirm that the same effect is observable at the low level of interaction required by the GOMS analysis, which is mostly based on the human motoric system.

The gender coefficient is significantly positive for point-and-click, drag-and-drop, menu selection and the total time, but is not significantly different from zero for text selection and editing. Thus women perform worse on point-and-click, drag-and-drop and menu selection. These results suggest that hypothesis *H3* (men's performance in drag-and-drop operations is better than women's) holds, corroborating the observations made by Inkpen (2001) regarding girls having difficulty doing drag-and-drop. However, our data do not support hypothesis *H2* (women's performance in point-and-click mouse operations is better than men's). This hypothesis was formulated based on Inkpen's (2001) results that were obtained from children. That leads us to conclude that Inkpen's results are more related to children's different learning styles than directly to gender.

Moreover, left-handed users perform neither significantly better nor significantly worse than right-handed.

In addition to model estimations we conducted some additional robustness tests with the aim to further check the soundness of our results. First, we computed Cook's D statistic in order to detect the presence of influential cases in the regressions. Ds are always lower than 1 for all the individuals in all the regression models. Second, we tested the existence of non-linear effects for the age variable (that is, whether middle-aged users perform better than both younger and older users). This was done by adding a quadratic term (Age_2) to the equations. None of these terms was found to be significant.

Finally, we re-estimated the models for three subsamples according to the browser used for the test (Chrome, Firefox and Internet Explorer; the number of people using other navigators was not enough to allow regression equation estimation) and the operating system (Windows, Linux and Mac). The results are qualitatively the same as those displayed in Table VII. For the sake of brevity we did not include in the paper the results of all these robustness checks. However, they are available from the authors on request.

Quantile regression results

As indicated above, heteroscedasticity in the data suggests that quantile regression could provide further understanding of the behaviour of data. We estimated conditional quantiles 5, 10, …, 95 per cent for all the six equations defined above. Due to space limitations we report only the results which add new information to the discussion in the comments on the OLS regression results (complete results are also available from the authors on request).

These results refer to the behaviour of the gender variable. Table VIII indicates the *p*-value of gender for each of the considered quantiles, and each of the regression models.

Percentile	Point-and-click	Drag-and-drop	Text selection	Text edition	Menu selection	$Time_{total}$
5	0.001	0.031	0.226	0.003	0.021	0.000
10	0.007	0.019	0.357	0.005	0.001	0.000
15	0.000	0.043	0.195	0.025	0.001	0.000
20	0.000	0.017	0.196	0.110	0.000	0.000
25	0.000	0.003	0.224	0.009	0.000	0.000
30	0.000	0.005	0.525	0.035	0.000	0.000
35	0.002	0.011	0.259	0.074	0.000	0.000
40	0.000	0.002	0.218	0.188	0.000	0.001
45	0.000	0.000	0.152	0.485	0.000	0.004
50	0.000	0.001	0.224	0.369	0.000	0.008
55	0.000	0.000	0.429	0.368	0.000	0.007
60	0.000	0.001	0.593	0.455	0.002	0.009
65	0.000	0.000	0.867	0.277	0.007	0.009
70	0.001	0.000	0.891	0.070	0.001	0.070
75	0.022	0.000	0.732	0.087	0.004	0.145
80	0.011	0.000	0.785	0.157	0.024	0.370
85	0.027	0.014	0.597	0.191	0.005	0.012
90	0.053	0.003	0.790	0.685	0.007	0.026
95	0.047	0.003	0.830	0.466	0.063	0.020

Table VIII. *P*-values for the quantile regression for gender

With regard to point-and-click tasks, it is remarkable that, for the case of the upper quantiles, the effect of gender is less significant (above Q70 gender is not significant at the 1 per cent level). In other words for low performers gender is not as influential as for intermediate or high performers. For the case of drag-and-drop, the results indicate the opposite, that is, for the best performers (quantiles under Q25) the effect of gender is less marked. In the case of text selection and editing tasks the coefficient of gender is not significant for the majority of the quantiles. This is in accordance with the results of the OLS models. The results for menu selection indicate that gender is significant at the 1 per cent level in all cases except some of the extreme quantiles (Q5, Q80, Q95). Finally, the results for the total time suggest that the influence of gender is more significant for intermediate and best performers (only quantiles below Q70 have a *p*-value which is under 1 per cent).

Conclusions and practical implications

The main goal of this work was to assess whether gender and age are sufficiently significant determining factors in mouse motion behaviour to support an automatic profiling system, as well as to evaluate the roles that laterality and user experience play in the overall performance.

Regarding age, the results obtained are consistent with previous research, especially with the work of Hill *et al.* (2011). The results obtained in tests 1-5 show a negative impact on the performance of the GOMS operators P, K, and D when age is increased. The increment in the overall execution time for each task proved to be much bigger in people aged over 40. This result makes us believe that it would be relatively easy to classify people above and below this age using data gathering agents.

With respect to gender the results obtained are consistent with the findings of Inkpen (2001) who found that girls perform worse with regard to drag-and-drop tasks. In our study males obtained better results when executing interaction operators based on pointing and dragging, which are required by the point-and-click, drag-and-drop

and menu selection tasks. Nevertheless, there were no significant differences in the typing operators required by the text selection and editing tasks. The differences were also so relevant that they can be used for the design of algorithms for automatic classification.

Surprisingly, no significant performance differences were detected between left- and right-handed users and we did not find any evidence regarding the influence of this factor on the overall performance.

The performance differences in gender and age are relevant enough to be gathered by a data gathering agent hidden in the user interface of the online information system. Since the main differences in performance are based on mouse pointing tasks (which are used more frequently than those based on text editing) the gender and age of the user could be estimated automatically shortly after user arrival on the web site.

This automatic profiling can be used to provide different content to different kinds of users and also to adapt the appearance and behaviour of different elements of the user interface. For example, the size of the clickable elements (buttons, links, menu items, etc.) could be dynamically increased when they are being used by elderly users. Since Fitts' function is logarithmic, a small increase in the size of the object would represent a drastic reduction in the time required by the user to click on it. This approach could not only increase the user's overall performance but could also significantly contribute to enhancing the user's experience on the site.

Once the feasibility of using age and gender in the automatic classification of users on certain parameters of web interaction performance has been demonstrated, the next step is determining the best strategy to implement a classification device. Several systems using different machine learning strategies (neural networks, regression trees, etc.) are suitable for such applications. A profitable avenue of research could be the comparison of the accuracy of different machine learning models. Although this research found significant differences in the studied variables regarding their influence on performance, there are still other important factors that could contribute to the development of an accurate automatic profiling system.

User accuracy is another important factor that can be measured in online information systems that could bring a different approach to classifying users. Therefore our next step will be the design of a similar case study in order to analyse the role that accuracy plays in classifying users. In this way it would be possible to design classifying systems based on a trade-off between speed and accuracy.

References

Abdulin, E. (2011), "Using the keystroke-level model for designing user interface on middle-sized touch screens", *Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems, ACM, New York, NY*, pp. 673-686.

Agudo, J.E., Sánchez, H. and Rico, M. (2010), "Playing games on the screen: adapting mouse interaction at early ages", *10th IEEE International Conference on Advanced Learning Technologies, IEEE, Los Alamitos, CA*, pp. 493-497.

Bestelmeyer, P.E.G. and Carey, D.P. (2004), "Processing biases towards the preferred hand: valid and invalid cueing of left- versus right-hand movements", *Neuropsychologia*, Vol. 42 No. 9, pp. 1162-1167.

Breusch, T.S. and Pagan, A.R. (1979), "A simple test for heteroscedasticity and random coefficient variation", *Econometrica*, Vol. 47 No. 5, pp. 1287-1294.

Chadwick-Dias, A., McNulty, M. and Tullis, T. (2002), "Web usability and age", *ACM SIGCAPH Computers and the Physically Handicapped*, Nos 73-74, June-September, pp. 30-37.

Cheong, Y., Shehab, R.L. and Ling, C. (2013), "Effects of age and psychomotor ability on kinematics of mouse-mediated aiming movement", *Ergonomics*, Vol. 56 No. 6, pp. 1006-1020.

Cook, R.D. and Weisberg, S. (1983), "Diagnostics for heteroscedasticity in regression", *Biometrika*, Vol. 70 No. 1, pp. 1-10.

Crook, C. (1992), "Young children's skill in using a mouse to control a graphical computer interface", *Computers & Education*, Vol. 19 No. 3, pp. 199-207.

Czaja, S.J. and Lee, C.C. (2006), "The impact of aging on access to technology", *Universal Access in the Information Society*, Vol. 5 No. 4, pp. 341-349.

De Andrés, J., Landajo, M. and Lorca, P. (2012), "Bankruptcy prediction models based on multinorm analysis: an alternative to accounting ratios", *Knowledge-Based Systems*, Vol. 30, June, pp. 67-77.

Dickinson, A., Arnott, J. and Prior, S. (2007), "Methods for human-computer interaction research with older people", *Behaviour & Information Technology*, Vol. 26 No. 4, pp. 343-352.

Fozard, J.L., Vercryssen, M., Reynolds, S.L., Hancock, P.A. and Quilter, R.E. (1994), "Age differences and changes in reaction time: the baltimore longitudinal study of aging", *Journal of Gerontology*, Vol. 49 No. 4, pp. 179-189.

Freudenthal, D. (2001), "Age differences in the performance of information retrieval tasks", *Behaviour & Information Technology*, Vol. 20 No. 1, pp. 9-22.

Guiard, Y., Olafsdottir, H.B. and Perrault, S.T. (2011), "Fitts's law as an explicit time/error trade-off", *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems - CHI '11, ACM Press, New York, NY*, pp. 1619-1628.

Haque, A., Sadeghzadeh, J. and Khatibi, A. (2006), "Identifying potentiality online sales in Malaysia: a study on customer relationships online shopping", *The Journal of Applied Business Research*, Vol. 22 No. 4, pp. 119-130.

Hill, R.L., Dickinson, A., Arnott, J.J., Gregor, P. and McIver, L. (2011), "Older web users' eye movements", *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems – CHI '11, ACM Press, New York, NY*, pp. 1151-1160.

Inkpen, K.M. (2001), "Drag-and-drop versus point-and-click mouse interaction styles for children", *ACM Transactions on Computer-Human Interaction*, Vol. 8 No. 1, pp. 1-33.

Joiner, R., Messer, D., Light, P. and Littleton, K. (1998), "It is best to point for young children: a comparison of children's pointing and dragging", *Computers in Human Behavior*, Vol. 14 No. 3, pp. 513-529.

Kambil, A., Eselius, E.D. and Monteiro, K.A. (2000), "Fast venturing: the quick way to start web businesses case solution and analysis, HBS case study solution & Harvard case analysis", *Sloan Management Review*, Vol. 41 No. 4, pp. 55-67.

Koenker, R., Bassett, G. and Jan, N. (1978), "Regression quantiles", *Econometrica*, Vol. 46 No. 1, pp. 33-50.

Kuhtz-Buschbeck, J.P., Stolze, H., Jöhnk, K., Boczek-Funcke, A. and Illert, M. (1998), "Development of prehension movements in children: a kinematic study", *Experimental Brain Research*, Vol. 122 No. 4, pp. 424-432.

Kutner, M.H., Nachtsheim, C.J. and Neter, J. (2004), *Applied Linear Statistical Models*, 4th ed., McGraw-Hill/Irwin, New York, NY.

Lee, M.-C. (2009), "Understanding the behavioural intention to play online games: an extension of the theory of planned behaviour", *Online Information Review*, Vol. 33 No. 5, pp. 849-872.

Lenhard, A. and Hoffmann, J. (2007), "Constant error in aiming movements without visual feedback is higher in the preferred hand", *Laterality*, Vol. 12 No. 3, pp. 227-238.

Liu, C. and Arnett, K.P. (2000), "Exploring the factors associated with web site success in the context of electronic commerce", *Information & Management*, Vol. 38 No. 1, pp. 23-33, available at: www.sciencedirect.com/science/article/pii/S0378720600000495 (accessed 10 June 2014).

Lynch, P.D. and Beck, J.C. (2001), "Profiles of internet buyers in 20 countries: evidence for region-specific strategies", *Journal of International Business Studies*, Vol. 32 No. 4, pp. 725-748.

MacKenzie, I.S., Sellen, A. and Buxton, W.A.S. (1991), "A comparison of input devices in element pointing and dragging tasks", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Reaching through Technology - CHI '91, ACM Press, New York, NY*, pp. 161-166.

Metter, E.J., Conwit, R., Tobin, J. and Fozard, J.L. (1997), "Age-associated loss of power and strength in the upper extremities in women and men", *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, Vol. A52 No. 5, pp. B267-B276.

Mieschke, P.E., Elliott, D., Helsen, W.F., Carson, R.G. and Coull, J.A. (2001), "Manual asymmetries in the preparation and control of goal-directed movements", *Brain and Cognition*, Vol. 45 No. 1, pp. 129-140.

Nap, H.H., De Greef, H.P. and Bouwhuis, D.G. (2005), "Access for all by cognitive engineering", *Gerontechnology*, Vol. 3 No. 4, p. 259.

Neely, K., Binsted, G. and Heath, M. (2005), "Manual asymmetries in bimanual reaching: the influence of spatial compatibility and visuospatial attention", *Brain and Cognition*, Vol. 57 No. 1, pp. 102-105.

Neerincx, M.A., Lindenberg, J., Rypkema, J.A. and van Besouw, N.J.P. (2000), "A practical cognitive theory of web-navigation: explaining age-related performance differences", position paper presented at Basic Research Symposium CHI 2000, The Hague, 2 April.

Oyewole, S.A. and Haight, J.M. (2011), "Determination of optimal paths to task goals using expert system based on GOMS model", *Computers in Human Behavior*, Vol. 27 No. 2, pp. 823-833.

Peters, M. and Ivanoff, J. (1999), "Performance asymmetries in computer mouse control of right-handers, and left-handers with left- and right-handed mouse experience", *Journal of Motor Behavior*, Vol. 31 No. 1, pp. 86-94.

Phillips, J.G. and Triggs, T.J. (2001), "Characteristics of cursor trajectories controlled by the computer mouse", *Ergonomics*, Vol. 44 No. 5, pp. 527-536.

Rikli, R. and Busch, S. (1986), "Motor performance of women as a function of age and physical activity level", *Journal of Gerontology*, Vol. 41 No. 5, pp. 645-649.

Rodrigues, P.C., Barbosa, R. and Carita, A.I. (2012), "Stimulus velocity effect in a complex interceptive task in right- and left-handers", *European Journal of Sport Science*, Vol. 12 No. 2, pp. 130-138.

Rohr, L.E. (2006), "Gender-specific movement strategies using a computer-pointing task", *Journal of Motor Behavior*, Vol. 38 No. 6, pp. 431-437.

Salthouse, T.A. (1984), "Effects of age and skill in typing", *Journal of Experimental Psychology, General*, Vol. 113 No. 3, pp. 345-371.

Sayers, H. (2004), "Desktop virtual environments: a study of navigation and age", *Interacting with Computers*, Vol. 16 No. 5, pp. 939-956.

Schneider, D.W. and Anderson, J.R. (2011), "A memory-based model of Hick's law", *Cognitive Psychology*, Vol. 62 No. 3, pp. 193-222.

Schrepp, M. (2010), "GOMS analysis as a tool to investigate the usability of web units for disabled users", *Universal Access in the Information Society*, Vol. 9 No. 1, pp. 77-86.

Sebora, T.C., Lee, S.M. and Sukasame, N. (2008), "Critical success factors for e-commerce entrepreneurship: an empirical study of Thailand", *Small Business Economics*, Vol. 32 No. 3, pp. 303-316.

Strom, P.S., Strom, R.D., Wingate, J.J., Kraska M.F. and Beckert T.E. (2012), "Cyberbullying: assessment of student experience for continuous improvement planning", *NASSP Bulletin*, Vol. 96 No. 2, pp. 137-153.

Tan, D.S., Czerwinski, M. and Robertson, G. (2003), "Women go with the (optical) flow", *Proceedings of the Conference on Human Factors in Computing Systems – CHI '03, ACM Press, New York, NY*, pp. 209-215.

Turban, E., Lee, J.K., King, D., McKay, J. and Marshall, P. (2000), *Electronic Commerce: A Managerial Perspective*, Prentice-Hall, Upper Saddle River, NJ.

Walker, N., Philbin, D.A. and Fisk, A.D. (1997), "Age-related differences in movement control: adjusting submovement structure to optimize performance", *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences*, Vol. 52 No. 1, pp. 40-52.

Weiser, E.B. (2000), "Gender differences in internet use patterns and internet application preferences: a two-sample comparison", *CyberPsychology & Behavior*, Vol. 3 No. 2, pp. 167-178.

Whittaker, J., Whitehead, C. and Somers, M. (2005), "The neglog transformation and quantile regression for the analysis of a large credit scoring database", *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Vol. 54 No. 5, pp. 863-878.

Xiang, L.X.L. and Xiaoli, C.X.C. (2010), "The research on performance of automobile human-machine interface based on BHR-GOMS behavior model", *Proceedings – 2010 IEEE International Conference on Intelligent Computing and Intelligent Systems, ICIS 2010, IEEE, Los Alamitos, CA*, pp. 174-178.

Xie, B. (2003), "Older adults, computers, and the internet: future directions", *Gerontechnology*, Vol. 2 No. 4, pp. 289-305.

Xu, H. (2006), "Consumers' attitudes of e-commerce in China", *Issues in Information Systems*, Vol. 7 No. 2, pp. 202-207.

Yan, P. and Guo, J. (2010), "The research of web usability design", *The 2nd International Conference on Computer and Automation Engineering, ICCAE 2010, IEEE, Los Alamitos, CA*, pp. 480-483.

Zhai, S. (2004), "Characterizing computer input with Fitts' law parameters – the information and non-information aspects of pointing", *International Journal of Human-Computer Studies*, Vol. 61 No. 6, pp. 791-809.

Zhou, W. and Duan, W. (2012), "Online user reviews, product variety, and the long tail: an empirical investigation on online software downloads", *Electronic Commerce Research and Applications*, Vol. 11 No. 3, pp. 275-289.

About the authors

Dr Javier De Andrés (PhD) is an Associate Professor of Accounting and Finance at the University of Oviedo. Dr De Andrés has taught financial analysis courses at the undergraduate, postgraduate and executive education levels. His research interests focus on artificial intelligence systems for the analysis of credit risk, ERP systems and XBRL. He has published more than 50 book chapters and research papers in such refereed scientific journals as the *European Journal of Operational Research*, the *Journal of the Royal Statistical Society (Series C – Applied Statistics)*, the *Journal of the American Society for Information Science and Technology*, etc.

Beatriz Pariente is a researcher and a PhD candidate in the Department of Computing at the University of Oviedo, where she is investigating usability engineering and web design strategy.

Dr Martin Gonzalez-Rodriguez (PhD) is a researcher and a consultant on Usability Engineering and Web Design Strategy. He works as a tenured Associate Professor in the Department of Computing at the University of Oviedo and as an independent Expert for the Spanish Assessment and Planning Agency. He is the co-founder of the International Conference on Web Engineering and the founder of *eMinds: International Journal on Human-Computer Interaction*. His research has received the ONCE international research and development award in new technologies for the blind and visually impaired (2008), the Vodafone research award for mobile technology (2008) and the TAW award for web accessibility (2005). Dr Martin Gonzalez-Rodriguez is the corresponding author and can be contacted at: martin@uniovi.es

Dr Daniel Fernandez Lanvin (PhD) is an Associate Professor in the Department of Computing at the University of Oviedo, and also an Adjunct Lecturer at the Cork Institute of Technology. Dr Fernández-Lanvin has taught computing logic, web development, software engineering and HCI at both undergraduate and postgraduate levels. His research interests principally involve HCI, web development and programming languages.

D Article 3: The dimension of age and gender as user model demographic factors for automatic personalization in e-commerce sites.

Computer Standards & Interfaces 59 (2018) 1–9

Contents lists available at ScienceDirect

Computer Standards & Interfaces

journal homepage: www.elsevier.com/locate/csi

The dimension of age and gender as user model demographic factors for automatic personalization in e-commerce sites

D. Fernandez-Lanvin, J. de Andres-Suarez, M. Gonzalez-Rodriguez, B. Pariente-Martinez*

Faculty of Computer Science, University of Oviedo, Spain

ARTICLE INFO

Keywords:
Personalization
User model
GOMS
Fitts' law
Hicks–Hyman's law
Salthouse' regularities

ABSTRACT

Personalization in e-commerce increases sales by improving customer perception of site quality. However, some demographic data about customers (crucial for the success of the personalization process) not always can be obtained explicitly, as is the case of anonymous web site visitors.

The paper describes a user study focused on determining whether it would be possible to categorize the age and gender of individual visitors of a web site through the automatic analysis of their behavior. Three tasks commonly found in e-commerce sites (*Point & Click, Drag & Drop* and *Item Selection*) were tested by 592 volunteers and their performance was analyzed using several different statistical methods. The study found consistencies in the execution times of individuals across the different tasks and revealed that age and gender are sufficiently determining factors to support an automatic profiling. Results also showed that relevant information about gender and age can be extracted separately through the individual analysis of each one of the mentioned interaction tasks.

1. Introduction

The success of online marketing is determined, among other factors by the level of *personalization* of the e-commerce sites. That is, the process of making a unique user experience for each customer. Personalization is a dominant business model in online marketing strategies [1] and is used to establish relationships between customers and sellers [2]. Its capability to provide recommendations to the customers is acknowledged to be an important feature of online shopping [3] as it enhances customer retention and increases sales [3].

Addressing similarities and differences among consumers is critical as differences in demographic factors may be associated to different tastes and therefore to different purchasing patterns [4].

There is strong empirical evidence showing that differences caused by gender and age, influence online shopping preferences [5,6]. Several authors pointed these two variables as key elements in the personalization process [7–11].

Personalization requires to collect relevant information about users and this information has a great relevance in the success of e-commerce [12–15]. However, data gathering is not a trivial issue. It can be explicitly made (e.g. getting this information through registration forms) or implicitly (e.g. monitoring customers purchase patterns) [16].

Although most of the demographic factors are explicitly collected through the registration process [16–18], this approach may result into biases and/or outdated data [16]. Users may withhold information due to privacy, social or cultural issues [18]. So for example, underage users may avoid controls in adult-oriented sites or social networks just providing fake data [19]. Even more important, the outcome of the data gathered explicitly may be limited given that most of the potential customers in online sites are anonymous and/or first visit users. The use of an implicit data gathering approach may help to surpass these limitations.

Building implicit data gathering systems able to estimate the age and/or gender of their users requires the prior identification of the interaction factors that make a user to be unique. This work explores the execution time required to perform basic interaction tasks as a candidate factor to build such kind of systems.

Although previous research efforts suggest an independent influence of both age and gender on the execution time [20–23], elegant literature lacks an evaluation of their combined effect in e-commerce applications. Therefore, this work attempts to assess the degree of association between gender and age on execution time conducting a joint evaluation of the effect of age and gender on interaction. It also introduces a combined analysis on how other variables such as the user's laterality (left handed or right handed) or the user's prior experience in the use of

* Corresponding author.
E-mail addresses: dflanvin@uniovi.es (D. Fernandez-Lanvin), jdandres@uniovi.es (J.d. Andres-Suarez), martin@uniovi.es (M. Gonzalez-Rodriguez), bparientem@gmail.com (B. Pariente-Martinez).
URL: http://www.hci.uniovi.es (M. Gonzalez-Rodriguez)

https://doi.org/10.1016/j.csi.2018.02.001
Received 7 September 2017; Received in revised form 11 December 2017; Accepted 2 February 2018
Available online 3 February 2018
0920-5489/

computers might influence the time required by the users to complete basic interaction tasks.

The study analyzes the performance of 592 volunteers executing three usability design patterns commonly found in the design of e-commerce interactive systems like Amazon, DeviantArt, Alibaba, etc. These patterns are: (i) *Point & Click*, (ii) *Drag & Drop* and (iii) *Item Selection*. Although this study is focused on the analysis of the users performance in the use of e-commerce usability patterns, its findings may impact other domains based on these patterns too (e.g. education and entertainment).

The goals of the research are to determine whether the influence of age and gender on the execution time is significant enough to infer its value through behavior analysis and to analyze which of the interaction tasks mentioned before would be the most appropriate to build such kind of personalization systems. If it is possible to infer the users age and/or gender through the quick inspection of the performance measured in the execution of basic interaction tasks, it would not only be possible to adapt marketing messages to a specific age and gender range, but also the development of tools targeted to prevent certain crimes, such as pedophilia, or illegal access to web sites. Adults pretending to be children in social networks could be detected through the analysis of their interactions with the user interface. A similar approach may be used to detect children or teenagers accessing adult web sites.

The remainder of the paper is structured as follows. Section 2 discusses prior literature about studies based on the influence of age and gender on user's performance. Section 3 designs the hypotheses, describes the empirical study used to test them and discusses the statistical methods employed. The fourth section is devoted to the presentation and discussion of the results. Finally, the main conclusions, practical implications and future research are presented.

2. Related work

Determining users' age and gender by observing their interactions with an e-commerce web site to update the site's user model dynamically, and therefore improving its marketing capabilities, is the overall goal of this work.

User modeling is the process of constructing user models [18]. A user model is an explicit representation of the properties of individual users or user classes. It allows the adaptation of the system to the user needs and preferences [24]. This process involves both static and dynamic user information. Static user information refers to basic characteristics (e.g. demographics) explicitly presented by the user during a registration procedure. On the other hand, dynamic user information is collected by observing users behavior and it is recorded in log files or in a list of objects visited by the web user [17].

Automated user profiling has been studied in previous works with different approaches and results. Most of these works deal with semantic information gathered from users interaction. For instance, Woo & Shirmohammadi [16] proposed an automatic user personality categorization model based on their digital personality. The authors collect information through the observation of user interacting with products. Yang & Claramunt [17] proposed a semantic user model that uses both static and dynamic user information to predict user features relevant for a specific application domain. Such et al. [25] analyzed the automated user profiling techniques and proposed an approach to control buyer profiling. However, their goal is to prevent users to be automatically identified, just the opposite to the motivation of this work. Fijalkowsky [26] proposes an e-commerce web system that collects data obtained from social network profiles and uses it to provide purchase recommendations to its users.

All these works are focused on user behavior information at the semantic level, under the assumption that demographic information will be explicitly collected and relying in users collaboration. Ghazarian and Noorhosseini [27] face the same problem from a lower abstraction level, using machine learning algorithms to adapt user interfaces to the needs of user groups with different levels of skills detected through the analysis of mouse motions. Garg et al. [28] also tries to identify users depending on interaction behavior like mouse movements and clicks, typing speed and system background processes. The authors extract relevant mouse related features such as average distance, speed, angles of movement and number of clicks during a session, and then they utilize Support Vector Machines (SVM) to learn the user specific feature sets. They conclude that this information is relevant enough to identify and/or differentiate users behind different aliases, but they use it only to prevent masquerading attacks to web sites.

None of these works are focused on the identification of the demographic factors, so their results do not analyze the feasibility of this approach. On the contrary, most of them assume that relevant data should be collected through the registration process, with the limitations that this strategy involves.

2.1. Ageing

Ageing negatively impacts the ability to use computers [20,29] and is typically defined in Human Computer-Interaction through an emphasis on declines in abilities and associated reductions in performance when using technology [30]. It produces a poorer motor control and sensory deficits [31]. Related studies show that older people have slower reaction times [32], delayed movements, a decline in motor skills [33]. Reduced mobility, caused by a loss in muscle strength [34], produces difficulties in the execution of movements [34]. This process of losing muscle strength begins in people aged over 40 [35].

Other studies revealed that ageing negatively influences the learning strategies used to operate online systems, as perception and cognition declines [36,37]. Senior users have been found to be slower than young adults when retrieving information [8,38], performing 3D navigation on desktop systems [39] or browsing the web [40].

Studies analyzing information search behavior [20,41] enforce the importance of ageing. The study of the behavior of expert older adults daily using the web, compared to their younger colleagues, concluded that age is a determining factor [20]. This work is particularly interesting because it is specifically focused in web interaction. However, the analyzed activities (search behavior and related) require different operators than those involved in mouse motion.

On the other hand, very young users reveal a poor execution time in the development of certain tasks. Basic interaction tasks like *Drag & Drop* are especially difficult for them as keeping the finger pressed down while controlling its trajectory requires a high demand of motor skills [42], perception and cognitive skills [43,44]. The execution time slows down if it is possible to replace *Drag & Drop* by *Point & Click*, as the operation can be easily resumed from the last pointing task in case of failure [45]. Attaching and lifting objects in the real world causes some difficulties to children under 8 years old as these tasks requires subtle hand-eye coordination [46]. At that age, the coordinate movements are further determined by cognitive factors rather than motor skills [10]. The speed of such coordinate movements evolves up to the age of 12 years [46].

Some authors reported how different interfaces influence the interaction of specific groups of users regarding their age [47], but no study was found about whether there are significant differences between the time required to execute different alternative interaction tasks conducting to the same result (e.g. *Point & Click* as an alternative to *Drag & Drop* to obtain the same result).

If these differences do not exist and the execution time keeps coherence in each basic interaction task, that is, if the time required by each group of users is similar in each task (*Point & Click, Drag & Drop* and *Item Selection*) it would not be necessary to analyze the three interaction tasks in the same user interface to detect the users age. It would be enough to analyze the users performance in only one of them. However, if those differences exist, it would be necessary to measure and to analyze the

users performance in all the three different proposed interaction tasks to categorize users according with their age.

2.2. Gender

Women process information in different ways than men [23]. Gender-associated differences in decision making, learning, and problem solving can be a determining factor in users effectiveness [48,49]. Even more, it has been observed that the self-perceptions concerning computer competence as well as the level of ICT-related social interactions is different for boys and girls [50].

It was observed that mens performance in navigating through virtual environments is better than women's when small displays are used. The use of larger displays reduces the gender performance gap since the women's performance improves while the men's performance is not negatively affected [51,52].

Inkpen [53] compared *Drag & Drop* tasks as opposed to *Point & Click* in children. Although there was no any significant gender difference in the overall movement time and/or general error rates, there were relevant differences in pickup and drop errors. The girls performed poorly when executing *Drag & Drop* tasks, as opposed to *Point & Click*. There were also performance correlation differences between gender and target size.

Rohr [22] evidenced that gender-specific movement biases emphasize speed for men and accuracy for women. Wahlstrom et al. [54] observed that when operating the mouse, women worked with greater extension and had a greater range of motion in the wrist when compared to men. This observation could explain Rohrs results regarding speed versus accuracy. They also found gender differences for musculoskeletal load. For most of the measured variables, women worked with higher loads than men. These differences are not limited to the low-level interaction. Collazos et al.[55] found significant differences in the way woman and men face collaborative work in computer-mediated communication.

3. Design of the empirical study

3.1. Hypothesis

The related previous studies evidence that there are significant differences between the times required by children and adults to execute different basic interactions tasks. However, to date, we found no studies evidencing these differences in adults, something that lead us to conjecture that the performance of one specific adult in these tasks could be correlated. If so, it would mean that the analysis of performance in one of them would be enough to identify adults, simplifying users age classification. On the other hand, even though there are not evidences of differences between genders in adults for these basic interaction tasks, some studies identified some differences between men and women in other activities that could determine the correlation between the performances in these basic interaction tasks. That lead us to wonder whether these correlations could be determined by users gender. Therefore, we formulate the following hypotheses to be verified/refuted by the empirical study:

- *(h_1) The execution time of the different tasks increases with the age of the subject under study*
- *(h_2) Womens execution time for the different tasks is longer than mens*
- *(h_2) The execution times of basic interaction tasks (Point & Click, Drag & Drop and Item Selection) are significantly correlated*

To assess whether the hypothesis formulated in the prior section hold, the performance of 592 individuals was analyzed in the execution of three basic interaction tasks.

3.2. Object of study

The tasks analyzed in this study were *Point & Click, Drag & Drop* and *Item Selection*. They were selected because of the crucial role they play in the usability patterns behind the design of successful e-commerce sites.

Point & Click is used to move the mouse pointer over an image or over CTA items (*Call To Action*) to click on it. It is commonly used by customers of online shopping sites like Amazon, eBay, ModCloth, Zappos, etc. to retrieve information about appealing products or to include them in the shopping basket.

Drag & Drop is mostly used to collect vast number of items to place them into the shopping basket. This task is commonly found as part of the usability patterns used in art/photo e-commerce sites like UXPin, DeviantArt, etc.

Finally, *Item Selection* is used to browse through small navigation bars or menus to select item categories. It is commonly used is popular sites like Alibaba, Walmart, Asos, Etsy, etc. Users were encouraged to complete these tasks achieving interaction goals in the minimum amount of time.

Their behavior was recorded by data gathering agents that measured the execution time required by everyone to complete every single task proposed he measurement of the users execution time in the different tests proposed was based on GOMS (Goals, Operators, Methods, and Selection rules). This analysis method was designed to estimate the users performance when they interact with different interfaces [56]. The method has been successfully used to estimate user performance in many different scenarios including interaction with automobiles [57], touch screens [58] and online web sites [59,60], among several others.

GOMS splits complex interaction tasks into low level components called *operators*. These operators include actions like mouse pointing (denoted as P), dragging (D), key typing (K), decision taking (M), etc. The execution of each operator requires a specific amount of time (denoted respectively as T_P, T_D, T_K, T_M, etc.), so GOMS estimates the execution time of complex interaction tasks as the sum of the execution times of the different operators required to complete the tasks. So, for example, the estimated execution time for a *Drag & Drop* interaction task would be $T_P + T_K + T_D + T_K$; that is, the time needed to move the mouse pointer over the movable object (T_P) plus the time required to press the mouse' button once the pointer is over the target (T_K) plus the time used to drag the object to a new position (T_D) plus the time required to release the mouse's button (T_K).

The execution time for each operator (T_P, T_D, T_K, T_M, etc.) is estimated using well-known psychological laws and regularities such as the Fitts' law (T_P and T_D), the Salthouses regularities (T_K), the Hicks–Hymans law (T_M), etc.

Fitts' law estimates the time needed to move a pointing object (the users' finger, the mouse pointer, a joystick, etc.) over a target as $a + b\ Log_2\ (D/S)$. Where D is the distance to the target, S represents the targets size and a and b are user dependent correction factors [61,62]. Salthouse's regularities predict the time required by different kind of users (ranging from novices to experts) to type texts of a known length [63]. The Hicks–Hymans law estimates the time required to take a decision (such as the selection of a menu item) as $a + b\ Log_2(n+1)$ where n represents the number of available options and a and b again are user dependent correction factors [64,65].

Although these laws help to estimate the execution time required by an average user, they have to be adapted to the specific needs of individual users. That is the case of the correction factors used by the Fitts' law and the Hicks–Hymans law which have to be obtained through the analysis of performance records previously obtained for specific users. The values for these correction factors rely on the external variables under analysis in this research, as it is the case of the age and gender.

The use of GOMS in this context has two main advantages. First, it helps to structure the study of the different interaction tasks using a common research framework to other similar studies. Second, the experimental measurement of the users' runtime for each specific task, to

a high degree of accuracy, facilitates a quick and accurate estimation of the global execution time for e-commerce sites whose user interfaces combine several of these interaction tasks.

Each test was designed to replicate the behavior of a real e-commerce application but hiding the features that might allow the user to identify it, thus avoiding the effect that the familiarity with the real product might have on the measurements to be obtained. Hence, the lexical and semantical levels (related to mouse movement and object recognition and perception) of the user interface of the corresponded e-commerce application were recreated in the most realistic way, while the semantic (iconic representation) and conceptual (final goal of the application) levels of the interface were ignored or hidden to avoid the mention familiarity.

The first test (task 1) was designed to analyze the behavior of users executing the *Point & Click* tasks required to select objects in web documents by moving the mouse pointer across the display to click on links, buttons, scrolling boxes, etc.

The test showed a sequential series of rectangles in different locations across the screen. Participants in the test had to click inside each rectangle to make it disappear before a new one appeared in a different location. The test used fourteen different rectangles distributed in positions that followed a Z pattern layout to keep a fair balance between left-handed and right-handed users. Along the test, the location of the different targets was changed using the horizontal (left to right, right to left) and vertical dimensions (top to bottom, bottom to top).

At the same time, Fitts' law was used to increase the difficulty of each interaction, increasing the distance to the target (D) and reducing its size (S), thus increasing the time required to click on the target by a factor of Log_2(D/S).

To click on the target, users had to use two GOMS operators: P and K (see Fig. 1). First, the users moved the mouse over the display to place the pointer over the square using T_P units of time (step 1 in Fig. 1). Next, users needed to click pressing the mouse button using a K operator (step 2 in Fig. 1). The time estimated by GOMS to complete each *Point & Click* action is therefore $T_P + T_K$. The time required to complete each point and click action ($T_P + T_K$) was recorded (in milliseconds) for each click interaction. The sum of the execution times required to complete the full test was recorded for later statistical analysis.

The second test (task 2) was designed to measure the time required to complete *Drag & Drop* tasks, commonly used to drag items into the shopping basket in electronic commerce applications.

In this second test users were asked to drag a red rectangle over a second one, which had a size two thirds bigger than the red one. Every time the user completed the task, both rectangles disappeared, and two new rectangles appeared in separate locations of the display. The process was repeated along fourteen interactions. Each time, the rectangles were distributed using a Z shaped layout to keep a fair balance between left-handed and right-handed users. The distance between objects was incremented and its size was reduced in each interaction, using the Fitts' law to increase the time required by the users to complete each interaction.

To drag the first rectangle over the second, the users had to select it first. Therefore, they needed to use the GOMS operators required in a *Point & Click* task. The P operator is required to point to the rectangle (step 1 in Fig. 2) and the K operator is needed to select it (step 2 in Fig. 2) clicking the mouse button. Next, users had to drag the rectangle using the dragging operator (D) until the first rectangle was over the second one (step 3 in Fig. 2) releasing it with a mouse button action (T_K). The time estimated by GOMS to complete each *Drag & Drop* action is therefore $T_P + T_D + 2T_K$. The time required to drag the object (T_D) was recorded (in milliseconds) to be used in the statistical analysis.

The third and last test (task 3) was designed to evaluate the user performance in the execution of *Item Selection* tasks, which are used to select items in a user interface (e.g. menus, combo boxes, radio button groups, etc.). In this test, users were asked to select a given color in a popup menu. To achieve this operation participants needed to execute

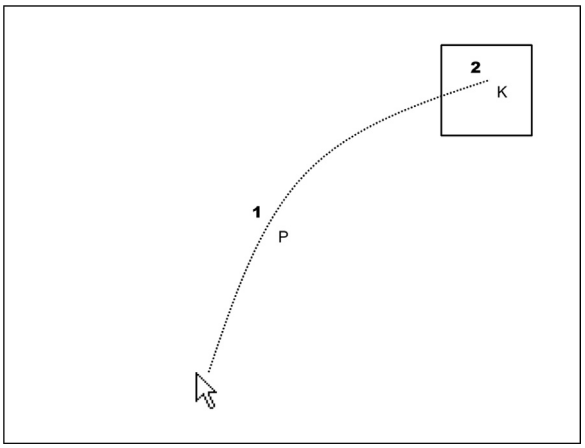


Fig. 1. GOMS operators required to complete a *Point & Click* task. Step 1: users move the mouse pointing (P) to the target. Step 2: the user clicks on the target Key pressing (K) the mouse button.

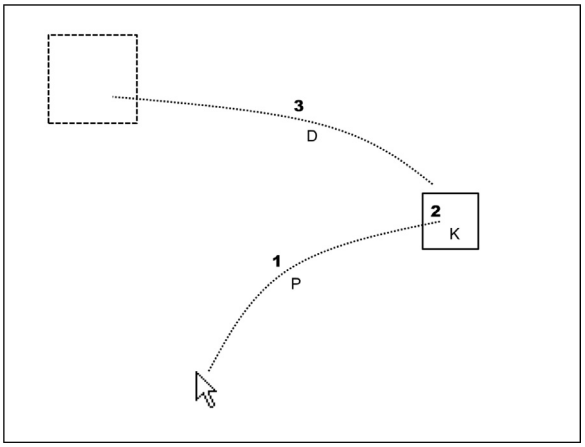


Fig. 2. GOMS operators required to complete a *Drag & Drop* task in the test application. Steps 1 and 2 are the same as in the *Point & Click* task described in Fig. 1. In step 3 users had to drag (D) the small rectangle over the big one.

a *Point & Click* task to display the menu items available clicking on the menus title. Then, users were asked to select a specific menu item whose name was displayed in the screen. Then, participants executed a second *Point & Click* task to click on the menu item corresponding to the asked color. The process was repeated ten times. In each interaction, the menu was placed in a different position using the Z shaped layout described before. Each menu contained five items. Each volunteer had to select each menu item two times across the interactions.

The time required to achieve the first *Point & Click* task (see Fig. 3) was denoted by $T_{P1} + T_{K1}$. It represents the Point (P_1) and Key pressing (K_2) operators required to activate the menu. The second runtime was denoted by $T_{P2} + T_{K2}$. Finally, the time needed by the mental operator M to take the decision (selecting which menu item satisfies the search constraints) was denoted by T_M. The resulting execution time predicted by GOMS for the entire test process, denoted by $T_{P1} + T_{K1} + T_M + T_{P2} + T_{K2}$ was recorded to later statistical analysis.

3.3. Subjects

GOMS assumes that the volunteers know how to use the web system under evaluation (either because they got some previous training or because they have used the system previously). GOMS also assumes that users will not commit any error during the process. Due to this high degree of expertise, users are supposed to interact as fast as possible.

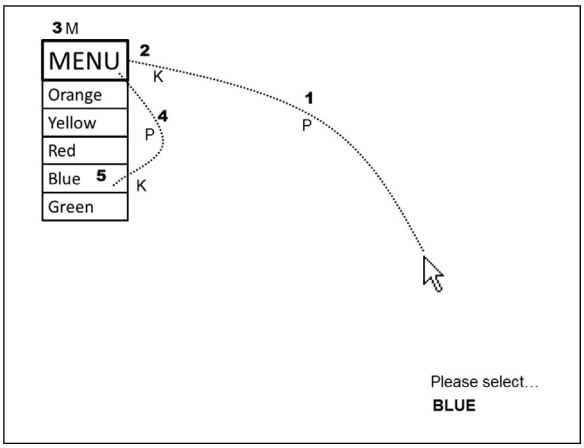


Fig. 3. GOMS operators required to complete an *Item Selection* task. Steps 1 and 2 define the *Point & Click* task required to activate the menu clicking on its title. Once the menu items are displayed, a Mental (M) operator is executed (step 3) to select the required color (displayed in the bottom-right corner of the screen). Steps 4 and 5 represent the Pointing (P) and Key pressing (K) GOMS's operators required to complete the *Point & Click* task needed to select the menu item.

Based on these precepts, GOMS is a reliable tool to estimate the user's effectiveness (execution time) instead of estimating the users efficiency (success/failure rate).

To meet these strong requirements, the 592 individuals participating in the case study were recruited through Twitter and *Foro Coches* (http://www.forococoches.com/), the most popular general purpose online community in Spain, thereby ensuring that participants were familiar with the basic interaction tasks frequently found in online systems. Therefore, participants could execute *Point & Click, Drag & Drop* and *Item Selection* tasks in a so natural way that they did not need to think about the steps needed to complete them.

This approach not only complied with the GOMS requirements but also allowed the participation of a high number of users. The sample used in this study include 592 individuals. It is large when compared with the samples used in the studies described in the *Related Work* section, which were mostly based on samples whose size ranges between 10 and 20 individuals.

This high number allowed the use of multivariate regression analysis to obtain more accurate results when compared with those of prior studies. In addition, it allowed the inclusion of some variables in the model that may bias the results if they are not adequately controlled for (handedness and prior experience with computers).

3.4. Variables of the study

Apart from the variables used to test our first two hypotheses (age and gender) and the execution times of the analyzed tasks we considered some additional variables for the testing of h_1 and h_2.

Specifically, we included handedness and previous user experience with computers. Several studies reported differences regarding movement between left and right handed individuals [66–68], movement preparation [69–72], stimulus velocity effect [73] and interactions between hand preference and hand performance [74]. Besides this, other studies suggest that skill performance and the amount of practice are correlated [75] following an exponential law [76].

These precedents suggest that users handedness and the users experience may have a sensible influence on the user behavior and therefore in their execution time. As a result, these two variables (handedness and amount of practice) were incorporated as control variables in the regression models that are explained in the next section.

As was noted above, the experiment was completed by 592 participants. As a summary, we indicate in Tables 1 and 2 the variables used in the study. Before participating in the tests, users were asked to fulfill a questionnaire to provide information about their age, gender, handedness (tendency to use either the right or the left hand) and experience in the use of computers. This last parameter was provided in terms of the number of weekly hours spent by the users interacting with computers. Some of the users were reluctant to provide their actual age (especially older users). As a result, we were forced to discretize the age value in ranges of 5 years. This way we sacrifice some of the statistical analysis to obtain this parameter from all the users participating in the tests.

Table 1
Dependent variables in the study.

Name	Definition
Point & Click	Time (T_P) required to pointing (P) each object during the test (measured in milliseconds)
Drag & Drop	Time (T_D) required to drag (D) each object during the test (measured in milliseconds)
Item Selection	Time required selecting each menu item during the test. It was calculated as $T_{P1} + T_{K1} + T_M + T_{P2} + T_{K2}$ (measured in milliseconds)

Table 2
Independent variables in the study.

Age	Age Group	Minimum Age	Maximum Age
	0	0	15
	1	16	20
	2	21	25
	3	26	30
	4	31	35
	5	36	40
	6	41	45
	7	46	50
	8	51	55
	9	56	60
	10	61	65
	11		>= 66
HoursUse	Weekly number of hours interacting with computers.		
Gender	1 Female, 0 Male		
Handedness	1 Left handed, 0 Right handed		

3.5. Statistical methods

First, we computed some descriptive statistics about both the dependent and independent variables. The exam of such data gives us a first idea of the features of the individuals in the sample and their behavior in the experiment.

Second, to test hypotheses h_1 and h_2, we estimated a Linear Regression model for each of the tasks. The regression equations have the following form:

$$Task_i = a_0 + a_1\ x\ Age + a_2\ x\ Gender + a_3\ x\ HoursUse + a_4\ x\ LeftHanded + \epsilon_i$$

Where $Task_i$ is the dependent variable in each one of the models, a_0 is the intercept term, a_1 to a_4 are the coefficients of the independent variables in the models and ϵ_i is the error term.

Regarding these equations, and as prior robustness checks, we tested for multicollinearity and heteroskedasticity. Multicollinearity tests were conducted through the calculation of the Condition Indices (CI) and the

Table 3
Descriptive statistics for the dependent variables in the study (execution times are measured in milliseconds).

	Mean	Std. Dev.	Min.	Max.
Point & Click	16,864.77	4294.92	9319	45,792
Drag & Drop	32,832.77	10615.61	19,595	159,867
Item Selection	61,139.34	14069.62	38,351	147,630

Variance Inflation Factors (VIF). To assess whether heteroskedasticity represents a problem we used the Cook and Weisberg test [77].

Furthermore, we also conducted some post-estimation additional tests which allow shedding light on specific concerns about whether (a) there are extreme values which have an abnormal influence on the results, (b) the model is not correctly specified and (c) results are sensitive about the browser/operating system used. First, to detect the presence of influential cases we computed Cooks D statistic for each data point in the regressions. Second, and regarding model specification, we tested for the existence of non-linear effects for the age variable (that is, whether middle-age users perform better than both younger and older users). This was done by adding a quadratic term (Age^2) to the equations and reestimating the models. Finally, we also re-estimated the models for different subsamples defined considering the browser used for the test (three subsamples: Chrome, Firefox, IExplorer, as the number of persons using other navigators was not enough to allow regression equation estimation) and the operating system (Windows, Linux and Mac).

Finally, and to know whether the execution time of individuals about one task is related to the performance in the other tasks (h_3) we conducted a correlation analysis. We computed Nonparametric correlation coefficients (Spearmans Rho) to avoid the problems caused by nonnormality of data. To test normality of data we used the Lilliefors test, and in all cases data distributions departed significantly from normality (results not reported due to space limitations). For the calculations of these statistics, as well as for all the other tests and equations indicated above, we used the statistical package STATA 11.

4. Results

4.1. Descriptive statistics

As we might expect the average execution time depends on the complexity of the test. As mentioned previously, some authors [42–44] reported a higher level of complexity in the execution of *Drag & Drop* tasks when compared with *Point & Click*. Thereby, Table 3 shows that *Drag & Drop* tasks required a higher amount of time that *Point & Click* tasks.

Furthermore, the runtime of menu selection tasks is higher than that of the other two. This result is consistent with predictions provided by GOMS analysis studied before. Notice that while the *Point & Click* and the *Drag & Drop* tasks required the execution of single P or D operator, the *Item Selection* tasks requires the execution of two P operators (one for menu activation and another one for item selection). Besides that, item selection requires the execution of a complex M operator to take the decision of what item to select.

With regard to the sample descriptive indicators (Tables 4–7) it is noticeable that the sample is mainly composed by individuals which are young, male and have intensive experience in the use of computers. However, the number of observations that correspond to the other types of web applications users (women, elder and low experienced users) is sufficient to conduct a valid statistical study. Furthermore, and regarding handedness, around 11% of the individuals in the sample are left-handed. This value is consistent with the global rate of left-handed people, that is estimated between 10% and 13% [78].

Table 4
Frequency distribution for Age.

Age	Number of observations
0	2
1	85
2	182
3	145
4	77
5	38
6	25
7	13
8	12
9	9
10	2
11	2
Total	592

Table 5
Frequency distribution for Gender.

Gender	Number of observations
0 (Male)	462
1 (Female)	130
Total	592

Table 6
Frequency distribution for HoursUse.

HoursUse	Number of observations
0	1
1	16
2	58
3	67
4	63
5	33
6	354
Total	592

Table 7
Frequency distribution for handedness.

Handedness	Number of observations
0 (Right Handed)	524
1 (Left Handed)	68
Total	592

4.2. Regression analysis results (hypotheses h_1 and h_2)

Table 8 indicates the main results of the three regression models and the related tests. Prior to the comment of the results we must highlight that all CIs of the different variables in the three regression models are below 15. In accordance to this, all VIFs are below 10. These values are common thresholds to discard the presence of significant multicollinearity among the variables of a linear regression model [79]. For the sake of clarity in the presentation of the results we do not include CI and VIF values in Table 8, but data are available from the authors upon request.

Results of the Cook–Weisberg test for heteroscedasticity are detailed in the last row of the table. We indicate the values of the chi-squared test statistic and the corresponding p value. As the null hypothesis for this test is that variance is constant we can conclude that such hypothesis is rejected in the three cases and heteroscedasticity is significant. So, we repeated the estimation of the regression equations using a robust estimation procedure, which consisted in the calculation of robust standard errors for the coefficients in the different regression equations

Table 8
Regressions results and related tests.

	Point & Click	Drag & Drop	Item Selection
Age Group			
Parameter stimate	500.74	1721.47	3077.63
t statistic	4.40	6.86	8.82
p-value	<0.001	<0.001	<0.001
Gender			
Parameter stimate	1742.03	5077.81	6293.93
t statistic	4.29	3.91	4.32
p-value	<0.001	<0.001	<0.001
HoursUse			
Parameter stimate	−394.89	−895.07	−1481.98
t statistic	−3.02	−2.46	−4.73
p-value	0.003	0.014	<0.001
Handedness			
Parameter stimate	707.98	2986.57	107.61
t statistic	1.18	1.53	0.07
p-value	0.238	0.127	0.946
Intercept			
Parameter stimate	16,759.52	30,363.66	57,363.51
t statistic	19.21	16.99	27.88
p-value	<0.001	<0.001	<0.001
F test			
F-statistic	17.48	13.46	33.27
p-value	<0.001	<0.001	<0.001
Adjusted R^2	11.41%	17.59%	25.49%
Cook–Weisberg test for heteroscedasticity			
Chi-squared	7.07	388.80	44.60
p-value	0.007	<0.001	<0.001

and robust p-values, including Whites correction [80]. Such results are those displayed in Table 8.

The layout of the rest of the table is as follows: in each column, we show the statistics for each one of the regression equations (where, the dependent variables are, respectively, time for completion of point & click, drag & drop, and item selection tasks). The first five cells of each column contain the estimates for each one of the independent variables and the intercept of each model. In each cell, the upper figure is the coefficient estimate, that in the middle is the robust t statistic (computed using the standard error that includes Whites correction) and the figure shown in the lower part of the cell is the robust p-value. In addition, the table displays for each model, the F statistic for the test of the joint significance of the coefficients and its p-value, as well as the adjusted R^2 and the results of the aforementioned Cook–Weisberg test for heteroscedasticity.

With regard to the results, we must first underline that although R^2s are not very high, conducted F tests evidence the jointly significance of the coefficients of the variables, that is, the set of variables, considered as a unit, influence the performance in all the tests.

Regarding the parameter estimates, coefficients for age are always positive and significant. These results give support to the first hypothesis (h_1) as it is evidenced that older users perform worse for all the tasks (needing more time to complete. So, the performance decline regarding the age is confirmed.

Furthermore, the gender coefficient is significantly positive in all cases. So, women perform worse on *Point & Click, Drag & Drop* and *Item Selection* tasks. These results suggest that hypothesis h_2 also holds corroborating the observations made by Inkpen [53] regarding girls having difficulty with *Drag & Drop* tasks. However, our data does not support Inkpens other observations related to the absence of any significant gender difference in the overall movement time. That leads us to conclude that Inkpens results are more related to children different learning styles than directly to the gender.

Table 9
Results of the correlation analysis.

	Point & Click	Drag & Drop	Item Selection
Point & Click			
Drag & Drop	0.731 0.000		
Item Selection	0.660 0.000	0.674 0.000	

Regarding the control variables in the model, it is first noticeable that prior experience with computers is significant in all cases. Coefficients for *HoursUse* are significantly negative in all cases, meaning that more hours of computer use always imply a better performance. These results are similar to those obtained by a prior study [21], which was focused on observations based on the execution of top level interaction tasks, mostly related with cognition and perception. Our findings confirm that the same effect is observable at the low level of interaction required by the GOMS analysis, which is mostly based on the human motoric system. With regard to the other control variable, handedness does not seem to have an influence, as left-handed users perform neither significantly best nor significantly worse than right-handed.

With respect to the additional post-estimation tests, we must first underline that Cooks D values are always lower than 1 for all the individuals in all the regressions so there are no influential cases in the models. Second, none of the quadratic terms (Age^2) that we included in alternative versions of the equations was found to be significant. So, we can reject the existence of non-linear effects for the age variable. Finally, the re-estimation of the models for different subsamples defined considering the browser used for the test (Chrome, Firefox, IExplorer,) and the operating system (Windows, Linux, Mac) produced results which are qualitatively the same as those displayed in Table 8. For the sake of brevity, we did not include in the paper the results. However, they are available from the authors upon request.

4.3. Correlation analysis results (hypothesis h_3)

The results of the correlation analysis we conducted to assess whether individuals that perform well in a certain task also perform well in the others (h_3) are shown in Table 9. In each of the cells we display the non-parametric Spearman correlation coefficient (upper figure) and the corresponding p-value (lower figure). Cells below the main diagonal contain the results of tests.

Data in Table 9 evidence that correlations are significant among all the tasks. This finding supports hypothesis h_3, suggesting that the execution time (performance) of an individual in a specific task, keeps its coherence in the other tasks as well. So, for example, if a person has superior performance in the execution *Point & Click* tasks, she/he is expected to also have superior performance in the execution of *Drag & Drop* and *Item Selection* tasks.

This finding may have a relevant impact in the future design of automatic user modeling algorithms. As the three proposed interaction tasks have the same usefulness in terms of user categorization, any of them can be used separately to achieve this goal. Moreover, the amount of data required to automatically infer the type of user may be notably reduced (as only one task is analyzed), which is crucial for the execution of real time algorithms.

5. Limitations, future directions, and conclusions

This work had two interrelated goals. First, we wanted to assess whether the gender and age are sufficiently significant determining factors to support an automatic profiling system based on the analysis of

mouse motion behavior when executing *Point & Click, Drag & Drop* and *Item Selection* tasks. Second, to figure out whether the individuals perform consistently across these basic interaction tasks, that is, if their performance in one of them are extrapolable (or not) to the others.

Regarding the first, the results of the empirical study reveal that both age and gender factors are significantly determinant. While older users performed worse than younger in each the interaction tasks, men obtained better results than women. On the other hand, in relation with the analysis of correlations between the execution times of the target basic interaction tasks, data gathered in the tests revealed consistencies in the execution times of individuals across them. User's performance measured in any of these tasks is coherent to their execution time in the other tasks.

These results open the door to implement a system that automatically classifies users in age and gender groups by observing the way they interact and perform in these basic interaction tasks with any web interface. However, these evidences must be taken carefully, given that the data was gathered through artificial and isolated ad-hoc tests, and not in a real web interface where the behavior of the user can differ from the one evidenced during the tests. On the other hand, the existing correlation between the way individuals perform across the different interaction tasks makes it more flexible not only to integrate the data gathering processes into the final system (since developers do not need to force the use of all of them), but also it expands the data gathering possibilities to a number of observations whose results could be combined in a hybrid voting algorithm or a machine learning based system.

The possible benefits of such a classification system are straightly applicable in e-commerce sites, the main target of this work, since the information architecture of the site (and the list of products or sales offered) could be adapted accordingly to the preferences of this target user. However, there are other possible applications like preventing some users to claim the identities of other users or from pretending being a different age and/or have a different gender. In addition, detecting old users would support the automatic adaptation of the interface to the specific features of this group, using for example bigger fonts and simpler interfaces.

Besides the design, implementation and evaluation of this system in a real environment, we consider there are other factors that could somehow determine users performance in basic interaction tasks and that should be considered to extend this study in the future. One of them is the cultural factor. The sample used is limited to western cultures individuals. Some studies suggest that the culture of an individual could determine his/her performance. Ford et al. [81] designed an experiment to evaluate if any of the Hofstedes cultural dimensions can affect human performance while interaction with computers. Even though their study did not provide sufficient evidence to reach any determining conclusion, we consider it would be interesting to extend this work to a multicultural sample of individuals to study such influence in these specific types of interaction.

Acknowledgments

This work has been funded by the Department of Science and Technology (Spain) under the National Program for Research, Development and Innovation: project TIN2009-12132, entitled SHUBAI: Augmented Accessibility for Handicapped Users in Ambient Intelligence and in Urban Computing Environments.

References

[1] P.-T. Chen, H.-P. Hsieh, Personalized mobile advertising: its key attributes, trends, and social impact, Technol. Forecast. Soc. Change 79 (3) (2012) 543–557.
[2] A. Shen, Recommendations as personalized marketing: insights from customer experiences, J. Services Market. 28 (5) (2014) 414–427.
[3] J. Choi, H.J. Lee, Y.C. Kim, The influence of social presence on customer intention to reuse online recommender systems: the roles of personalization and product type, Int. J. Electron. Commerce 16 (1) (2011) 129–154.
[4] H. Xu, X. Zou, H. Wang, Consumers attitudes of e-commerce in china, Issues Inf. Syst. 7 (2) (2006) 202.
[5] A. Haque, J. Sadeghzadeh, A. Khatibi, Identifying potentiality online sales in malaysia: a study on customer relationships online shopping, J. Appl. Business Res. (JABR) 22 (4) (2011).
[6] M.-C. Lee, Understanding the behavioural intention to play online games: an extension of the theory of planned behaviour, Online Inf. Rev. 33 (5) (2009) 849–872.
[7] E.B. Weiser, Gender differences in internet use patterns and internet application preferences: a two-sample comparison, CyberPsychol. Behavior 3 (2) (2000) 167–178.
[8] D. Freudenthal, Age differences in the performance of information retrieval tasks, Behaviour Inf. Technol 20 (1) (2001) 9–22.
[9] H.-J. Kim, K. Bae, H.-S. Yoon, Age and gender classification for a home-robot service, in: Robot and Human interactive Communication, 2007. RO-MAN 2007. The 16th IEEE International Symposium on, IEEE, 2007, pp. 122–126.
[10] J.E. Agudo, H. Sánchez, M. Rico, Playing games on the screen: Adapting mouse interaction at early ages, in: Advanced Learning Technologies (ICALT), 2010 IEEE 10th International Conference on, IEEE, 2010, pp. 493–497.
[11] Y. Cheong, R.L. Shehab, C. Ling, Effects of age and psychomotor ability on kinematics of mouse-mediated aiming movement, Ergonomics 56 (6) (2013) 1006–1020.
[12] L. Ardissono, A. Goy, Tailoring the interaction with users in web stores, User Model User-adapt. Interact. 10 (4) (2000) 251–303.
[13] S.R. Alpert, J. Karat, C.-M. Karat, C. Brodie, J.G. Vergo, User attitudes regarding a user-adaptive ecommerce web site, User Model User-adapt. Interact. 13 (4) (2003) 373–396.
[14] T.C. Sebora, S.M. Lee, N. Sukasame, Critical success factors for e-commerce entrepreneurship: an empirical study of thailand, Small Business Econ. 32 (3) (2009) 303–316.
[15] E. Turban, D. King, J. McKay, P. Marshall, J. Lee, L. Viehland, Electronic commerce 2008 a managerial perspective(2008).
[16] N.H. Woo, S. Shirmohammadi, Modeling and measurement of personality for e–commerce systems, in: Instrumentation and Measurement Technology Conference Proceedings, 2008. IMTC 2008. IEEE, IEEE, 2008, pp. 787–792.
[17] Y. Yang, M.-A. Aufaure, C. Claramunt, Towards a dl-based semantic user model for web personalization, in: Autonomic and Autonomous Systems, 2007. ICAS07. Third International Conference on, IEEE, 2007, p. 61.
[18] J. Zhang, A.A. Ghorbani, Gumsaws: A generic user modeling server for adaptive web systems, in: Communication Networks and Services Research, 2007. CNSR'07. Fifth Annual Conference on, IEEE, 2007, pp. 117–124.
[19] P.S. Strom, R.D. Strom, J.J. Wingate, M.F. Kraska, T.E. Beckert, Cyberbullying: assessment of student experience for continuous improvement planning, NASSP Bull. 96 (2) (2012) 137–153.
[20] R.L. Hill, A. Dickinson, J.L. Arnott, P. Gregor, L. McIver, Older web users' eye movements: experience counts, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2011, pp. 1151–1160.
[21] J. De Andrés, B. Pariente, M. Gonzalez-Rodriguez, D. Fernandez Lanvin, Towards an automatic user profiling system for online information sites: identifying demographic determining factors, Online Inf. Rev. 39 (1) (2015) 61–80.
[22] L.E. Rohr, Gender-specific movement strategies using a computer-pointing task, J. Mot. Behav. 38 (6) (2006) 431–437.
[23] L. Beckwith, M. Burnett, V. Grigoreanu, S. Wiedenbeck, Gender hci: what about the software? Computer 39 (11) (2006).
[24] W. Liu, F. Jin, X. Zhang, Ontology-based user modeling for e-commerce system, in: Pervasive Computing and Applications, 2008. ICPCA 2008. Third International Conference on, vol. 1, IEEE, 2008, pp. 260–263.
[25] J.M. Such, A. Garcia-Fornes, V. Botti, Automated buyer profiling control based on human privacy attitudes, Electron. Commer. Res. Appl. 12 (6) (2013) 386–396.
[26] D. Fijałkowski, R. Zatoka, An architecture of a web recommender system using social network user profiles for e-commerce, in: Computer Science and Information Systems (FedCSIS), 2011 Federated Conference on, IEEE, 2011, pp. 287–290.
[27] A. Ghazarian, S.M. Noorhosseini, Automatic detection of users skill levels using high--frequency user interface events, User Model User-adapt. Interact. 20 (2) (2010) 109–146.
[28] A. Garg, R. Rahalkar, S. Upadhyaya, K. Kwiat, Profiling users in gui based systems for masquerade detection, in: Proceedings of the 2006 IEEE Workshop on Information Assurance, vol. 2006, 2006, pp. 48–54.
[29] A.D. Fisk, S.J. Czaja, W.A. Rogers, N. Charness, J. Sharit, Designing for Older Adults: Principles and Creative Human Factors Approaches, CRC press, 2009.
[30] J. Vines, G. Pritchard, P. Wright, P. Olivier, K. Brittain, An age-old problem: examining the discourses of ageing in hci and strategies for future research, ACM Trans. Comput.-Human Interact. (TOCHI) 22 (1) (2015) 2.
[31] A. Dickinson, J. Arnott, S. Prior, Methods for human–computer interaction research with older people, Behaviour Inf. Technol. 26 (4) (2007) 343–352.
[32] J.L. Fozard, M. Vercruyssen, S.L. Reynolds, P. Hancock, R.E. Quilter, Age differences and changes in reaction time: the baltimore longitudinal study of aging, J. Gerontol. 49 (4) (1994) P179–P189.
[33] N. Walker, D.A. Philbin, A.D. Fisk, Age-related differences in movement control: adjusting submovement structure to optimize performance, J. Gerontol. Series B 52 (1) (1997) P40–P53.
[34] N.B. Stubbs, J.E. Fernandez, W.M. Glenn, Normative data on joint ranges of motion of 25-to 54-year-old males, Int. J. Ind. Ergon. 12 (4) (1993) 265–272.
[35] E.J. Metter, R. Conwit, J. Tobin, J.L. Fozard, Age-associated loss of power and strength in the upper extremities in women and men, J. Gerontol. Series A 52 (5) (1997) B267–B276.
[36] R.W. Morrell, Older Adults, Health Information, and the World Wide Web, Psychology Press, 2001.

[37] B. Xie, Older adults, computers, and the internet: future directions, Gerontechnology 2 (4) (2003) 289–305.
[38] H.P. Nap, H.P. De Greef, D. Bouwhuis, Access for all by cognitive engineering, Gerontechnology 3 (4) (2005) 258.
[39] H. Sayers, Desktop virtual environments: a study of navigation and age, Interact. Comput. 16 (5) (2004) 939–956.
[40] M.A. Neerincx, J. Lindenberg, J. Rypkema, S. Van Besouw, A practical cognitive theory of web-navigation: Explaining age-related performance differences, Position Paper CHI 2000 Workshop Basic Research Symposium, 2000.
[41] T. Tullis, Older adults and the web: lessons learned from eye-tracking, in: Universal Acess in Human Computer Interaction. Coping with Diversity, 2007, pp. 1030–1039.
[42] I.S. MacKenzie, A. Sellen, W.A.S. Buxton, A comparison of input devices in element pointing and dragging tasks, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 1991, pp. 161–166.
[43] A. Chadwick-Dias, M. McNulty, T. Tullis, Web usability and age: how design changes can improve performance, in: ACM SIGCAPH Computers and the Physically Handicapped, 73–74, ACM, 2003, pp. 30–37.
[44] S.J. Czaja, C.C. Lee, The impact of aging on access to technology, Univ. Access Inf. Soc. 5 (4) (2007) 341.
[45] R. Joiner, D. Messer, P. Light, K. Littleton, It is best to point for young children: a comparison of children's pointing and dragging, Comput. Human Behav. 14 (3) (1998) 513–529.
[46] J.P. Kuhtz-Buschbeck, H. Stolze, K. Jöhnk, A. Boczek-Funcke, M. Illert, Development of prehension movements in children: a kinematic study, Exp. Brain Res. 122 (4) (1998) 424–432.
[47] D.C.M. Carvalho, M.E.C. Bessa, L.G.M. Magalha es, E. Carrapatoso, Interaction paradigms versus age-related user profiles: an evaluation on content selection, IEEE Lat. Am. Trans. 13 (2) (2015) 532–539.
[48] L. Beckwith, Gender hci issues in end-user software engineering, in: Human Centric Computing Languages and Environments, 2003. Proceedings. 2003 IEEE Symposium on, IEEE, 2003, pp. 273–274.
[49] L. Beckwith, M. Burnett, Gender: An important factor in end-user programming environments? in: Visual Languages and Human Centric Computing, 2004 IEEE Symposium on, IEEE, 2004, pp. 107–114.
[50] G. Christoph, F. Goldhammer, J. Zylka, J. Hartig, Adolescents' computer performance: the role of self-concept and motivational aspects, Computers Education 81 (2015) 1–12.
[51] M. Czerwinski, D.S. Tan, G.G. Robertson, Women take a wider view, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2002, pp. 195–202.
[52] D.S. Tan, M. Czerwinski, G. Robertson, Women go with the (optical) flow, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2003, pp. 209–215.
[53] K.M. Inkpen, Drag-and-drop versus point-and-click mouse interaction styles for children, ACM Trans. Computer-Human Interact. (TOCHI) 8 (1) (2001) 1–33.
[54] J. Wahlström, J. Svensson, M. Hagberg, P.W. Johnson, Differences between work methods and gender in computer mouse use, Scandinavian J. Work Environ. Health (2000) 390–397.
[55] C. Collazos, L.A. Guerrero, M. Llaña, J. Oetzel, Gender: An influence factor in the collaborative work process in computer-mediated communication, in: International Conference on Nanotechnology and Environment Engineering, (June 2014), 2016.
[56] S.K. Card, A. Newell, T.P. Moran, The psychology of human-computer interaction(1983).
[57] L. Xiang, C. Xiaoli, The research on performance of automobile human-machine interface based on bhr-goms behavior model, in: Intelligent Computing and Intelligent Systems (ICIS), 2010 IEEE International Conference on, volume 2, IEEE, 2010, pp. 174–178.
[58] E. Abdulin, Using the keystroke-level model for designing user interface on middle–sized touch screens, in: CHI'11 Extended Abstracts on Human Factors in Computing Systems, ACM, 2011, pp. 673–686.
[59] M. Schrepp, Goms analysis as a tool to investigate the usability of web units for disabled users, Univ. Access Inf. Soc. 9 (1) (2010) 77–86.
[60] S.A. Oyewole, J.M. Haight, Determination of optimal paths to task goals using expert system based on goms model, Comput. Human Behav. 27 (2) (2011) 823–833.
[61] S. Zhai, Characterizing computer input with fittslaw parametersthe information and non-information aspects of pointing, Int. J. Human Comput. Stud. 61 (6) (2004) 791–809.
[62] Y. Guiard, H.B. Olafsdottir, S.T. Perrault, Fitt's law as an explicit time/error trade-off, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2011, pp. 1619–1628.
[63] T.A. Salthouse, Effects of age and skill in typing., J. Exp. Psychol. 113 (3) (1984) 345.
[64] L. Rosati, How to design interfaces for choice: Hick-hyman law and classification for information architecture, in: Proceedings of the International UDC Seminar, 2013, pp. 121–134.
[65] D.W. Schneider, J.R. Anderson, A memory-based model of hicks law, Cogn. Psychol. 62 (3) (2011) 193–222.
[66] A. Lenhard, J. Hoffmann, Constant error in aiming movements without visual feedback is higher in the preferred hand, Laterality 12 (3) (2007) 227–238.
[67] P.E. Mieschke, D. Elliott, W.F. Helsen, R.G. Carson, J.A. Coull, Manual asymmetries in the preparation and control of goal-directed movements, Brain Cogn. 45 (1) (2001) 129–140.
[68] J.-L. Velay, S. Benoit-Dubrocard, Hemispheric asymmetry and interhemispheric transferin reaching programming, Neuropsychologia 37 (8) (1999) 895–903.
[69] W.F. Helsen, J.L. Starkes, D. Elliott, M.J. Buekers, Manual asymmetries and saccadic eye movements in right-handers during single and reciprocal aiming movements, Cortex 34 (4) (1998) 513–530.
[70] P. Boulinguez, S. Barthélémy, B. Debu, Influence of the movement parameter to be controlled on manual rt asymmetries in right-handers, Brain Cogn. 44 (3) (2000) 653–661.
[71] P.E. Bestelmeyer, D.P. Carey, Processing biases towards the preferred hand: valid and invalid cueing of left-versus right-hand movements, Neuropsychologia 42 (9) (2004) 1162–1167.
[72] K. Neely, G. Binsted, M. Heath, Manual asymmetries in bimanual reaching: the influence of spatial compatibility and visuospatial attention, Brain Cogn. 57 (1) (2005) 102–105.
[73] P.C. Rodrigues, R. Barbosa, A.I. Carita, J. Barreiros, O. Vasconcelos, Stimulus velocity effect in a complex interceptive task in right-and left-handers, Eur. J. Sport Sci. 12 (2) (2012) 130–138.
[74] M. Peters, J. Ivanoff, Performance asymmetries in computer mouse control of right-handers, and left-handers with left-and right-handed mouse experience, J. Motor Behav. 31 (1) (1999) 86–94.
[75] R.W. Howard, Learning curves in highly skilled chess players: a test of the generality of the power law of practice, Acta Psychol. 151 (2014) 16–23.
[76] A. Heathcote, S. Brown, D.J. Mewhort, The power law repealed: the case for an exponential law of practice, Psychon. Bull. Rev. 7 (2) (2000) 185–207.
[77] R.D. Cook, S. Weisberg, Diagnostics for heteroscedasticity in regression, Biometrika 70 (1) (1983) 1–10.
[78] M. Raymond, D. Pontier, A.-B. Dufour, A.P. Moller, Frequency-dependent maintenance of left handedness in humans, Proc. R. Soc. London B 263 (1377) (1996) 1627–1633.
[79] M.H. Kutner, C. Nachtsheim, J. Neter, W. Li, Applied Linear Statistical Models, McGraw-Hill Irwin, 2005.
[80] H. White, A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, Econometrica (1980) 817–838.
[81] G. Ford, H. Gelderblom, The effects of culture on performance achieved through the use of human computer interaction, in: Proceedings of the 2003 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on Enablement through Technology, South African Institute for Computer Scientists and Information Technologists, 2003, pp. 218–230.

Bibliography

[1] Abdulin, E. Using the keystroke-level model for designing user interface on middle-sized touch screens. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. 2011, pp. 673–686.

[2] Agudo, J. E., Sánchez, H., and Rico, M. Playing games on the screen: Adapting mouse interaction at early ages. In *2010 10th IEEE International Conference on Advanced Learning Technologies* (2010), IEEE, pp. 493–497.

[3] Alfaro, E., García, N., Gámez, M., and Elizondo, D. Bankruptcy forecasting: An empirical comparison of adaboost and neural networks. *Decision Support Systems 45*, 1 (2008), 110–122.

[4] author. Forocoches. `https://www.forocoches.com`.

[5] Bauer, E., and Kohavi, R. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine learning 36*, 1-2 (1999), 105–139.

[6] Becerra-Fernandez, I., Zanakis, S. H., and Walczak, S. Knowledge discovery techniques for predicting country investment risk. *Computers & Industrial Engineering 43*, 4 (2002), 787–800.

[7] Beckwith, L. Gender hci issues in end-user software engineering. In *Human Centric Computing Languages and Environments, 2003. Proceedings. 2003 IEEE Symposium on* (2003), IEEE, pp. 273–274.

[8] Beckwith, L., and Burnett, M. Gender: An important factor in end-user programming environments? In *Visual Languages and Human Centric Computing, 2004 IEEE Symposium on* (2004), IEEE, pp. 107–114.

[9] Beckwith, L., Burnett, M., Grigoreanu, V., and Wiedenbeck, S. Gender hci: What about the software? *Computer 39*, 11 (2006).

[10] Bestelmeyer, P. E., and Carey, D. P. Processing biases towards the preferred hand: valid and invalid cueing of left-versus right-hand movements. *Neuropsychologia 42*, 9 (2004), 1162–1167.

[11] Bhargava, N., Sharma, G., Bhargava, R., and Mathuria, M. Decision tree analysis on j48 algorithm for data mining. *Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering 3*, 6 (2013).

[12] Boinee, P. Insights into machine learning: Data clustering and classification algorithms for astrophysical experiments.

[13] Boinee, P., De Angelis, A., and Foresti, G. L. Meta random forests. *International Journal of Computational Intelligence 2*, 3 (2005), 138–147.

[14] Boulinguez, P., Barthélémy, S., and Debu, B. Influence of the movement parameter to be controlled on manual rt asymmetries in right-handers. *Brain and cognition 44*, 3 (2000), 653–661.

[15] Breiman, L. Heuristics of instability in model selection. technique report. statistics department. *University of California at Berkeley* (1994).

[16] Breiman, L. Bagging predictors. *Machine learning 24*, 2 (1996), 123–140.

[17] Breiman, L. Bias, variance, and arcing classifiers. Tech. rep., Tech. Rep. 460, Statistics Department, University of California, Berkeley ..., 1996.

[18] Breiman, L. Prediction games and arcing algorithms. *Neural computation 11*, 7 (1999), 1493–1517.

[19] Breiman, L. Random forests. *UC Berkeley TR567* (1999).

[20] Breiman, L. Random forests. *Machine learning 45*, 1 (2001), 5–32.

[21] Breiman, L., et al. Arcing classifier (with discussion and a rejoinder by the author). *The annals of statistics 26*, 3 (1998), 801–849.

[22] Breusch, T. S., and Pagan, A. R. A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society* (1979), 1287–1294.

[23] Çakir, A., Çalış, H., and Küçüksille, E. U. Data mining approach for supply unbalance detection in induction motor. *Expert systems with applications 36*, 9 (2009), 11808–11813.

[24] Card, S. K., Moran, T. P., and Newell, A. The psychology of human-computer interaction. 1983, 1983.

[25] Carvalho, D., Bessa, M., Magalha es, L. G. M., and Carrapatoso, E. Interaction paradigms versus age-related user profiles: an evaluation on content selection. *IEEE Latin America Transactions 13*, 2 (2015), 532–539.

[26] Cawley, G. C., and Talbot, N. L. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research 11*, Jul (2010), 2079–2107.

[27] Chadwick-Dias, A., McNulty, M., and Tullis, T. Web usability and age: how design changes can improve performance. *ACM SIGCAPH Computers and the Physically Handicapped*, 73-74 (2002), 30–37.

[28] Chaparro, A., Bohan, M., Fernandez, J., Choi, S. D., and Kattel, B. The impact of age on computer input device use:: Psychophysical and physiolo-

gical measures. *International Journal of Industrial Ergonomics 24*, 5 (1999), 503–513.

[29] Chawla, N. V., Japkowicz, N., and Kotcz, A. Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter 6*, 1 (2004), 1–6.

[30] Chawla, N. V., Lazarevic, A., Hall, L. O., and Bowyer, K. W. Smoteboost: Improving prediction of the minority class in boosting. In *European conference on principles of data mining and knowledge discovery* (2003), Springer, pp. 107–119.

[31] Chen, J., Yin, J., Zang, L., Zhang, T., and Zhao, M. Stacking machine learning model for estimating hourly pm2. 5 in china based on himawari 8 aerosol optical depth data. *Science of The Total Environment 697* (2019), 134021.

[32] Cheong, Y., Shehab, R. L., and Ling, C. Effects of age and psychomotor ability on kinematics of mouse-mediated aiming movement. *Ergonomics 56*, 6 (2013), 1006–1020.

[33] Christoph, G., Goldhammer, F., Zylka, J., and Hartig, J. Adolescents' computer performance: The role of self-concept and motivational aspects. *Computers & Education 81* (2015), 1–12.

[34] Collazos, C., Guerrero, L. A., Llaña, M., and Oetzel, J. Gender: An influence factor in the collaborative work process in computer-mediated communication. In *International Conference on Nanotechnology and Environment Engineering, (June 2014)* (2016).

[35] Cook, R. D., and Weisberg, S. Diagnostics for heteroscedasticity in regression. *Biometrika 70*, 1 (1983), 1–10.

[36] Crook, C. Young children's skill in using a mouse to control a graphical computer interface. *Computers & Education 19*, 3 (1992), 199–207.

[37] Czaja, S. J., and Lee, C. C. The impact of aging on access to technology. *Universal Access in the Information Society 5*, 4 (2007), 341.

[38] Czerwinski, M., Tan, D. S., and Robertson, G. G. Women take a wider view. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (2002), ACM, pp. 195–202.

[39] Davis, J., and Goadrich, M. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning* (2006), pp. 233–240.

[40] de Andrés, J., Landajo, M., and Lorca, P. Bankruptcy prediction models based on multinorm analysis: An alternative to accounting ratios. *Knowledge-Based Systems 30* (2012), 67–77.

[41] De Andrés, J., Pariente, B., Gonzalez-Rodriguez, M., and Fernan-

dez Lanvin, D. Towards an automatic user profiling system for online information sites: Identifying demographic determining factors. *Online Information Review 39*, 1 (2015), 61–80.

[42] Dickinson, A., Arnott, J., and Prior, S. Methods for human–computer interaction research with older people. *Behaviour & Information Technology 26*, 4 (2007), 343–352.

[43] Dietterich, T. G. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning 40*, 2 (2000), 139–157.

[44] Domingo, C., Watanabe, O., et al. Madaboost: A modification of adaboost. In *COLT* (2000), Citeseer, pp. 180–189.

[45] Domingos, P. Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (1999), pp. 155–164.

[46] Donker, A., and Reitsma, P. Aiming and clicking in young children's use of the computer mouse. *Computers in Human Behavior 23*, 6 (2007), 2863–2874.

[47] Donker, A., and Reitsma, P. Young children's ability to use a computer mouse. *Computers & Education 48*, 4 (2007), 602–617.

[48] Dubitzky, W., Granzow, M., and Berrar, D. P. *Fundamentals of data mining in genomics and proteomics*. Springer Science & Business Media, 2007.

[49] Efron, B., and Tibshirani, R. J. *An introduction to the bootstrap*. CRC press, 1994.

[50] Fernandez-Lanvin, D., de Andres-Suarez, J., Gonzalez-Rodriguez, M., and Pariente-Martinez, B. The dimension of age and gender as user model demographic factors for automatic personalization in e-commerce sites. *Computer Standards & Interfaces 59* (2018), 1–9.

[51] Fisk, A. D., Czaja, S. J., Rogers, W. A., Charness, N., and Sharit, J. *Designing for older adults: Principles and creative human factors approaches*. CRC press, 2009.

[52] Flach, P. A. The geometry of roc space: understanding machine learning metrics through roc isometrics. In *Proceedings of the 20th international conference on machine learning (ICML-03)* (2003), pp. 194–201.

[53] Ford, G., and Gelderblom, H. The effects of culture on performance achieved through the use of human computer interaction. In *Proceedings of the 2003 annual research conference of the South African institute of computer scientists and information technologists on Enablement through technology* (2003), South African Institute for Computer Scientists and Information Technologists, pp. 218–230.

[54] Fozard, J. L., Vercruyssen, M., Reynolds, S. L., Hancock, P., and Quilter, R. E. Age differences and changes in reaction time: the baltimore longitudinal study of aging. *Journal of gerontology 49*, 4 (1994), P179–P189.

[55] Freudenthal, D. Age differences in the performance of information retrieval tasks. *Behaviour & Information Technology 20*, 1 (2001), 9–22.

[56] Freund, Y. Boosting a weak learning algorithm by majority. *Information and computation 121*, 2 (1995), 256–285.

[57] Freund, Y., Schapire, R. E., et al. Experiments with a new boosting algorithm. In *icml* (1996), vol. 96, Citeseer, pp. 148–156.

[58] Friedman, J., Hastie, T., Tibshirani, R., et al. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics 28*, 2 (2000), 337–407.

[59] Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[60] Frøkjær, E., Hertzum, M., and Hornbæk, K. Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (2000), pp. 345–352.

[61] Geisser, S. The predictive sample reuse method with applications. *Journal of the American statistical Association 70*, 350 (1975), 320–328.

[62] Goh, K.-S., Chang, E., and Cheng, K.-T. Svm binary classifier ensembles for image classification. In *Proceedings of the tenth international conference on Information and knowledge management* (2001), pp. 395–402.

[63] Guiard, Y., Olafsdottir, H. B., and Perrault, S. T. Fitt's law as an explicit time/error trade-off. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2011), ACM, pp. 1619–1628.

[64] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. The weka data mining software: an update. *ACM SIGKDD explorations newsletter 11*, 1 (2009), 10–18.

[65] Haque, A., Sadeghzadeh, J., and Khatibi, A. Identifying potentiality online sales in malaysia: A study on customer relationships online shopping. *Journal of Applied Business Research (JABR) 22*, 4 (2006).

[66] Hastie, T., Tibshirani, R., and Friedman, J. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[67] Heathcote, A., Brown, S., and Mewhort, D. J. The power law repealed: The case for an exponential law of practice. *Psychonomic bulletin & review 7*, 2 (2000), 185–207.

[68] Helsen, W. F., Starkes, J. L., Elliott, D., and Buekers, M. J. Manual asymmetries and saccadic eye movements in right-handers during single and reciprocal aiming movements. *Cortex 34*, 4 (1998), 513–530.

[69] Hernández-Orallo, J., Flach, P., and Ferri, C. A unified view of performance metrics: translating threshold choice into expected classification loss. *Journal of Machine Learning Research 13*, Oct (2012), 2813–2869.

[70] Hill, R. L., Dickinson, A., Arnott, J. L., Gregor, P., and McIver, L. Older web users' eye movements: experience counts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2011), pp. 1151–1160.

[71] Ho, T. K. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence 20*, 8 (1998), 832–844.

[72] Holte, R. Summary of the workshop, 2003.

[73] Howard, R. W. Learning curves in highly skilled chess players: a test of the generality of the power law of practice. *Acta psychologica 151* (2014), 16–23.

[74] Hu, W., Hu, W., and Maybank, S. Adaboost-based algorithm for network intrusion detection. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 38*, 2 (2008), 577–583.

[75] Hurst, A., Mankoff, J., and Hudson, S. E. Understanding pointing problems in real world computing environments. In *Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility* (2008), pp. 43–50.

[76] Hwang, F., Keates, S., Langdon, P., and Clarkson, J. Mouse movements of motion-impaired users: a submovement analysis. *ACM SIGACCESS Accessibility and Computing*, 77-78 (2003), 102–109.

[77] Inkpen, K. M. Drag-and-drop versus point-and-click mouse interaction styles for children. *ACM Transactions on Computer-Human Interaction (TOCHI) 8*, 1 (2001), 1–33.

[78] Japkowicz, N. The class imbalance problem: Significance and strategies. In *Proc. of the Int'l Conf. on Artificial Intelligence* (2000).

[79] Japkowicz, N. Class imbalances: are we focusing on the right issue. In *Workshop on Learning from Imbalanced Data Sets II* (2003), vol. 1723, p. 63.

[80] Japkowicz, N., and Stephen, S. The class imbalance problem: A systematic study. *Intelligent data analysis 6*, 5 (2002), 429–449.

[81] Joiner, R., Messer, D., Light, P., and Littleton, K. It is best to point for young children: A comparison of children's pointing and dragging. *Computers in Human Behavior 14*, 3 (1998), 513–529.

[82] Kambil, A., Eselius, E. D., and Monteiro, K. A. Fast venturing: the

quick way to start web businesses. *MIT Sloan Management Review 41*, 4 (2000), 55.

[83] Keates, S., Hwang, F., Langdon, P., Clarkson, P. J., and Robinson, P. Cursor measures for motion-impaired computer users. In *Proceedings of the fifth international ACM conference on Assistive technologies* (2002), pp. 135–142.

[84] Khan, S. A., and Rana, Z. A. Evaluating performance of software defect prediction models using area under precision-recall curve (auc-pr). In *2019 2nd International Conference on Advancements in Computational Sciences (ICACS)* (2019), IEEE, pp. 1–6.

[85] Khoshgoftaar, T. M., Golawala, M., and Van Hulse, J. An empirical study of learning from imbalanced data using random forest. In *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)* (2007), vol. 2, IEEE, pp. 310–317.

[86] Koenker, R., and Bassett Jr, G. Regression quantiles. *Econometrica: journal of the Econometric Society* (1978), 33–50.

[87] Kohavi, R., Sommerfield, D., and Dougherty, J. Data mining using a machine learning library in c++. *International Journal on Artificial Intelligence Tools 6*, 04 (1997), 537–566.

[88] Korting, T. S. C4. 5 algorithm and multivariate decision trees.

[89] Kuhtz-Buschbeck, J., Stolze, H., Jöhnk, K., Boczek-Funcke, A., and Illert, M. Development of prehension movements in children: a kinematic study. *Experimental Brain Research 122*, 4 (1998), 424–432.

[90] Kutner, M. H., Nachtsheim, C., Neter, J., and Li, W. *Applied linear statistical models*. McGraw-Hill Irwin, 2005.

[91] Le Cessie, S., and Van Houwelingen, J. C. Ridge estimators in logistic regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 41*, 1 (1992), 191–201.

[92] Lee, M.-C. Understanding the behavioural intention to play online games. *Online information review* (2009).

[93] Lenhard, A., and Hoffmann, J. Constant error in aiming movements without visual feedback is higher in the preferred hand. *Laterality 12*, 3 (2007), 227–238.

[94] Leshem, G. Improvement of adaboost algorithm by using random forests as weak learner and using this algorithm as statistics machine learning for traffic flow prediction. research proposal for a ph. d. *Research proposal for a Ph. D. thesis, the Hebrew university of Jerusalem* (2005).

[95] Leshem, G. *Traffic Flow Prediction and Minimization of Traffic Congestion Using Adaboost-Random Forests Algorithm*. Hebrew University of Jerusalem,

2007.

[96] Ling, C. X., Huang, J., and Zhang, H. Auc: a better measure than accuracy in comparing learning algorithms. In *Conference of the canadian society for computational studies of intelligence* (2003), Springer, pp. 329–341.

[97] Ling, C. X., and Li, C. Data mining for direct marketing: Problems and solutions. In *Kdd* (1998), vol. 98, pp. 73–79.

[98] Liu, C., and Arnett, K. P. Exploring the factors associated with web site success in the context of electronic commerce. *Information & management 38*, 1 (2000), 23–33.

[99] Liu, W., and Chawla, S. A quadratic mean based supervised learning model for managing data skewness. In *Proceedings of the 2011 SIAM International Conference on Data Mining* (2011), SIAM, pp. 188–198.

[100] Lv, Y., Shi, X., Ran, L., and Shang, M. Random forest-based ensemble estimator for concrete compressive strength prediction via adaboost method. In *The International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery* (2019), Springer, pp. 557–565.

[101] Lynch, P. D., and Beck, J. C. Profiles of internet buyers in 20 countries: Evidence for region-specific strategies. *Journal of International Business Studies 32*, 4 (2001), 725–748.

[102] MacKenzie, I. S., Sellen, A., and Buxton, W. A. A comparison of input devices in element pointing and dragging tasks. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (1991), pp. 161–166.

[103] Maclin, R., and Opitz, D. An empirical evaluation of bagging and boosting. *AAAI/IAAI 1997* (1997), 546–551.

[104] Metter, E. J., Conwit, R., Tobin, J., and Fozard, J. L. Age-associated loss of power and strength in the upper extremities in women and men. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences 52*, 5 (1997), B267–B276.

[105] Mieschke, P. E., Elliott, D., Helsen, W. F., Carson, R. G., and Coull, J. A. Manual asymmetries in the preparation and control of goal-directed movements. *Brain and cognition 45*, 1 (2001), 129–140.

[106] Morrell, R. W. *Older adults, health information, and the World Wide Web.* Psychology Press, 2001.

[107] Muntean, M., Vălean, H., Ileană, I., and Rotar, C. Improving classification with support vector machine for unbalanced data. In *2010 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR)* (2010), vol. 3, IEEE, pp. 1–6.

[108] Nap, H., De Greef, H., and Bouwhuis, D. Access for all by cognitive

engineering. *Gerontechnology 3*, 4 (2005), 258.

[109] Neely, K., Binsted, G., and Heath, M. Manual asymmetries in bimanual reaching: The influence of spatial compatibility and visuospatial attention. *Brain and Cognition 57*, 1 (2005), 102–105.

[110] Neerincx, M., Lindenberg, J., Rypkema, J., and Van Besouw, S. A practical cognitive theory of web-navigation: Explaining age-related performance differences. In *Position Paper CHI 2000 Workshop Basic Research Symposium* (2000).

[111] Ngo, T. Data mining: practical machine learning tools and technique, by ian h. witten, eibe frank, mark a. hell. *ACM SIGSOFT Software Engineering Notes 36*, 5 (2011), 51–52.

[112] Oyewole, S. A., and Haight, J. M. Determination of optimal paths to task goals using expert system based on goms model. *Computers in Human Behavior 27*, 2 (2011), 823–833.

[113] Pariente-Martinez, B., Gonzalez-Rodriguez, M., Fernandez-Lanvin, D., and De Andres-Suarez, J. Measuring the role of age in user performance during interaction with computers. *Universal Access in the Information Society 15*, 2 (2016), 237–247.

[114] Peters, M., and Ivanoff, J. Performance asymmetries in computer mouse control of right-handers, and left-handers with left-and right-handed mouse experience. *Journal of Motor Behavior 31*, 1 (1999), 86–94.

[115] Phillips, J. G., and Triggs, T. J. Characteristics of cursor trajectories controlled by the computer mouse. *Ergonomics 44*, 5 (2001), 527–536.

[116] Platt, J. Fast training of support vector machines using sequential minimal optimization. advances in kernel methods—support vector learning (pp. 185–208). *AJ, MIT Press, Cambridge, MA* (1999).

[117] Provost, F., Fawcett, T., and Kohavi, R. The case against accuracy estimation while comparing induction algorithms. In *ICML Conference* (1998).

[118] Quinlan, J. R. Induction of decision trees. *Machine learning 1*, 1 (1986), 81–106.

[119] Quinlan, J. R. Improved use of continuous attributes in c4. 5. *Journal of artificial intelligence research 4* (1996), 77–90.

[120] Ramprasath, L., and Singh, K. Statistical options: Crash resistant financial contracts based on robust estimation. *Statistics & probability letters 77*, 2 (2007), 196–203.

[121] Raymond, M., Pontier, D., Dufour, A.-B., and Moller, A. P. Frequency-dependent maintenance of left handedness in humans. *Proceedings of the Royal Society of London B: Biological Sciences 263*, 1377 (1996), 1627–1633.

[122] Rijsbergen, V., and Joost, C. K. Information retrieval butterworths london.

[123] Rodrigues, P. C., Barbosa, R., Carita, A. I., Barreiros, J., and Vasconcelos, O. Stimulus velocity effect in a complex interceptive task in right-and left-handers. *European Journal of Sport Science 12*, 2 (2012), 130–138.

[124] Rohan, T. I., Siddik, A. B., Islam, M., Yusuf, M. S. U., et al. A precise breast cancer detection approach using ensemble of random forest with adaboost. In *2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)* (2019), IEEE, pp. 1–4.

[125] Rohr, L. E. Gender-specific movement strategies using a computer-pointing task. *Journal of motor behavior 38*, 6 (2006), 431–137.

[126] Rosati, L. How to design interfaces for choice: Hick-hyman law and classification for information architecture. In *Proceedings of the International UDC Seminar* (2013), pp. 121–134.

[127] Salthouse, T. A. Effects of age and skill in typing. *Journal of Experimental Psychology: General 113*, 3 (1984), 345.

[128] Salzberg, S. L. C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993, 1994.

[129] Sayers, H. Desktop virtual environments: a study of navigation and age. *Interacting with Computers 16*, 5 (2004), 939–956.

[130] Schapire, R. E. The strength of weak learnability. *Machine learning 5*, 2 (1990), 197–227.

[131] Schneider, D. W., and Anderson, J. R. A memory-based model of hick's law. *Cognitive psychology 62*, 3 (2011), 193–222.

[132] Scholkopf, B., Burges, C., and Smola, A. Advances in kernel methods: Support vector learning cambridge.

[133] Schrepp, M. Goms analysis as a tool to investigate the usability of web units for disabled users. *Universal Access in the Information Society 9*, 1 (2010), 77–86.

[134] Sebora, T. C., Lee, S. M., and Sukasame, N. Critical success factors for e-commerce entrepreneurship: an empirical study of thailand. *Small Business Economics 32*, 3 (2009), 303–316.

[135] Slegers, K., Van Boxtel, M. P., and Jolles, J. Computer use in older adults: determinants and the relationship with cognitive change over a 6 year episode. *Computers in Human Behavior 28*, 1 (2012), 1–10.

[136] Stone, M. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological) 36*, 2 (1974), 111–133.

[137] Strom, P. S., Strom, R. D., Wingate, J. J., Kraska, M. F., and Beck-

ert, T. E. Cyberbullying: Assessment of student experience for continuous improvement planning. *NASSP Bulletin 96*, 2 (2012), 137–153.

[138] Stubbs, N. B., Fernandez, J. E., and Glenn, W. M. Normative data on joint ranges of motion of 25-to 54-year-old males. *International Journal of Industrial Ergonomics 12*, 4 (1993), 265–272.

[139] Tan, D. S., Czerwinski, M., and Robertson, G. Women go with the (optical) flow. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (2003), ACM, pp. 209–215.

[140] Tang, Y., Zhang, Y.-Q., Chawla, N. V., and Krasser, S. Svms modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 39*, 1 (2008), 281–288.

[141] Tetko, I. V., Livingstone, D. J., and Luik, A. I. Neural network studies. 1. comparison of overfitting and overtraining. *Journal of chemical information and computer sciences 35*, 5 (1995), 826–833.

[142] Thomas, G. D. Machine learning research: Four current directions. *Artificial Intelligence, Magazine 18*, 4 (1997), 97–136.

[143] Tullis, T. Older adults and the web: lessons learned from eye-tracking. *Universal Access in Human Computer Interaction. Coping with Diversity* (2007), 1030–1039.

[144] Turban, E., King, D., Lee, J., and Viehland, D. Electronic commerce: A managerial perspective 2002. *Prentice Hall: ISBN 0 13*, 975285 (2002), 4.

[145] Velay, J.-L., and Benoit-Dubrocard, S. Hemispheric asymmetry and interhemispheric transferin reaching programming. *Neuropsychologia 37*, 8 (1999), 895–903.

[146] Vines, J., Pritchard, G., Wright, P., Olivier, P., and Brittain, K. An age-old problem: Examining the discourses of ageing in hci and strategies for future research. *ACM Transactions on Computer-Human Interaction (TOCHI) 22*, 1 (2015), 2.

[147] Viola, P., and Jones, M. Fast and robust classification using asymmetric adaboost and a detector cascade. In *Advances in neural information processing systems* (2002), pp. 1311–1318.

[148] Wahlström, J., Svensson, J., Hagberg, M., and Johnson, P. W. Differences between work methods and gender in computer mouse use. *Scandinavian journal of work, environment & health* (2000), 390–397.

[149] Walker, N., Philbin, D. A., and Fisk, A. D. Age-related differences in movement control: Adjusting submovement structure to optimize performance. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences 52*, 1 (1997), P40–P53.

[150] Wang, J., Liu, D., Jusoff, K., Zafar, N. A., Memon, M. S., Zhou, J., Chaudhuri, S. B., Mahadevan, V., Ferdinando, H., Kang, J., et al. 2010 the 2nd international conference on computer and automation engineering, iccae 2010: Welcome. In *2010 The 2nd International Conference on Computer and Automation Engineering, ICCAE 2010* (2010), vol. 2, p. 5451830.

[151] Weiser, E. B. Gender differences in internet use patterns and internet application preferences: A two-sample comparison. *Cyberpsychology and behavior 3*, 2 (2000), 167–178.

[152] White, H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society* (1980), 817–838.

[153] Whittaker, J., Whitehead, C., and Somers, M. The neglog transformation and quantile regression for the analysis of a large credit scoring database. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 54*, 5 (2005), 863–878.

[154] Wolpert, D. H. Stacked generalization. *Neural networks 5*, 2 (1992), 241–259.

[155] Wu, M., and Zuo, Y. Trimmed and winsorized means based on a scaled deviation. *Journal of Statistical Planning and Inference 139*, 2 (2009), 350–365.

[156] Xiang, L., and Xiaoli, C. The research on performance of automobile human-machine interface based on bhr-goms behavior model. In *2010 IEEE International Conference on Intelligent Computing and Intelligent Systems* (2010), vol. 2, IEEE, pp. 174–178.

[157] Xie, B. Older adults, computers, and the internet: Future directions. *Gerontechnology 2*, 4 (2003), 289–305.

[158] Xu, H., Zou, X., and Wang, H. Consumers' attitudes of e-commerce in china. *Issues in Information Systems 7*, 2 (2006), 202.

[159] Xu, Q.-S., and Liang, Y.-Z. Monte carlo cross validation. *Chemometrics and Intelligent Laboratory Systems 56*, 1 (2001), 1–11.

[160] Yadav, S., and Shukla, S. Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In *2016 IEEE 6th International conference on advanced computing (IACC)* (2016), IEEE, pp. 78–83.

[161] Zajicek, M. Designing for dynamic diversity-interfaces for older people.

[162] Zhai, S. Characterizing computer input with fitts' law parameters—the information and non-information aspects of pointing. *International Journal of Human-Computer Studies 61*, 6 (2004), 791–809.

[163] Zhang, G., and Fang, B. Logitboost classifier for discriminating thermophilic and mesophilic proteins. *Journal of biotechnology 127*, 3 (2007), 417–424.

[164] Zhang, Y., and Yang, Y. Cross-validation for selecting a model selection procedure. *Journal of Econometrics 187*, 1 (2015), 95–112.

[165] Zhang, Z., and Xie, X. Research on adaboost. m1 with random forest. In *2010 2nd International Conference on Computer Engineering and Technology* (2010), vol. 1, IEEE, pp. V1–647.

[166] Zhou, W., and Duan, W. Online user reviews, product variety, and the long tail: An empirical investigation on online software downloads. *Electronic Commerce Research and Applications 11*, 3 (2012), 275–289.

[167] Ziolko, B., Manandhar, S., Wilson, R. C., and Ziolko, M. Logitboost weka classifier speech segmentation. In *2008 IEEE International Conference on Multimedia and Expo* (2008), IEEE, pp. 1297–1300.