Towards a Framework for Usability Testing of Interactive Touchless Applications

SAŠA KUHAR and KRISTJAN KOŠIČ, University of Maribor

Interaction with user interfaces only with usage of hands and bodies was a few years back still science fiction, but now present reality. Touchless interfaces are slowly becoming mainstream and therefore it is of crucial importance to address them in a concise and standardized way. Usability is an important factor in all software quality models and a key factor during development of interactive applications. The objective of this paper is to address usability challenges we are facing during touchless application design. A conceptual usability case study is proposed with new human-computer interaction factors in mind. Factors like efficiency, ease-of-use, pleasure, fatigue, naturalness, smoothness, responsiveness and accuracy are identified and related to usability scenarios.

Categories and Subject Descriptors: H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Evaluation/ methodology*; H.1.2 [**Models and Principles**]: User/Machine Systems—*Human Information Processing*

General Terms: Human factors, Design, Measurement

Additional Key Words and Phrases: User experience, UX Metrics, Touchless User Interfaces, Human-Computer Interaction, ADORA

1. INTRODUCTION

ISO 9241-210 [2010] defines User Experience (UX) as ”a person's perceptions and responses, that result from the use or anticipated use of a product, system or service”. It does not include merely the interaction that happens during the use, but all the users' emotions, beliefs, preferences, perceptions, physical and psychological responses, behaviours and accomplishments that occur before, during and after use. Tullis and Albert [2013] extend the definition by adding another characteristic that is: ”the users' experience is of interest, and observable or measurable”. So according to that definition it is not enough for the user to interact with an interface, but this action has to be measured in some way.

In this paper we elaborate on usability issues with touchless interfaces and review evaluation scenarios as well as UX metrics. We describe the characteristics of touchless UIs in section 3 and introduce a case study on measuring usability of gesture interfaces in section 4. Finally, section 5 presents conclusions and discusses future directions of this work.

2. USER EXPERIENCE AND USABILITY

UX must not be confused with Usability as usability is a narrower term. Usability is considered the ability of the user to use the thing to carry out a task successfully: “The extent to which a product can be used by specified users to achieve specified goals with **effectiveness, efficiency** and **satisfaction** in a specified context of use” [9241-11 1998]. UX looks at the individual's entire interaction with the

Author's address: S. Kuhar, Faculty of electrical engineering and computer science, University of Maribor, Smetanova ul. 17, 2000 Maribor, Slovenia; email: sasa.kuhar@um.si; K. Košič, Faculty of electrical engineering and computer science, University of Maribor, Smetanova ul. 17, 2000 Maribor, Slovenia; email: kristjan.kosic@um.si;

In: Z. Budimac, T. Galinac Grbac (eds.): Proceedings of the 3rd Workshop on Software Quality, Analysis, Monitoring, Improvement, and Applications (SQAMIA), Lovran, Croatia, 19.-22.9.2014, published at http://ceur-ws.org.

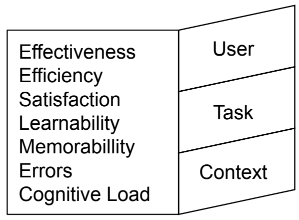


Fig. 1. PACMAD Usability model [Harrison et al. 2013]

thing, as well as thoughts, feelings, and perceptions, that result from that interaction [Tullis and Albert 2013].

When discussing usability, it is important to distinguish between summative and formative usability evaluation. Formative usability has strong ties to the practice of iterative design — building something, checking to see where it could be improved, improving it, and trying again. Summative evaluations emphasize the importance of effectiveness and efficiency in the context of use and the subjective metric of satisfaction [Lewis 2014]. The main goal of summative evaluation study is to evaluate whether people can use a product for its intended purpose effectively, efficiently, and with a feeling of satisfaction. The formative evaluation study reveals the presence of usability when there is absence of usability problems.

Nielsen [1993] defined five attributes of usability: (i) efficiency (relative to the accuracy and completeness with which users achieve goals), (ii) satisfaction (freedom from discomfort and positive attitudes towards the use of the product), (iii) learnability (the system should be easy to learn so that the user can rapidly start getting work done with the system), (iv) memorability (the system should be easy to remember so that the casual user is able to return to the system after some period of not having used it without having to learn everything all over again) and (v) errors (the system should have a low error rate, so that users make few errors during the use of the system and that if they do make errors they can easily recover from them).

Harrison et al. [2013] combined the attributes from ISO 9241-11 [1998] and Nielsen [1993] and added another interesting factor, that is **cognitive load**. Cognitive load refers to the amount of cognitive processing required by the user to use the application. Harrison's PACMAD model (figure 1) derives from mobile devices, where he argues that users of mobile applications may be performing additional tasks, such as walking, while using the mobile device. In our opinion the model could be well applied to touchless user UIs as those require from a user to combine usual interaction tasks with speaking and moving. In touchless UIs user needs to use predefined gestures and voice control simultaneously while executing wanted tasks and this can raise the mental load. Therefore it is important that the aspect of cognitive load is included in usability studies of touchless interfaces. More about touchless UIs is described in section 3.

All previously mentioned models recognise three factors that can affect usability of an application. These are: (i) user (users physical limitations, their knowledge and previous experience), (ii) task (goal the user is trying to accomplish) and (iii) context of use (environment in which the user will use the application including physical location and interaction with other people and objects). All of these are eligible for usability of touchless UIs.

2.1 How to evaluate UX

Use experience has to be measured in order to evaluate whether it is good or bad, to discover problems and opportunities for improvement, or to compare different UIs. Many metrics are available for the

Table I. Eleven common usability scenarios and the metrics that may be most appropriate for each.

Usability Study Scenario	**Task Success**	**Task Time**	**Errors**	**Efficiency**	**Learn-ability**	**Issues-based Metrics**	**Self-reported Metrics**	**Behavioral & Physiological Metrics**	**Combined & Comparative Metrics**	**Live Website Metrics**	**Card-Sorting Data**
1. Completing a transaction	x			x		x	x			x	
2. Comparing Products	x			x			x		x		
3. Evaluating frequent use of the same product	x	x		x	x		x				
4. Evaluating navigation and/or information architecture	x		x	x							x
5. Increasing awareness							x	x		x	
6. Problem discovery						x	x				
7. Maximizing usability for a critical product	x		x	x							
8. Creating an overall positive experience							x	x			
9. Evaluating the impact of subtle changes										x	
10. Comparing alternative designs	x	x				x	x		x		
11. Cognitive load	x	x					x	x			

The table is adapted from [Tullis and Albert 2013] and is supplemented with cognitive load scenario.

evaluation of UX. Metrics add structure to the design and evaluation process, give insight into the findings and provide information to the decision makers, they offer a way to estimate the number of users likely to experience a problem, are a key ingredient in calculation of ROI as well as can reveal patterns that are difficult or even impossible to see. For that UX metrics need to be observable (directly or indirectly), quantifiable (turned into a number or counted in some way), and they have to measure some aspect of the user experience [Tullis and Albert 2013].

Many different metrics are available for different evaluations. When evaluating UX one has to consider the goals of the study, the technology that is available to collect the data, and budget as well as time that are available for conducting evaluation. In table I we present ten common usability scenarios and the metrics that may be most appropriate for each, adapted from [Tullis and Albert 2013]. We added 11^{th} scenario: cognitive load, as we believe it is important for evaluation of touchless UIs.

For details about the scenarios and metrics see [Tullis and Albert 2013]. At this point we describe only the newly added scenario.

2.2 Cognitive Load measurement for evaluation of UX

Cognitive load theory (CLT) describes the relationship between the capacity of working memory and the cognitive demands of a particular task [Anderson 2012]. It is based on an idea, that cognitive capacity in working memory is limited and that if a mental task requires too much capacity knowledge acquisition and reasoning will be hindered [Jong 2009]. One solution to this is to design a UX that optimizes the use of working memory capacity and avoids cognitive overload.

CLT distinguishes between three types of cognitive load: (i) intrinsic (the inherent difficulty of the problem at hand), (ii) extraneous (generated by the representation of the content presented to the user for interpretation and action) and (iii) germane (imposed by learning a new task) [Anderson 2012]. Extraneous, intrinsic and germane cognitive load are modelled to be additive: a reduction of extraneous cognitive load frees working memory capacity that can be used for germane learning processes [Hollender et al. 2010].

The amount of extraneous load due to software use is influenced by the complexity of the software, a suboptimal software design according to traditional usability goals, and the expertise of the learner with regard to the use of the software. Load can be lowered by designing highly usable software applications and by training learners to use the software [Hollender et al. 2010].

Cognitive load can be measured in different ways. Most common measurements include task completion time and accuracy, NASA-TLX test (a survey with subjective responses), EEG-based measurement (determining cognitive load magnitude by analysing the temporal, spectral, and spatial patterns of brain activity), pupil dilation, eye tracking and blinking measurement, galvanic skin response, and heat flux [Anderson 2012; Chen et al. 2011; Haapalainen et al. 2010].

3. TOUCHLESS USER INTERFACES

Touchless user interfaces require devices, that can either execute or sense interactive behaviour, where the interaction happens without mechanical contact between the human and any part of artificial system. Touchless interaction can be multimodal, in which case the interactive behaviour produces simultaneous events in the visual modality (colour, form, or position change), in the auditory modality (speech, sounds), or in the olfactory modality (odors) [de la Barré et al. 2009]. Voice control can be realized by recording the voice with a microphone and processing it through dedicated algorithms. Body gestures can be detected in different ways from using wearable sensors to environmental sensors [Jalaliniya et al. 2013].

Touchless UIs are suppose to remove the burden of physical contact with an interactive system and make interaction pleasurable [de la Barré et al. 2009]. To achieve that one must carefully combine characteristics of physical and digital world and must try to produce a solution with "natural" Human-Computer Interaction (HCI). One must also consider the fact that hand movement is not equal to gesture, as gesture is a body movement which is being performed with the perceivable intention to express something [de la Barré et al. 2009]. For the user to learn all the gestures and voice controls, that are implemented in a certain UI, a built-in tutorial is a welcomed feature in touchless UIs.

4. CASE STUDY DESIGN: MEASURING USABILITY OF GESTURE INTERFACES DURING SURGERY WITH ADORA

According to Madan and Dubey [2012] usability is the most widely used concept in software engineering field and defines the software system's demand and use. Demand for software quality and usability is increasing and there are several usability models, that can be used and tailored to our needs (for more on usability models see Madan and Dubey [2012]).

4.1 Usability and gesture interfaces

The specifics of HCI need additional factors that will help us to successfully implement a usability study with gesture interaction in mind. Touchless HCI software interacts with users by using gestures and voice commands that are tied to gesture recognition engine. Factors like efficiency, ease-of-use, pleasure, fatigue, naturalness, smoothness, responsiveness, and accuracy should be investigated in detail. Investigation should be based through simple and complex tasks. According to Farhadi-Niaki et al. [2013] gesture based systems cause more fatigue and appear less natural then finger gestures, however factors such as time, overall satisfaction, and easiness were not affected.

Gesture recognition is a rather new field of HCI, so it is normal that problems are present. The main issues are usually: (i) lack of standardization gestures, (ii) lack of cues, (iii) inability to discover operations and (iv) requiring memorization of the player (head memory). These problems were classified by

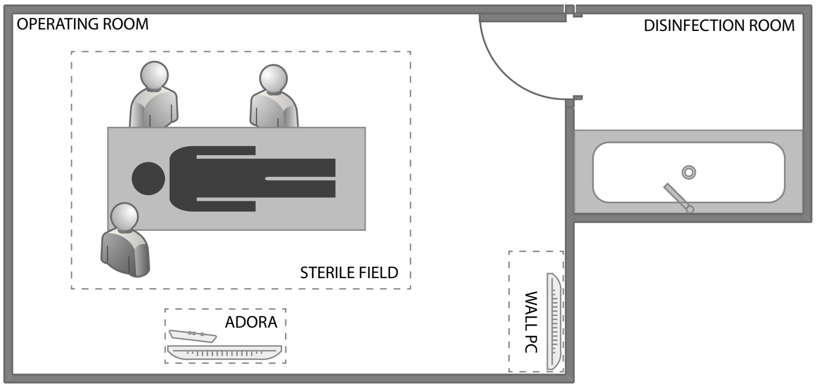


Fig. 2. Touch-less surgery with ADORA. Placement in an operating room. Problem of accessing data and renewed disinfection.

Norman et al. [2010] where he also argued that gesture based interfaces are a **step backwards in usability** due to lack of:

(1) **Visibility** - available gestures at each moment are not clear, neither speeds or precision of the movement.
(2) **Feedback** - there is not enough feedback for the user during the gesture. The user does not know if the action was because of the correct gesture or something failed in the detection.
(3) **Consistency and Standards** - there are no standards that would define consistency of HCI menus or standardized gestures.
(4) **Discoverability** - the user must know all the gestures in advance (head memory), or there is no feedback that would help him to connect gestures with application outcomes.
(5) **Reliability** -

> "...Accidental gesture activation is common in gestural interfaces, as users happen to touch something they didn't mean to touch. Conversely, frequently users do intend to touch a control or issue a gestural command but nothing happens because their touch or gesture was a little bit off. Since gestures are invisible, users often do not know that they made these mistakes" [Norman et al. 2010].

This is especially the case with Microsoft Kinect camera.

Therefore all the above mentioned issues need to be properly addressed in UX evaluation after they have been carefully studied, planned, and implemented in a system.

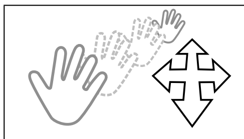
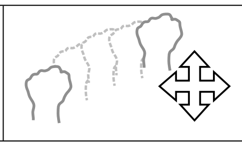
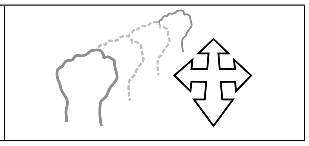
4.2 ADORA solution

ADORA[1] is an interactive physician's assistant, that enables presentation of information about a patient before and during surgical procedures. It offers a comprehensive and integrated natural user interface experience for physicians. With its use of contact-free interaction it shortens the duration of surgeries and indirectly affects the environmental and economic aspects of healthcare.

Touchless methods of HCI (gestures and voice support) have been integrated into ADORA solution during development. Before using touchless assistant physicians had to leave the patient and the sterile field in order to access critical patient data, that was accessible through the wall PC. In order to

[1] http://www.adora-med.com.

Table II. Most common used gestures in ADORA

Push	**Grab and move**	**Grab and move in depth**
		

continue with the surgery they had to disinfect, change gloves, gown and mask), memorize the picture and then return to the patient (see figure 2). This takes time and causes additional stress for the doctor. ADORA delivers patient data in a touchless manner, so that the surgeon can stay by the patient while viewing the data. Touchless interaction was developed with the help of a Microsoft Kinect sensor.

The solution was designed together with the users, physicians. One of the main requests was a minimal set of gestures, so that they can use the solution in the most possible natural way. As the surgeon has always active hands during surgery, all the gestures and their sets are designed in a one-handed way. Table II shows most common used gestures in ADORA.

4.3 Experiment design and procedure

The most important thing during planning a usability study is understanding users and the goals they are trying to accomplish. The user and his expectations from the solution define core building blocks of the study. In our case study a user is a physician in an operating room who is performing long complex surgeries. His personal goal is to successfully complete a surgery with as little distraction possible.

ADORA solution is already implemented and finalized, therefore a summative usability study is in order. Tullis and Albert [2013] define summative usability as answers to the following questions: (i) Did we meet the usability goals of the project?, (ii) What is the overall usability of the product?, (iii) How does our product compare against the competition? and (iv) Have we made improvements from one product release to the next?. Main goal of summative usability study is to evaluate how well a product or piece of functionality meets its objectives.

Experiment design reflects the scenarios selected from table I and the factors that are specific for HCI and gesture interfaces. The experiment will consists of three parts. Pre-test survey: to gather participant's data that will be used during the main part of experiment. Practice and test sessions: the core of the experiment will consist of tasks completion rate, possible errors and factors related to them. The final part will be executed with the post-task questionnaire.

In the main part of the evaluation, each participant will be instructed to perform a defined task using gestures and voice commands to control ADORA solution with the help of Kinect sensor. Tasks will include normal operations that surgeons are otherwise used carrying out on a wall PC.

Task were based on the Ux scenarios in table I. Scenario one, four and seven were addressed with selected tasks and post-task questionnaires. At this point cognitive load scenario is not included.

Tasks will be split into simple and advanced sets. For example a simple set of tasks would be selecting a surgery (pushing a button with gesture) and loading a set of medical images for the patient (a combined set of gestures). Advanced set of tasks will include changing a view, loading multiple image series into different views, combining the views and manipulation of medical images such as point based zoom and image adjusting (combined operation of brightness and contrast merged into one gesture). Table III list a set of tasks surgeon will be performing during experiment.

There will be two groups of participants (surgeons). First group will be educated about functionalities and gestures that are available in ADORA, while the second group will identify gestures with the help of a tutorial that is available in the solution. Group data will be used to determine if pre-education of

Table III. Taks a surgeon will be performing during second phase of study

Task	**Gesture**	**Voice**	**Level**
Login	Push button	Y	basic
Select operation	Drag to selected surgery and push	N	advanced
Change view	A combined set	Y	basic
Load a set of images	Interactive push with feedback	N	basic
Combine set of images	Advanced set	N	basic
Point zoom to a defined spot	Grab and pull/push	N	advanced
Adjust image to specified level	Grab and move	N	advanced
Lock position	Move and push	Y	basic

Advanced gestures are combined with domain knowledge that surgeons need during task execution.

the users affects the overall experience, or if the built-in tutorial is enough to gain needed operational knowledge.

During task execution data will be gathered according to Fitts' Law Test [Zhai 2004] in terms of completion time, errors and throughput. Completion time will be measured in three iterations and then average will be calculated of the multi-directional Fitt's law tasks. A post analysis of measured times will be done with Scheffe criterion for significance [Fleiss 1999]. Similar analysis will be made for throughput and error.

In the final phase, after task completion, users will be asked to complete a survey containing a Device Assessment Questionnaire suggested by ISO 9241-9 [2000]. All questions will be ranked with a 7-point Likert scale, from strongly agree to strongly disagree and will be fine-tuned with inclusion of HCI factors mentioned above. Only a combination of questionnaires, performance metrics, and inclusion of HCI factors can give a comprehensive and valid picture for several reasons: (i) users might be influenced not to honestly report their experience, (ii) a combination will give us more insight what system factors influenced noteworthy user ratings, (iii) for most of the uality of experience aspects described, there exist no valid and reliable metrics for the case of gesture interfaced systems, so a mixture can help to interpret the results better [Wechsung et al. 2012].

5. CONCLUSION

In this paper terms and definitions of Usability, User experience and Human Computer Interaction were challenged all together. There are already many models that define usability and user experience, but none of them has yet been tailored to the needs and challenges of gestural interfaces. While gestural interfaces are yet to become mainstream, a lot of them can be found in the gaming world - where touchless devices have already been well accepted.

Having a natural mapping between the body actions and the reactions on the screen gives very positive reactions and in order to keep the positive interaction flow with the users, further research is needed in the gestural interfaces domain. A standardized framework is needed, that will define consistency and standards for evaluation of gestural interfaces and that will include the aspect of cognitive load.

REFERENCES

ISO 9241-11. 1998. ISO 9241-11:1998 - Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability. (1998). http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=16883

ISO 9241-210. 2010. ISO 9241-210:2010 - Ergonomics of human-system interaction – Part 210: Human-centred design for interactive systems. (2010). http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=52075

ISO 9241-9. 2000. ISO 9241-09:2000 - Ergonomic requirements for office work with visual display terminals (VDTs) – Part 9: Requirements for non-keyboard input devices. (2000). http://www.iso.org/iso/catalogue_detail.htm?csnumber=30030

Erik W. Anderson. 2012. Evaluating visualization using cognitive measures. In *Proceedings of the 2012 BELIV Workshop on Beyond Time and Errors - Novel Evaluation Methods for Visualization - BELIV '12*. ACM Press, New York, New York, USA, 1–4. DOI:http://dx.doi.org/10.1145/2442576.2442581

Siyuan Chen, Julien Epps, and Fang Chen. 2011. A comparison of four methods for cognitive load measurement. In *Proceedings of the 23rd Australian Computer-Human Interaction Conference on - OzCHI '11*. ACM Press, New York, New York, USA, 76–79. DOI:http://dx.doi.org/10.1145/2071536.2071547

R de la Barré, Paul Chojecki, and Ulrich Leiner. 2009. Touchless interaction-novel chances and challenges. In *Proceedings of the 13th International Conference on Human-Computer Interaction*, Julie A. Jacko (Ed.). Springer-Verlag Berlin, Heidelberg, 161–169. DOI:http://dx.doi.org/10.1007/978-3-642-02577-8_18

Farzin Farhadi-Niaki, S.Ali Etemad, and Ali Arya. 2013. Design and Usability Analysis of Gesture-Based Control for Common Desktop Tasks. In *Human-Computer Interaction. Interaction Modalities and Techniques*, Masaaki Kurosu (Ed.). Lecture Notes in Computer Science, Vol. 8007. Springer Berlin Heidelberg, 215–224. DOI:http://dx.doi.org/10.1007/978-3-642-39330-3_23

Joseph L. Fleiss. 1999. *The Parallel Groups Design*. John Wiley & Sons, Inc., 46–90. DOI:http://dx.doi.org/10.1002/9781118032923.ch3

Eija Haapalainen, SeungJun Kim, Jodi F. Forlizzi, and Anind K. Dey. 2010. Psycho-physiological measures for assessing cognitive load. In *Proceedings of the 12th ACM international conference on Ubiquitous computing - Ubicomp '10*. ACM Press, New York, New York, USA, 301. DOI:http://dx.doi.org/10.1145/1864349.1864395

Rachel Harrison, Derek Flood, and David Duce. 2013. Usability of mobile applications: literature review and rationale for a new usability model. *Journal of Interaction Science* 1, 1 (Dec. 2013), 1–16. DOI:http://dx.doi.org/10.1186/2194-0827-1-1

Nina Hollender, Cristian Hofmann, Michael Deneke, and Bernhard Schmitz. 2010. Review: Integrating Cognitive Load Theory and Concepts of Human-computer Interaction. *Comput. Hum. Behav.* 26, 6 (Nov. 2010), 1278–1288. DOI:http://dx.doi.org/10.1016/j.chb.2010.05.031

Shahram Jalaliniya, Jeremiah Smith, Miguel Sousa, Lars Büthe, and Thomas Pederson. 2013. Touch-less interaction with medical images using hand & foot gestures. In *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication - UbiComp '13 Adjunct*. ACM Press, New York, New York, USA, 1265–1274. DOI:http://dx.doi.org/10.1145/2494091.2497332

Ton Jong. 2009. Cognitive load theory, educational research, and instructional design: some food for thought. *Instructional Science* 38, 2 (Aug. 2009), 105–134. DOI:http://dx.doi.org/10.1007/s11251-009-9110-0

James R. Lewis. 2014. Usability: Lessons Learned . . . and Yet to Be Learned. *International Journal of Human-Computer Interaction* 30, 9 (June 2014), 663–684. DOI:http://dx.doi.org/10.1080/10447318.2014.930311

Ankita Madan and Sanjay Kumar Dubey. 2012. Usability evaluation methods: a literature review. *International Journal of Engineering Science and . . .* 4, 02 (2012), 590–599. http://www.ijest.info/docs/IJEST12-04-02-143.pdf

Jakob Nielsen. 1993. *Usability Engineering*. Morgan Kaufmann. 362 pages. http://www.nngroup.com/books/usability-engineering/

Donald A Norman, Jakob Nielsen, and Nielsen Norman Group. 2010. Gestural Interfaces : A Step Backward In Usability. (2010).

Thomas Tullis and William Albert. 2013. Measuring the User Experience, Second Edition: Collecting, Analyzing, and Presenting Usability Metrics. (July 2013). http://dl.acm.org/citation.cfm?id=2526258

Ina Wechsung, Klaus-Peter Engelbrecht, Christine Kühnel, Sebastian Möller, and Benjamin Weiss. 2012. Measuring the Quality of Service and Quality of Experience of multimodal human–machine interaction. *Journal on Multimodal User Interfaces* 6, 1-2 (2012), 73–85. DOI:http://dx.doi.org/10.1007/s12193-011-0088-y

Shumin Zhai. 2004. Characterizing computer input with Fitts' law parameters—the information and non-information aspects of pointing. *International Journal of Human-Computer Studies* 61, 6 (Dec. 2004), 791–809. DOI:http://dx.doi.org/10.1016/j.ijhcs.2004.09.006