Empirical usability studies on user interface-modules and elements: A prerequisite of usable applications specifically tailored to different mobile devices

Manfred Tscheligi, Verena Giller, Reinhard Sefelin, Claus Lamm, Rudolf Melcher Johann Schrammel

CURE (Center for Usability Research & Engineering)

A-1110 Vienna, Austria

Tel: +43 (1)743 54 51-11, Fax: +43 (1) 743 54 51-30

Email: cure@cure.at

Abstract

Mobile devices are increasingly playing an important role in our daily private and business lives. Today they are not only tools for one to one communication but also platforms for applications, which enable users to access and to enter information from and to a broad variety of different sources. Two crucial questions need to be addressed in order to make these applications useful and usable: (1) which user interface elements are ideal for a certain task performed with a certain device, and (2) which level of detail is appropriate for which kind of device. This paper will present an approach of how to answer these questions with empirical studies. An overview of the development of a device classification, which is the basis for such studies will be presented as well. Furthermore we will discuss two example studies which will show how empirical data can inform the design of applications which are tailored to users' needs and to their devices. Finally an outline on the research issues of the future based on our analysis of the present will be given.

I. Introduction

At the time we are writing this paper we are conducting a high amount of studies whose results are essential for the development of application programming guidelines for mobile business applications. These studies are part of an EU funded project (CONSENSUS [2]) which develops a mark-up language supporting the automatic adaptation of user interfaces for mobile devices. The mark-up language development process, however, will not be discussed in this paper. This paper will show which kinds of questions can be answered by empirical studies and will present two example studies and the results, which can be deduced from them.

Every application has to be tailored to a special user group and to the tasks of these user groups. Moreover a mobile application must also be tailored to the devices with which this application is used. Whereas the internet is often used for exploration, users of mobile services generally try to accomplish specific, small tasks that are of high relevance in a commonly very specific context [6]. In general we may assume that the smaller the device the more specific will be the task. Often the same user group may use one and the same application with different devices. Depending on the devices used different levels of

interaction possibilities and different levels of detail of the information that is presented with these devices are appropriate. Therefore the development of mobile applications must ensure that the data presentation and the input mechanisms are adapted to the device in use. For these reasons a device classification clustering the available mobile devices into a manageable number of groups is necessary. Otherwise developers would have to produce an application for each device, which would lead to an incredibly high amount of sub-applications. The device classification that was developed for the conduction of our empirical studies will be discussed briefly in the second chapter.

Empirical studies, which are informing an application tailoring process, have two main goals:

1. **Definition of input mechanisms and output mechanisms, which are optimal for a certain class of devices.** Questions to be answered are e.g. which classes of devices shall display text inside a scrollable field and which classes shall enable the user to page through texts, or: which device classes shall enable users to enter search commands to select from a list and which device classes are ideal for a step wise navigation through lists.
2. **Definition of maximal detail levels for both input and output** Example questions: How long may be forms which have to be filled out with different classes of devices, or: How long may be alphabetical or numerical lists which are used to enter names or e.g. product numbers.

These two questions can be summarised as (1) how shall users interact with an application depending on the device class? and (2) what shall users be able to do with an application depending on the device class?. Chapter 3 will discuss the methodological challenges of how these two types of questions can be answered by empirical studies. Furthermore example studies will be presented. Chapter 4 of this paper will present an overview of how these studies should be continued in the future in order to gain a pool of data which is relevant for both future devices and future application domains.

II. Device classification

A crucial prerequisite of empirical studies as described above is an appropriate set of device classes. These device classes have to mirror devices, which are currently available but have to be defined in a way that further classes can be derived from them when new types of devices are entering the market.

We started the clustering process with an analysis of the devices which are currently available and which will be available in the next future. The main focus of this evaluation was on devices with different input and output modalities. We targeted mainly high-end business devices, with at least browser abilities and beyond.

The evaluation was not representative from a market point of view, but a valid summary of all different kinds of mobile devices. The analysis and subsequent classification of the devices focused on characteristics that have an influence on the (optimal) user interface and on the users' behaviour. This analysis led to three main dimensions, which are defining the eight device classes, which we finally developed. Van Welie and de Groot [5], who also developed a device classification, used a very similar approach, which was also based on the three dimensions described below.

The first and most crucial dimension is represented by the possible **Presentation Structure**, which strongly influences the features, constraints and appearances of the user interfaces. It is a combination of the display properties resolution and physical screen size.

The second dimension represents the supported **Input Modality.** Modalities to enter data are e.g. keypad, pen, voice, small keyboards. Based on the selected input modality the selection of interaction elements, their behaviour and their appearance may have to be varied.

The third dimension is the **Mark-Up Language (which determines the Widget Set)**. The utilizable ML (and the available Widgets) has a big impact on the look and feel of mobile applications. Voice applications (which are representing class 0 of our classification) are also using a kind of "widget set" which is defined by the capabilities of the mark-up language (VXML, SALT).

The following list shows examples of the eight device classes which were defined on the basis of these three dimensions: The list also includes examples of devices which are part of these classes. The empirical studies, which are discussed in the next chapter, are conducted with these devices and with one additional representative of device class 3.

Class 1: Screen size: approx. 64x96pixel, input: T9, WAP; Example: SIEMENS, ME45

Class 2: Screen size: approx. 176x220pixel, input: T9, WAP, Example: NOKIA, 7650

Class 3: Screen size: approx. 160x160/240x320pixel, input: pen, HTML; Example: COMPAQ, iPAQ Pocket PC

Class 4: Screen size: approx. 640x320pixel, input: QUERTY, HTML; Example: NOKIA, Communicator 9210i

Certainly a classification like this can never be exhaustive and without any ambiguity. So we are aware that there are devices available on the market, which do not fit into our device classification. So e.g. the Nokia 5510 has a rather small mobile-phone-like display but includes a PDA-like QUERTY-keyboard. However, we made sure that the results of our empirical studies can also be applied for such "out -siders". In these cases we can define these devices as part of two devices classes. So the input capabilities of such a device are part of device class 4 but its output capabilities are part of device class 1.

III. Empirical studies

Set up and methodological approach

In this chapter we will explain how we set up and conducted empirical studies to answer the two kinds of questions, which we described in section I. Whereas in the first case (definition of input and output mechanisms) different kinds of alternatives have to be compared in the second case (definition of maximal detail levels) users' "thresholds of pain" have to be assessed. Therefore different methodological approaches are needed which we will discuss below.

Definition of input mechanisms and output mechanisms which are optimal for a certain class of devices

To optimise both input and output mechanisms for a certain device class different alternatives have to be designed and chosen which then can be compared using empirical data.

Often two or more alternatives are obvious. So e.g. in the case of the display of longer texts there are two evident possibilities of going through the pages page by page and of scrolling down the whole text. However, there are other forms of interaction where the different alternatives are not as obvious. Often alternatives which are needed from a mobile usability point of view have not been developed yet or are not available as standard elements. In these cases alternatives have to be designed and implemented as testable prototypes. An example for such a case are tables, where solutions for WAP are still under development (see e.g. [4]) and where HTML-elements still are poorly adapted to the requirements of the different device classes.

Tests that compare alternative input and output mechanisms should at least measure three variables:

- Efficiency
- Error rate
- Subjective user satisfaction.

All of these factors are highly interrelated with the overall usage costs of an application (especially of a business application). Based on experience we know that these variables are highly correlated. If this, for some reasons, is not the case they have to be prioritised or weighted. Depending on the context other variables like trust can also play an important role and have to be considered. Take for example the input of a credit card number with a voice driven application. Again different alternatives have to be compared (feedback after each chunk of digits or feedback only after the whole number was said) and trust will be one of the major factors, which defines which alternative to recommend.

The efficiency of an element mostly is measured by time measurements of subtasks. So e.g. one subtask of the main task of ordering a product may be the selection of the product form a list. The time a user needs to operate the list and to finish the selection process then is separated from the overall task and can be measured. Errors which have to be defined clearly before the test is conducted are counted. The subjective user satisfaction usually is measured with ratings on five- or seven-point scales. An example of a test, which is comparing two different ways of user navigation through longer texts will be given below.

Definition of maximal detail levels for both input and output

Often interaction mechanisms which are optimal for a certain device are sub-optimal for another one. So e.g. users will have little difficulties to enter full product names when they can use a QUERTY or a virtual keyboard but they will struggle to perform the same task with a keypad of a mobile phone. Users of mobile applications are operating within different contexts and with devices, which are differently usable for reading, entering, searching and for “pocketing” tasks. Therefore it is necessary to define maximum complexity levels beyond which the complexity of a task performed with a certain class of devices is not reasonable anymore.

To define such levels it is necessary to confront users with different complexity levels and to define the variables which are influencing the levels of complexity. So it does not help just to know that the maximum-complexity-border lies between alternative A and B, you must also

be able to name the variable which is responsible for the difference between these two alternatives.

So e.g. when we assessed the maximum number of list entries which may be presented to users operating a voice system we confronted users with four different lengths (4, 8, 15 and 21 entries). Then we asked them to express their subjective level of effort which was necessary to do this selection. The way of how we encouraged users to express these levels is described in one of the example studies below. Another example is the assessment of forms to enter database entries. In these studies we compared the threshold of pain of users who were operating forms on different devices. We could show that users who could operate html-forms that included different kinds of elements like check boxes and radio buttons with direct manipulation using a pen had higher thresholds of pain regarding the number of possible inputs than e.g. users who had to navigate through wap-lists. From these thresholds we could derive maximum complexity levels for forms used with different device classes.

The goal of studies like these is to define clear benchmarks of maximal complexity for different device classes. As Bill Buxton put it: Developing things which are like the swiss army knife (it can do everything but nothing really well) will lead to a world full of functionality but with a big lack of usability. Each device has it advantages and drawbacks and each application must be developed in a way that it is not overloading its users. Therefore the applications have to be tailored with regard to maximum complexity levels which are proven by empirical data. How such data can be collected can be seen below in our second example study.

Example studies

Text reading

The first study was a test which compared two ways of navigation through longer texts. The first possibility forced users to go through a text page by page and to press a “next” button in order to go to the next page. This way of navigation was compared to a presentation where users had to scroll down the whole text. The test set up was designed in a way that order effects could not occur and a special index (see Dickes and Steiwer [3]) made sure that the complexity and readability as well as the number of words of both texts was equal. The number of words and characters of the texts, which we used for the tests with mobile phones, was smaller than these numbers of texts used for tests with PDAs (mobile phones: approximately 400 words, PDAs: approximately 550 words). For each device class we used two different alternative texts. 10 paid subjects (6 males and 4 females, ranging from 22-33 years of age) were recruited for the purposes of the study.

Figure 1 shows the subjects’ average reading speeds per device class and Figure 2 shows their average ratings of their reading comfort. Only the differences measured with devices of class 1 were statistically significant (t-test for paired samples, $p < .05$).

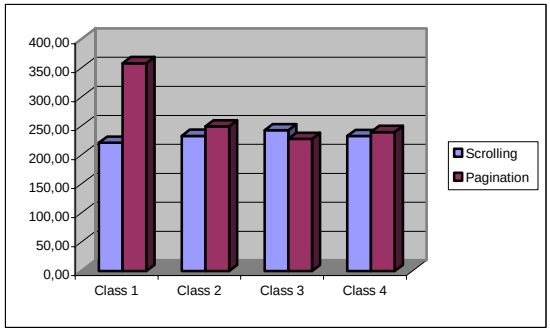

Figure 1: Average reading speed

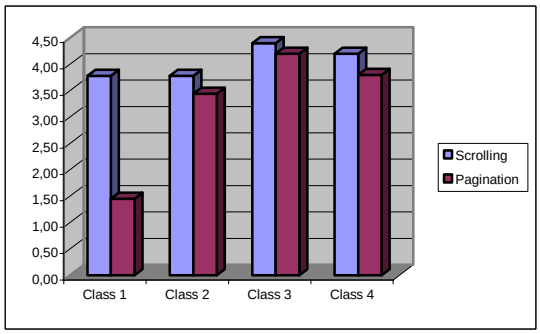

Figure 2 Average reading comfort expressed with 5 point scales (1: uncomfortable, 5 comfortable)

The results of this study lead to the conclusion that in general a designer should have good reasons to use a pagination mechanism instead of a scrolling one. This is especially true when he/she is designing for devices of class 1. The tests were conducted with typical content pages, which did not contain interaction elements. Our hypothesis was that possibly users prefer to page through the text displayed on mobile devices because the handling of elements that are supporting scrolling may be more difficult than the handling of these elements with a typical desktop PC. Another advantage of pagination could have been that scrolling easily could lead to a loss of the orientation between the lines. On the other hand the pagination mechanism provides users with pages whose content does not move and where the reading process is much more like reading a book.

However, the empirical data gathered could not prove that these two possible advantages outperform the disadvantages of long loading times and of an interaction mechanism, which is not consistent with today's web sites. That means that as long as loading times are at today's level there seem to be no clear reasons to implement pagination. Some user statements collected during our tests also support this. So a lot of users complained that they are often loosing the plot due to the long loading times. Another advantage of the scrolling concept, which was often mentioned by our subjects, was that scrolling gives them more control over the text. So one user said, "When I am scrolling I can control the speed. On the other hand the pagination thing gives me the feeling that the computer controls how fast I am allowed to read."

List lengths

In this study we confronted users with different lists showing alphabetical ordered names and asked them to select target names from these lists. The target names' positions differed between the lists. We measured the time users needed to perform these tasks and we asked users to draw lines which should represent their subjective rating of the effort which was needed in order to perform this task. So we gave them sheets where they could see the

different list lengths and which also showed a line, which was representing their threshold of pain. Users should draw the lines in such a way that their lengths expressed the closeness of this kind of selection to their personal threshold.

For this study again 10 paid subjects (8 males and 2 females, ranging from 20-34 years of age) were recruited. Again the test was set up in a way that order effects could not appear. Figure 3 shows the results of these tests. The average lengths of the lines are represented by the y-axis. Note that the line lengths are standardised by dividing the length of the line drawn by the users through the length of the difference between the given starting point of the line and the predefined line, which was marking the threshold. Therefore every line in Figure 3, which is longer than one, represents average line lengths whose lengths were above the threshold. The x-axis in Figure 3 shows the position of the target name in relation to the screen size. So 2.5 means that users had to scroll down two and a half screens in order to select the name. We could show that there is a high correlation between the time users needed to perform such a selection task and the lengths of the lines which they drew to express their subjective validation of it ($r = .53$, $p < .01$).

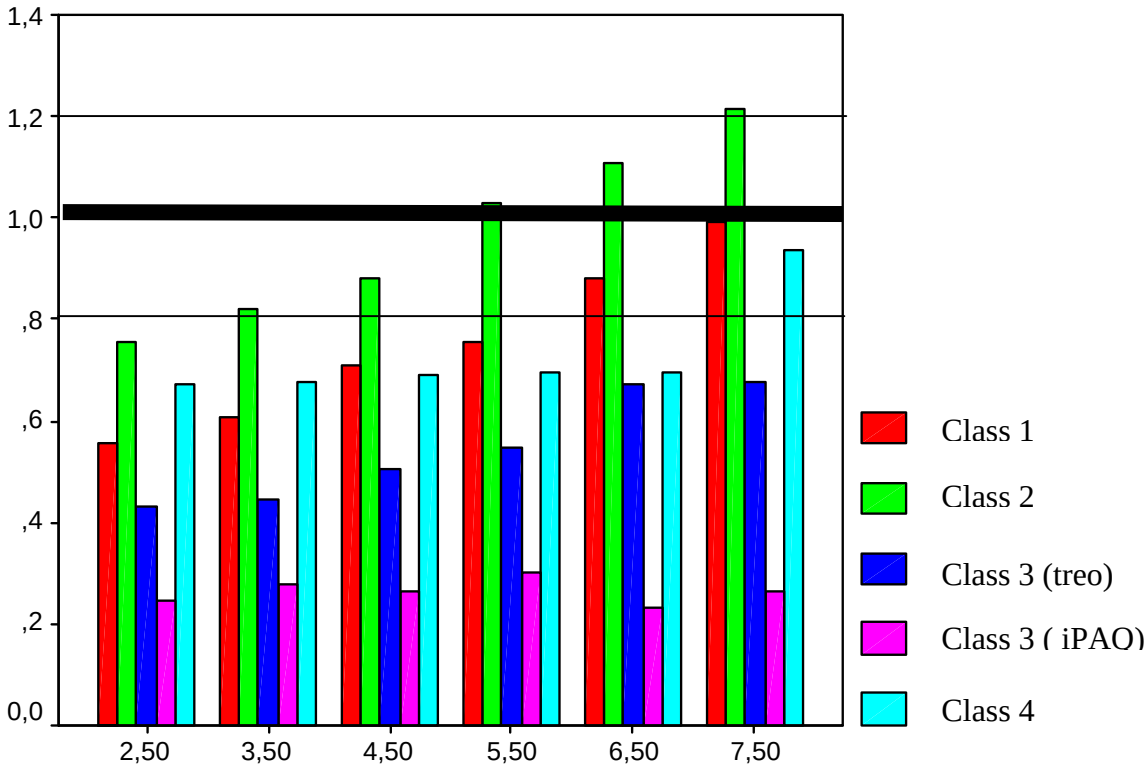

Figure 3 Average line lengths

For our tests we used two different devices of device class 3: A COMPAQ: iPAQ Pocket PC (browser type: Konqueror) and a handspring treo-Communicator (browser type: Blazer). The differences between these two devices occurred due to a different look and feel of list boxes displayed with the two browsers. The differences between device class 1 and 2 occurred because users had difficulties to operate the joystick of the representative of device class 2 in a way, which allowed them to increase the speed of scrolling.

However, for the definition of maximum complexity levels such differences due to sub-optimal interface characteristics, which are not described by the three dimensions defining the device classification (see section II), have to be neglected. So until a device classification exists which also mirrors such hardware differences, which are currently not covered by our classification, we have to mask these results. In section IV we will discuss the needs of a

further development of our classification. Therefore at the moment our recommendation for the maximal list lengths for devices of class 2 lies above the number of entries allowed for device class 1 although the results shown in Figure 3 seem to point into another direction.

Furthermore the measured differences have to be analysed with care because the threshold, which is measured in a lab environment under almost perfect conditions (office lighting, dry environment, no noise and a sound user posture) will always define the absolute upper limit. So if designers assume that their applications are used in contexts, which are not as perfect as our lab conditions, they have to subtract a context depending number of list entries from our maximum values. For these reasons and because of statistical blurring we defined the general maximum level 20% below the one, which on average was chosen by our subjects. Special context conditions, however, still may force designers to under-run this limit.

Discussion

These two examples show how empirical studies can be conducted and which data can be collected. They also show that studies like these can only answer very detailed questions and that there is an almost infinite number of further open questions. You cannot assume to cover a broader range of questions with one study because studies like these include a lot of devices and require carefully chosen set ups.

We want also to emphasise that because of differences inside the device classes the results have to be analysed very carefully. Often these differences are not obvious before the actual test has been conducted. Therefore it is important that appropriate conclusions are drawn from this data. In the last section only two of our studies were reported. The overall approach of the CONSENSUS project is to develop a mosaic of empirical tests, which are well fitting together. In synergy with the already available empirical and qualitative data they will lead to an overall picture of do´s and don't´s.

IV. Future challenges

Next generation of mobile user tasks

Users' tasks in the field of mobile communication are growing and changing fast. Some issues which will be important in the future and which need to be researched are:

- Picture communication
- Communication inside of groups and knowledge sharing
- Mobile commerce
- Context awareness
- …

This list is not meant to be exhaustive. It just shows that a lot of tasks are existing which go beyond e.g. basic list selection or text reading. In the case of mobile commerce trust issues have to be considered as well. In addition to the analysis of basic interaction techniques empirical studies have to focus also on higher-level tasks.

Empirical studies on the tasks of the mobile future could help to avoid the mistakes of the past. They are a chance to make it right the first time when it comes to the launch of mobile applications of the second or third generation.

Further user groups

The book of vision of the WWRF 2001 (Draft version) [7] states under Task 1.4 that objectives of proposed research should be to "develop methods and concepts for the design of adaptive mobile applications and services which are most appropriate for realising universal access, regarding especially the limitations of people with special needs." This definitely means that future empirical work should also include other types of user groups, rather than focusing only on mobile applications for business users.

Empirical studies are needed to evaluate upcoming new interaction paradigms for the different device classes (like higher degrees of multi-modality), which are especially supporting non-business user groups.

Furthermore the future tailoring process must also include the different user groups and not only the devices (and device classes), which are used by them. Children, users with special needs, inexperienced users etc. need different input and output mechanisms and require different maximum levels of complexity. That means that in the long run the goal must be to develop different recommendations for one and the same device class depending on the target group for which an application is developed.

Further development of device classes

In section III we saw that devices have special characteristics, which are not defined by the three dimensions, defining the device classes, which we discussed in section II. That means that users' performance with two devices of one and the same device class may differ because of hardware- or browser-specific differences, which are not part of the device class specification. Therefore future research should focus on the definition of classes, which are minimising these problems. As long as this concerns the browser capabilities the number of differences, which have to be considered, may be manageable. However, when it comes to hardware differences the number of differences is almost infinite and steadily growing.

So the challenge of the future development of device classes will be to include such differences and to define them and their effects on the efficiency and user satisfaction of each task. Furthermore new classes will have to be developed because of new products and types of interface designs that are entering the market. Therefore the three dimensions, which currently are used for the definition of device classes, will have to be evaluated on a current basis to ensure that they fulfil the requirements of a device classification including the latest developments.

Open pool of data

In this paper we tried to emphasise that empirical data is an essential prerequisite of the development and production of applications that are tailored to the devices on which they are running. To ensure that these sets of data are available to everybody who is developing mobile applications it will be necessary to establish an open data pool. This pool shall include the results as well as the detailed set up of empirical studies, which focus on the comparison of input and output mechanisms and on the definition of complexity thresholds.

So we have to facilitate the development of such an open data pool. This is an evident goal on an international level to support the development of useable applications in an organised way. The current situation at which these data are not available to a general public is sub-optimal for everybody. For the users, because their devices show applications which cannot be used with the device they hold in their hands, for the producers of mobile devices because the demand of their products partly depends on the usability and on the usefulness of the applications running on them and for the developers of mobile applications because they cannot base their development on solid knowledge to increase the quality of their work.

For the development of such a pool, however, a broad agreement of all players in the mobile field is required. An organisation like the WWRF certainly could be an ideal platform for its set up.

V. Conclusion

This paper discussed the need for empirical studies and their importance for the development of applications tailored to the devices on which they are running. In the CONSENSUS project the empirical studies are the basis for the development of application programming guidelines, which will be used to develop building blocks for an automatic adaptation process that considers usability constraints for the targeted devices. Empirical studies and the guidelines, which are based on them, however, can and should inform the development process of every application independent from how it is produced.

We showed also that a classification of devices is necessary in order to reduce the field of devices to a manageable number. Although a proper classification is a prerequisite of empirical studies and of the conclusions to which they are leading it is important to keep in mind that this classification has to be updated steadily. New devices entering the market as well as results gained by empirical studies may lead to a redefinition of the dimensions, which are describing these classes.

Empirical studies are answering two types of questions. We summarised them as (1) how shall users interact with an application depending on the device class? and (2) what shall users be able to do with an application depending on the device class? In our outlook on future challenges we argued that the answers to these two questions might be different depending on the target group of an application. Furthermore it will be necessary to broaden the research to higher-level tasks and to tasks, which are currently not supported by the available applications. Finally we recommended the development of an open data pool, which is available to the developers of mobile applications. Only such a data pool will ensure that the latest empirical results are always considered during the development process of the mobile applications of the next generation.

VI. Literature

[1] Buxton, W. (2001). Less is more (more or less). In The invisible future. (Denning, J.P. ed) pp. 145 – 180. McGraw-Hill

[2] **CONSENSUS:** IST PROGRAMM/KA4/AL:IST-2001-4.3.2/CONSENSUS/CN:IST-2001-32407 3G Mobile Context Sensitive Adaptability User Friendly Mobile Work Place for Seamless Enterprise Applications. (www.consensus-online.org)

[3] Dickes, P. and Steiwer, L. (1977). Ausarbeitung von Lesbarkeitsformeln für die deutsche Sprache. In Zeitschrift für Entwicklungspsychologie and Pädagogische Psychologie. Band IX, 1/1977, 20-28 (in German)

[4] Korte, R.P. (2001). WML-Tutorial. http://www.wml-tutorial.de/inhalt.html

[5] Van Welie, M. and de Groot, B. (2002). Consistent multi-device design using device categories. In Proceedings of the 4th International Symposium, Mobile HCI 2002, Pisa, Italy

[6] Van Welie, M. and de Ridder, G. (2001). Designing for Mobile Devices: a Context-Oriented Approach, Satama Interactive, www.satama.com

[7] Wireless World Research Forum (2001). Book of visions (draft version). http://www.wireless-world-research.org/