Kaisa Rolig

Feasibility of mobile reaction time measurement technology for neurocognitive assessment

Faculty of Information and Natural Sciences

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 20.5.2010

Thesis supervisor:

Prof. Mikko Sams

Thesis instructor:

Dr.Phil. Ville Ojanen

Aalto University
School of Science
and Technology

AALTO UNIVERSITY
SCHOOL OF SCIENCE AND TECHNOLOGY

ABSTRACT OF THE
MASTER'S THESIS

Author: Kaisa Rolig

Title: Feasibility of mobile reaction time measurement technology for neurocognitive assessment

Date: 20.5.2010 Language: English Number of pages:10+44

Faculty of Information and Natural Sciences

Department of Biomedical Engineering and Computational Science

Professorship: Cognitive neuroscience Code: S-114

Supervisor: Prof. Mikko Sams

Instructor: Dr.Phil. Ville Ojanen

Mobile phones of today are highly sophisticated small scale computers. They offer a novel cost-efficient way to measure reaction times. However, for the measurements to be feasible to neurocognitive assessment, they have to fulfill certain requirements.

Variances of mobile- and traditional computer-based measurement method were compared to study the measurement error caused by the mobile system. The repeatability of the measurements were also studied to ensure their reliability. Mobile and computer measurements were compared to determine the agreement of the two methods and lastly we conducted series of power analyses to determine the number of subjects needed to detect a statistically significant effect using either of the measurement method.

We found that the mobile measurement method does not increase variance in reaction time measurements when compared to computer measurements. They are also as reliable as the computer measurements and their statistical power is marginally stronger. However, agreement of the two methods is quite poor, which would lead to difficulties, were the two methods used interchangeably. Nevertheless, the results are rather promising and seem to confirm the feasibility of the mobile reaction time measurement method for neurocognitive assessment.

Keywords: Reaction time, neurocognitive assessment, mobile measurement method, assessing feasibility

AALTO-YLIOPISTO
TEKNILLINEN KORKEAKOULU

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Kaisa Rolig

Työn nimi: Mobiilin mittausteknologian soveltuvuus neurokognitiviiseen arviointiin

Päivämäärä: 20.5.2010 Kieli: Englanti Sivumäärä:10+44

Informaatio- ja luonnontieteiden tiedekunta

Lääketieteellisen tekniikan ja laskennallisen tieteen laitos

Professuuri: Kognitiivinen neurotiede Koodi: S-114

Valvoja: Prof. Mikko Sams

Ohjaaja: FT Ville Ojanen

Nykyiset matkapuhelimet ovat kehittyneitä, pienen mittakaavan tietokoneita. Ne tarjoavat uuden ja kustannustehokkaan tavan mitata reaktioaikoja. Niiden täytyy kuitenkin täyttää tietyt vaatimukset, jotta ne soveltuisivat neurokognitiiviseen arviointiin.

Mobiililaitteen aiheuttaman mittavirheen tutkimiseksi mobiililla ja perinteisellä tietokonepohjaisella mittaohjelmistolla tehtyjen mittausten variansseja verrattiin keskenään. Myös mittausten toistettavuuksia tarkasteltiin niiden luotettavuuden varmentamiseksi. Kahdella mittaohjelmistolla saatuja mittaustuloksia verrattiin keskenään yhtenevyyden arvioimiseksi. Lisäksi mittauksille suoritettiin tilastollinen voima-analyysi, jotta voitiin verrata tilastollisesti merkittävän efektin tunnistamiseksi vaadittavaa koehenkilöiden lukumäärää molemmilla mittaustavoilla.

Tulosten perusteella voidaan sanoa, että mobiili mittaustapa ei lisää merkittävästi mittausten varianssi verrattuna tietokonemittauksiin. Mittaustapa on myös yhtä luotettava kuin tietokonemittaukset ja sen tilastollinen voima on hivenen suurempi. Sen sijaan kahdella tavalla saadut mittaukset eivät ole yhtenevät, mikä aiheuttanee ongelmia, jos kahta mittaustapaa on tarkoitus käyttää samassa tutkimuksessa. Kuitenkin voidaan sanoa, että tulokset ovat melko lupaavia ja vaikuttavat vahvistavan oletuksen mobiilin mittalaitteen soveltuvuudesta neurokognitiiviseen arviointiin.

Avainsanat: Reaktioaika, neurokognitiivinen arviointi, mobiili mittaustapa, soveltuvuuden arviointi

Preface

Well, here we are. Who would have thought we see the day that I write the preface of my Master's Thesis? — Quite a lot of people, actually, and I would like to take this opportunity and thank them.

First of all I want to thank my supervisor Mikko Sams, who has been most gracious about the delays and last minute hassle. To my instructor Ville Ojanen I would like to express my gratitude, not only for the great work he has done as my instructor, but for taking the chance in hiring me when Aivokunto was little more than an idea. Look at us now!

My humble thanks goes also to my parents, Eeva and Heikki Malkamäki, who have relentlessly pushed me towards the goal. Their efforts made sure I never lost sight of the prize at the end of the road.

I would like to send my love to my former co-workers in Laboratory of Computational Engineering: Jaakko Kauramäki, Virpi von Alfthan and Iina Aaltonen. You guys have been there for me for a very long time and I wont ever forget that. I hope we will stand beside each other for many years to come. Katri Koskentalo and Hanna Puharinen: from my former co-workers to my current co-workers. We have shared a lot together and I treasure each day I get to work with you. Most of all I feel privileged to call you my friends and confidants. I must not forget my newest co-workers Toni Huttunen and Toni Toivanen. It feels great to come home from work with my jaws and stomach sore from all the laughing. Special thanks to Toni T. for being the best boss ever and ignoring the times I have spend with my Master's Thesis instead of my job.

Last I would like to thank the most important person in my life: my lover, my best friend, my husband, my rock. Tomi: you complete me.

Helsinki, May 18, 2010

Kaisa Rolig

Contents

Abstract ii

Abstract (in Finnish) iii

Preface iv

Contents v

Symbols and abbreviations x

1 Introduction 1

2 Background 4

2.1 History of reaction time measurements 4

2.2 Current applications of reaction time measurements 6

2.3 Mobile measurements 7

2.3.1 Benefits of mobile measurements 7

2.3.2 Challenges of mobile measurements 8

2.4 Sources of variability in reaction time measurements 9

2.4.1 Variability in reaction times 9

2.4.2 Error in measured reaction times 10

3 Theory behind the methods used to assess feasibility 11

3.1 Analysis of variance 11

3.1.1 When to use one-way ANOVA 11

3.1.2 Basic concepts 11

3.2 Requirement 1: Data variance 12

3.2.1 Test of equal variances 13

3.3 Some critique of Pearson's ρ 13

3.3.1 Reliability and agreement 15

3.4 Requirement 2: Reliability of the measurements 16

3.4.1 Intraclass correlation 16

3.4.2 Repeatability coefficients 17

3.5 Agreement of methods . . . 17
3.5.1 Limits of agreement . . . 18
3.5.2 Power analysis . . . 18

4 Methods **20**
4.1 Study protocol . . . 20
4.2 Instrumentation . . . 21
4.2.1 Mobile measurements . . . 21
4.2.2 Computerized measurements . . . 22
4.2.3 Laboratory settings . . . 22
4.3 Subtest description . . . 23
4.3.1 Choice Reaction Time . . . 23
4.3.2 Flanker Interference . . . 24
4.3.3 Delayed Matching to Sample . . . 25
4.4 Subjects . . . 26
4.5 Data analysis . . . 26
4.5.1 Structure of the data . . . 26
4.5.2 Test of equal variances . . . 27
4.5.3 Intraclass correlation . . . 28
4.5.4 Repeatability coefficients . . . 28
4.5.5 Limits of agreement . . . 29
4.5.6 Power Analysis . . . 29

5 Results **32**
5.1 Requirement 1: Data variance . . . 32
5.1.1 Test of equal variances . . . 32
5.2 Requirement 2: Reliability . . . 33
5.2.1 Repeatability coefficients . . . 33
5.2.2 Intraclass correlations . . . 34
5.3 Requirement 3: Agreement of methods . . . 34
5.3.1 Limits of agreement for computer and laboratory measurement 34
5.3.2 Power Analysis . . . 35

6 Conclusions and discussion **38**

6.1 Test of equal variance 38

6.2 Agreement of computer and mobile reaction time measurements . . . 39

6.3 Repeatability coefficients 40

6.4 Intraclass correlation 40

6.5 Power Analysis 41

6.6 Summary 41

References **42**

List of Figures

1 Early reaction time measurement setup 4
2 Personal computer of the 70's 5
3 Pearson's ρ 14
4 Study protocol 20
5 Mobile device — E60 21
6 Computer settings 22
7 Block design - CRT 23
8 Block design - FI 24
9 Block design - DMS 25
10 Structure of the data 27
11 Repeatability coefficients 33
12 Intraclass correlation 34
13 Agreement 35

List of Tables

1 Feasibility requirements 2
2 Intersubject SSD 32
3 Equality of variances 32
4 Intrasubject SSD 33
5 Repeatability coefficients 33
6 Intraclass correlation coefficients 34
7 Agreement 35
8 Number of subjects 36
9 Number of repetitions 37

Symbols and abbreviations

Symbols

σ^2	Variance
ρ	Correlation coefficient
p	Probability

Abbreviations

rt	Reaction Time
SD	Standard Deviation
SSD	Sample Standard Deviation
ANOVA	Analysis of Variance
CRT	Choice Reaction Time
FI	Flanker Interference
DMS	Delayed Matching to Sample
ICC	Intraclass Correlation
RC	Repeatability Coefficient
AC	Agreement Coefficient

1 Introduction

Many traditional neuropsychological tests are implemented with pencil and paper. This kind of implementation most often requires a one-to-one administration and manual analysis of the data by a neuropsychologist. With most records held nowadays on computers, the process often involves typing in the results into a computer.

Automating the test with computer seems logical. The analysis stage is much simplified and the actual administration of the test is also more structured and often fully automated. However, the computer interface has is limitations. Computer is never as flexible as a human being and can not recognize the subtle differences in the person's behavior often important in the testing. Also, the input methods provided by the computer are somewhat limited. Heavy amendments have to be made, sometimes changing the test altogether.

On the other hand, computer provides novel ways to measure the performance of an individual. Highly accurate timing and recording provide the possibility to measure the time of cued activity. Alongside the development of the personal computers, reaction time measurements ensured a strong foothold as a widely used dependent variable in neurocognitive and clinical research.

With more and more possible applications of reaction time measurements we find ourselves facing similar challenges as arose in the early days of these measurements: unless we can find more cost-efficient and easy-to-use ways to measure reaction times, the benefits of measuring them are limited to specialized fields of study. With enough computing power and larger screens, the mobile phones of today and tomorrow provide hopefully a suitable — and reliable — method to conduct screening of various diseases and concussions, and to monitor one's cognitive performance on day-to-day basis.

The purpose of this Thesis is to study feasibility of measuring reaction times using mobile phone. To do so, we need to set some basic requirements for the mobile measurement method. We divided the question of feasibility to the three aspects listed below:

1. The mobile measurement method does not significantly increase measurement

error when compared to traditional computer-based measurements.

2. Mobile measurements needs to be reliable.

3. Data collected with the two methods should agree and produce the same effects.

Table 1: Analyses used to assess feasibility. On the left are the three requirements that mobile measurements have to meet. On the right are the corresponding analyses.

Requirement	**Analysis**
1: Variance	Test of equal variances
2: Reliability	Intraclass correlation Repeatability coefficient
3: Method agreement	Limits of agreement Power Analysis

First of the requirements sets constraints on the variance of mobile measurements. As discussed in section 2.4, reaction time measurements have several sources of error that lead to relatively large variance. We do not want to introduce another *significantly large* source of error when measuring reaction times with a mobile device. The variances of mobile and computer measurements are therefore compared by Levene's test of equal variances to ensure that the mobile measurement method fulfills requirement 1.

Second requirement ensures that the mobile measurement method produces the same results from measurement to measurement if the settings remain constant. This is verified by comparing repeatability coefficients which quantifies the interval in which difference between two measurements is most likely to reside. Also, since most of the cases reaction times are used to study phenomenon on larger population, the intraclass correlation of the two different measurement methods are also compared.

Third requirement has to do with the intended use of mobile reaction time measurement method. If it is used in conjunction with computer measurement, the measured reaction times need to be identical between the two methods. This is studied with Bland and Altman's limits of agreement. The other part of the requirement is studied with power analysis. It reveals the number of subjects and repetitions needed to detect statistically significant effects in reaction times. The

mobile measurement method should not increase the number of subjects or the number of repetitions needed to detect the studied effects.

This Thesis is divided into five parts. At first we take a more thorough look into the background of this study. Second part concentrates on the theory behind the analysis performed to study feasibility. Third part describes the study paradigm and other methods used in the study. Fourth part describes the results of the study and last part discusses the obtained results further.

2 Background

2.1 History of reaction time measurements

The following section is largely based on Luce (1991).

In 1868 F. C. Donders suggested that the time needed for a simple detection task consists of the time it takes to perceive the stimulus plus the time it takes to generate the response. He then used a *subtractive method* to infer how much time was needed for intervening tasks, such as identification, comparison, or other higher-level judgments. This can be, in a way, considered as the first reported use of reaction times in psychology.

Joseph Jastrow, in 1890, stated another major argument for examining reaction times. If the processing of information by the mind is highly structured, then different paths through that structure will lead to different time courses, and those differences will be reflected in the reaction times.

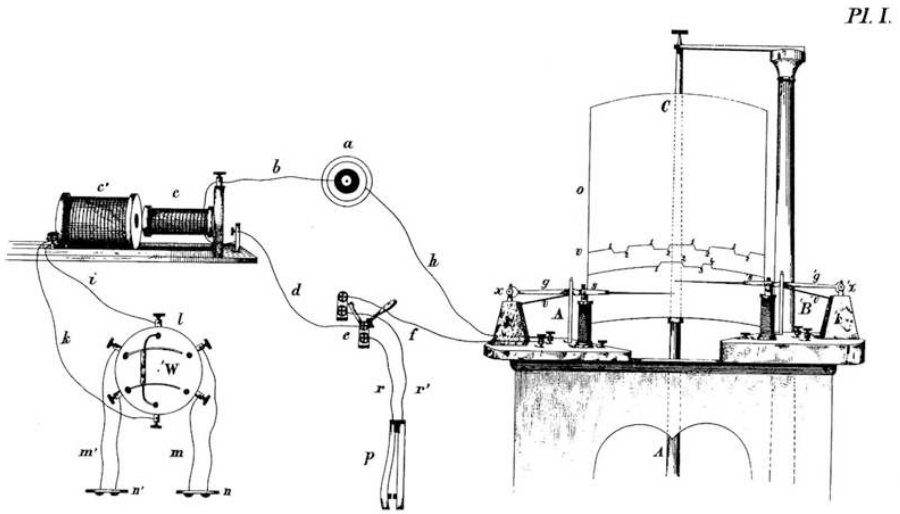


Figure 1: Illustration of Donder's setup for studying reaction times. (Adapted from de Jaager (1865))

From the first attempts to use reaction times until the growth of modern cognitive psychology beginning in the mid 1950s, reaction times were largely the focus of specialists and were not usually recorded by others. One important reason for this was the sheer technical difficulty of carrying out the measurements with the

equipment of the time.

On the other hand, a major change after World War II was the shift from a strongly behaviorist-operational orientation to a cognitive one. This philosophical adjustment in the 1950s and ease of measurement emerging with the rise of personal computer has led to very extensive use of reaction times as a crucial dependent variable.

70's and 80's can — in a sense — be described as the golden years of reaction time measurements. Before the sophisticated imaging methods used today, the reaction times were the most reliable way to study human behavior.

Figure 2: A personal computer system of the late 70's

Indeed, most of the studies of late 70's and early 80's concentrated on basic functions of visual (Navon, 1977) and auditory perception, properties of lingual processes (Tallal, 1980), as well as on hand-eye-coordination (Anzola et al., 1977) and other behavior unique for human beings.

Richard Shiffrin's and Walter Schneider's work on Automatic/controlled processing theory is one of the most notable work of the 70's utilizing reaction times. They outline their theory in Schneider and Shiffrin (1977) and Shiffrin and Schneider (1977) as follows: Automatic/controlled processing theory assumes that human performance is the result of two qualitatively different processes; automatic and controlled processing. Automatic processing is a fast, parallel process not limited by short term memory which uses little subject effort, permits little direct subject control, but requires extensive and consistent training to develop. Controlled processing is a comparatively slow, serial process limited by short term memory which requires subject effort, permits a large degree of subject control, but needs little

training to develop.

2.2 Current applications of reaction time measurements

In the first decade of the 21st century the reaction times are measured regularly in many different fields of medicine and psychology.

In clinical research, the reaction times are used to study the effects of drugs and experimental medication (Bolla et al., 2002). Also brain injury (Hetherington et al., 1996), various diseases (Elsass and Hartelius, 2009) and other disorders are studied by these means. Most of these studies are aimed to increase the understanding of the effects of lesions or disorders and ultimately creating a medication or other remedy. Some are focused in detecting the early symptoms of, for example, Alzheimer's disease. However, work on disease detection has not yet been the focus of wide interest since the screening of larger populations has not been cost-efficient.

In neurocognitive research the reaction times are most commonly measured in conjunction with EEG, MEG, PET, fMRI or other imaging method. While reaction times represent the actual outcome of a decision, these imaging methods concentrate on what goes on in the brain electrically or metabolically. Their function is to locate the source of activity whereas reaction times help to make sure the study design measures the intended activity.

The more infield-applications of reaction time measurements are still somewhat limited by the cost of conducting full neurophysiological studies. There is a clear demand for a simple and cost-efficient way to measure reaction times. Recently, just such a device has been developed (Kim et al., 2009). However, despite the many great qualities of the device, it is limited to measure only certain reaction times (simple and choice reaction time). And since the device has no other function than measuring reaction times, marketing it may be difficult.

2.3 Mobile measurements

2.3.1 Benefits of mobile measurements

Almost every Fin has a mobile phone. It is as essential as the keys or the wallet and goes with them everywhere. Today the mobile phone is so much more than just a phone: it is a radio and a MP3-player, it helps to kill time with games and quizzes, it gives access to the Internet, and allows to check and read email.

To allow gaming, the phones must have quite powerful processors and enough memory to run graphics. On the other hand, wireless connection to network is needed for the email and Internet access. In a way, the mobile phone has become a computer in a pocket with most of the current models supporting a Java platform.

Since the mobile phone is always nearby, it is quite easy to do the measurements. Say a researcher is interested in the effects of migraine attack on cognitive functions. It is often difficult if not impossible to anticipate the attacks. And when the attack hits, the subject would have to rush to a laboratory to do the tests, no matter what the time is.

When the test battery is installed to the subjects phone, he can do the measurements whenever is appropriate and in his own surroundings. The data are sent to a main server, from where the researcher can analyze them when it is appropriate for him. This makes the mobile measurements cost-efficient and allows the design of more complex paradigms.

Since the mobile measurements reduce the costs and effort per subject, the number of subjects in a study can be larger than normally. This increases the statistical power of the study and enables detection of smaller effects. This is useful when the researcher is interested in effects in a population, as is the case in clinical and neurocognitive research.

On the other hand, since a subject can easily repeat the measurements, mobile measurements also benefit fields where the interest is on case studies, such as sport concussion research.

2.3.2 Challenges of mobile measurements

When the subject does the measurements independently, there is no way for the researcher to ensure proper settings. Even though the subject is instructed to do the measurements in an optimal environment, in practice the settings very rarely match those in a laboratory. This increases the external error of the reaction times, as described in section 2.4. Also, input from the measurement administrator has been shown to have a positive impact on the subject's performance.

In the commercial software the accuracy of the computer system is usually well known. However, this is not the case with mobile technology. We do not know, if the system causes systematic error, which can make the reaction times incomparable to computer measurements, as discussed in section 2.4. Also, the variance added by the mobile measurement technology might significantly differ from the computer system.

When designing a test battery for a mobile device, the limitations of the user interface has to be taken into consideration. Even though the layouts of different mobile phones are highly alike, there can be significant differences in the placement and number of available buttons. Also, the number of buttons is significantly smaller than of a computer keyboard.

Another drawback of mobile phones is the lack of a mouse and other custom input devices. The tests must be designed in a way that they utilize only the push buttons. Also, the screen on a mobile phone is substantially smaller than regular computer displays, which leads to a lot smaller view angle. Tests taking advantage of the peripheral vision are impossible to implement on a mobile phone.

Currently auditory stimuli are difficult to produce with a mobile phone. This might change in the future, but for now the mobile phone is suitable for only visual stimuli.

2.4 Sources of variability in reaction time measurements

2.4.1 Variability in reaction times

Reaction time rt to a constant stimulus is normally distributed around a true value μ with some variance σ^2, as stated in equation 1. The variance σ^2 consist of several independent sources of variability, both internal and external (equation 2).

$$rt = \mu + \sigma^2 \tag{1}$$

$$\sigma^2 = \sigma^2_{in} + \sigma^2_{ex} \tag{2}$$

While intending to find a global, true reaction time, the internal sources of variability can be further divided roughly into two source: *intra-individual variability and inter-individual variability.*

Intra-individual variability is the variability that occurs naturally in each individual's performance. Studies have shown that especially inter-trial variability can be explained by attention and anticipation. Also, a recent study shows that a portion of this variability can be explained by intrinsic fluctuations within cortical systems (Fox et al., 2007).

From previous studies we know that the reaction times tend to vary, for example, due to age (Deary and Der, 2005), gender (Adam et al., 1999) and clinical disorders. These variations in the performance lead into inter-individual variability. When we have a group of subjects, this variation constitutes for major part of the variation in the data.

The internal sources of variability are always present in reaction time data. The statistical power of a test design determines, how well the design functions despite this variability. For more detailed discussion on statistical power, see section 3.5.2.

Few examples of the external error sources are varying measurement conditions (e.g. lighting or background noise) and the time of day. Variation caused by mea-

surement conditions can be notably reduced by doing the measurements in a constrained laboratory environment. Time of day variation can naturally be reduced by doing all the measurements at a certain time of day.

2.4.2 Error in measured reaction times

The process of measurement itself causes error in *the measured reaction times*. In addition of adding a new source of variance σ^2_{meas}, the measurement device can also add a constant shift x_0 to the measurements as stated in equation 3.

$$measured\ rt = rt + x_0 + \sigma^2_{meas} \tag{3}$$

By repeating the reaction time measurement we can estimate the value and variance of a measured reaction time with its *mean* and *sample variance*. Equations 4 and 5 illustrates the relations between the estimates and the sources of error.

$$\bar{rt} = \mu + x_0 \tag{4}$$

$$s^2 = \sigma^2_{in} + \sigma^2_{ex} + \sigma^2_{meas} \tag{5}$$

3 Theory behind the methods used to assess feasibility

3.1 Analysis of variance

In neurocognitive research *analysis of variance* or ANOVA is usually performed to find out whether a group or groups deviate significantly from a common mean. However, this is only the tip of an iceberg in statistical usage of ANOVA.

For example test of equal variances (see 3.2.1) and calculation of intraclass correlation (see 3.4.1) are based on one-way ANOVA. One-way ANOVA is also used in calculating repeatability coefficients (see 3.4.2) and performing power analysis (see 3.5.2).

3.1.1 When to use one-way ANOVA

One-way ANOVA is used on two dimensional datasets of observations. This means, that each observation can be described with two independent variables such as *car model* and *average consumption*, *home town* and *political party* or *measurement conditions* and *subject identity*.

And as mentioned before, we want to answer the following question: *does a group deviate significantly from a common mean?* Or by using our examples: does one car model's average consumption differ from the rest, are people in some town politically biased or does the measurement condition affect the performance of a group of subjects.

3.1.2 Basic concepts

The basic concept in one-way ANOVA is that the total variance of a dataset — as the dataset itself — can be explained with two components: *differences among the group means* and *differences within a group*. The ratio of the the two determine whether there is statistical difference between the groups. This ratio is known as

F-statistics (equation 6).

$$F = \frac{variance\ of\ the\ group\ means}{mean\ of\ the\ within\ group\ variances} = \frac{VMg}{MVg} \tag{6}$$

Variance of the group means or VMg is calculated using equation 7, where k is the number of groups, N_i are the number of observations in a group, $\bar{x}_{ix}$ are group means and $\bar{x}_{xx}$ is the mean of *all* observations. For dataset with equal number N of observations in each group, the VMg can be calculated using equation 8.

$$VMg = \frac{\Sigma_{i=1}^{k} N_i(\bar{x}_{ix} - \bar{x}_{xx})^2}{k-1} \tag{7}$$

$$s_{\bar{g}} = N \times Var[Mean_j[x_{ij}]] \tag{8}$$

Mean of the within-group variances or MVg is calculated using equation 9, where k is the number of groups, N_i are the number of observations in a group, x_{ij} are the observations and $\bar{x}_{ix}$ are group means.

$$MVg = \frac{\Sigma_{i=1}^{k}\Sigma_{j=1}^{N_i}(x_{ij} - \bar{x}_{ix})^2}{\Sigma_{i=1}^{k}(N_i - 1)} = Mean[Var_j[x_{ij}]] \tag{9}$$

3.2 Requirement 1: Data variance

As discussed in section 2.4, the variance of measured reaction times can roughly be divided in three components: variance due to internal sources of error (σ_{in}^2), variance due to external sources of error (σ_{ex}^2) and variance due to used measurement method (σ_{meas}^2). The first of the three can be considered constant when the group of subjects remain unchanged. The latter two, however, may change when using a new measurement method.

To assess change in variance due to measurement method, we need to compare mobile measurements and the computer measurements obtained in similar settings.

For the mobile measurements to fulfill requirement of equal variances, variances of the measurements done in a laboratory environment using the two measurement methods should not be significantly different.

As discussed in section 2.3.2, doing the measurements outside laboratory environment can potentially increase the external error source of reaction times. So for the mobile measurements to fulfill requirement of equal variances, the data variance in unconstrained environment should not significantly differ from the measurements done in laboratory environment.

3.2.1 Test of equal variances

The following is largely based on Brown and Forsythe (1974).

The two test commonly used to test equality of variances across multiple groups are Bartlett's and Levene's test. However, Bartlett's test is quite sensitive to deviations from normality and outliers. For data prone to outliers, the more robust Levene's test of equal variances is better suited.

The null hypothesis of Levene's test is that the variances are equal. Levene's statistic is obtained from an one-way ANOVA between groups, where each observation has been replaced by its absolute deviation from its group mean. Critical value for the Levene's statistic is obtained from the F-distribution as the upper critical value.

3.3 Some critique of Pearson's ρ

Pearson product-moment correlation coefficient or Pearson's ρ is a common measure of the correlation between two random variables X and Y. The two random variables are assumed unequal both in metrics and variance and so Pearson's ρ is a measure of *interclass correlation*.

In general, correlation measures the strength and direction of linear relationship between X and Y. Correlation coefficient ranges from -1 (perfect decreasing

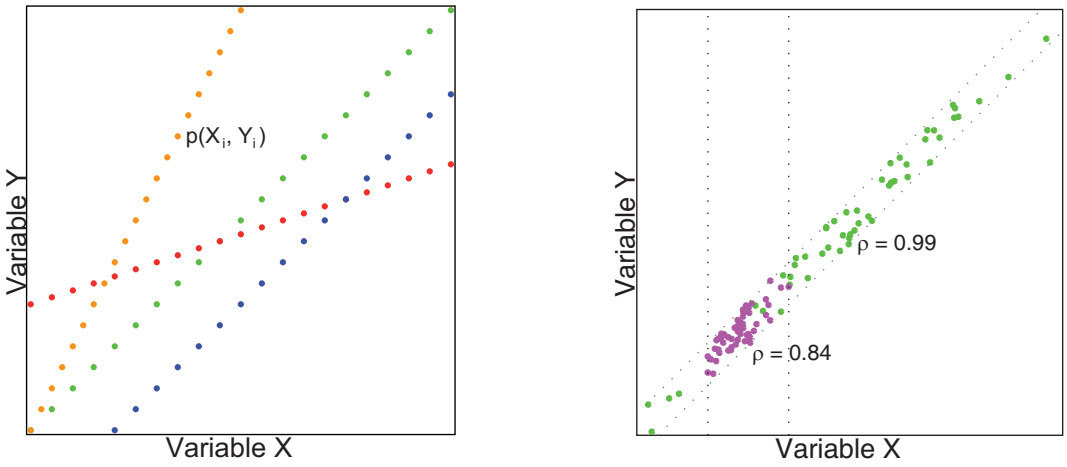
relationship) to 1 (perfect increasing relationship) with 0 for no linear relationship.

A linear relationship between two random variables X and Y can be described with equation 10, where a is the slope between Y and X, b is a constant shift between the two variables and ε is some random error.

$$Y = aX + b + \varepsilon \tag{10}$$

When $\varepsilon = 0$ in equation 10, Pearson's ρ will be 1.0 for all positive a and -1.0 for all negative a. Constant shift b has no effect on the correlation. This is illustrated in the left side of figure 3 with four datasets having $\rho = 1.0$. As can be interpreted from these datasets, as long as the points $p(X_i, Y_i)$ fall on a straight line, the correlation is 1.0.

The right side of figure 3 illustrates a situation where two artificial datasets are derived from same X and Y. Both datasets have 50 data points but the green dataset has five times the range of the purple dataset. Even though the linear relationship between X and Y is identical in both cases, Pearson's ρ is substantially larger for the green dataset (0.99 versus 0.84).



(a) Four datasets with $\rho = 1.0$

(b) Two datasets with unequal ρ

Figure 3: The box on the left shows four artificial dataset having linear relationships with correlation of 1.0. The box on the right shows two datasets with random variation. Both the green and purple datasets have equal number of data points and they follow equation 10 with $a = 1$ and $b = 0$.

3.3.1 Reliability and agreement

If we are studying the reliability of a method, we can assume that the metrics and variance of the method are constant. Indeed, changing variance would indicate very poor reliability. Also, when we are measuring reaction times with two methods, the metrics are by definition equal in both methods. And as stated in requirement 1, the variances of the two methods should also be (nearly) identical.

Heavily simplifying we can say that when studying reliability and agreement, all the observations come from same distribution. Mathematically this means that the linear relationship of the variables can be described with equation 11: two sets of observations from one distribution are equal with small random error between the two. The green datasets in figure 3 illustrate equation 11.

$$X_1 = X_2 + \varepsilon \tag{11}$$

As can be seen from equations 10 and 11, the scale factor a should be very close to one. If a was for example two, observations in X_1 would always be twice the size of observations in X_2. Orange dataset in the left side of figure 3 illustrates a change in scale.

Quite intuitively such differences are unexceptable when there is only one metric like reaction time. But as discussed above, Pearson's ρ produces the same result for all positive values of a. This means that the scale between the observations has to be determined by other means.

The constant shift b in equation 10 should be zero for the two equations to be identical. This means there should be no constant difference between measurements from different methods and at least not within a (reliable) method. Blue dataset in the left side of figure 3 illustrates a non-zero shift between variables.

As discussed above, Pearson's ρ is unaffected by the value of b. This means that Pearson's ρ fails to recognize any constant difference and such a difference should be determined by other means.

3.4 Requirement 2: Reliability of the measurements

In everyday life reliability means that things work how they are supposed to and when they are supposed to. Reliability of reaction time measurements is quite similar: we need to be sure, that doing the measurements in constrained, unchanging settings will provide unchanged results.

As discussed in section 2.4, the results of reaction time measurements will hardly ever be identical. Still, we can expect them to reside within boundaries defined by the sample variance s^2.

Requirement 1 states, that the variances under different measurement conditions must be equal. For the mobile measurements to fulfill requirement of reliability, the variance of the measurements have to be small enough to produce reliable, unchanging results for a constant group of subjects.

3.4.1 Intraclass correlation

Correlation coefficient is often reported as a measure of reliability. This is because correlation tells us how well we can predict the behavior of data from group of subjects over repetitions.

Although Pearson's ρ is sometimes used as a measure of reliability (Tornatore et al., 2005), it does not describe the correct relation. Pearson's ρ is by definition a measure of *inter*class correlation describing the linear relationship between two measures with different metric and variance (McGraw and Wong, 1996). For more detailed discussion on Pearson's ρ, see section 3.3.

To represent reliability of a given measurement with unchanging metric and variance, we need to calculate some *intra*class correlation coefficient, or *ICC*. As noted by McGraw and Wong (1996) there is a variety of different *ICC* used in psychology. Cronbach's alpha is often reported in studies on questionnaire reliability, but many studies do not report the type of the ICC used (Kane et al., 2005; Wilk et al., 2002).

Since methods for calculating ICCs use analysis of variance (Harris, 1913), we

must specify a model for the sample data in order to know which analysis to perform (McGraw and Wong, 1996). Selecting the wrong model to represent the data can have a dramatic effect on the numerical value of ICC which might result in incorrect interpretation of the method reliability.

3.4.2 Repeatability coefficients

Repeatability coefficient or rc gives us the limits within which difference between two repeated measurements reside for 95 % of the subjects. As such, it quantifies the amount of difference that is reasonable to expect within a subject's measurements.

Repeatability coefficient can be calculated with equation 12 where $t_{\alpha/2}$ is the critical t-value from normal distribution ($\alpha = 0.05$) and s_{intra} is average within-subject sample standard deviation calculated with equation 9. As can be seen from equation 12, rc is directly proportional to average within-subject SSD.

$$rc = \sqrt{2}\, t_{\alpha/2} \times s_{intra} \tag{12}$$

3.5 Agreement of methods

Even though ultimately we wish to use a mobile device to measure certain cognitive functions, we are using an indirect measure of reaction times. As a result, we have to evaluate the new mobile method by comparison with an established technique rather than with the true quantity. And when two methods are compared neither provides an unequivocally correct measurement.

We may either compare the mobile measurements with psychological questionnaires or computer measurements. In a validation study the former would be used, since we would need to show that our selection of tests truly measures the quantities they are supposed to. But since this Thesis is interested in the feasibility of using mobile device to measure reaction times, we concentrate solely on comparison with the computer measurements.

3.5.1 Limits of agreement

In medical field many studies give the Pearson's ρ between the results of the two measurement methods as an indicator of agreement. But as discussed in section 3.3, simply calculating Pearson's rho between two variables with common metrics and variance does not provide enough information on their relationship.

Bland and Altman (1986) suggest a method that quantifies the extent to which two methods agree. Instead of merely looking at the linear relationship between measurements obtained with two methods, we are interested in the difference between the measurements and how this difference behaves across the spectrum of values.

First step of the method is to check whether the difference of the measurements correlates with the actual measurements. This would indicate a difference in scale between the methods.

Second step of the method is to calculate the average difference $\bar{d}$ and the standard deviation of the differences sd_d of the two methods. The limits $\bar{d} \pm 1.96 \times sd_d$ are the 95 % confidence intervals for the difference, also called *the limits of agreement*. These limits gives us an estime of how steady the difference of the two methods is. If the limits are small enough, the two methods agree and they can be used interchangeably. The question of what is small enough depends on the quantity measured and also on the intended use of the new method.

Studying repeatability coefficients is relevant to the agreement study since the repeatabilities of the two methods of measurement limit the amount of agreement which is possible. If one method has poor repeatability the agreement between the two methods is bound to be poor too. When the old method is the more variable one, even a new method which is perfect will not agree with it. If both methods have poor repeatability, the problem is even worse.

3.5.2 Power analysis

The following section is largely based on Bausell and Li (2002).

In most reaction time studies the goal is to demonstrate a difference between two conditions. First of all — for this to happen — there has to be some *real difference* deriving from the conditions. And if the difference exist, it has to be *statistically significant*.

But even though we do not get a statistically significant difference, it does not mean that there is no difference. Sometimes study's *statistical power*, or the probability that statistical significance will be obtained, might be too small to detect the difference. To ensure that a real difference is not overlooked due to apparent lack of statistical difference, the study should be designed in a way that guarantees sufficient statistical power.

When calculating the number of subjects or repetitions needed to detect an effect, the statistical power should be 0.8 and the significance level of testing should be $p \leq 0.05$. The statistical power is determined by the ratio between the difference between conditions and disturbing factors, called *effect size*, and the statistical test used in data analysis.

4 Methods

4.1 Study protocol

The study was a counterbalanced cross-over study. The study had three different measurement conditions or *blocks*: computer measurements, laboratory measurements and home measurements. The different blocks are further discussed in the following sections.

The subjects were first randomly divided into six groups with equal number of subjects in each group. Each group had an unique order in which the three measurement blocks were performed. After completing the measurement phase, the subjects filled an user questionnaire and were given a chance to give verbal feedback. Figure 4 illustrates the study time line.

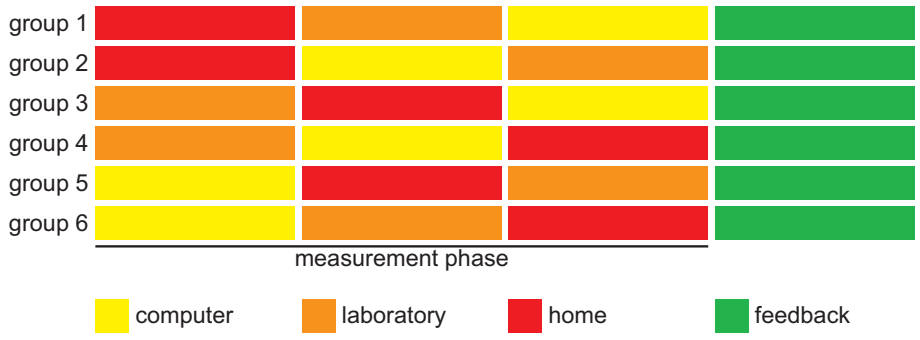


Figure 4: Illustration of the study protocol and time line. The subjects were randomly divided into six groups. Each group completed three measurement blocks and gave feedback after the measurement phase.

Computer measurements were used as a control in the study. The subjects did two measurement using a computer in a constrained laboratory environment (see section 4.2.3). Subjects did the computer measurements on a single occasion with at least 15 minute brake between the two measurements.

Mobile measurements were also done in a constrained laboratory environment to estimate the source of error due to mobile measurement method. Subjects did total of seven measurements on seven consecutive workdays.

Mobile measurements were done in an unconstrained environment to estimate the external sources of error. The subjects were instructed to do the measurements

alone in a quite environment of their choice. The subjects did total of fourteen measurements with two measurements a day on seven consecutive days. The subjects were instructed to do the first measurement of the day in the morning, before 12 AM and the second measurement in the evening, after 8 PM.

4.2 Instrumentation

4.2.1 Mobile measurements

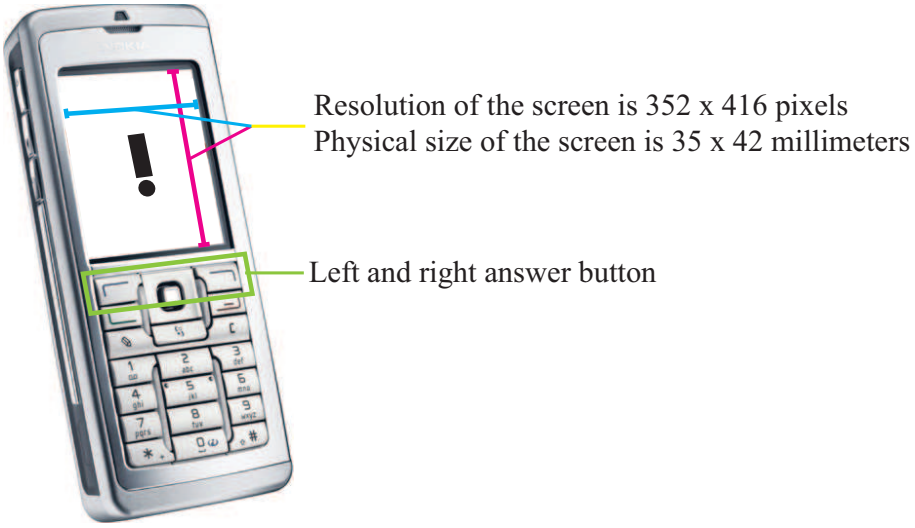


Figure 5: The mobile device in the study was a Nokia E60 cellular phone. The used answer buttons are just below the screen highlighted with green. The size of the screen is 352×416 *pixels* with 100 $pixels/mm^2$.

Software used for the mobile measurements was a mobile reaction time measurement battery called *Mindex*. Mindex is implemented with Java (for MIDP 2.0 and CLDC 1.1) and during the measurements only beta-version of the software was available.

The mobile part of the study was conducted using Nokia E60 mobile phone illustrated in figure 5. The screen on the phone has a resolution of 352×416 px with 100 px/mm^2 and is able to reproduce full RGB-space (16 million colors). *The answer buttons* used in the reaction time tests are located just below the screen.

4.2.2 Computerized measurements

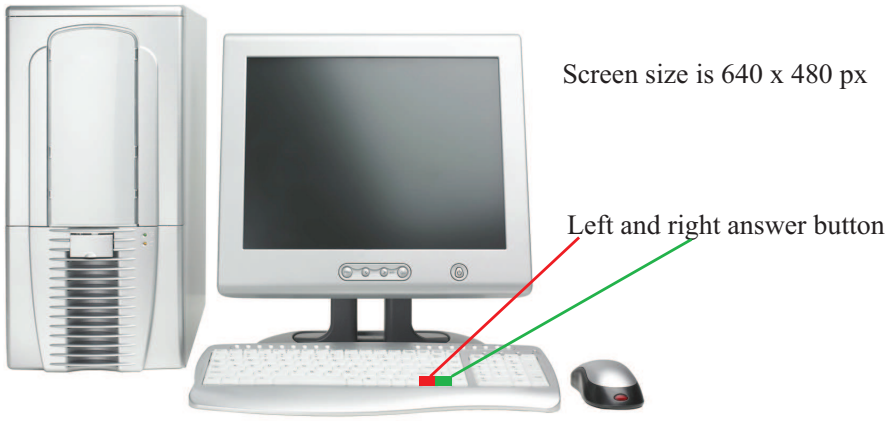


Figure 6: Illustration of the computer settings. The screen resolution was 640 × 480 *pixels*. The answer buttons are *left and down arrow buttons* and are highlighted with blue. Buttons used to select the answer in initial and evaluation questionnaire are *number buttons from one to five* and are highlighted with red.

The computerized reaction time measurements were done with Presentation. The used setting is illustrated in figure 6.

The resolution of the computer screen was 640 × 480 *pixels* and it was located approximately 30 cm from the front end of the table. A keyboard located in front of the screen was used for answering. The answer buttons used in the reaction time tests are *left and down arrow buttons*. Buttons used to select the answer in initial and evaluation questionnaire are *number buttons from one to five*.

4.2.3 Laboratory settings

Part of the mobile measurements and the computer measurements were done in a study laboratory in a TKK facility. The laboratory had two identical cubicles that were soundproofed and had a constant lighting. The computer used resided in one of the cubicles and was only used for this particular study at the time.

4.3 Subtest description

This section describes the different subtests used in both mobile and computerized measurements. All the tests are implemented in a way that only two buttons are required for answering. With mobile device both hands were used for answering. With computer only the primary hand was used.

4.3.1 Choice Reaction Time

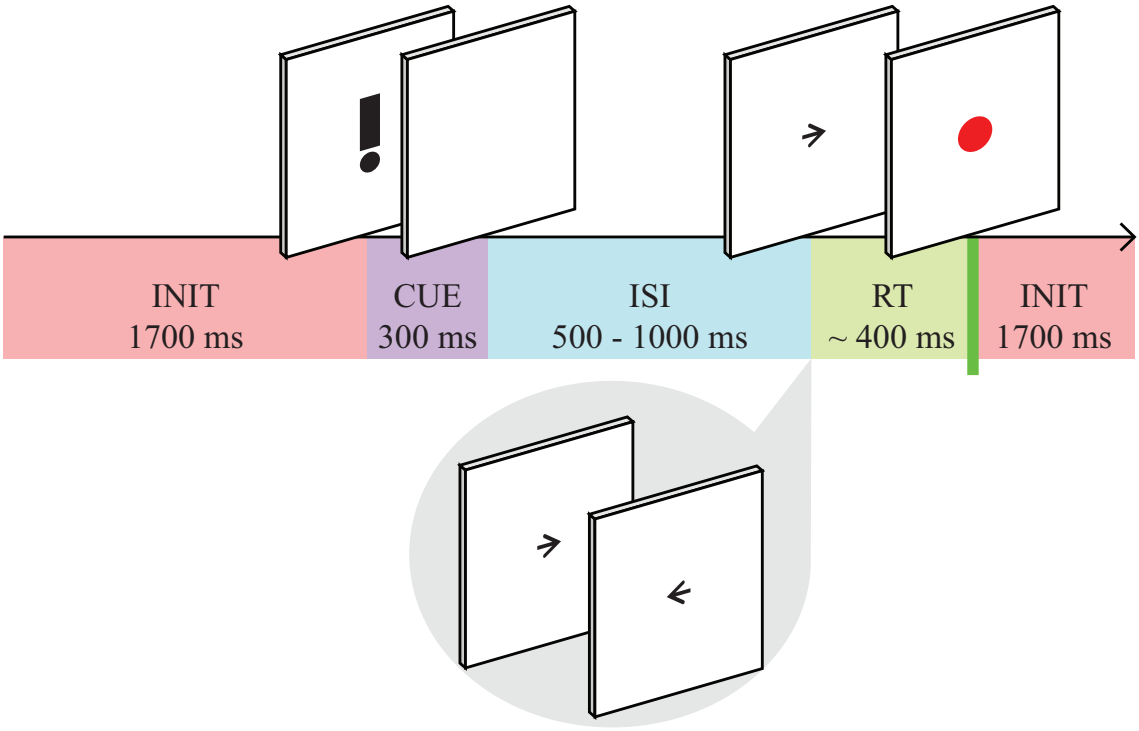


Figure 7: Block design of CRT. Each block starts with an initial pause. After the pause, a cue is shown followed by a random pause. After the pause a single arrow pointing either left or right is presented. The subject reacts to the direction of the arrow. If the response is incorrect, an error cue is visible during the initial pause of the next block.

In the choice reaction time test or *CRT* the subject reacts to *a single arrow* pointing either *left* or *right*.

Single test instance consists of 30 blocks. Each block starts with *an initial pause* of 1700 *ms*. After the pause, *a cue* is visible for 300 *ms* followed by *a random pause* ranging from 500 to 1000 *ms*. After the pause subject is presented with a single

arrow pointing either *left* or *right*.

The task is to press the answer button corresponding to the direction of the arrow. If the response is incorrect, *an error cue* is visible during the initial pause of the next block.

4.3.2 Flanker Interference

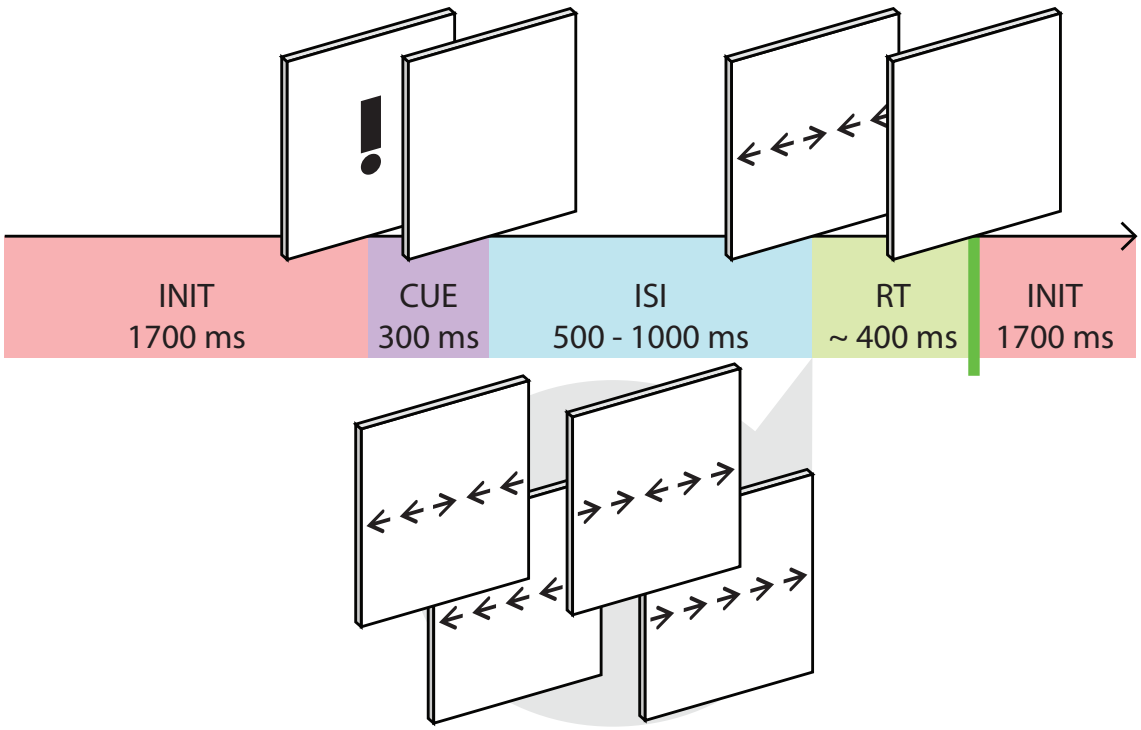


Figure 8: Block design of FI. Each block starts with an initial pause. After the pause, a cue is shown followed by a random pause. After the pause an array of arrows is presented. The subject reacts to the direction of the middle arrow. If the response is incorrect, an error cue is visible during the initial pause of the next block.

In the flanker interference test or *FI* the subject reacts to *an array of arrows* with either *all arrows* pointing to *the same direction* or *the middle arrow* pointing to *the opposite direction* compared to the trailing arrows.

Single test instance consists of 60 blocks (30 congruent and 30 incongruent). Each block starts with *an initial pause* of 1700 *ms*. After the pause, *a cue* is visible for 300 *ms* followed by *a random pause* ranging from 500 to 1000 *ms*. The stimulus

presented after the pause is either five arrows pointing to the left or five arrows pointing to the right in *the congruent condition*. In *the incongruent condition* the stimulus is one of the previous two with the direction of the middle arrow reversed.

The task is to press the answer button corresponding to the direction of the middle arrow and discarding the two trailing arrows on either side. If the response is incorrect, *an error cue* is visible during the initial pause of the next block.

The flanker interference test is designed to measure the ability to concentrate on a target stimulus with interfering stimuli. The reaction times under incongruent conditions are longer than under congruent condition. The difference in the reaction times is known as the *flanker effect*. The size of the effect depends on the subject's ability to concentrate and on the overall cognitive performance level.

4.3.3 Delayed Matching to Sample

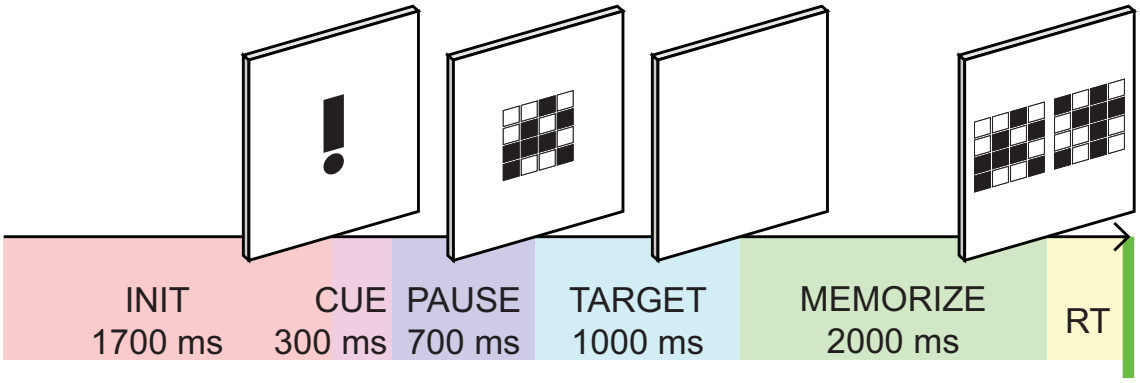


Figure 9: Block design of DMS. Each block starts with an initial pause. After the pause, a cue is shown followed by a pause. A checkerboard pattern is presented for a while after which the subject has time to memorize the pattern. After the memorization pause two checkerboard patterns are presented. The subject reacts according to which side the target resides. If the response is incorrect, an error cue is visible during the initial pause of the next block.

In the delayed matching to sample or *dms* the subject memorizes a checkerboard pattern and then selects the matching pattern from two possibilities.

Single test instance consists of 30 blocks. Each block starts with *an initial pause* of 1700 *ms*. After the pause, *a cue* is visible for 300 *ms* followed by *a pause* of 700 *ms*. *The target pattern* is visible for 1000 *ms* followed by *a memorization pause*

of 2000 ms. After the pause two checkerboard patterns are presented side by side.

The task is to press the answer button on the same side of the screen that the target pattern resides. If the response is incorrect, *an error cue* is visible during the initial pause of the next block.

4.4 Subjects

The subjects were recruited by placing an add in a news group of Helsinki University of Technology (HUT). The first 30 applicants were selected, of which 26 followed the study protocol and are included in the study.

The subject were aged from 19 to 26 with one 48 year-old. All of the subjects are right-handed and 13 are male. None have diagnosed brain-disorders possibly having an effect on the results.

Most of the subjects were students in the Helsinki University of Technology. They were paid 150 euros on completion of all the phases.

4.5 Data analysis

The data collected with the mobile device were first preprocessed using Matlab. The data from the computer measurements were preprocessed using Excel. These data were then combined to create the dataset described in section 4.5.1. All the analyses were done using Matlab.

4.5.1 Structure of the data

Figure 10 illustrates the structure of our dataset. On the first level we have the different subjects. Each subject has data on three different measurement conditions, which is represented by the second level of dataset.

On the third level we have all the repetitions of the blocks. As described in section

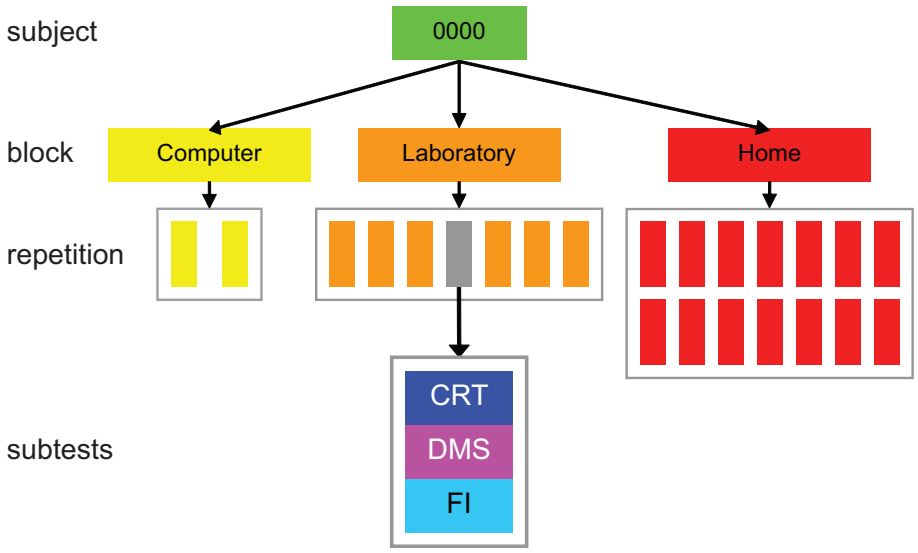


Figure 10: Illustration of the structure of the data. First level of the dataset represents the different subjects in a group. On the second level are different measurement conditions. Third level represents the different repetitions of the measurement under a condition. On the fourth level is the different subtests used.

4.1, the number of repetitions depends on the measurement conditions. Fourth level represents the different subtests described in section 4.3.

In all the analyses, data from different subtests are studied separately. This means that each analysis is done separately on every subtest and the fourth level of dataset is in a sense omitted in the actual analyses.

Since the flanker interference test has two conditions with unequal reaction times (see section 4.3.2), we divided the test into two parts for the analyses: *incongruent* flanker interference and *congruent* flanker interference. Also *the flanker effect* discussed in section 4.3.2 is used in assessing agreement.

4.5.2 Test of equal variances

To minimize the internal sources of error in reaction times, the data was averaged over repetitions (third level of dataset, see section 4.5.1) for each block.

First we tested the equality of variances of computer and laboratory measurements to asses whether the source of error due to measurement method has a sig-

nificant effect on the variance.

Second we tested the equality of variances of laboratory and home measurements to asses whether the source of error due to unconstrained environment has a significant effect on the variance.

Third the overall effect of mobile measurements on the data variance was tested by comparing computer and home measurements.

4.5.3 Intraclass correlation

The method to calculate intraclass correlations with one-way ANOVA is described in detail in McGraw and Wong (1996).

The data from different measurement conditions and subtests were analysed separately resulting in two-dimensional datasets. The correlation coefficients were calculated between repetitions with subjects as a group of observations.

The correlation coefficients were tested against Cohen's large effect size criteria of $\rho = 0.50$ (Cohen, 1988).

4.5.4 Repeatability coefficients

The data from different measurement blocks and subtests were analysed separately resulting in two-dimensional datasets (see figure 10).

First we calculated the within-subject sample standard deviation using equation 9. Then the repeatability coefficients were calculated using equation 12.

95 % confidence intervals for the rcs were derived from a Chi-square distribution with $(number\ of\ repetitions - 1)$ degrees of freedom.

4.5.5 Limits of agreement

We compared the computer and laboratory measurements to find out how well the two agreed. The following analysis was repeated for each subtest.

First we averaged the data over repetitions. Second we calculated both the average of the two methods and the difference *computer* − *laboratory* for each subject.

Third we calculated Spearman's ρ to asses whether the differences correlate with the mean reaction time. A high correlation would indicate that the difference depends (linearly) on the reaction time and that the two methods can not be used interchangeably.

Since there was no significant correlation as can be seen in table 7, fourth step was to calculate the average difference and standard deviation of the differences. From these two we can calculate the 95 % confidence interval for the difference using equation 13, where $\bar{d}$ is the average difference, sd_d is the SSD of the difference and $t_{\alpha/2} \times sd_d$ is *agreement coefficient ac*.

$$limits\ of\ agreement = \bar{d} \pm t_{\alpha/2} \times sd_d = \bar{d} \pm ac \tag{13}$$

4.5.6 Power Analysis

The following equations are from Bausell and Li (2002).

In most reaction time studies the researcher is interested ultimately in difference between mean reaction times of two conditions. In such case, the used statistical test is paired t-test.

The effect size between two independent means is calculated with equation 14, where $M_1 - M_2$ is the difference between the mean reaction times and SD_{pooled} is the pooled (aka combined) standard deviation.

$$ES = \frac{M_1 - M_2}{SD_{pooled}} \tag{14}$$

With paired t-test the ES has to adjusted because the correlation between the paired observations directly impacts the error term. ES_{adj} is calculated with equation 15, where r is the projected correlation between pairs of observations.

$$ES_{adj} = \frac{ES}{\sqrt{1-r}} \tag{15}$$

From ES_{adj} we calculate the t-value obtainable from the study t_{hyp} using equation 16, where N is the number of subjects. We also need to find the critical t-value t_{cv} with $N-1$ degrees of freedom and the desired significance level.

$$t_{hyp} = \frac{ES_{ad}}{\sqrt{2/N}} \tag{16}$$

To calculate the actual power, we simply subtract t_{cv} from t_{hyp} and 'pretend' that the difference is a z-statistic. Then we ascertaining what proportion of the normal curve is to the left of the z-score and get the propability. Since the t_{hyp} is not normally distributed, we use a correction term. The power can be calculated from equation 17, where df is the degrees of freedom $(N-1)$.

$$power = p\left(z \leq \frac{t_{hyp} - t_{cv}}{\sqrt{1 + t_{cv}^2/2df}}\right) \tag{17}$$

Calculating the N from the above equations is quite laborious but can be done recursively using computer, if we know the desired power, significance level, correlation between paired observations and the effect size. In our analyses, we used power of 0.8 and a significance level of 0.05 as suggested by Bausell and Li (2002).

We calculated the number of subjects needed to prove a difference between two conditions with no repetitions. The intraclass correlations calculated in section 5.2.2 were used as the correlation between paired observations and the *inter*subject

variability was used as SD_{pooled}.

We also calculated the number of repetitions a single subject would have to make on average to prove a difference. Again, the ICC was used as the correlation between paired observation and now the *intra*subject variability was used as SD_{pooled}.

5 Results

5.1 Requirement 1: Data variance

Sample standard deviations between subjects or *within group* were calculated with one-way ANOVA as describe in section 3.1 and these results are listed in table 2.

Table 2: Average intersubject sample standard deviations for each subtest under each condition.

	crt	**dms**	**fiC**	**fiI**
computer	22.8 *ms*	71.2 *ms*	25.0 *ms*	40.7 *ms*
laboratory	34.8 *ms*	76.9 *ms*	30.9 *ms*	43.5 *ms*
home	27.6 *ms*	65.7 *ms*	32.5 *ms*	40.2 *ms*

5.1.1 Test of equal variances

The equality of variances was checked pairwise between computer and laboratory measurements, between computer and home measurements and between laboratory and home measurements. The p-values are shown in table 3 for each subtest and each combination mentioned above. The null hypothesis of equal variances is rejected with $p \leq 0.05$.

Table 3: The p-values of the Levene's test. The columns of the table represent the different subtests and the rows represent different comparisons. The equality of variances was checked pairwise between computer and laboratory measurements, between computer and home measurements, and between laboratory and home measurements.

	crt	**dms**	**fiC**	**fiI**
computer vs. laboratory	0.29	0.96	0.08	0.47
laboratory vs. home	0.36	0.32	0.94	0.21
computer vs. home	0.34	0.47	0.12	0.99

5.2 Requirement 2: Reliability

5.2.1 Repeatability coefficients

Average within-subject sample standard deviations were calculated for each subtest and each measurement block. The results are listed in table 4. Repeatability coefficients rc were calculated from the SSDs as described in 4.5.4 and are listed in table 5.

Table 4: Average intrasubject sample standard deviation for each subtest and each measurement block.

	crt	**dms**	**fiC**	**fiI**
computer	14.9 *ms*	43.4 *ms*	12.0 *ms*	28.8 *ms*
laboratory	21.1 *ms*	33.9 *ms*	17.5 *ms*	21.5 *ms*
home	19.2 *ms*	36.2 *ms*	18.9 *ms*	22.2 *ms*

Table 5: Repeatability coefficients for each subtest and each measurement block.

	crt	**dms**	**fiC**	**fiI**
computer	41 *ms*	120 *ms*	33 *ms*	80 *ms*
laboratory	59 *ms*	94 *ms*	48 *ms*	60 *ms*
home	53 *ms*	100 *ms*	52 *ms*	61 *ms*

Figure 11 illustrates the repeatability coefficients with 95 % confidence intervals. Agreement coefficients ac between computer and laboratory measurements listed in table 7 are also added to figure 11 for visual comparison.

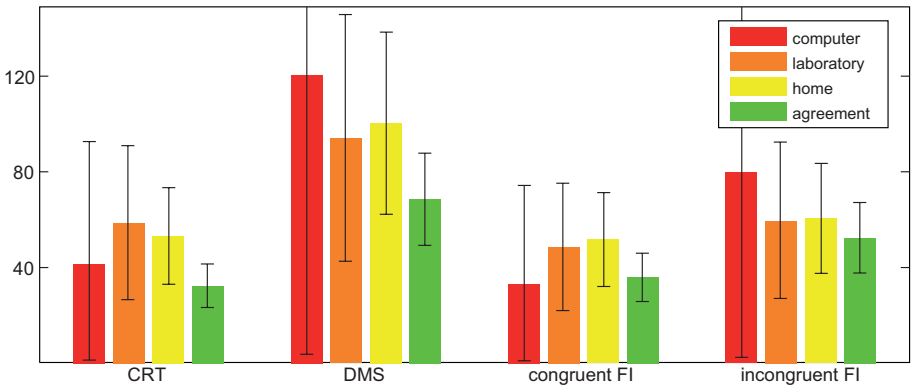


Figure 11: Illustration of the repeatability coefficients. Each group of bars represents one subtest and different colors represents different measurement blocks. The green bars represent the agreement coefficients listed in table 7.

5.2.2 Intraclass correlations

ICCs between repetitions are listed in table 6. Figure 12 illustrates the ICCs of table 6 with 95 % confidence intervals.

Table 6: Intraclass correlation coefficients between repetitions for each subtest under each condition. Correlation coefficients significantly ($p \leq 0.05$) greater than 0.5 are marked with an asterisk.

	crt	**dms**	**fiC**	**fiI**
computer	0.68	0.63	0.79*	0.47
laboratory	0.67*	0.81*	0.72*	0.77*
home	0.65*	0.73*	0.72*	0.72*

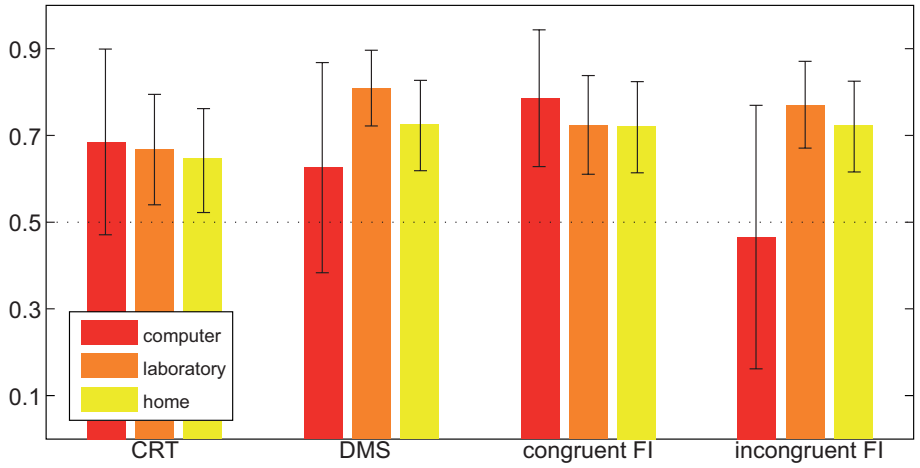


Figure 12: Intraclass correlation coefficients with 95 % confidence intervals. Each group of bars represents one subtest and different colors represents different measurement conditions.

5.3 Requirement 3: Agreement of methods

5.3.1 Limits of agreement for computer and laboratory measurement

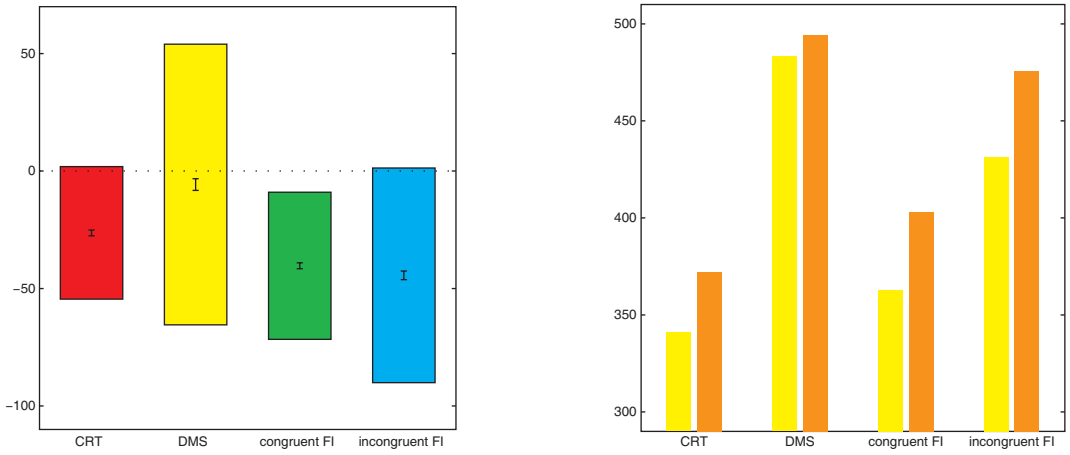
The agreement between computer and laboratory measurements was calculated according to section 4.5.5. The rows of table 7 list the associated measures and figure 13 illustrates the analysis.

First two rows of table 7 are the correlation between the difference in the two measurements and mean of the two measurements as Spearman's ρ values, and the

Table 7: Agreement of computer and laboratory measurement. It lists the mean of the differences between the two measurements and the agreement coefficients, all in milliseconds. The last row shows the overall mean reaction times of both measurements for comparison.

	crt	**dms**	**fiC**	**fiI**
Spearman's ρ	-0.25	-0.07	-0.32	-0.15
p_ρ	0.23	0.74	0.12	0.46
$\bar{d}$	-26 ms	-6 ms	-40 ms	-44 ms
$\pm ac$	32 ms	69 ms	36 ms	52 ms
$\bar{rt}$	357 ms	489 ms	383 ms	454 ms

associated probability of the null hypothesis of no correlation. Next two rows of table 7 show the mean of the differences $\bar{d}$ between the two measurements and the agreement coefficients ac, all in milliseconds. The last row of table 7 shows the overall mean reaction times of both measurements for comparison.



(a) Limits of agreement

(b) Average reaction times for the two methods

Figure 13: Graphical representation of the calculated differences between mobile and computer measurements with the agreement coefficients. The colored boxes in the left picture indicate the limits in which the difference in measurements is likely to reside with 95 % of the cases. The picture on the right shows the average reaction times for computer and mobile measurements.

5.3.2 Power Analysis

Tables 8 and 9 list the results of power analysis described in section 4.5.6.

Table 8: Number of subjects needed in a study to achieve statistical power of 0.8. $N_{subjects}$ is calculated for each subtest under each condition for differences of size 10, 20, 30, 40 and 50 milliseconds.

Difference		**crt**	**dms**	**fiC**	**fiI**	**fiEf**
10 ms	*computer*	30	258	27	128	81
	laboratory	36	132	34	55	39
	home	41	140	42	65	36
20 ms	*computer*	9	66	9	34	22
	laboratory	11	35	10	15	12
	home	12	37	12	18	11
30 ms	*computer*	6	31	5	16	11
	laboratory	6	17	6	8	7
	home	7	18	7	9	6
40 ms	*computer*	4	18	4	10	7
	laboratory	5	10	4	6	5
	home	5	11	5	6	5
50 ms	*computer*	4	13	4	7	6
	laboratory	4	8	4	5	4
	home	4	8	4	5	4

Table 8 lists the needed number of subjects in a study to achieve statistical power of 0.8 ($N_{subjects}$). It is calculated for each subtest under each condition for differences of size 10, 20, 30, 40 and 50 milliseconds.

Table 9 lists the number of repetitions a subject on average has to make in order to achieve statistical power of 0.8 (N_{rep}).

Table 9: Number of repetitions a subject on average has to make in order to achieve statistical power of 0.8. N_{rep} is calculated for each subtest under each condition for differences of size 10, 20, 30, 40 and 50 milliseconds.

Difference		**crt**	**dms**	**fiC**	**fiI**	**fiEf**
10 ms	*computer*	11	98	8	69	61
	laboratory	11	28	11	14	21
	home	15	38	13	19	21
20 ms	*computer*	5	26	4	19	17
	laboratory	5	9	5	5	7
	home	6	11	5	7	7
30 ms	*computer*	4	13	3	10	9
	laboratory	4	5	3	4	5
	home	4	6	4	4	5
40 ms	*computer*	3	8	3	7	6
	laboratory	3	4	3	3	4
	home	3	5	3	4	4
50 ms	*computer*	3	6	3	5	5
	laboratory	3	4	3	3	3
	home	3	4	3	3	3

6 Conclusions and discussion

6.1 Test of equal variance

Overall the results seem to confirm our hypothesis that the mobile reaction time measurements are as reliable as the computerized measurements. Levene's test of equal variances did not find significant differences between computer and laboratory measurements. Also, the unconstrained environment did not cause significant change in the variance of the measurement.

Even though the unconstrained - and unoptimal - environment most likely affects the measurement process, our results seem to indicate that the effect is relatively small compared to the overall variance of reaction time measurements. The intraindividual variance is quite large and multiple repetitions are strongly recommendable regardless of the measurement method. As can be seen from tables 4 and 2, there is no increase in intraindividual or interindividual variance due to unconstrained environment. Also, there is no significant difference in the actual reaction times between measurements done in constrained and unconstrained environment.

These results are quite promising when considering the possible implementations of mobile reaction time measurements. Knowing that the absence of a researcher or proper laboratory settings does not decrease the reliability of the data allows actual in-field studies of reaction times and also more extensive population studies.

For example routine scanning of the elderly population for Alzheimer's and other dementing diseases allows the detection of a disease in its early stage or *a mild cognitive impairment*. Mobile reaction time measurements could provide a cost-efficient and effortless way to conduct wide scans. Detecting the mild cognitive impairment early on and delaying the progression to dementia even by a year may result in significant savings for the public healthcare system.

6.2 Agreement of computer and mobile reaction time measurements

The results of the Bland and Altman's method, however, reveal a difference between the computer and laboratory measurements. There is a constant shift in reaction times between the two measurement methods. What we find particularly surprising is the fact that the shift is not constant throughout our test battery varying from -6 ms to -44 ms as can be seen in table 7. This finding led us to think that the difference is not caused by the measurement system, i.e. the used mobile device, and is more likely to originate from a different source.

One potential source for the difference in reaction times could be the quite different scale of screen and input device between the two measurement methods. Mobile device and its screen are very compact. This results in a small visual angle, since the normal distance of viewing is approximately at an arm's length. Especially in the flanker interference task the target stimuli are small and closely spaced when viewed on mobile device. This in turn makes the correct identification of the stimuli much harder than on the computer screen.

Since the difference is quite similar for both choice reaction time task and flanker interference task - which two shares an identical arrow as the target stimulus - we can assume that the difference in reaction times is due to a longer identification period of the stimulus.

In delayed matching to sample task the difference in reaction times is virtually non-existent. If indeed the explanation to the difference is stimulus identification, this would suggest that the checkerboard stimulus type used in DMS is not as prone to changes in visual angle as the arrow.

The role of the different input device must also be taken into consideration. In mobile measurements the subjects were instructed to use both thumbs to perform the task whereas in computer measurements they used the index and middle finger of their primary hand. This difference in input setup might have an effect on the reaction times, but different study setup would be needed to give insight on the matter.

As long as the difference is acknowledged and the study paradigm is devised in a

matter that do not require comparison between measurements done using computer and mobile device, this difference in reaction times does not present any real problem for using a mobile device for reaction time measurements. Even so, a further study concentrating on locating the source of this difference is highly recommendable.

6.3 Repeatability coefficients

As can be seen from figure 11, the repeatability coefficients are quite similar in every situation, which is due to equality of variances. When we compare these to the agreement coefficients in the same figure, we can see that the limits of agreement are quite small when compared to the internal repeatability of reaction times measured with any of the methods. This is quite a promising result since it would indicate that the agreement coefficient could no be any smaller. So even though there is a difference in reaction times measured with different methods, it is relatively constant across the subjects.

6.4 Intraclass correlation

Intraclass correlation coefficients of both the mobile laboratory measurements and unconstrained mobile measurent are rather good. They demonstrate correlation significantly stronger than Cohen's large effect size criteria. However, the computer measurements fail this test in most of the subtests. We believe that this is not actually due to a poorer correlation but reflects the lesser number of repetitions in the computer measurements. Increasing the number of repetitions might also increase the correlation.

These results seem to confirm the need to have several repetitions when measuring reaction times. As discussed in section 6.5, quite small number of repetitions is sufficient. However, if the number of repetitions is too small, we can not trust that the averaged times represent the true reaction times.

6.5 Power Analysis

One of the criteria we set for the mobile measurements was the power to produce same statistical effects as the computer measurements. Tables 8 and 9 list the needed number of subjects and repetitions to detect an effect of a given size. Our results indicate that the statistical power of mobile measurements is at least as good as that of the computer measurements.

One interesting fact seen from these results is that with relatively small number of repetitions we can reliably detect quite small changes in reaction times within an individual. Say an individual with migraine is interested in seeing how their cognitive functions are impaired in different phases of the headache. After they have established a reliable baseline, it takes only four to six migraine attacks to detect the impaired reaction times in given tasks.

6.6 Summary

There is a difference in the reaction times measured with computer and mobile device in laboratory conditions. This difference has probably more to do with the changed visual angle and not properties of the mobile technology. Nevertheless, this difference in reaction times should be kept in mind when designing study paradigms taking advantage of the mobile technology.

Regardless of the difference in reaction times, Our results seem to confirm the feasibility of mobile reaction time measurement technology for neurocognitive assessment. The measurements are reliable and the mobile technology does not significantly increase the measurement error.

References

Adam, J., Paas, F., Buekers, M., Wuyts, I., Spijkers, W., and Wallmeyer, P. (1999). Gender differences in choice reaction time: evidence for differential strategies. *Ergonomics*, 42(2):327–335.

Anzola, G., Bertoloni, G., Buchtel, H., and Rizzolatti, G. (1977). Spatial compatibility and anatomical factors in simple and choice reaction time. *Neuropsychologia*, 15(2):295–302.

Bausell, R. B. and Li, Y.-F. (2002). *Power Analysis for Experimental Research: A Practical Guide for the Biological, Medical and Social Sciences*. Cambridge University Press.

Bland, J. M. and Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 1:307–310.

Bolla, K., Brown, K., Eldreth, D., Tate, K., and Cadet, J. (2002). Dose-related neurocognitive effects of marijuana use. *Neurology*, 59:1337–1343.

Brown, M. B. and Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, 69(346):364–367.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 2nd edition.

de Jaager, J. (1865). *De physiologische tijd bij psychische processen*. PW Van de Weijer.

Deary, I. and Der, G. (2005). Reaction time, age, and cognitive ability: Longitudinal findings from age 16 to 63 years in representative population samples. *Aging, Neuropsychology, and Cognition (Neuropsychology, Development and Cogniti*, 12(2):187–215.

Elsass, P. and Hartelius, H. (2009). Reaction time and brain disease: relations to location, etiology and progression of cerebral dysfunction. *Acta Neurologica Scandinavica*, 71(1):11–19.

Fox, M., Snyder, A., Vincent, J., and Raichle, M. (2007). Intrinsic fluctuations within cortical systems account for intertrial variability in human behavior. *Neuron*, 56(1):171–184.

Harris, J. A. (1913). On the calculation of intra-class and inter-class coefficients of correlation from class moments when the number of possible combinations is large. *Biometrika*, 9(3):446–472.

Hetherington, C. R., Stuss, D. T., and J.Finlayson, M. A. (1996). Reaction time and variability 5 and 10 years after traumatic brain injury. *Brain Injury*, 10(7):473–486.

Kane, R. L., Short, P., Sipes, W., and Flynn, C. F. (2005). Development and validation of the spaceflight cognitive assessment tool for windows (WinSCAT). *Aviation, Space, and Environmental Medicine*, 76(6):183–191.

Kim, H., Eckner, J., Richardson, J., and Ashton-Miller, J. (2009). A novel portable visuomotor manual reaction time test. *American Society of Biomechanics Annual Assembly, State College, Pennsylvania*.

Luce, R. D. (1991). *Response Times: Their Role in Inferring Elementary Mental Organization*. Oxford University Press.

McGraw, K. O. and Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1):30–46.

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive psychology*, 9(3):353–383.

Schneider, W. and Shiffrin, R. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological review*, 84(1):1–66.

Shiffrin, R. and Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological review*, 84(2):127–190.

Tallal, P. (1980). Auditory temporal perception, phonics, and reading disabilities in children* 1. *Brain and language*, 9(2):182–198.

Tornatore, J. B., Hill, E., Laboff, J. A., and McGann, M. E. (2005). Self-administered screening for mild cognitive impairment: Initial validation of a computerized test battery. *Journal of Neuropsychiatry and Clinical Neurosciences*, 17:98–105.

Wilk, C. M., Gold, J. M., Bartko, J. J., Dickerson, F., Fenton, W. S., Knable, M., Randolph, C., and Buchanan, R. W. (2002). Test-retest stability of the repeatable

battery for the assessment of neuropsychological status in schizophrenia. *American Journal of Psychiatry*, 159:838–844.