Multimodal Technologies and Interaction

MDPI

Article

Human Perception Measures for Product Design and Development—A Tutorial to Measurement Methods and Analysis

Christian Hatzfeld [1,*], Manuel Kühner [2], Stefan Söllner [3], Tran Quoc Khanh [3] and Mario Kupnik [1]

1 Measurement and Sensor Technology, Technische Universität Darmstadt, Merckstr. 25, 64289 Darmstadt, Germany; kupnik@emk.tu-darmstadt.de
2 Mechatronics and Micro Systems Technology, Heilbronn University, Max-Planck-Str. 39, 74081 Heilbronn, Germany; manuel.kuehner@gmail.com
3 Lightning Technology, Technische Universität Darmstadt, Hochschulstraße 4a, 64289 Darmstadt, Germany; stefan-soellner@gmx.de (S.S.); khanh@lichttechnik.tu-darmstadt.de (T.Q.K.)
* Correspondence: c.hatzfeld@emk.tu-darmstadt.de; Tel.: +49-6151-16-23884

Received: 28 July 2017; Accepted: 27 October 2017; Published: 31 October 2017

Abstract: This tutorial describes the necessary steps for designing and conducting a perception experiment in order to obtain design parameters for human–machine interactions. It is intended for engineers and product designers, which require design parameters not included in the current state of the art. Topics addressed are the preposition of hypotheses, the selection of parameters, psychophysical measurement procedures and the calculation of sample sizes. Relevant steps for data analysis from psychology and social sciences are applied to the engineering and design context and guidelines for reporting results are given. The required steps are illustrated with an example experiment assessing detection thresholds of damping parameters of haptic automotive rotary controls with regard to parameters like knob diameter and distraction. Results imply significant effects of knob diameter on both absolute and differential thresholds, but no effect of distraction, implying a good transferability of laboratory results to real-world applications.

Keywords: psychophysics; design of experiments; data analysis

1. Introduction

The properties of human perception are a crucial design constraint for every technical system intended for human–machine interaction [1]. Depending on the application, the perceptual properties determine whether a technical system properly serves its purpose in terms of a defined information transfer and the tolerable level of nuisance and unwanted effects. Studies with human subjects have to be conducted to determine such properties. These studies are normally made by scientists with various backgrounds: while psychologists use such studies to gain insight into the perceptual processes of the sensual modality under investigation, engineers need reliable results for requirement engineering or system evaluation.

In this paper, we outline the crucial steps in the design of a perceptual study intended to be used in an engineering and design context. Compared to similar studies in medicine and social science, such studies help to gain insight into a special part of the perception process, for example the minimum strength of a stimulus to be surely distinguished from another stimulus or the tolerable amount of light before a test person experiences subjective glare.

1.1. Perception Studies for Engineers

The difficulty of perception studies for engineers lies in the difference of basic assumptions about the "system" to be measured. Humans are generally the opposite of technical systems: They are nonlinear, time-variant and depend on a priori knowledge or experience.

To be able to describe human perception with technically usable parameters, measurement methods from psychophysics and statistical methods that are normally not necessarily used in technical measurements have to be employed. In this tutorial, we focus on the the basic concepts from experimental psychology and statistics that are needed to design a perception study with meaningful output. However, the theoretical basis of the employed concepts often has a wider scope and more applications beyond the aspects discussed here. These other applications are mentioned shortly as a starting point for refined measurements with a more specific focus.

This tutorial is intended for researchers with an engineering background and supposed to provide a guide to a successful study design for the analysis of human perception measures. Furthermore, the interested reader should find relevant references for further, in-depth analysis of certain aspects of the perception process.

1.2. Measuring Perception

The analysis of the relation of objective measurable physical measures (stimulus) and the perceived sensation of the human subjected to this stimulus is topic of psychophysics, a sub-branch of experimental psychology. G.H. Weber is attributed to be one of the first psychophysicists, investigating the perception of weight, back in the end of the 19th century [2]. Nowadays, psychophysics addresses four general perception primitives [3]:

- detection—detecting, whether there is a stimulus present or not;
- discrimination—detecting, whether two stimuli are different in one or more parameters;
- identification—identify an unknown stimulus from a given set of stimuli;
- scaling—relation of the size of two or more stimuli (or their parameters).

In this list, detection and discrimination primitives are largely depending on the capabilities of the sensory system. For scaling and identification, further aspects like memory and personal experience of the subject play a much larger role. Furthermore, these primitives build on one another: if a subject cannot detect stimuli, no discrimination can be performed and without a discrimination of different stimuli, no scaling or identification data can be obtained.

An important description of the human sensory system's capabilities is the psychometric function (Figure 1), which is mostly used to describe detection and discrimination capabilities. This function relates a stimulus Φ to detection probabilities P_Ψ and is normally defined as in Equation (1)

$$\Psi(\Phi;\alpha,\beta,\gamma,\lambda) = \gamma + (1-\gamma-\lambda)\cdot f(\Phi;\alpha,\beta), \tag{1}$$

as for example described by Wichmann et al. [4]. In Equation (1), Φ denotes the stimulus, α and β describe the location and sensitivity parameters of the base function f, respectively. These parameters are related to the perception threshold and the sensitivity of the subject (or observer) in a psychophysical experiment [4]. The base function f commonly is either a cumulative normal distribution (Equation (2)) or a logistic function (Equation (3)), rarely a Weibull distribution:

$$f(\Phi;\alpha,\beta) = \frac{\beta}{\sqrt{2\pi}} \int_{-\infty}^{\Phi} \mathrm{e}^{-\left(\frac{\beta^2}{2}\left(\hat{\Phi}-\alpha\right)^2\right)} \mathrm{d}\hat{\Phi}, \tag{2}$$

$$f(\Phi;\alpha,\beta) = \frac{1}{1+\mathrm{e}^{-\beta(\Phi-\alpha)}}. \tag{3}$$

The parameters γ and λ relate to the guess and lapse rates in the experiment, i.e., cases where the observer will give a positive answer for an undetectable stimulus or a negative answer for a stimulus well above the threshold.

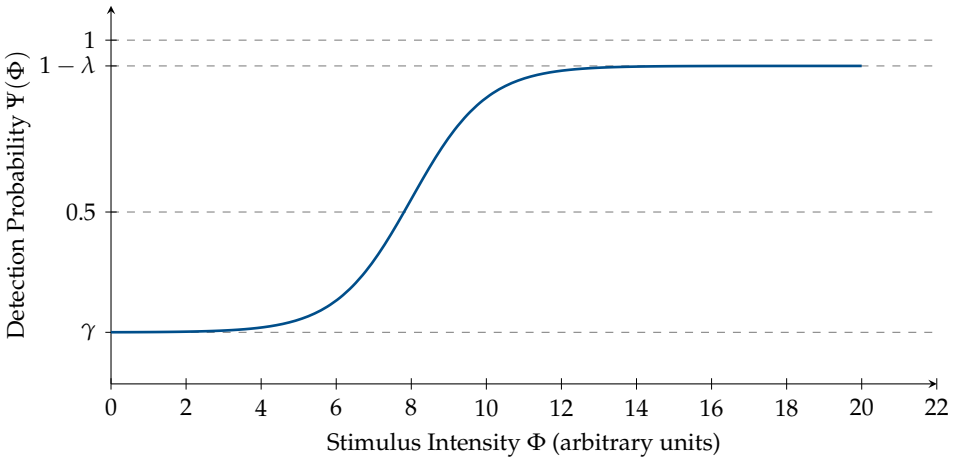


Figure 1. Sample psychometric function $\Psi(\Phi)$ with location parameter $\alpha = 8$, sensitivity parameter $\beta = 1$, corrected guess rate $\gamma = 0.15$ and lapse rate $\lambda = 0.075$. A stimulus with intensity $\Phi = \alpha = 8$ is detected in 50 % of all trials. For discrimination tasks, the relation between test and reference stimulus is often used as stimulus Φ.

Today, there is a large number of psychophysical measurement methods and well-established perception measures, which can be used to evaluate the psychometric function and the above-mentioned perception primitives. A closer look on the formulation of the psychometric function also can reveal different approaches of psychometric measurements from different disciplines: many engineering problems can be answered with a single parameter of the psychometric function (in most cases, the location parameter α) in order to derive design parameters. Studies with a mainly psychological background also investigate other parameters of the psychometric function in order to focus on the general perception process.

In both cases, experiments will normally consist of a repeated presentation of stimuli to a human observer with a single stimulus parameter changing in the course of the experiment. The observer has to decide whether a stimulus is present or different from a reference stimulus depending on the investigated measure. Based on the answers of the observer, the psychometric function and/or psychometric measures (see Section 2.1.1) are calculated from the experiment. The rules for changing stimulus parameters and for calculating the measure are defined by the psychophysical measurement procedure used.

1.3. Structure of This Paper

Despite the selection of the measurement procedure, the experimental design and the statistical analysis is crucial for conducting successful perception studies. Based on the authors' own experiences with visual and haptic perception studies as well as formal training in measurement and instrumentation, this paper is intended to give a focused insight into all steps required to design, conduct and analyze a perception study. It will address questions regarding the design of a perception study, the formulation of a hypothesis and the selection of the measurement procedure in Section 2, the measurement setup and the associated errors in Section 3 and the data analysis in Section 5. To illustrate the procedures described in this work, we present an example from the work of the second author [5,6] measuring the absolute detection threshold for the damping of rotary switches in automotive applications in Section 6.

2. Study Design

The design of a study includes the formulation of a hypothesis, the selection of a psychophysical measure and other parameters to be included in the study, the recruiting of test subjects, and the selection of a measurement procedure.

2.1. Hypothesis

Establishing well-defined hypotheses is crucial for obtaining meaningful results. In most cases, several hypotheses are needed to confirm theories and assumptions from the design process. In general, a large number of hypotheses requires an increased number of experiments (This is just a statistical property: if a large number of hypotheses is tested on the same experimental data, the chance of a false-positive result increases.). Therefore, a small number of hypotheses is advisable and their formulation requires care to include all necessary parameters, constraints and conditions. The basic building blocks for a hypothesis are defined measures for perception properties, the intended information to be drawn from the experiment and known influence from external parameters on perception.

The following examples illustrate typical hypotheses and experimental questions regarding perception parameters for design and development:

- Light emitting diode (LED) technology allows for control of not only brightness and color of white light consisting of different LED colors, but also has an impact on the color rendering of illuminated objects. The energy consumption of LED lightning can be lowered at the expense of the quality of color rendering. Thompson investigates in [7] whether color rendering quality loss is detectable in peripheral and central vision and proposes an optimized dynamic lightning scheme that reduces energy consumption in peripheral vision without affecting the brightness of the scene.
- New automotive lightning systems promise increased sight and visibility, but potentially increase glare of other road users. In [8], Zydek investigates an improved glare-free lightning system and does not find an increased glare of other road users, but a better sight compared to conventional systems.
- The wide use of touch-based interfaces in consumer electronics triggers new applications with haptic feedback in professional contexts. Personal protection gear such as gloves is more wide-spread in professional applications. The question arises of whether these protectional measures have to be considered in the design of professional haptic interfaces. Two recent studies by Seeger et al. [9] and Hatzfeld et al. [10] investigate absolute and differential thresholds for protective and surgical gloves, respectively, and find different parameters with and without gloves, but not necessarily a need to consider these in the design of haptic interfaces.
- The example in Section 6 investigates the perceptual thresholds of damping in rotary controls in combination with other parameters such as detent or user distraction. It aims to find design parameters for clearly distinguishable controls as well as acceptable tolerance values.

2.1.1. Perception Measures

For the use of perception measures in the product development process, it is probably advisable to define hypotheses in terms of well-established psychophysical properties. Five common perception measures that can be used for hypothesis formulation are listed as follows:

Absolute Threshold (AT): The stimulus magnitude that is required for a positive detection of the stimulus [11]. It is the primary measure for the detection primitive and is sometimes also termed *Detection Threshold*. If ATs are directly obtained (not calculated from a psychometric function), the detection probability depends on the psychometric procedure used.

Just Noticeable Difference (JND): The difference in stimulus magnitude that is required to distinguish two stimuli. It is the primary measure for differentiation capabilities [11,12]. There are two common definitions of the JND:

With a direct comparison of test and reference stimulus, the JND is defined as the nominal difference in stimulus intensity $\Delta\Phi = \Phi_1 - \Phi_2$ that is detected in 50 % of all trials. This difference $\Delta\Phi$ is sometimes also called *Differential Threshold (DT)*. In some cases, other convergence levels are also used in this definition, depending on the psychometric procedure used for the experiment.

In the second definition, the JND is calculated from the psychometric function as the difference given in Equation (4):

$$\text{JND} := \Phi|_{\Psi=0.75} - \Phi|_{\Psi=0.25}\,. \tag{4}$$

In many cases, relative values are used to describe the subject's ability to distinguish different stimuli. These relative values are calculated from the JND with respect to a reference stimulus Φ_0 as given in Equation (5), also known as *Weber Fraction*:

$$\text{Weber Fraction} := \frac{\Delta\Phi}{\Phi_0 + a}. \tag{5}$$

The parameter a denotes an additional term that is commonly associated with sensory noise in the human perception system and has to be considered for low stimulus levels Φ_0 near the absolute perception threshold (i.e., some dBs higher than AT).

Point of Subjective Equality (PSE): The configuration of two stimuli that are considered as equal by an observer. This measure belongs to the identification primitive [13].

Just Tolerable Difference (JTD): The perceived difference between two stimuli that is still tolerable with respect to the intended usage. This measure is also known as *Quality JND*. Like the above-mentioned PSE, it is a measure of identification, but has much larger importance for the design of technical systems [14,15].

Power Function Exponent (a): This measure is derived from Steven's Power Law (Equation (6)) and is used to describe the correlation between the objective stimulus intensity Φ above the absolute threshold Φ_0 and the subjective perceived magnitude M:

$$M = k \cdot (\Phi - \Phi_0)^a. \tag{6}$$

The parameter k is a scaling parameter, which is fitted to the experimental data, but has no prominent meaning for describing scaling parameters. The exponent a is used for this purpose and known for a variety of scaling tasks [16].

Psychological and Psychophysical Measures: The above-mentioned measures are closely related to the perception primitives given in the introduction (Section 1.2). They are used to quantify the capabilities of sensory and memory processes in the test person.

Sometimes, a stronger consideration of the individual assessment by the test subject is of interest. An example is *visual glare*, which includes two aspects: *physiological glare* refers to the impairment of visual perception and is easily measured by asking a glared test subject whether it is able to see or identify a defined visual structure (optotype). By comparing detection probabilities of the optotype with respect to different glaring strengths, a quantifiable measure of physiological glare can be constructed.

The aspect of the subjective sensation of glare, i.e., *psychological glare*, is not subject to any quantification. It is only measurable in terms of *more* or *less* and varies largely over any population. Representational assessments such as *unnoticeable* or *disturbing* help to describe the sensation, but do not solve the problem satisfactorily [17]. Because of the diffuse answering options, more test subjects should be considered in studies investigating such psychological attributes compared to studies investigating psychophysical parameters.

2.1.2. Typical Experiments for Product Development

There are two typical experiment types used in the product development process:

Assessment of a Psychophysical Measure: In this experiment type, the value of the psychophysical measure itself is of interest. The experiment is conducted with respect to the parameters of the stimulus (for example, absolute threshold of a vibration with respect to frequency) and the most relevant external parameters are considered with respect to typical configurations for the intended application.

Quantification of Influencing Factors: In this case, the assessment of external parameters is the focus of the psychophysical test. One wants to know whether and to what extent external parameters have a measurable impact on the human perception.

In most cases, a real experiment is somewhere in between these general experiment types. The main differences are the type of classification of external parameters; experimental setup and data analysis are basically the same for both of these types.

2.1.3. External Influences

External influences on perception can manipulate the perception process itself (i.e., altering mechanical or optical properties of receptors), introduce disturbance signals (i.e., glare, unwanted auditory components of a vibration) or affect the cognitive processing of the test person (i.e., disturbances in the lab or experimental setting, mood, etc.). Typical influences on perception can be found in summary works (see, for example, [18–20] for haptic, visual and acoustic perception, respectively) or can be derived from the intended experimental setup. In most cases, pre-tests with a naïve test person can reveal external influences that were not thought of at first.

All of these influences have to be gathered in the planning of the experiment and considered as possible parameters in the experiment.

2.2. Classifying Parameters

Parameters for psychophysical experiments are derived from the stimulus properties, the properties of the psychophysical procedure and external influences. In most cases, the number of possible parameters is too large to consider all of them. Therefore, a classification of parameters is useful to reduce experimental effort and ensure selective and useful results.

The first selection is made for *dependent variables*. These values are the primary outcome of the experiment and are normally constituted by the psychophysical measures investigated. Sometimes, other values like reaction times or subjective assessments of the test persons are also of interest. Only a few parameters can be chosen to be the *independent variables*, i.e., parameters that will be varied during the experiment because of the experimental effort. The majority of parameters constitutes the *controllable variables*, which can be measured or at least closely watched (keeping the test setup at a constant temperature, pre-selection of test persons based on age, body length and weight, etc.). If the values of controllable variables vary throughout the experiment or between subjects, they can be considered as covariates in the analysis of the results (see Section 5). All other parameters have to be considered as *confounding variables*, i.e., unwanted and uncontrollable influences on the experimental results.

Typical examples for parameter classification are given in Table 1. The distinction between independent variables and controllable variables is somewhat arbitrary: depending on the hypothesis, controllable variables from Table 1 can become independent and vice versa. Each independent variable v_i in an experiment is set to a discrete value $v_i \in \{v_{i,1}, v_{i,2}, \ldots v_{i,k}\}$ with normally $k_i = 2, \ldots, 7$ elements. Since the number n_T of variable combinations (or *treatments*) in the experiment is given by $n_\mathrm{T} = \prod_i k_i$, the exact variable definition has an impact on the length of the experiment. The exact values of $v_{i,k}$ have to be chosen with care—if $v_{i,k}$ is too close to $v_{i,k+1}$, one might underestimate the amount of the overall effect, if they are spaced too far, one might just miss significant effects. Selection of starting values can begin with the standard value of the parameter and find other values at the end of the acceptable range of the user or in the technically reasonable range. Since human perception can approximately be described on a logarithmic scale in most cases, values of obviously psychophysical parameters are supposed to be spaced logarithmically as well.

Table 1. Examples for possible variables in haptic and visual perception.

Type of Variable	Haptics	Vision
Dependent variable	Psychophysical construct in terms of JND, PSD etc.	
Independent variable	Stimulus parameter (frequency, intensity etc.), contact area [21], contact force, masking stimuli	Stimulus parameter (size, intensity), time of adaption
Controllable variable	Skin moisture, skin temperature [22], test person's age [23], test person's sex	Amblyopia, spectral power distribution (SPD) of the stimulus, adaption field
Confounding variables	Fatigue, experience of test person, other modalities, change of experimenter, non-thought-of variables	

Despite the above-mentioned classification of variables with respect to the hypothesis, there is another classification with respect to the test subject: *within subject* variables are variables that can be assessed in all values v_i without a dependence on the human subject in the experiment. *Between subject* variables are variables that are depending on the human subject, for example sex or training status. This classification of variables is needed for a thorough calculation of sample sizes, i.e., the number of test persons as outlined in the next section. Furthermore, this classification is also relevant for the selection of a proper data analysis procedure.

2.3. Measurement Procedure

In general, psychometric procedures assess the psychometric function (Figure 1) as a whole or a single point of it defined by a given detection probability Ψ. Different procedures are known in psychophysics research. Despite the classical works of Weber [2] and Fechner [24] and the measurement methods developed by them, modern psychophysics is further influenced by *Signal Detection Theory* [25,26], dealing with human decision processes in the brain and advances in statistical modeling. These works result in a variety of different psychometric measurement procedures that differ in their pre-requisites and capabilities.

Many procedures have already been analyzed for strength and weaknesses. We therefore suggest to choose a procedure from these well-known variants instead of developing their own measurement protocols as described in Section 2.3.3.

After a short introduction, some basic selection criteria are given in this chapter. In general, a psychometric measurement procedure consists of a psychometric method describing the presentation of the stimuli used in the experiment and a response paradigm that defines possible responses for the subject.

2.3.1. Psychometric Methods

A psychometric method will define the stimulus placement, i.e., which stimulus to use next, and the calculation rule for obtaining a threshold or other parameter from the responses of the test subject. There are three classes of these methods:

Classic Psychometric Methods were first developed by Fechner in the end of the 19th century. A straightforward approach is to repeatedly present a number of different stimuli in the expected range of perceivable stimulus intensities, normally about 100 to 200 presentations for a stable result. Based on the perceived stimuli, a psychometric function can be fitted with a number of different fitting algorithms [4,11]. This method is called the *Method of Constant Stimuli* and one of the methods with the least requirements on a priori knowledge about the psychometric function.

An adaption of this method is the *Method of Limits* aiming to focus on stimuli intensities in the middle part of the psychometric function. Test stimuli are not presented randomly, but in ascending or descending order. If the subject changes their response from "perceived" to "not perceived" (or vice versa), the descending presentation of stimuli will be stopped and an ascending order will start at a lower stimulus level until a positive response is recorded. The location parameter α of the

psychometric function is calculated as the mean of the stimuli at the end-points. Compared to the Method of Constant Stimuli, the Method of Limits increases efficiency, but delivers only a single point of the psychometric function.

The last classic method is termed the *Method of Adjustment*. When using this method, test persons adjust a freely controllable test stimulus to a given reference stimulus until both are perceived equally.

A drawback of these methods is the limited consideration of the responses of the subject during the course of the experiment. Classic psychometric methods are therefore prone to misplacement of stimuli and tend to be less efficient compared to the other classes.

Today, the Method of Constant Stimuli is used regularly in experiments where stimulus intensity cannot be adjusted continuously, which is a pre-requisite for the more sophisticated adaptive methods described below. Furthermore, new technologies such as additive manufacturing of haptic specimen allow new kind of experiments that are more easily conducted using the Method of Constant Stimuli.

To assess properties of stimulus identification and scaling behavior, the Method of Adjustment provides an intuitive approach to collect such kind of data. Because of the easy adaption of the subject's task, different hypotheses can be assessed with very similar experimental setups.

Non-Parametric Adaptive Methods are based on a given set of rules for stimulus placement and result calculation. The stimulus placement rule depends on the response of the subject in the course of the experiment. The most prominent non-parametric method is the *Staircase Method*, which is based on the Method of Limits. The current stimulus level is increased for each negative response of the test person and reduced for each positive response, leading to a threshold with a detection probability of $\Psi = 0.5$. Normally, the stepsizes are fixed. Extensions to this placement rule led to the nowadays frequently used *Transformed Up-Down-Staircase* [27]. This method changes the stimulus level only after a certain number of consecutive correct or false responses and therefore targets a threshold with a predefined detection probability of $P_\Psi \neq 0.5$. The number of selectable detection probabilities, however, is limited and the number of trails is slightly dependent on the chosen detection probability. A staircase-variation by *Kaernbach* introduces variable step sizes (*Weighted Up-Down-Procedure*, WUD, [28]) with good performance measures and a freely selectable detection probability.

Other adaptive methods such as *Parametric Estimation of Sensory Thresholds* (PEST, [29]) are based on statistical tests to decide which stimulus to place next, the number of sign changes between consecutive stimuli presentations (*Accelerated Statistical Approximation*, ASA, [30]), or a simple bracketing technique to minimize the interval containing the threshold (*Modified Binary Search*, MOBS, [31]). All adaptive methods are credited with a better performance than non-adaptive classical methods [32]. There are differences with respect to the possible convergence levels (fixed or arbitrary) and requirements on stimulus availability (fixed set or arbitrary adjustable). Performance metrics [33,34] show marginal differences only between these methods and a strong dependency on the experimental conditions.

Parametric Adaptive Methods do not focus on the determination of a threshold with a certain detection probability but on the identification of the parameters of a model of the psychometric function. Pre-defined parameter sets (priors) are tested for their ability to account for the responses of the test person. First approaches were made with a Maximum-Likelihood-Estimation of the priors [35], and further developments of this method led to the *Updated Maximum Likelihood Procedure* (UML) [36]. Another approach was made by *Kontsevich and Tyler* [37] by developing the Ψ-Method based on conditional probabilities for each of the prior distributions.

Parametric methods incorporate all trial results in their calculations and usually require less trials compared to non-parametric methods because of that. They also offer the possibility to investigate other parameters of the psychometric function more easily than other methods and have no requirements regarding the availability of the stimuli. They can be used in experiments with fixed stimuli size, i.e., a real specimen as haptic test object.

2.3.2. Response Paradigms

Paradigms define the way a subject has to respond to a presented stimulus. The easiest paradigm is the Yes-No-Paradigm (YN). With this paradigm, the subject states whether a stimulus has been detected or not. The experimenter, however, has no control about the decision criterion of the subject, i.e., the strength of a stimulus to be perceived as strong enough to justify a positive response.

The Signal Detection Theory (SDT) introduced by Green [25] deals with these decision processes and provides several means to deal with fluctuating decision criteria of the subject. SDT assumes a certain amount of sensory noise to be always present in the subject's neural system that has to be overcome by additional neural activity arising from the sensory processes involved in perceiving the stimulus. The theoretical framework of Signal Detection Theory provides several other interesting methods for use in psychophysical research, for example means to assess the performance of observers. For the scope of this paper, however, the development of so-called Forced-Choice-Paradigms is of central importance, since these paradigms were developed to overcome the problems with varying decision criteria of the subject. They are often referred to as *nAFC paradigms*, i.e. n Alternative Forced Choice Paradigms depending on the number n of possible responses. With these paradigms, the test person has to denote which of n alternatives contain a stimulus instead of just reporting a positive or negative response. Since the experimenter can control the placement of the stimulus, the correctness of the response is no longer depending on the decision criterion of the subject. These paradigms overcome the problem with changing decision criteria, at the cost of prolonged experimental time or more complex setups. They can generally be combined with all of the above-mentioned methods.

In the works of Kaernbach [38], the option of an unforced-choice paradigm was developed for the WUD method. In this case, test persons have another response alternative "I do not know" in addition to forced choice alternatives. This paradigm is supposed to alleviate the experimental procedure for untrained observers and shows a slight performance gain compared to the standard method. The concept was transferred by the first author of this work to Ψ and UML methods with an increase in performance [33].

2.3.3. Selection of a Procedure

Because of the significant better performance of adaptive procedures, new studies should use non-adaptive procedures only in case of fixed stimuli parameters (like gratings on real objects). Prins and Kingdom give a good insight in modern psychophysical techniques [11]. Furthermore, there are several freely available software packages to simulate and analyze psychophysical measurements, see [33,39] for an overview. Agreement of simulations and human behavior is demonstrated in several studies by Kaernbach [38], Karmali et al. [40], and Madigan and Williams [41]. Based on literature reviews and own simulations, we can give the following general recommendations for selecting a psychometric procedure:

- Parametric methods like Ψ [37] or the Updated-Maximum-Likelihood-Method (UML) [36,42] provide performance benefits for assessing a complete psychometric function. However, they require certain assumptions concerning the psychometric function that may not be available for every kind of experiment [35]. An alternative is the use of the classic Method of Constant Stimuli and finding an acceptable trade-off between accuracy and duration of the experiment.
- For the approximation of the $\Psi = 0.5$ point of the psychometric function, parametric methods are the best choice as well, if an assumption about the form of the function can be made. If that is not the case, one can use the wide-spread adaptive staircase methods, which are easy to implement and only rely on some weak assumptions. An alternative is using approximation methods for stochastic processes such as the ASA procedure [43] or the WUD method [28,38]. These are based on a more complex mathematical basis and can be set to an arbitrary detection probability. Furthermore, they rely on weak assumptions about the psychometric function only and are

therefore suitable for experiments with little a priori knowledge about the form and parameters of the psychometric function of the subject.

- For psychometric measures that cannot be described as a parameter of a psychometric function, other types of psychometric procedures must be used. Stevens gives several examples for magnitude estimation tasks [16], in addition to the Method of Adjustment described above.

Regarding the choice of response paradigms, several aspects have to be taken into account. The usage of forced-choice paradigms is supposed to minimize the effect of varying response criteria and ultimately ensures the validity of the obtained detection probability. This can increase the validity of experiments assessing design parameters. On the other hand, the usage of nAFC paradigms increases the duration of detection experiments with temporal order of the stimulus presentation by a factor of n. This is especially important for haptic and (to a lesser degree) for acoustic experiments, since stimuli in these domains can only be presented in temporal order, not in a spatial arrangement. Next to the duration of the experiment, observers will affect the choice of the number of alternatives: according to Jäkel et al. [44], naïve observers perform better with a four alternative forced choice paradigm compared to experienced observers that perform best with a two alternative forced choice paradigm.

Computer simulations of psychometric procedures show some general effects of nAFC paradigms [33]: using a small number of alternatives n and as such increasing the guess rate of the experiment, results in decreased efficiency, i.e., more experimental trials have to be conducted for a certain accuracy of the threshold estimation. On the other hand, accuracy increases for larger values of n. Several simulations identify the combinations of Ψ with a YN-paradigm and Transformed Up-Down-Staircase with an 1up-3down progression rule and a 3AFC forced choice paradigm [45,46] as preferable combinations for efficient and reliable estimation of perception parameters. For shallow psychometric functions, unforced choice paradigms show beneficial behavior in Monte Carlo simulations. A test of suitable configurations of the procedure with a pre-test or a simulation can be used to minimize the study duration (Figure 2).

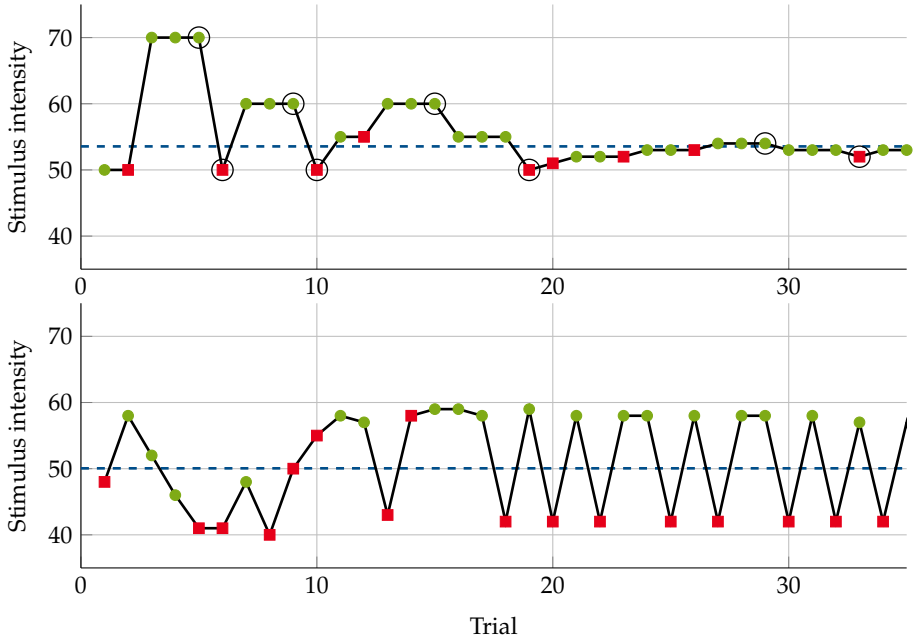


Figure 2. Examples of psychometric methods: Transformed Up-Down-Staircase with 1up-3down progression rule (**top**) and Ψ (**bottom**). Graphs are taken from a simulation and show the tested stimulus intensity for each trail. Red squares denote negative, green dots denote positive responses of the test persons to the stimulus. The dotted line shows the calculated threshold, and reversals of the staircase method are denoted by larger circles.

2.4. Subject Selection

After establishing a hypothesis, selecting the study variables and the psychometric procedure, the selection of subjects addresses the question of how much and what kind of subjects will take part in the study. While social studies and other scientific disciplines refrain from a pre-selection of subjects to obtain results with maximum generality, in engineering, one can narrow down the relevant target group in almost any case because of the intended application. Therefore, the authors conclude that a well-founded pre-selection of subjects is helpful for the derivation of design parameters.

The pre-selection only should be performed with regard to known external parameters affecting the psychophysical parameter in investigation. Normally, one will define the target group—for example well-suited persons in the age group of 40 to 65 years as buyers of a upper-class car—and derive requirements on the subjects from this point. In this example, one could reason that perception thresholds can be obtained from a subject group aged 30 years and younger, since the capabilities of haptic perception degrade with age (see [18]). If more psychological assessments such as subjective quality are under investigation, the age of the target group should be met more exactly, since social and psychological factors will probably have an influence on the results.

A pre-selection of subjects can also be performed to rule out known influences on the perceptual capabilities. This could include a limit on artificial sensory support systems such as the strength of glasses or contact lenses as well as the type of a hearing aid. In the case of haptics, several medical conditions are known to alter certain aspects of haptic perception. This can be used to screen subject bodies for these conditions, as for example shown for autism by Cascio et al. [47] or anorexia nervosa by Grunwald et al. [48].

It is advisable to compensate subjects monetarily or with other incentives to generate a higher motivation for test participation. Experience shows that motivated subjects will be more focused and produce better results.

Regarding the required **number of subjects**, no clear answer can be given. In general, one can assume, which the investigation of mainly psychological properties requires more subjects than the investigation of mainly psychophysical properties. A formal deduction of the required number of subjects can be done by considering the allowable β-error (type II-error, i.e., the probability of retaining a false hypothesis, sometimes also named the power of a study with Power $= 1 - \beta$) and the expected effect size of the influencing parameters. This measure is, for example, given by Cohens d in the case of experiments, where means of groups are compared or the population effect size ω^2 for ANOVA-type analysis. Keppel and Wickens give calculation examples for different types of experiments in [49]. These examples are, however, based on an estimation of the effect size expected, which is difficult to determine without prior knowledge from other experiments.

The authors recommend to focus on the investigation of medium effects. On the one side, most effects measurable with psychometric procedures are in fact medium and large effects. On the other side, this size is of practical importance, i.e., justifies consideration in the development process of a product. Despite the effect size, the form of the experiment and desired type I and type II errors will influence the number of subjects needed for a study. Tables to determine the correct number can be found for example in [49] or statistics software can be used. The *G*Power* software package by Faul et al. [50] is freely available for such power analysis of psychophysical experiments (Figure 3).

Two things have to be noted regarding the approach outlined above. First, sample size calculations do not take into account that (modern) psychophysical measurements will normally contain some type of averaging when calculating the threshold from the run data. This could be a basic approach to reduce the sample size of psychophysical experiments, but no thorough statistical consideration for this has been done so far. Second, the increase of treatment groups will reduce the required sample size for a given type II error. Hence, it can be wise to consider further external parameters in the design of a study (when considering the effect on the type I error).

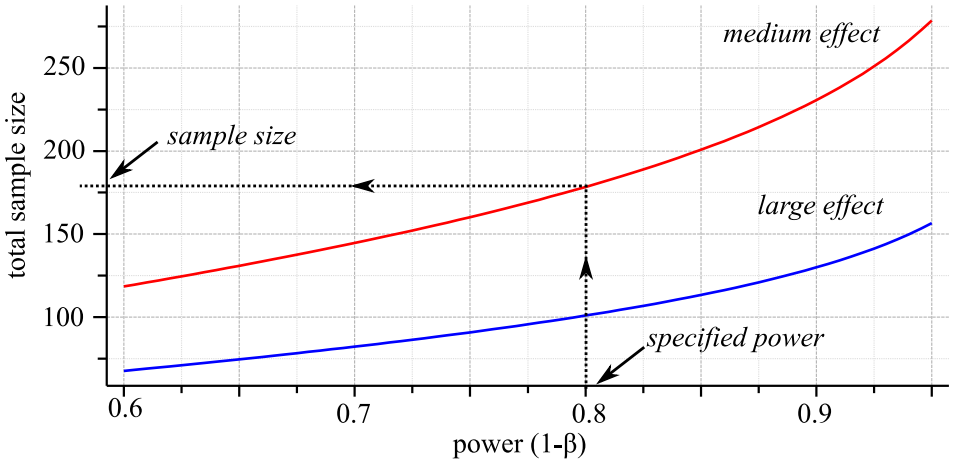


Figure 3. Calculation of sample sizes using *G*Power 3.19* for a two-factor, within-subject design with four treatment groups and a single dependent variable.

3. Measurement Setup and Errors

The **measurement setup** of a perception study has to fit well for the investigated perception parameter. This means, for example, that all parts of the measurement setup exhibit adequate frequency response, rated ranges and sampling rates for the expected values in the experiment. The design and construction of the setup has to be neat to prevent unwanted effects and errors such as errors induced by inadequate electromagnetic shielding. We recommend using fully automated measurement setups to prevent errors induced by the experimenter, i.e., for example by an automated data acquisition instead of reading the results manually from a faltering digital display. The setup has to be documented including all procedures and measurements of systematic and random errors (see below).

Usually, there are two types of **errors** in psychophysical experiments: errors in the perception of the test subject and errors of the measurement setup producing stimuli and measuring signals and answers. The former errors are subject of investigation of the experiment and can be minimized by carefully considering the perception-influencing variables as listed in Section 2.2. This will not yield an error-free measurement, but minimizes the uncertainty of the experimental result.

Errors of the measurement setup also diminish the explanatory power of a study and have to be analyzed thoroughly to identify the minimum amount of uncertainty that is associated with the measurement results. An analysis of systematic error propagation should be conducted as well as a calibration documentation of the setup and its components with known input signals and a null signal. Preferably, long time stability, reproducibility and random errors are also analyzed and documented. This can be done by applying the *Guide to the Expression of Uncertainty in Measurement (GUM)* [51] to the test setup. If an unacceptable large error is derived from this analysis, a larger sample size (i.e., more test subjects) can be used to increase the explanatory power of the experiment.

Systematic errors have to be analyzed and corrected before the experiment is conducted. A common source of such a systematic error in haptics is the moving mass of a test setup after the force measuring point. Other external sources of interference should be analyzed and considered in the analysis, if possible.

4. Conducting the Test

In most cases, a **pre-test** is recommended to verify assumptions about the psychometric procedures used and the statistical parameters used for the sample size calculation. The researchers working on the study will take part in the pre-test, but at least one person naive to the test should also take part in it to check the understandability of the test instructions.

Test instructions should be given in written form to the subjects to avoid unwanted distractions and priming of the subject. Despite this, one should keep in mind to assess the socio-demographic

data of the test-subject (if that is necessary for the data analysis) and get a formal consent about the data usage from the test subject. Although intended for medical studies, the *Declaration of Helsinki* [52] is a de-facto standard for experiments with human test persons and can be used as a gold standard for these procedures.

During all testing, **interactions with other sensual modalities** should be kept in mind and eventually controlled, for example by ear plugs and masking noise for haptic experiments. For vision, stray lights and insufficient adaption of the observer to the illumination of the setup are undesirable and have to be considered by technical means and adequate adaptation times.

When conducting experiments with different treatments, the **order of the treatment** for each test subject is supposed to be chosen in such a way, for which learning and habituation as well as systematic effects are minimized. Known methods from Design of Experiments (DoE) [53], i.e., randomization and blocking, as well as Latin Square Designs [54], which will propose a treatment order for each subject, can be used to minimize such effects over the entire sample. This is of utter importance, if the experiment is not fully automated, but relies on interaction with the experimenter. In this case, the experimenter has to be taken into account as a confounding factor.

5. Data Analysis

5.1. Checking the Data

All acquired data should be analyzed with proper statistical means, i.e., methods exceeding the calculation of a simple average. This implies that the type of statistical distribution of the results has to be tested. Data sets not included in the analysis have to be addressed and the criteria for this decision (i.e., measurement errors, dropout of the participant, etc.) have to be reported.

Commonly, data analysis based on the Nyman–Pearson model is used in psychophysics, as opposed to Fisher or Bayesian models. A further discussion about the different capabilities of theses methods is out of scope of this paper, but can be found for example in [55]. Most of these statistical analysis methods rely on normal distributed data, sometimes other requirements like equal variances for all treatments are required. There are a large number of different tests for normal distribution (χ^2, *Kolmogorow–Smirnow*, *Lillefors*, *Shapiro–Wilk*), which differ in their properties and capabilities. Since psychophysical studies often incorporate small sample sizes, the *Shapiro–Wilk* is a good choice with good test power and does not require prior knowledge of the parameters of the tested distribution [56]. If tests cannot assure the normal distribution of the data, transformations help to assure this requirement. Typical transformations include the logarithm, the square root and the reciprocal of data points [57,58].

The general location parameters of the results of the experiment are the main outcomes for experiments in the context of product design and engineering. These parameters (i.e., mean and standard error for normal distributed results, and median and inter-quartile range for non-normal distributed results) are reported in tabular form for the complete data set and for the sub-sets defined by the values of the main independent variables $v_{i,k}$. In most cases, other representations like boxplots or histograms can give a more intuitive insight into the data and do not rely on a specific kind of data distribution.

5.2. Checking the Hypothesis

As mentioned in Section 2.1.2, there are two basic experiment types in psychophysical research. For the first type, the assessment of a psychophysical measure, one will primarily use post hoc tests analyzing the difference between two individual treatments n_i. In this case, one will answer the question whether the results (i.e., the psychophysical measures) with treatment n_i are significantly different to the results obtained with treatment n_{i+1}. When the results of the different treatment groups are normally distributed, *Student's t-test* is the method of choice. The tests assess whether two samples are drawn from the same population or from different ones and does not rely on further pre-requisites

other than normality. If the results are not normally distributed, the *Wilcoxon rank sum test* can test for group differences. This test does not rely on a distribution, but only on the symmetry of the data compared to the distribution's median.

These pairwise tests work well for small number of treatment groups. If a large number of external parameters is included in the study, an *analysis of variance* (ANOVA) type test should be conducted. This group of statistical analysis assesses whether the variability of a result can be determined by the variability of treatment factors. Pre-requisites for ANOVA analysis are equal variances for all treatments and a normal distribution of the residuals, which also can be tested by means like Levene's test for example. Depending on the data structure, the number and types of parameters and the existence of covariates, there are different types of ANOVA analyses. The authors recommend a look into a textbook about statistics like, for example, [58] for the selection of the correct type of analysis.

ANOVA-type tests assure whether different levels of a parameter have a significant influence on the dependent variable, but do not show which factor levels have significantly different outcomes. Therefore, so called post hoc tests can be used to identify treatment groups with different outcomes. The selection of a suitable post hoc test also depends on the data structure and the number of treatment groups—even more than the ANOVA type tests. All tests will be conducted with specific software. While "all-purpose" engineering tools will include the necessary functions, the usage of a specialized statistics program is advisable—especially because their workflow includes the formalized tests for data pre-requisites and more convenient options for exporting the data. Common software packages are GNU R (R Foundation for Statistical Computing, Vienna, Austria), IBM SPSS™(IBM Corp., Armonk, NY, USA) or Minitab™(Minitab Inc., State College, PY, USA), the selection is based on individual preferences of the user as well as pricing and license issues.

5.3. Reporting the Results

Besides an exact outline of the study design and its methods, a description of the test setup and test persons (population and sample) are desirable. Based on this information, the study has to be repeatable. The steps undertaken for data analysis should be described briefly in the study report. If possible, a discussion of the results with respect to other studies with similar setups and intention should be made. When large differences occur, a detailed discussion of these differences and suggestions for further studies is advisable.

In addition to the location parameters of the data, the test output of all statistical tests has to be included in the study report. This includes the significance value (p in most cases), the appropriate test statistic and the effect size for ANOVA type analysis. Regarding the effect sizes, the reporting of ω^2 (estimator of the effect on the population) or η^2 (relative effect size in the sample including unaccounted variance) is preferred to the report of partial η_p^2 (relative effect size in the sample without unaccounted variance) [59,60]. The report of all of these values is beneficial to use the data for sample size calculations of subsequent studies [49] and for meta-analyses [61]. As many scientific journals allow digital supplements to publications, the supply of anonymous experimental data in such a form has to be considered.

6. Example: Haptic Perception of Viscous Damping of Rotary Switches

This chapter gives an example from haptic psychophysics that is placed in the context of automotive user interfaces. With increasing functionality of modern cars as well as a larger demand of entertainment functions, automotive user interfaces have to be able to provide an increased amount of functions while maintaining safe operating conditions for the driver. This leads to an demand of clearly distinguishable parts of the user interface, for example switches or rotatory controls. These can be defined by their mechanical parameters, i.e., reaction forces, travel or damping parameters.

A perceptual analysis of these parameters allows the definition of well distinguishable properties, when values are selected well above the differential threshold. Similarly, absolute thresholds can be used to define perceptually reasonable error margins for design optimizations.

The main purpose of the experiment is to measure thresholds for a specific mechanical parameter of rotatory switches–damping. The damping torque T_d as a function of the angular velocity $\dot{\varphi}$ and the linear damping coefficient d of a rotatory knob can be described as in Equation (7):

$$T_d(\dot{\varphi}) = d \cdot \dot{\varphi}. \tag{7}$$

Dependent parameters are the psychophysical parameters absolute threshold (AT) and differential threshold (DT) for the parameter d (Equation (7)). With regard to the automotive context of the experiment, typical parameters such as knob diameter, distraction of the subjects and detents (i.e., angle-periodic increase of reaction torque) in the characteristic torque-angle-curve of the control are selected as independent parameters. From prior research, an effect on the dependent parameters is expected from both the diameter of the knob and the adding of detents to the angle-torque-characteristic of the rotary knob. Whether observer distraction has an effect on the measured psychophysical parameters is one of the questions of this experiment: similar threshold values for cases with and without distraction of the observer for experiments like this allow for a direct transfer of laboratory measurements to real environmental conditions.

6.1. Measurement Setup and Stimuli

The measurement setup consists of a haptic rotary control simulator, which is built from a DC motor (model RE60, Maxon Motor AG, Sachseln, Switzerland), an incremental encoder (model 2RMHF, Scancon Encoders AS, Hillerød, Danmark) and a torque sensor (model DRFL-I-0,2, Wintec GmbH, Schorndorf, Germany). It is integrated in a real-time system (xPC, The Mathworks, Natick, MA, USA) with a data acquisition card (model NI PCI-6221, National Instruments, Austin, TX, USA) and a loop rate of 20 kHz.

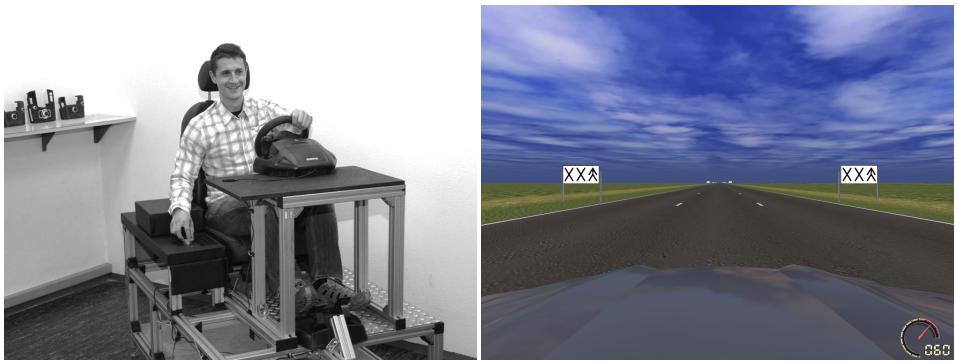
A nonlinear impedance control scheme is employed to display stimuli with adjustable detents and damping. This control scheme measures the actual angular position and velocity of the rotary knob and calculates the necessary motor output torque to resemble the impedance of the current state of the control element according to the angle-torque-characteristic [62] superimposed with the desired damping characteristic investigated in this experiment. In addition to the impedance control, an internal feedback loop minimizes friction and damping of the setup. With this control scheme, an inertia of $J = 28 \times 10^{-6}\,\mathrm{kg\,m^2}$ and a slide friction torque of $|T_{\mathrm{fric}}|$ = 3–5 mN is achieved.

The damping intensity under investigation is added to the static properties of the setup, thus comprising the stimulus used in the experiment. Other independent parameters are selected as follows:

- Knob diameter: $D_1 = 20\,\mathrm{mm}$ and $D_2 = 38\,\mathrm{mm}$.
- Detent profile: a high-grade detent profile with a maximum torque of 25 mN m, a slope proportion of 1:5 (rise to fall) and a spatial period of 18° can be superimposed on the stimulus. Experimental conditions are activated detent or no detent.
- Distraction: thresholds are either determined as a primary task (no distraction condition), or as a secondary task to a standardized Lane Change Test (LCT) [63] (distraction condition).

The detent condition is only tested in experiments with the LCT condition to reduce the effort of the experiments. With this limitation, parameter factors yield six different stimuli conditions for both absolute and differential thresholds. A value of $d_0 = 2\,\mathrm{mN\,m\,s}$ is selected as a reference stimulus for the differential threshold. This value corresponds to a clearly perceivable damping effect, comparable to stirring a pencil in honey.

The haptic simulator is integrated in a seatbox, including a force-feedback steering wheel (model Momo, Logitech, Apples, Switzerland) and a 21.5″ monitor (Figure 4). The latter are used for the lane change test.



(**a**) (**b**)

Figure 4. Seatbox with steering wheel (**a**) and screenshot of the lane change test (**b**). Monitor, headphones for acoustic shielding and haptic simulator are not shown in (**a**).

6.2. Subjects

The experiment was conducted with $n = 26$ subjects. For detecting medium effects, a test power of $1 - \beta = 0.8$, and a significance level of $\alpha = 0.05$ can be calculated. Subjects were recruited from students at the Hochschule Heilbronn. Four subjects were female, and all except one subject were right-handed. Median age was 25.5 years with a range of 23 to 67 years.

6.3. Measurement Procedure

A simple staircase procedure with a yes/no paradigm is used in this experiment [64]. This procedure converges at a detection probability of $\Psi = 0.5$, and results were calculated from six reversals of the staircase. Starting values for absolute thresholds were placed above the threshold as estimated in a pre-test. Starting test stimuli for differential thresholds were smaller than the reference threshold.

All subjects started with a training session for the lane change test. Half of the subjects started with experiments in the LCT condition. Further randomization is achieved by changing the order of control knob diameters. They were instructed to use a typical angular velocity for probing the rotary controls. To minimize unwanted distraction in the non-LCT condition, test persons wore headphones with music playing. Absolute detection thresholds were measured first for all test persons and conditions.

6.4. Results

Measurement results for absolute detection thresholds and Weber fractions are given as minimum, maximum, 0.25 and 0.75 quartiles, median, mean $\hat{\mu}$ and empirical standard deviation $\hat{\sigma}$ (Tables 2 and 3). Weber fractions are based on a direct comparison of the stimulus with a reference stimulus of 2 mN m s. Absolute thresholds are larger for larger knob diameters and increase slightly with the presence of distraction and with a detent in the characteristic curve (Figure 5).

Table 2. Absolute detection thresholds for damping in different conditions with knob diameter (D1/D2), distraction (LCT/-), and detent (D/-). Values are given in Section 6.1.

Condition	*Absolute Threshold* (mN m s)						
	Min.	$Q_{0.25}$	**Median**	$Q_{0.75}$	**Max.**	$\hat{\mu}$	$\hat{\sigma}$
D1	0.015	0.046	0.105	0.164	0.455	0.148	0.135
D1.LCT	0.005	0.07	0.135	0.225	0.305	0.149	0.087
D1.LCT.D	0.065	0.255	0.355	0.495	0.87	0.386	0.216
D2	0.035	0.135	0.235	0.35	0.455	0.245	0.128
D2.LCT	0.025	0.161	0.298	0.439	0.745	0.312	0.184
D2.LCT.D	0.065	0.386	0.54	0.703	1.395	0.598	0.349

Table 3. Weber fractions for damping in different conditions with knob diameter (D1/D2), distraction (LCT/-), and detent (D/-). Values are given in Section 6.1.

Condition	*Weber Fraction* (%)						
	Min.	$Q_{0.25}$	**Median**	$Q_{0.75}$	**Max.**	$\hat{\mu}$	$\hat{\sigma}$
D1	0.40	10.00	17.90	30.85	50.40	19.79	13.95
D1.LCT	0.85	17.10	22.30	31.56	44.15	22.84	10.72
D1.LCT.D	1.25	22.20	29.80	38.11	52.90	28.86	14.57
D2	1.65	15.00	18.75	27.90	48.75	20.91	10.96
D2.LCT	1.25	13.96	21.88	27.40	48.75	22.00	11.63
D2.LCT.D	1.65	20.01	29.58	41.03	70.00	30.98	17.36

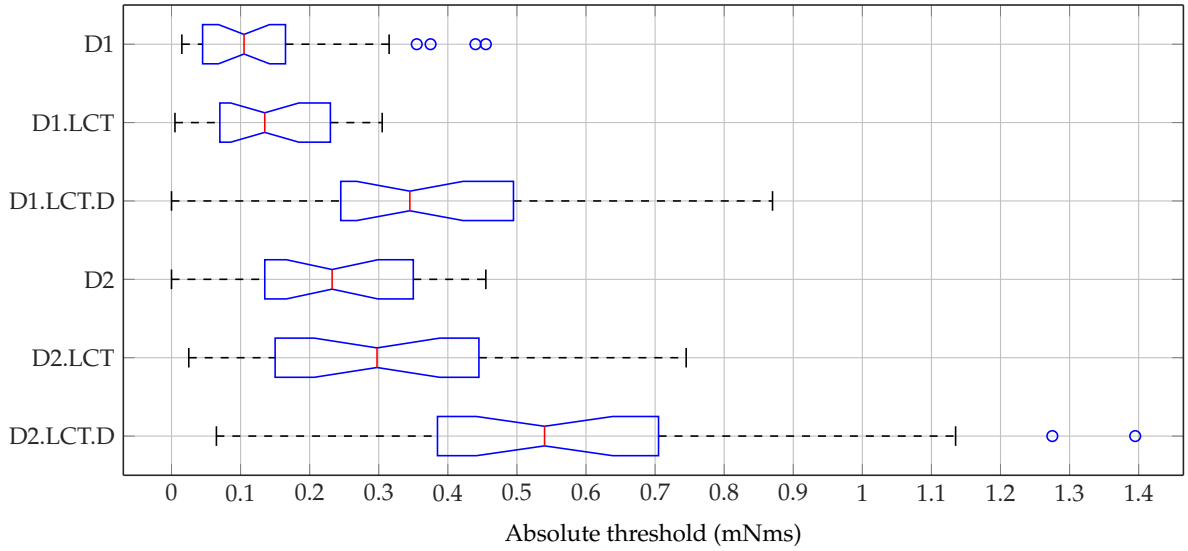


Figure 5. Boxplots of obtained absolute thresholds. Boxplots denote median (thick red vertical line) and interquartile range (IQR) between 0.25 and 0.75 quantile (blue box). Notches denote the confidence intervals ($\alpha = 0.05$) of the median values. Outliers (circles) are defined as data points with a distance of more than 1.5 IQR from 0.25 or 0.75 quantile, respectively, as indicated by the horizontal dashed lines. Stimulus conditions are coded with diameter (D1/D2), distraction (LCT/-) and detent condition (D/-). Boxplots show an increase in data span for distraction as well as detent condition. Non-overlapping notches of the central boxes hint at significant median differences of the data sets.

To assess the effect of independent parameters, the differences of the factor levels are calculated and analyzed. The differences for the distraction condition (Figure 6) are small, a *t*-test showed no significant differences of the threshold means of experiments with and without distraction ($p_{\mathrm{D1}} = 0.106$, $p_{\mathrm{D2}} = 0.989$). Threshold means for experiments with different knob diameter (Figure 7) are significantly different ($p = 0.003$, $p_{\mathrm{LCT}} = 0.001$, $p_{\mathrm{LCT,D}} = 0.001$). Further analysis of the data shows no significant difference of the relation of the thresholds compared to the relation of the knob diameter. In this case, a non-parametric Wilcoxon-Test is used, since the threshold relations are not normally distributed. The effect of adding a detent to the stimulus is significant (Figure 8, $p_{\mathrm{D1}} = 0.001$, $p_{\mathrm{D2}} = 0.001$).

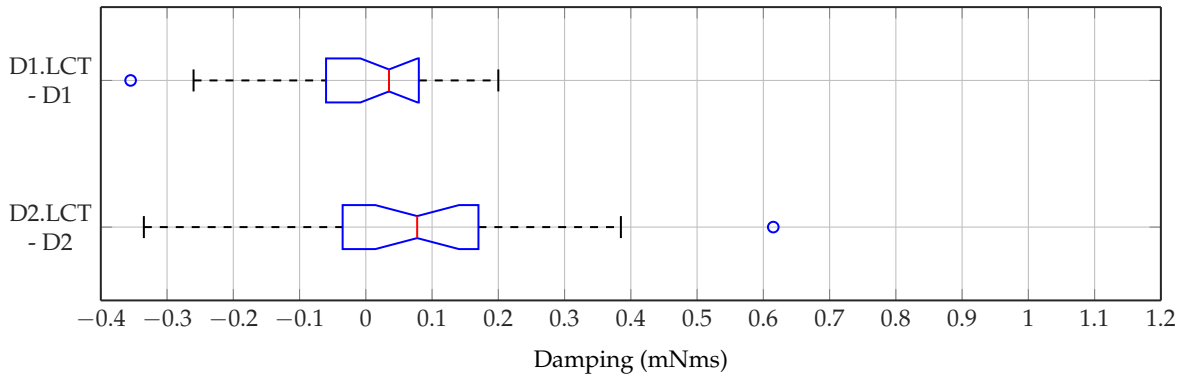


Figure 6. Effect of distraction on detection thresholds for damping with respect to knob diameter. Differences are not significant.

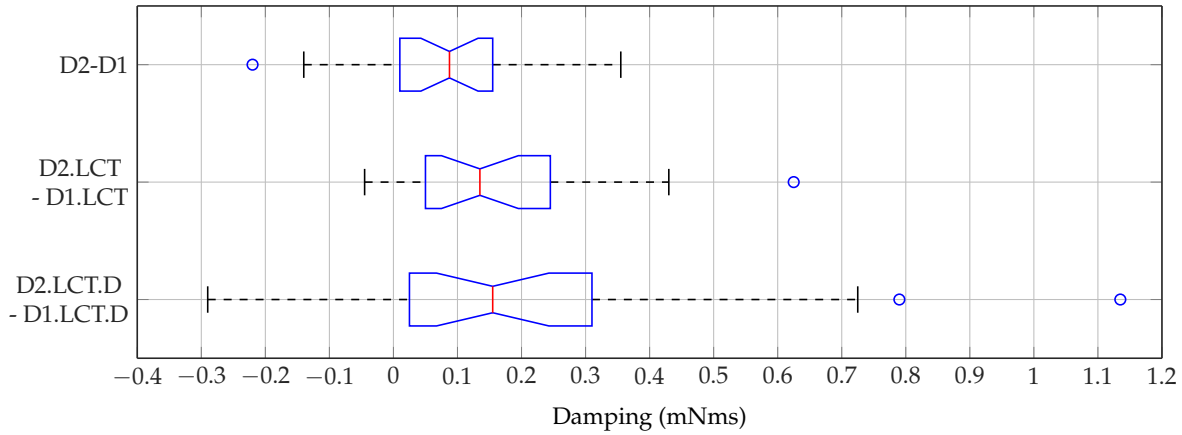


Figure 7. Effect of knob diameter on detection thresholds with respect to distraction and detent. Differences are significantly different from zero and are correlated to knob diameter.

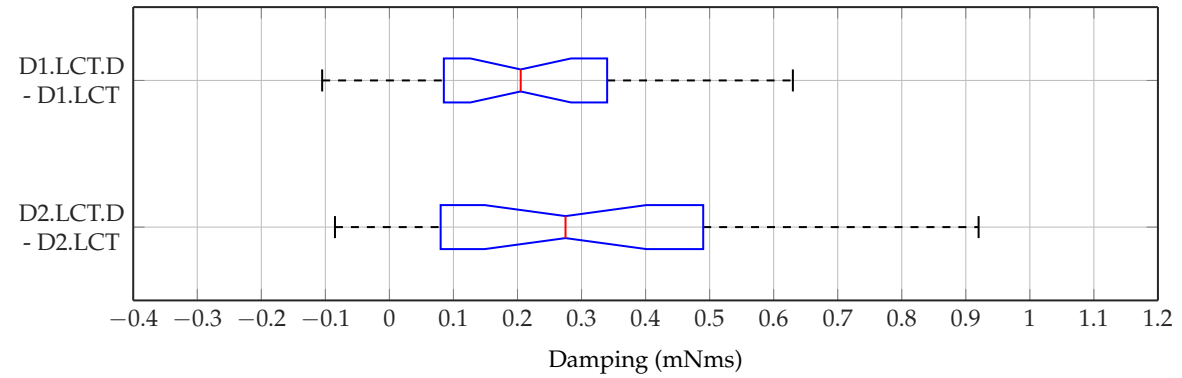


Figure 8. Effect of detent condition on detection thresholds with respect to knob diameter. Differences are significantly different from zero.

Results of differential thresholds (Figure 9) are evaluated the same way as absolute thresholds. Numerical values of the distribution are shown in Tables 2 and 3, a graphical representation is given by the boxplots in Figure 9. The effects of different experimental conditions are calculated for further analysis, but not shown here for brevity. They can be found in [6] (Chapter 7).

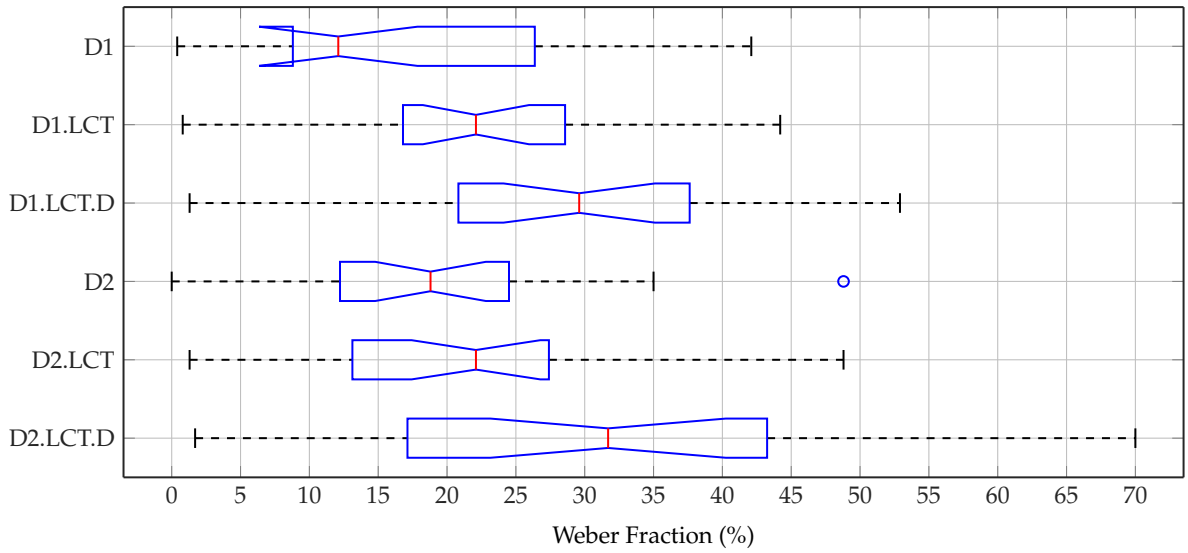


Figure 9. Boxplots of obtained differential thresholds given as Weber fractions.

A summary of the significance of effects with corresponding effect sizes is given in Table 4. In this experiment, only paired, within-subject tests are conducted, i.e., the same person will conduct the experiment with all of the different conditions. For such pairwise comparisons of condition A and condition B, the effect measure Cohen's d'' is calculated according to Equation (8):

$$d'' = \frac{\hat{\mu}_D}{\hat{\sigma}_D}, \tag{8}$$

with $\hat{\mu}_D$ and $\hat{\sigma}_D$ as mean and empirical standard deviation of the distribution of the differences $x_D = x_B - x_A$. This difference x_D is calculated for each participant and each comparison. As a convention, values of $|d''| \geq 0.14$ constitute a small effect, $|d''| \geq 0.35$ a medium effect and $|d''| \geq 0.57$ a large effect in this case [65]. Effect sizes in Table 4 are given as means over all experiment parts, i.e., for the effect of knob diameter as mean of the effects of the pairwise comparisons shown in Figure 7.

Table 4. Summary of effects and mean effect size.

Effect	Significance	Mean Effect Size d''
Absolute Threshold		
Distraction	no	0.172
Detent	yes	1.126
Knob diameter	yes	0.796
Differential threshold		
Distraction	no	0.164
Detent	yes	0.571
Knob diameter	no	0.098

The significant effects of detent and knob diameter for absolute thresholds and detent for differential thresholds are large effects in this experiment. The non-significant effect of distraction calculates to a small effect, but the experimental design does not include enough samples to assert this property with the given requirements on significance and test power.

6.5. Discussion

The experiments yield absolute thresholds for damping of rotary knobs ranging from 0.105 mN m s for knobs with small diameter to 0.540 mN m s for knobs with larger diameter, superimposed detent and a distracting primary task.

The effect of the *knob diameter* and the relation between the calculated absolute threshold from these experiments imply that tangential forces at the fingertip are the relevant parameters for the detection of damping properties. Absolute thresholds will therefore scale with the diameter of the knob, as confirmed by the significant effect of the knob diameter on the absolute damping threshold. No effect of knob diameter on the differential thresholds is assured in the experiments, which is expected based on Weber's Law for super-threshold reference stimuli.

Only *detents* have an effect on both absolute and differential thresholds, leading to higher thresholds in both cases. Absolute thresholds are almost doubled and differential thresholds increased by 50 %. These results imply a masking effect of detents on mechanical properties like damping and are further evaluated in [6].

The experiments further show that there is no considerably medium or large effect of *distraction* on the threshold values for damping. The transfer of experimental results for similar measures from a laboratory setting to real-world applications is hence considered valid.

Results of this experiment can be used to determine suitable or acceptable levels of damping in rotary knobs from absolute thresholds as well as limits for production variances from differential thresholds. The effect of knob diameter on absolute thresholds introduces a new design parameter and allows for trade-offs between knob size and the detection threshold of damping.

7. Conclusions

This tutorial outlines principles to design, conduct, analyze and report measurements of perception parameters for application in design and engineering purposes. It proposes orientations and guidelines rather than precise instructions, since each psychophysical experiment is different. Therefore, critical discussion of all parts of the experiment are needed during the planning. Precise and understandable instructions have to be given to the test persons for conducting the experiment and a statistical sound data analysis has to be conducted in order to obtain meaningful results. Finally, detailed descriptions of an experimental setup and reliable results are required in publishing any psychophysical study.

The authors are confident that following these recommendations will lead to results that are more reliable and therefore can improve technical designs and processes. Furthermore, the re-usability of reports and publications in research and application is increased beyond the originally planned purpose of the study.

Acknowledgments: This work was partly funded by the Deutsche Forschungsgemeinschaft (DFG) under Grant No. HA 7165/1-1.

Author Contributions: Christian Hatzfeld, Stefan Söllner, and Manuel Kühner designed the tutorial section; Manuel Kühner designed, performed and analyzed the experiments; Christian Hatzfeld, Stefan Söllner, Manuel Kühner, Tran Quoc Khanh and Mario Kupnik wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Kern, T.A.; Hatzfeld, C. The User's Role in Haptic System Design. In *Engineering Haptic Devices*, 2nd ed.; Hatzfeld, C., Kern, T.A., Eds.; Springer: London, UK, 2014.
2. Weber, E.H. *Tastsinn und Gemeingefühl*; Engelmann: Leipzig, 1905.
3. Gall, S.; Beins, B.; Feldman, J. Psychophysics. In *The Gale Encyclopedia of Psychology*; Gale: Detroit, Michigan, 2001.

4. Wichmann, F.; Hill, N. The psychometric function: I. Fitting, Sampling, and Goodness of Fit. *Percept. Psychophys.* **2001**, *63*, 1293–1313.
5. Kühner, M.; Wild, J.; Bubb, H.; Bengler, K.; Schneider, J. Haptic Perception of Viscous Friction of Rotary Switches. In Proceedings of the 2011 IEEE World Haptics Conference (WHC), Istanbul, Turkey, 21–24 June 2011; pp. 587–591.
6. Kühner, M. Haptische Unterscheidbarkeit Mechanischer Parameter bei Rotatorischen Bedienelementen. Ph.D. Thesis, Technische Universität München, Munich, Germany, 2014.
7. Do Thompson, M.R. Psychophysical Evaluations of Modulated Color Rendering for Energy Performance of LED-Based Architectural Lighting. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2007.
8. Zydek, B. Blendungsbewertung von Kraftfahrzeugscheinwerfern unter Dynamischen Bedingungen. Ph.D. Thesis, Technische Universität Darmstadt, Darmstadt, Germany, 2014.
9. Seeger, M.; Stein, T.; Schmidt, L. Vibrotaktile Wahrnehmung bei der Verwendung von Handschuhen. In *Mensch und Computer 2017—Tagungsband*; Gesellschaft für Informatik: Regensburg, Germany, 2017.
10. Hatzfeld, C.; Dorsch, S.; Neupert, C.; Kupnik, M. Influence of surgical gloves on haptic perception thresholds. *Int. J. Med. Robot. Comput. Assist. Surg.* **2017**, doi:10.1002/rcs.1852.
11. Prins, N.; Kingdom, F.A.A. *Psychophysics: A Practical Introduction*; Academic Press: Maryland Heights, MO, USA, 2010.
12. Macmillan, N.; Creelman, C. *Detection Theory: A User's Guide*; Lawrence Erlbaum: London, UK, 2005.
13. Colman, A. *A Dictionary of Psychology*, 3rd ed.; Oxford University Press: Oxford, UK, 2009.
14. Conner, M.; Booth, D.; Clifton, V.; Griffiths, R. Individualized optimization of the salt content of white bread for acceptability. *J. Food Sci.* **1988**, *53*, 549–554.
15. International Organization for Standardization (ISO). *ISO 20462-3: Photography—Psychophysical Experimental Methods for Estimating Image Quality*; ISO: Geneva, Switzerland, 2012.
16. Stevens, S.S. *Psychophysics*; Transaction Books: Piscataway, NJ, USA, 1975.
17. Theeuwes, J.; Alferdinck, J.W.; Perel, M. Relation between glare and driving performance. *Hum. Factors J. Hum. Factors Ergonom. Soc.* **2002**, *44*, 95–107.
18. Hatzfeld, C. Haptics as an Interaction Modality. In *Engineering Haptic Devices*, 2nd ed.; Hatzfeld, C., Kern, T.A., Eds.; Springer: London, UK, 2014.
19. Zwicker, E.; Fastl, H. *Psychoacoustics*; Springer: Heidelberg, Germany, 1999.
20. Drasdo, N. Vision Research: A Practical Guide to Laboratory Methods. *Brain* **1999**, *122*, 2000–2002.
21. Peters, R.; Hackeman, E.; Goldreich, D. Diminutive Digits Discern Delicate Details: Fingertip Size and the Sex Difference in Tactile Spatial Acuity. *J. Neurosci.* **2009**, *29*, 15756–15761.
22. Verrillo, R.; Bolanowski, S. The Effects of Skin Temperature on the Psychophysical Responses to Vibration on Glabrous and Hairy Skin. *J. Acoust. Soc. Am.* **1986**, *80*, 528–532.
23. Gescheider, G.A.; Bolanowski, S.J.; Hall, K.L.; Hoffman, K.E.; Verrillo, R.T. The Effects of Aging on Information-Processing Channels in the Sense of Touch: I. Absolute Sensitivity. *Somatosens. Mot. Res.* **1994**, *11*, 345–357.
24. Fechner, G.T. *Elemente der Psychophysik*; Breitkopf und Härtel: Leipzig, Germany, 1860.
25. Green, D.M.; Swets, J.A. *Signal Detection Theory and Psychophysics*; Wiley: New York, NY, USA, 1966.
26. Wickens, T.D. *Elementary Signal Detection Theory*; Oxford University Press: Oxford, UK, 2002.
27. Levitt, H. Transformed Up-Down Methods in Psychoacoustics. *J. Acoust.Soc. Am.* **1971**, *49*, 467–477.
28. Kaernbach, C. Simple adaptive testing with the weighted up-down method. *Atten. Percept. Psychophys.* **1991**, *49*, 227–229.
29. Taylor, M.; Creelman, C. PEST: Efficient Estimates on Probability Functions. *J. Acoust. Soc. Am.* **1967**, *41*, 782–787.
30. Kesten, H. Accelerated Stochastic Approximation. *Ann. Math. Stat.* **1958**, *29*, 41–59.
31. Tyrrell, R.A.; Owens, D.A. A rapid technique to assess the resting states of the eyes and other threshold phenomena: the modified binary search (MOBS). *Behav. Res. Methods Instrum. Comput.* **1988**, *20*, 137–141.
32. Leek, M.R. Adaptive Procedures in Psychophysical Research. *Percept. Psychophys.* **2001**, *63*, 1279–1292.
33. Hatzfeld, C.; Kupnik, M.; Werthschutzky, R. Performance simulation of unforced choice paradigms in parametric psychometric procedures. In Proceedings of the 2015 IEEE World Haptics Conference (WHC), Evanston, IL, USA, 22–26 June 2015; pp. 475–481.

34. Otto, S.; Weinzierl, S. *Comparative Simulations of Adaptive Psychometric Procedures*; Jahrestagung der Deutschen Gesellschaft für Akustik: Berlin, Germany, 2009; pp. 1276–1279.
35. Harvey, L.O. Efficient Estimation of Sensory Thresholds. *Behav. Res. Methods* **1986**, *18*, 623–632.
36. Shen, Y.; Richards, V.M. A maximum-likelihood procedure for estimating psychometric functions: Thresholds, slopes, and lapses of attention. *J. Acoust. Soc. Am.* **2012**, *132*, 957–967.
37. Kontsevich, L.; Tyler, C. Bayesian Adaptive Estimation of Psychometric Slope and Threshold. *Vis. Res.* **1999**, *39*, 2729–2737.
38. Kaernbach, C. Adaptive Threshold Estimation with Unforced-Choice Tasks. *Percept. Psychophys.* **2001**, *63*, 1377–1388.
39. Prins, N.; Kingdom, F.A.A. Palamedes: Matlab Routines for Analyzing Psychophysical Data. Available online: http://www.palamedestoolbox.org/ (accessed on 28 July 2017).
40. Karmali, F.; Chaudhuri, S.E.; Yi, Y.; Merfeld, D.M. Determining thresholds using adaptive procedures and psychometric fits: Evaluating efficiency using theory, simulations, and human experiments. *Exp. Brain Res.* **2016**, doi:10.1007/s00221-015-4501-8.
41. Madigan, R.; Williams, D. Maximum-Likelihood Psychometric Procedures in Two-Alternative Forced-Choice: Evaluation and Recommendations. *Atten. Percept. Psychophys.* **1987**, *42*, 240–249.
42. Green, D. Stimulus selection in adaptive psychophysical procedures. *J. Acoust. Soc. Am.* **1990**, *87*, 2662–2674.
43. Anderson, A.J.; Johnson, C.A. Comparison of the ASA, MOBS, and ZEST threshold methods. *Vis. Res.* **2006**, *46*, 2403–2411.
44. Jäkel, F.; Wichmann, F.A. Spatial four-alternative forced-choice method is the preferred psychophysical method for naïve observers. *J. Vis.* **2006**, *6*, 1307–1322.
45. Kollmeier, B.; Gilkey, R.H.; Sieben, U.K. Adaptive Staircase Techniques in Psychoacoustics: A Comparison of Human Data and a Mathematical Model. *J. Acoust. Soc. Am.* **1988**, *83*, 1852–1862.
46. Schlauch, R.; Rose, R. Two-, Three-, and Four-Interval Forced-Choice Staircase Procedures: Estimator Bias and Efficiency. *J. Acoust. Soc. Am.* **1990**, *88*, 732.
47. Cascio, C.; McGlone, F.; Folger, S.; Tannan, V.; Baranek, G.; Pelphrey, K.A.; Essick, G. Tactile Perception in Adults with Autism: a Multidimensional Psychophysical Study. *J. Autism Dev. Disord.* **2008**, *38*, 127–137.
48. Grunwald, M.; Ettrich, C.; Krause, W.; Assmann, B.; Dähne, A.; Weiss, T.; Gertz, H.J. Haptic perception in anorexia nervosa before and after weight gain. *J. Clin. Exp. Neuropsychol.* **2001**, *23*, 520–529.
49. Keppel, G. *Design and Analysis: A Researcher's Handbook*; Pearson Education: Old Tappan, NJ, USA, 1991.
50. Faul, F.; Erdfelder, E.; Lang, A.G.; Buchner, A. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* **2007**, *39*, 175–191.
51. International Organization for Standardization (ISO). *ISO/IEC Guide 98-3: Uncertainty of Measurement—Part 3: Guide to the Expression of Uncertainty in Measurement*; ISO: Geneva, Switzerland, 2008.
52. Declaration of Helsinki—Ethical Principles for Medical Research Involving Human Subjects. Technical Report, World Medical Association, 2013. Available online: http://www.wma.net/en/30publications/10policies/b3/ (accessed on 28 July 2017).
53. Montgomery, D.C. *Design and Analysis of Experiments*; John Wiley & Sons: Hoboken, NJ, USA, 2009.
54. Ai, M.; Li, K.; Liu, S.; Lin, D.K. Balanced incomplete Latin square designs. *J. Stat. Plan. Infer.* **2013**, *143*, 1575–1582.
55. Christensen, R. Testing Fisher, Neyman, Pearson, and Bayes. *Am. Stat.* **2005**, *59*, 121–126.
56. Razali, N.M.; Wah, Y.B. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *J. Stat. Model. Anal.* **2011**, *2*, 21–33.
57. Wallenstein, S.; Zucker, C.L.; Fleiss, J.L. Some statistical methods useful in circulation research. *Circ. Res.* **1980**, *47*, 1–9.
58. Field, A. *Discovering Statistics Using IBM SPSS Statistics*, 4th ed.; Sage Publications: New York, NY, USA, 2014.
59. Levine, T.R.; Hullett, C.R. Eta squared, partial eta squared, and misreporting of effect size in communication research. *Hum. Commun. Res.* **2002**, *28*, 612–625.
60. Olejnik, S.; Algina, J. Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychol. Methods* **2003**, *8*, 434–447.
61. Nitsch, V.; Färber, B. A Meta-Analysis of the Effects of Haptic Interfaces on Task Performance with Teleoperation Systems. *IEEE Trans. Hapt.* **2012**, *6*, 387–398.

62. Hogan, N. Impedance Control: An Approach to Manipulation. *ASME J. Dyn. Syst. Meas. Control* **1985**, *107*, 1–24.
63. International Organization for Standardization (ISO). *ISO 26022: Road Vehicles—Ergonomic Aspects of Transport Information and Control Systems—Simulated Lane Change Test to assess in-Vehicle Secondary Task Demand*; ISO: Geneva, Switzerland, 2010.
64. Kühner, M.; Bubb, H.; Bengler, K.; Wild, J. Adaptive Verfahren in der Psychophysik. In *Ergonomie*; Lehrstuhl für Ergonomie, Technische Universität München: München, Germany, 2012; p. 26.
65. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*; Academic Press: Cambridge, MA, USA, 1988.