On estimating the Difference Limen:
A comparison of the 2AFC and the reminder task

Dissertation
der Fakultät für Informations- und Kognitionswissenschaften
der Eberhard-Karls-Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

Vorgelegt von
MSc. Einat Lapid
aus Tel-Aviv

Tübingen
2008

Tag der mündlichen Qualifikation:	20.02. 2008
Dekan:	Prof. Dr. Michael Diehl
1. Berichterstatter:	Prof. Dr. Rolf Ulrich
2. Berichterstatter:	Prof. Dr. Thomas Rammsayer (Universität Bern)

Acknowledgements

Although the current study started as a simple comparison of psychophysical methods and procedures to measure the difference limen, as the work progressed, it turned to a voyage into the abstract field of discrimination strategies, which may or may not be employed by humans. And there is nothing simple about that...of course changing subjects from biology, in which 'what you see is what you get' to psychophysics, one must learn how to think in concepts instead of functional thinking. To think abstract instead of concrete, after all, the mind is the biggest mystery of them all, isn't it?

First and foremost I would like to thank Professor Rolf Ulrich who had the courage to take an inexperienced someone like me along to the ride and believe in me right from the very start. And what a ride it was. It would be an understatement to compare this to an emotional rollercoaster. Sometimes drowning under the 'sea of terminology' and concepts, and ideas. But there was Prof. Ulrich, who is always willing to answer any question, redundant or important. I really am thankful for his help and patience. Indeed, I wish to all doctorates a supervisor who both has huge knowledge, but most important has not less a passion to share this knowledge and to teach it.

Special thank to all the residents of the 4th floor. To Dr. Bettina Rolke, Dr. Hannes Schröter, Karin Bausenhart, Daniel Bratzke, Michael Steinborn, Tanja Seifried, Anja Fiedler, Tanja Leonard, Susana Ruiz Fernández, Verena Seibold , Judit Nitschke and Frau Monika Freitag. All of you have experienced my endless questions of just about everything and patiently answered them. Thanks for the fruitful discussions, the help with MrF, the help with MATLAB, comments on manuscripts, of course, the most needed help through German bureaucracy, the stories about East Germany, testing my experiments, and the willingness to help in general. And to make me feel welcomed in a foreign cold land.

I would also like to thank my HiWis, Andy Kramer and Marie-Luise Zeitler who have done a great job under my pressuring demands and to Roland Hirsch and Alexander Braun for technical support.

Of course to my family Aharon, Geula, Hadar, Lidor and Maymuy that always support me in my decisions how crazy and far out they may seem and last but not least to my husband Björn which deserves all the thanks in the world for everything, couldn't have done it without you.

Table of contents

1 Introduction

1.1 Basic concepts in psychophysics

The main concern of psychophysics is the relation between a stimulus and the associated sensation that this stimulus elicits in an organism. Understanding this connection is an important step in a continuous effort to reveal the mechanisms that underlay the mind. Within this effort, researchers are constantly in a search of fast, accurate and reliable methods that enable them to evaluate ones *performance*. Performance is a quantified response to a stimuli and it provides a mathematical connection between the inner (psychic) to the outer world (physical) (Corso, 1963; Gescheider, 1997; Guilford, 1954; Marvit, Florentine & Buus, 2003; Treutwein, 1995; Ulrich & Miller, 2004).

Performance reflects on one's *sensitivity*, that is, high performance is correlated with high sensitivity. Sensitivity in turn is defined as the ability of a person to detect a stimulus and discriminate between different stimuli. Sensory events are integral parts of the everyday life, thus the main application of psychophysics is to determine and compare the sensitivity of people across various stimulus conditions. Examples of sensitivity measures are abundant and include brightness, loudness, pitch and scent. Those sensations are correlated with light intensity, auditory intensity, auditory frequency and different concentration of smell molecules, respectively (Buss, Hall, Grose & Dev, 2000; Dunn, 2001; Linschoten, Harvey, Eller, & Jafek, 2001; Macmillan, & Creelman, 1991; Vogels & Orban, 1986).

In fact, psychophysics provides tools that could be used in several ways in everyday activities. Examples range from leisure activities, such as helping to improve and understand the acoustics of concert halls, so that sound is perceived more superiorly and enjoyable by the audience (Witew, Behler & Vorländer, 2005), to importance in diagnostic medicine. For example, measuring hearing and the efficiency of hearing aides after cochlear implantation (Cao & Wang, 2006; Donaldson, Viemeister & Nelson, 1997), or measuring the efficiency of treatment for impaired eye sight (Fronius, Cirinia, Cordey & Ohrloff, 2005). Further, in occupational therapy the therapists are interested in the person's sensory processing and experience and their influence on the person's decision making. Later they try to translate these data into

improving the life of the person by trying to lead him/her to a more satisfactory life (Dunn 2001).

Of those tools *sensory threshold* is a fundamental concept. It refers to the amount of stimulus that is needed in order to create a **conscious** experience in a subject. This term differs from the term *neurological threshold*, which is the amount of stimuli that is required for a neuron or a network of neurons to respond, but is not necessarily noticed by a subject. Sensory threshold include two of the key measurable parameters in psychophysics, the *absolute threshold* and the *difference threshold*, also termed the *difference limen* (DL) or the just noticeable difference (jnd). The absolute threshold refers to the smallest quantity of stimulus energy that could still create a sensation in a subject. The DL refers to the minimum change in the stimulus level (e.g intensity, frequency) that would result in a different sensation, that is to say, how much two stimuli must differ on some physical scale, for an subject to be able to discriminate between them. These thresholds are assumed to reflect one's sensitivity, so lower thresholds are connected with higher sensitivities (Gescheider, 1997; Macmillan, & Creelman; 1991, Hill; 2001). For example, the absolute thresholds for hearing are frequencies between a minimum of 20 Hz to maximum of 20,000 Hz. That means that a normal person will not hear the sound created by the vibrations below 20Hz or above 20000Hz regardless to how loud (dB) they are.

When measuring the difference threshold for loudness, researchers use a pair of stimuli that one of them refers to as the *standard* and the other as a *comparison*. For instance, if the standard has a loudness of 20dB and the subject can only tell the difference, that is, to say louder when the comparison is 22dB the DL will be 2dB (Gescheider, 1997; Birnbaum, 1994). If another person judge the comparison to be louder only when it is presented in 24dB his DL would be 4dB. These examples can illustrate differences in sensitivity between people, and as stated before, the lower DL reflects higher sensitivity and thus higher performance.

Measuring the DL in discrimination tasks enables the investigation of another interesting topic in sensory processing, namely, the relation between the DL to the stimulus size (e.g intensity, length, orientation etc.). Much literary work is dedicated to asses this connection to see if and how the DL is dependent on the stimulus size (e.g., Ekman, 1959; Getty, 1975; Killeen & Weiss, 1987). For example, if the DL is 4 units for a standard stimulus with a size of 16 units is it still 4 units when the

standard stimulus is higher, say 20 units? Or vice versa when the standard stimulus is lower, for example 10 units?

The ratio between the DL (also noted $\Delta\Phi$) in a specific discrimination task to the standard stimulus level of the task (Φ) is called the *Weber fraction* (WF) named after the German psychophysiologist Ernst Heinrich Weber. The Weber's law is depicted in equation as follow:

$$\Delta\Phi/\Phi = c \qquad (1.1)$$

where c symbolizes a constant. Thus, Weber's law suggests that the DL is linearly coupled to the size of standard stimulus and therefore the DL grows as the stimulus level grows. For example, if the law is applicable, and c is 10% for an auditory duration discrimination task, we would expect the DL to be 10 msec when the standard is 100 msec and 50 msec when the standard is 500 msec in duration.

One can also notice that small Weber fractions are associated with higher sensitivity. Thus, Weber fraction is a practical measure because it can be objectively compared across modalities such as visual, auditory, and somatosensory or conditions such as various stimulus intensities. Consequently, the relative sensitivity of the different sensory systems can be evaluated, as Weber fraction serves as an index for the discrimination power (Guilford, 1954; Kling & Riggs, 1971). It is broadly known today that the Weber fraction differs widely from sense to sense (modality), for example, duration discrimination of auditory intervals is much better than duration discrimination of visual intervals, demonstrated by lower Weber fraction for auditory modality (e.g., Grondin, 2001; Grondin, 2003; Grondin, Meilleur-Wells, Ouellette & Macar, 1998; Ulrich, Nitschke & Rammsayer, 2006) and also within modality when different methodologies are employed (e.g., Lapid, Ulrich & Rammsayer, 2008).

There is abundant evidence that the Weber Law is valid in wide range of stimuli intensities and modalities (e.g., Grondin, Ouellet & Roussel, 2001; Zeng & Shannon, 1999), however there are also indications of violation of the law. For example, it is well known that at low intensities of stimulus (e.g. for very short standard durations, very low light intensities) the Weber fraction is liable to increase considerably (Kling & Riggs, 1971; Grondin, 2003; Dawis, 1979). As well, in the field of time perception, it was reported by Drake and Botte (1993) that there is a maximum sensitivity range for intervals of 300-800

msec in which the Weber fraction is the lowest, and thus c is actually not constant. There are indications that the Weber fraction may be effected by the task it self. For example the Weber Fraction is lower for auditory tempo sensitivity. That is, discrimination is better with increasing number of intervals in the standards and comparisons to be judged. (Drake & Botte, 1993).

1.2 Classical psychophysical methods

Fechner (1860) was the first to suggest practical methods to estimate both the absolute threshold and the jnd (DL). Those methods are nowadays called the *Classical psychophysics methods* and although they were the base for numerous methods, they are still vastly employed today in psychophysics. Three methods exist, the *method of limits*, the *method of adjustment,* and *method of constant stimuli* (in Farell & Pelli, 1999; Gecsheider, 1997; Graham, 1950; Green & Swets, 1966; Kling & Riggs, 1971; Treutwein, 1995).

1.2.1 The method of limits

This method is quite often used for measuring the sensory threshold, for example the hearing threshold (audiometry) or the threshold of smelling (Linschoten et al., 2001; Gelfand, 1990). In this method the experimenter uses two different, often interleaved tracks, one descending and one ascending. The ascending track starts with a stimulus that is much below the subject's threshold, and the subject is asked to indicate whether he perceives the stimulus or not. In each trial the stimulus level is increased by small amounts, until the subject reports that the stimulus is detected. Detecting the stimulus terminates the ascending track. The descending track starts with a stimulus much higher then the threshold, and in each trial the size of the stimulus is being reduced by small amounts, until the subject report that he can no longer perceive the stimulus. When the sensation disappears, the descending track terminates. Each of the stimulus levels in which a track was terminated at can be used as a threshold estimator. Conventionally, several ascending and descending tracks are performed and the threshold is calculated as the average of the values that are found to terminate the tracks. These values are the transition points between detecting and not detecting a stimulus in the descending track and vice versa for the ascending track.

When estimating the difference threshold, there is an additional reference stimulus in each trial, referred to as the standard stimulus. In the descending track, the comparison stimulus is decreased in each trial by a small amount, until the comparison stimuli is no longer perceived as larger then the standard stimulus. In the ascending track the comparison stimulus is increased in each trial by a small amount, until it is no longer perceived as smaller then the standard stimulus. The DL is again calculated as an average across several tracks' termination values of stimuli.

1.2.2 The method of adjustments

This method can be applied for continuous or quasi-continuous stimuli only. This method is quite similar to the method of limits but the subject is much more active and controls the level of the stimulus him/herself. For example, changing a level of a pure tone (Hesse, 1986) or changing the level of contrast between visual stimuli (Smith, 1971). For the absolute threshold, the subject is asked to adjust the stimulus so it is just noticeable for him when the track is ascending, or until the sensation disappears if the track is descending. This procedure is repeated for a fairly large amount of adjustments, and the absolute threshold taken as the average of the settings. More often this method is used for estimating another important parameter in psychophysics, the *point of subjective equality* (PSE), in that case, the subject is asked to 'match' a comparison stimulus with a specific standard stimulus, until they are perceived as equal.

A popular 'matching' instruction is 'nulling' (Farell & Pelli, 1999). It is based on the assumption that the subject knows what a specific stimulus should be like and is asked to bring the comparison stimulus to this state, for example, adjusting a line to be straight or to null a motion of a luminance grating by adjusting the contrast (Cavanagh & Anstis, 1991). This procedure is repeated numerous times. Therefore the estimation will on some occasions be higher and on some occasions lower than the standard. This process will result in a group of estimations around the standard which is approximately normally distributed. In this case, the mean of the distribution will specify the PSE, and the measure of dispersion of the estimations around the standard, like the standard deviation is used for calculating the DL. If

the discrimination is good the estimations will be close together, and will be much more variable in the case of poor discrimination.

1.2.3 The method of constant stimuli

In contrast to the above described methods, this method consists of a limited number of values of the stimulus to be presented to the subject (e.g., several lengths, several frequencies). Those values are predetermined and are repeatedly presented to the subject throughout the experiment and therefore the method is named constant stimuli. Ideally the levels range between a stimulus that is hardly ever detected to a stimulus that is always detected. The rest of the levels are placed with equal gaps between the two. When estimating the absolute threshold the subject is asked to indicate whether he perceives the stimuli (yes) or not (no). When the difference threshold is estimated those predetermined levels are now referred to as the *comparison stimuli*. Those levels are presented against a specific stimulus that is referred to as *the standard stimulus* and are chosen to be placed around the physical size of the standard stimulus, both below and above it. The subject is asked to indicate whether these levels are different then the standard stimulus ('yes' response) or the same ('no' response), in other versions the subject may be asked to indicate whether these levels are smaller or larger in magnitude (e.g., shorter, longer) than the standard.

When assessing the absolute threshold, the researcher is searching for the stimulus level that was detected (answer 'yes') 50% of the times it was presented. When assessing the difference threshold the researcher looks for the specific comparison stimulus level that was judged as different from or larger than the standard stimulus ('yes' responses) 75% of the times it was presented.

1.3 General classification of the methods

The three classical methods provided a base from which vast amount of methods and experimental designs (procedures) were developed and are widely used these days. The various methods can be categorized according to several not mutually exclusive partitions.

1.3.1 Yes-No vs. Forced-choice designs

One distinctive partition is between the *yes-no* designs, that are derivates from the classical methods, and the *forced-choice* category of designs. Traditionally the main difference between those methods is the number of intervals presented in each trial. In the yes/no design normally one stimulus in one interval is presented in each trial, and the subject is asked to determine whether it is the target stimulus or not, or whether he detected the stimulus or not. The proportion of **positive** answers that are given by the subject is the response variable that correlates the desired percentage of 'yes' answers (an arbitrary level that is desired by the experimenter, for example, 50%) with a specific value.

In contrast, the *forced-choice* design commonly employs several presentation intervals in each trial. The intervals may be simultaneously presented in different locations and thus spatially separated, or they can be sequentially presented and thus temporally separated. In either case, only one interval contains the target stimulus, and the subject is asked to determine which interval it is. The order of the presentation of the intervals is random. In a less common version, one of n stimuli is shown, and the subject choose which stimuli it was by indicating one of the numbers between 1 to n. In contrast to the yes-no design, the percentage of **correct** response is now the response variable and the threshold is a value which correlates with a specific desired percent of correct responses. The number of alternatives is determined by the experimenter but most commonly there are two alternatives (Hill, 2001; Macmillan & Creelman, 1991; Treutwein, 1995). Originally the forced-choice methods were evolved in order to prevent the bias in response that could rise from habituation or the expectation of the subject regarding the stimuli such as in descending order of stimuli. Additionally the tendency of subjects to report 'yes' even though they have not actually perceived the stimulus was noticed (termed *response bias*). Forcing an answer on random interval was believed to control this response bias. However, forced-choice may also create a bias from a different direction, such as a tendency of subjects toward a specific interval (Gescheider, 1997; Green & Swets, 1966).

1.3.2 Discrimination vs. Detection design

Another distinct partition is between *discrimination* and *detection* designs. The core difference between detection and discrimination is

based on whether the reference stimulus is null (zero) or not. In fact, detection is a special case of discrimination when the reference stimulus is zero. For example if an experiment involves discriminating a tone from a background noise it will be called detection, however, if an experiment involves discrimination of a tone from another standard tone it will be termed discrimination (Klein, 2001; Kling & Riggs, 1971; Macmillan & Creelman, 1991, 2005; Treutwein, 1995).

1.3.3 Non-adaptive vs. Adaptive designs

The central characteristic to the non-adaptive design is that the stimuli levels are predetermined by the experimenter and are repetitively presented to the subject throughout the experiment, regardless of the subject's responses. However, in the adaptive design, the levels of the stimuli are entirely dependent on the subject's responses and will be changed accordingly. This third partition will be discussed in details in section 1.7.

Those partitions are rather artificial and nonetheless they are not firm. In fact, an actual experimental design is most likely to be a combination between two, or even all of the categories (McKee, Klein & Teller, 1985). For example an experiment can consist of 2 intervals in each presentation (2AFC), 1 interval is the standard and the other is a comparison. The comparison stimulus may have several predetermined values that are repeatedly presented to the subject in random order (constant). The subject is asked to report which interval contain the louder tone (discrimination). Alternatively, the subject may be asked to indicate whether the two tones that are presented to are identical ('yes') or different ('no').

In short, all the methods share the aim to measure the threshold regardless their name or classification and a single experimental design may be under several categorization. For illustration see figure 1.1.

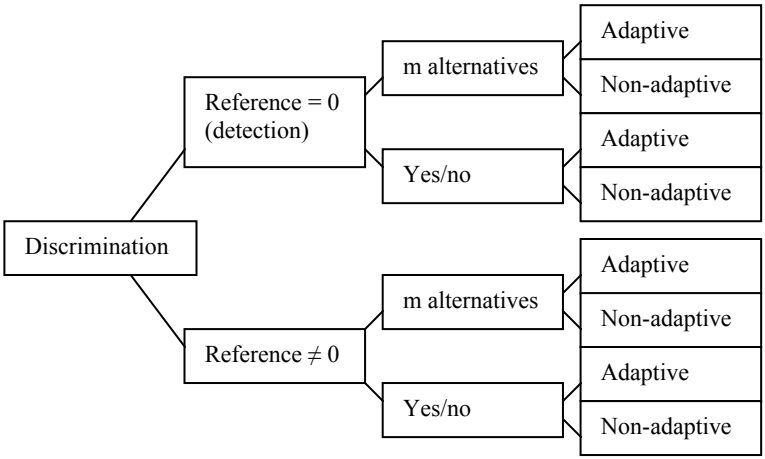

Figure 1.1: Illustration of the different partitions under which an experiment can be classified. m denotes the number of alternatives in each trial.

1.4 The classical threshold theory

The primary concern of the classical psychophysics, as mentioned before, was to establish the connection between the body and the mind, and especially to reveal the underlying relations between the intensity of the internal physical activity caused by external stimuli and their influence on the 'mental' activity. The conceptual tool that was, and still is, used to address these questions is the *threshold* (Guilford, 1954).

The previous section dealt with experimental methods to measure the *threshold* in order to transfer it to a quantitative mean. This section, however, will present the threshold in a somewhat more theoretical way, and will try to explain the essence behind this concept. For that matter the *absolute threshold* and the DL will not be treated as distinct concepts. In this case the absolute threshold may be regard as a special case of the difference threshold (Luce, 1963; Macmillan & Creelman, 1991).

The implicit assumption that stands in the base of psychophysics is the existence of at least two continua. One continuum is physical and the other psychological. The physical continuum represents any physical stimuli such as wavelength of light, frequency of a sound wave, weight, etc., that can be measured with physical units. The psychological continuum however is more complicated and abstract. It may include a physiological process, that is, the amount of neural excitation corresponding with a specific stimulus, a mental process that is sensory based and that corresponds to the neural excitation, and the judgment reported by the subject. The psychological continuum is also referred to as the response continuum and it is the sum of the above mentioned processes (Corso, 1963; Guilford, 1954; Thurstone, 1927). It is intuitive then, that the response continuum is narrower then the stimuli continuum, matched with the fact that some stimuli are too small to be perceived or to induce any response, while on the other side, some stimuli are too large and will not induce further response on a specific sensation scale, but transform to another kind of sensation. For example temperature receptors when a temperature exceed a certain magnitude (e.g., too hot or too cold) will transfer pain signals instead (Allchorne, Broom & Woolf, 2005).

Firstly it was believed that the threshold will be a sharp transition between a sensation to non-sensation within the subject. That would suggest a single boundary stimulus value, which below it a response will never occur and above it a response will always occur (i.e, the absolute threshold), or alternatively, that below it one response is elicited and above it another response is elicited (i.e., difference threshold) (Gescheider, 1997; Kling & Riggs, 1971). A stimulus that stimulates the receptors initiate a cascade of neural activity in the brain centres that increases with the intensity of the stimulus. Still, already then it was recognized that there is always some level of neural activity in the brain and therefore in order for a stimulus to be detected it had to create high excitation levels that surpass the already existing level of excitation. It was believed that for each sense a certain barrier should be overcome in order for the stimulus to be perceived consciously and from that view the threshold was named the *sensory threshold* (Corso, 1963; Green & Swets, 1966).

However quite early in empirical experimental results it was evident that the threshold is not a rigid fixed value on the stimulus continuum. That means that the transition from, for example, not perceiving a tone

to perceiving it is not a sharp cut-off that corresponds to a specific value. Moreover, it was apparent that the same stimulus presented to the same subject in different occasion did not yield the same response, for example 'yes I perceived it' or vice verse (Corso, 1963; Guilford, 1954; Green & Swets, 1966). Alternatively a subject is not even consistent in his judgment while comparing the same pair of stimuli repeatedly (Thurstone, 1927).

Therefore it is of primary assumption of classical threshold theory, that even if on certain moment the threshold is a sharp boundary, it will change from one moment to the other due to random fluctuations. That is to say, that for each point of time, there is a *momentary threshold* that a specific stimulus should exceed if it is to be detected. In each trial when a stimulus is presented to the subject it creates neural activity that depends among others also on the readiness and sensitivity of the receptors, on the intensity of the stimulus, state of adaptation and the background level of activity. If the neural activity induced by the presented stimuli exceeds this *momentary threshold* it will create a response in the subject. In consequence that stimulus level (when elicit a response) will now represent the *momentary threshold* (Boring, 1917; Dunn, 2001; Gescheider, 1997; Kling & Riggs, 1971).

In addition, it was also evidently clear that non-sensory factors affect the threshold as well. Among those factors are psychological and physiological features in the subject such as, tiredness, his readiness and attitude towards the experiment, and his understanding of the task. Also physical factors such as the delay in which a stimulus starts, the probability of the stimulus, as well as methodological factors may affect threshold measurements (Bausenhart, Rolke & Ulrich, 2007; Corso, 1963; Green & Swets, 1966).

This nature of the threshold explains why the same stimulus presented to the same subject will not bring out the same response, and for that reason why no single value of stimulus can be referred to as the threshold. From this follows, that an estimation of the threshold is determined over large amount of observations in each stimulus level. The value that is finally assigned to assess the threshold is **statistically** computed and represents the central tendency (e.g., a mean) of the distribution of the randomly fluctuating *momentary thresholds*. These fluctuations are assumed to be normally distributed. To put it simple, the threshold is based on the **probability** of the response. The value that is determined to be the threshold is a value that corresponds to an arbitrary

percentage of the times in which the stimulus was indeed perceived, conventionally 50% for *absolute threshold*. Traditionally, 75% of the times when two stimuli that are to be discriminated are perceived as different for the *difference threshold* but other percentages are used as well (Corso, 1963; Gescheider, 1997; Guilford, 1954; Kling & Riggs, 1971). The methods that were described in the previous section are all designed to statistically compute the value of the threshold and the variation of its distribution. The next chapter will elaborate on psychometric functions, the tool that is used for this computation.

To sum up this section it will be emphasized that both the *absolute threshold* and the *DL* in the classical psychophysics, are neither scaled nor measured as a magnitude of sensation that is created by a specific stimulus on the response continuum. They are however measured on the stimulus continuum such as the amount of energy of stimulus that could be detected or discriminated. This amount corresponds to a specific land mark on the response continuum such as specific percentage of positive answers. Still, although those measurements do give important information on the senses, they do not give a whole view on the sensory system (Gescheider, 1997; Guilford, 1954).

1.5 Psychometric functions

Psychophysics main focus is the relation between the stimuli to the impression it induces within the subject and his response to it (Boring, 1917; Klein, 2001). The wide variety of psychophysical methods and tasks all share the common goal of measuring the subject's performance. The connection between this focus and this goal is largely achieved by a basic and important tool of psychophysics, namely, the *psychometric function* (Hall, 1981; Miller & Ulrich, 2004). The *psychometric function* is in fact an analytic function that is assumed to describe the correlation between the probability of a certain response P(c) and some physical aspect of the stimuli such as the intensity of the stimulus, or its length. An observed psychometric function plot illustrates the probability of a certain response (e.g., 'yes') on the y-axis, as a function of the intensity of the stimulus on the x-axis (abscissa) (Klein, 2001; Miller & Ulrich, 2001; Wichman & Hill, 2001). Figure 1.2 is a hypothetic example of an experiment that employs a detection task.

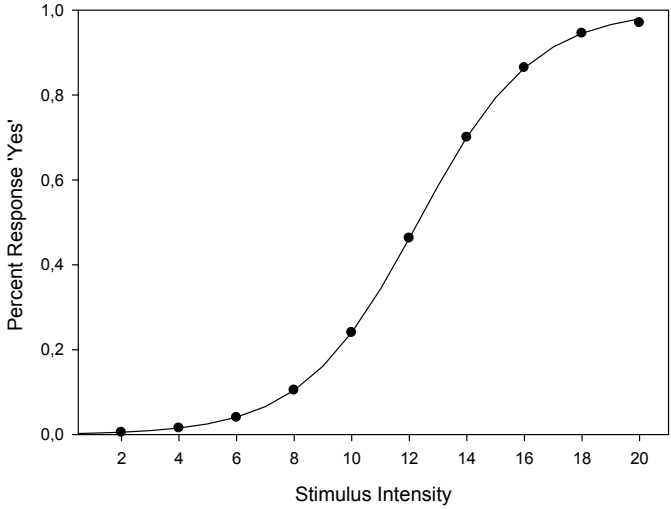

Figure 1.2: Typical but hypothetical psychometric function that is obtained when employing the yes- no task with a constant stimuli method in order to measure the absolute threshold. The black dots represent the percent of the response 'yes I detect it' by a subject as a function of the stimulus intensity. A function then is fitted to the dots.

The subject is presented with one of ten different stimulus intensities in each trial and is asked to indicate whether he detects the stimuli (the response is 'Yes') or he does not detect it (the response is 'No'). Each of the stimulus intensities is presented to the subject several times. The graph shows only the probability of the response 'Yes' that was given by the subject for each of the intensities of the stimulus. A psychometric function is then fitted to the observed data. Typically the psychometric function increases with the intensity of the stimulus in a detection task and the intensity of the comparison stimuli relative to the standard stimulus in a discrimination task (Hall, 1981; Leek, 2001; Miller & Ulrich, 2001). As mentioned before the conventional detection threshold is that stimulus value that is detected 50% of the times. In figure 1.2 the threshold is the intensity value that corresponds to 12.3 units.

As is depicted in figure 1.2, the fitted psychometric functions when measuring performance normally follow s-shape called an ogive. This shape is the cumulative density function stemming from the underlying probability distribution and it is assumed by most models used to describe or predict the psychometric function (Gescheider, 1997; Levitt, 1970; McKee et al., 1985; Simpson, 1988; Wichman & Hill, 2001). When specific stimulus intensity is presented to the subject it creates a certain magnitude of sensory reaction. This magnitude is changing from trial to trial according to some probability distribution. This reaction is in turn compared to some criterion and the percentage of positive response in each of the stimulus intensities, is the probability that this magnitude exceeded the criterion (Miller & Ulrich, 2001).

The idea that supports this notion can be clarified by using the threshold concept as an example and the phi-gamma hypothesis as one of the explanations and justification of using a cumulative distribution. The phi-gamma hypothesis is based on the curve of error (Boring, 1917; Kling & Riggs, 1971). If figure 1.2 shows a measurement of the detection threshold of a given stimulus, each of the intensities on the x axis will be detected only at those times that the magnitude of the sensory reaction they elicited is larger then the momentary threshold (Gescheider, 1997). The factors that affect the evoked magnitude fluctuate randomly from moment to moment and therefore, as previously stated, the same stimulus intensity will not elicit the same reaction repeatedly. Namely the same stimulus intensity once will be detected and once not.

For simplicity, Figure 1.3 shows an assumed situation that on any given time for any stimulus intensity, there are six factors that effect the stimulus impression. Each of the factors can either dispose against or towards making an impression, in this case detection. There is a single possibility that all of them work towards (+3) and a single possibility that they all work against detection (-3). There are six possibilities that one factor will be against while the other five factors are towards detection (+2) and six possibilities of the exact opposite, that one factor dispose towards while the other five factors dispose against detection (-2). There are fifteen possibilities that two factors will dispose against detection while four factors dispose towards detection (+1), and vice versa that four factors will dispose against and two factors towards detection (-1). Finally, there are 20 possibilities that three factors dispose against and three factors dispose towards detection (0). There

are all together 63 possibilities (cases) in this example (Boring, 1917). In Figure 1.3 the bars represents the frequency of each case from the general number of possible cases. The bell-shaped curve, A, shows that the chance or random dispositions distribute more or less normally and that the chance level is highest when the number of factors that dispose against detection is equal to the number of factors that dispose towards detection. Curve B shows the cumulative form of curve A, which is actually the frequency of cases that occurred until a certain value. Each point on the curve includes the frequencies of the preceding points.

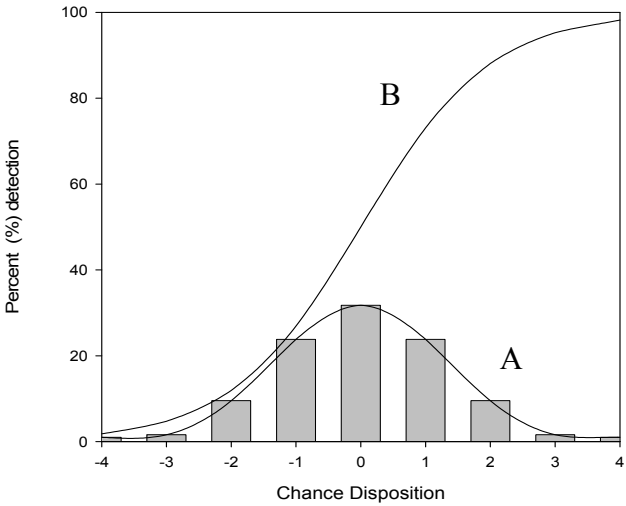

Figure 1.3: Distribution of six chance factors that either work against or towards detection (Boring 1917). Each grey bar represents the frequency of occurrence of a specific number of factors that either dispose against or towards detection. For example -3 shows the frequency of occurrence when all 6 factors work against detection, while +3 shows the frequency of occurrence when all 6 factors work towards detection. The rest of the bars show the frequency of occurrence of the cases in between. Curve A represents the fact that chance factors are normally distributed and curve B is the cumulative form of curve A.

One can notice that the curve in figure 1.2 is identical to curve B in figure 1.3 with the exception that in figure 1.2 the x-axis represents the intensity values of the stimulus instead of chance dispositions. If we now superimpose figure 1.2 on figure 1.3B then the stimulus value 12.3 of figure 1.2 corresponds to the 0 on curve B in figure 1.3. The stimulus values between 14 and 20 correspond to +1, +2,+3 and +4 respectively, and the stimulus values between 10 and 2 correspond to -1,-2,-3 and -4 respectively. It is now clear why the frequencies of detection as a function of every level of the stimulus, or simply the *psychometric function*, is referred to the phi-gamma hypothesis where phi refers to the probability of response and gamma to the stimulus intensity (Gescheider, 1997).

To conclude this section, as stated before, the threshold is traditionally taken as the stimulus value that corresponds to 50% on the curve. This is also the maximum point on the curve of errors. Simply put, the value that corresponds to this point has the maximum chance to either be detected or not, it is the transition value.

1.5.1 Psychometric function as a model

The psychometric function can not be observed directly but can be indirectly deduced from the experimental data. The common way is to assume an analytical specific relationship between the underlying probability of the response 'yes' or any positive response, to the stimulus intensity (Miller & Ulrich, 2001; Wichman & Hill, 2001). Often when modelling the psychometric function it has the following form:

$$P(x) = \gamma + (1-\lambda-\gamma)\, F(x) \qquad (1.2)$$

In which, λ represents the rate of lapses that is independent of the stimuli intensity, γ represents the lower asymptote of the function or simply the guess rate, $(1-\lambda)$ represents the upper asymptote. $F(x)$ denotes the underlying detection function that is independent of λ and γ and which determines its shape. The detection function ranges from 0% to 100% (Hill, 2001; Klein, 2001; Strasburger, 2001).

In practice, only one or two parameters that are in the main of interest are computed from the observed psychometric function ($F(x)$). These parameters are aspects of the psychometric function and they summarize the subject's performance. The parameters are the location

(threshold) and the scale (slope) of the psychometric function (Hill, 2001; Leek, 2001; Treutwein & Strasburger, 1999). For example, in both detection and discrimination tasks the location of the psychometric function refers to the value which yields 50% of positive responses. However, in the detection task this value commonly specifies the absolute threshold while in discrimination task this value specifies the PSE. Measuring the PSE can give information about the perception of a certain stimuli and measure how certain design (Meese, 1995) or experimental manipulation can affect (e.g., elongate, shorten) the stimulus perception (McAuley & Kidd, 1998; Tse, Intriligator, Rivest & Cavanagh, 2004). The scale, namely the slope parameter, refers to the steepness of the function and it is the inverse of the spread of the function. It is most important in discrimination tasks where it is a measure of the DL. Thus, the steeper the function and less spread, the smaller the DL is.

The previous example (in Figure 1.2) however is true when the function that is used to estimate performance is full, that is, it ranges from 0% to 100% and the chance level at the lower asymptote is zero. There are cases which will be described next in which the psychometric function ranges differently.

It was previously stated, that out of the four parameters that assume to describe the underlying psychometric function, only the threshold and sometimes the slope are estimated in practice. The parameters that are not estimated λ and γ are predetermined to have a certain value. The lapsing rate λ is usually set at zero, however, when investigating its influence on slope estimation, Wichman and Hill (2001) allowed it to vary but restricted it to low values between 0 and 0.05. The guessing parameter γ is dependent on excepted chance level of performance. Consequently, in the yes-no design in which the expected chance level is zero, γ equals zero, whereas in the forced-choice designs the expected chance level is dependent on the number of the alternatives, namely, number of intervals that are present in each trial. Accordingly, in the forced-choice design γ is equal to 1/m, with m represents number of alternatives (Hall 1981; O'regan & Humbert, 1989; Treutwein & Strasburger, 1999). As an example, for the two alternative forced-choice (2AFC) design $\gamma = ½$. It is straightforward to conclude then that the function of the 2AFC is not a full psychometric function and that it ranges between ½ and 1, in contrast to the full psychometric function show in Figure 1.2 for the yes-no design that ranges from 0 and 1. In

fact the 2AFC function corresponds to the upper part of the full yes-no psychometric function, hence the mid point of this function corresponds with 75% of correct responses which in turn specify the absolute threshold in detection tasks and the difference limen (DL) in discrimination tasks. Figure 1.4 depicts theoretical psychometric function in the two alternative forced-choice designs.

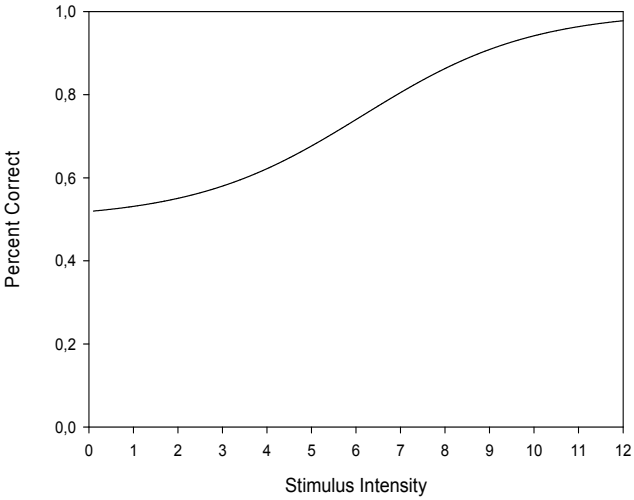

Figure 1.4: Form of the psychometric function employed in the 2AFC design. The function shows the percent of correct answers as a function of the stimulus intensity. It is notable that the psychometric function ranges from ½ to 1.

1.5.2 The underlying detection function

In the literature a wide variety of two parametric functions are employed in order to estimate the threshold and the slope. The three most common ones are the cumulative normal function used by for example McKee et al. (1985), Foster & Bischof (1987), Lages & Treisman (1998), the logistic function that is actually an approximation of the normal function but easier to compute has been use for example by Hall (1981), Linschoten et al. (2001), Marvit, et al. (2003), O'regan & Humbert (1989), Simpson (1988) , and Treutwein & Strasburger

(1999), and finally the Weibull function that was employed by Simpson (1995), Watson & Fitzhugh, (1990) and Wichman & Hill (2001) among others. However, the theoretical reasons to use one or the other function to fit the data are not of importance to the current study and will not be discussed. It is important to note that in practice the decision to use any one of the sigmoidal functions has little affect on the both slope and threshold estimations, with smaller influence on the threshold estimation (Wichman & Hill, 2001). The current study employed the logistic function and it will be formulated in the method section of experiment 1.

1.6 2AFC versus reminder task

As previously stated several methods are available in order to estimate the DL. The term method will be replaced now with the term task for the rest of the thesis. Two of the most common tasks are the *reminder task* and the *two-alternative forced choice* (2AFC) task (Macmillan & Creelman, 2005). The reminder task is often called the *method of constant stimuli*. The term "method of constant stimuli" (MCS) may create some confusion. As mentioned above, this term originates from the fact that an experimenter determines the set of comparison stimuli before running an experiment and uses this set throughout the experiment (Woodworth & Schlosberg, 1954, p. 200). In fact, this definition would also apply to the 2AFC method when the set of stimuli is predetermined or even to the method of single stimuli (MSS) which is a version of the MCS that is also commonly used. This method is identical to MCS only the reference (standard) stimulus is omitted from each trial and mostly is presented only at the beginning of an experiment (Grondin, 1993; Lages & Treisman, 1998; N'Diaye, Ragot, Garnero & Pouthas, 2004)

For this reason, throughout this thesis the term "constant stimuli" will not be used. Instead I will use the term "non-adaptive" when the set of stimuli is predetermined and the term "adaptive" when the stimulus set is determined by the response history of a subject during an experiment. These latter terms, however, will be discussed in details in the next section.

Both in the reminder task and the 2AFC task the subject is presented with a standard and a comparison stimulus on each trial. Moreover, several comparison durations are employed. In the reminder task, the standard is always presented first, followed by the comparison.

In fact the name reminder comes from the fact that the first interval i.e. the standard serves as a reminder to the subject of the interval to be judged (Macmillan & Creelman, 1991; Morgan, Watamaniuk & McKee, 2000). On each trial, the magnitude of the comparison can be smaller, equal to, or larger than the magnitude of the standard (e.g., the duration of the comparison may be shorter, equal to, or longer than the duration of the standard). At the end of each trial, the subject is asked to report whether the comparison was longer or shorter than the standard by responding "longer" or "shorter", respectively. Alternatively the subject may be asked to indicate which of the two intervals contain the larger stimuli (e.g., the longer duration). In both cases, the result generated by this reminder task can be displayed in the form of a psychometric function, by plotting the proportion of "longer" responses against comparison duration. Typically, this function is full and resembles the shape of an ogive curve that is zero at small comparison values, and approaches one for large values. This function is demonstrated in figure1.2. A psychometric function from a certain parametric family (e.g., logistic function) is then fitted to the observed data points. Finally, the DL is estimated as half the interquartile range of this fitted function, that is, $DL=(x_{.75}-x_{.25})/2$, where $x_{.25}$ and $x_{.75}$ denotes the value of the comparison that yields 25% and 75% "longer"-responses, respectively. The steeper this psychometric function is, the smaller is DL, and thus the higher is the differential sensitivity of the subject (see Luce & Galanter, 1967). The values $x_{.25}$ and $x_{.75}$ are usually used and the most common alternatives are $x_{.29}$ and $x_{.71}$, which arises in transformed up-down procedures (see section 1.7) (e.g., Bode & Carhart, 1973; Levitt, 1970

Although, the 2AFC task also employs a standard and a comparison on each trial, this task for estimating DL differs greatly from the reminder task. There are two subtle but crucial differences between these two tasks. First, whereas with the reminder task, the standard always precedes the comparison, the standard and the comparison are presented in random order across trials in 2AFC. Second, in the reminder task, the comparison can be smaller, equal to, or larger than the standard. By contrast, in the 2AFC task, the comparison is always larger than or at least equal to the standard and the subject is asked to indicate whether the first or second stimulus is larger (e.g., in the case of a duration discrimination task, the subject indicates whether the first or second interval appeared longer). In contrast to the reminder task, in the

2AFC task the experimenter simply notes whether the response is correct. The proportion of correct and not proportion of 'longer' responses is displayed as a function of the comparison stimulus intensity on the psychometric function. As appose to the reminder task, however, this proportion increases from the chance level of .5 when the value of the comparison is equal or very close to the value of the standard up to 1.0 for very large comparisons. This psychometric function was demonstrated in figure 1.4. As before, a psychometric function from a parametric family is fitted to the observed data points (e.g., a scaled logistic function). Based on this function, one determines the difference threshold, which is usually defined as the comparison value at which the proportion of correct responses is .75 (e.g., McKee et al., 1985). The DL is finally obtained by subtracting the value of the standard from this estimated threshold value. That is, the DL indicates the value of the comparison (i.e. DL + value of standard) at which the comparison is correctly detected in 75% of the time.

In psychophysical research, one or the other of these two tasks is exclusively employed to measure the DL. For example in research on time perception and tempo-sensitivity, DL is sometimes estimated with the 2AFC task (e.g., Ahmed, Lewis & Maurer, 2004; Drake & Botte, 1993; Grondin, 1993; Grondin et al., 2001; Karmarkar & Buonomano, 2003; McAuley & Kidd, 1998; Nagarajan et al.,1998; Rammsayer & Lima, 1991; Wright, Buonomano, Mahncke, & Merzenich, 1997; Wright & Sabin, 2007) and sometimes with the reminder task (e.g., Getty, 1975; Grondin, 2001; Jones & McAuley, 2005; Miller & McAuley, 2005; Rammsayer & Ulrich, 2005; Tse, et al., 2004; Ulrich et al., 2006).

This practice in research may reflect a common implicit assumption that the two tasks should result in more or less the same estimates of DL. For example Wright et al., 1997 directly compare the Weber fraction they estimated with the 2AFC method to the Weber fraction reported in Getty 1975. The later was estimated by employing the reminder method. In fact the Weber fraction reported in Wright et al. 1997 is slightly more then double as the one reported in Getty's study. Literature search on different studies with temporal discrimination of filled intervals task[1],

[1] Duration discrimination has been studied with empty and filled intervals. Empty interval is a quiet duration marked in the beginning and the end of the interval, while filled interval means that the duration of the interval is filled with the stimulus.

employing those different methods revealed a potential effect on the Weber fraction. Since the WF is often calculated as the DL divided by standard stimuli, disagreement of WF might suggest a discrepancy between DL estimations. For example Grondin (1993), Grondin et al. (1998) and Grondin et al. (2001) found the WF to be approximately 13%, 9% and 11% respectively. In contrast the WF was approximately estimated to be 5% and 7% by Getty, (1975) and by Rammsayer & Ulrich, (2005) respectively. The former WFs are in fact achieved by using the 2AFC task and the later by the reminder task. Additionally in tempo sensitivity Miller & McAuley (2005) used the reminder task and compared their results to Drake & Botte's (1993) which used the 2AFC task. In this case the results were similar, namely, the two tasks yielded Weber fraction of about 5%. However, the stimuli employed in those studies were empty intervals.

In general many studies in psychophysics are compared regardless the method that is used in them, but giving the inconsistencies mention above together with the fact that the 2AFC and the reminder are two tasks that enjoy great popularity in the literature, it is important to find out whether these method are indeed comparable, that is, do they provide similar estimation of the DL or not. Alternatively is this disagreement a result of the different subjects which participated on the various studies? Since the author is not aware of a previous empirical study that has systematically addressed this matter, the main goal of the present study was to provide data for assessing this issue. Hence, in the experiments reported below, both the reminder task and the 2AFC task were employed in order to estimate the DL within the same group of subjects to evaluate whether this implicit assumption holds. Moreover, in order to assess the stability of the DL obtained with each task, I tested the subjects on two occasions and computed the test-retest reliability (cf. Linschoten et al., 2001).

For measuring the DL, I used a duration discrimination tasks with two modalities, auditory (experiments 1 -5) and visual (experiment 6). In addition, two visual discrimination tasks were employed, random-dot pattern (experiment 7) and line-discrimination task (experiment 8).

1.7 Adaptive vs. non-adaptive procedures

A yet another important issue arises when an experimenter is about to design an experiment. Except from the decision of whether he/she will use the reminder task or the 2AFC task for measuring DL, additional essential decision needs to be made whether the data should be collected by means of an adaptive or a non-adaptive psychophysical procedure. With a typical non-adaptive procedure, five or more levels of the comparison value are predetermined by the experimenter around the threshold region and administered to the subject several times and in random order. With adaptive procedures, however, levels of the comparison are not predetermined but governed by the participant's response history. For example, the levels may change according to a pre specified rule after specific number of wanted answers (e.g., right or wrong) depending which point or points on the psychometric function are to be estimated (Levitt 1970).

In psychophysics literature both of the procedures are widely in use regardless the method employed in the study and examples are abundant for data collected adaptively (e.g., Ahmed et al., 2004; Buss, et al., 2001; Drake & Botte, 1993; Grondin et al., 2001; Karmarkar & Buonomano, 2003; McAuley & Kidd, 1998; Nagarajan et al., 1998; Rammsayer & Lima, 1991; Stellmack, Viemeister, & Byrne, 2004; Ulrich et al., 2006) as well as for data collected with a non-adaptive procedure (e.g., Berens & Pastore, 2005; Grondin, 2001; Jones & McAuley, 2005; Mcgavren, 1965; Miller & McAuley, 2005; N'Diaye et al., 2004; Schwartz, 1990; Thompson, Schiffman & Bobko,1976, van Oeffelen & Voss 1982)

Adaptive procedures are essentially a version of the classical method of limits discussed above. However, unlike the method of limits, the adaptive procedures are not terminated after a reversal in the subject's response, but instead the direction (e.g., intensity) of the stimulus level is reversed (Leek, 2001; Levitt, 1970; Treutwein, 1995). For example in discrimination task a correct answer can decrease the comparison stimulus intensity towards the standard stimulus level in order to make it difficult for the subject to discriminate between them, and a wrong answer will increase its intensity relative to the standard stimulus, and therefore the discrimination will be facilitated. These increments or decrements in the stimulus level are called *step size* and they might be equal in size for the regular up-down, and the transform

up-down (TUD) (Levitt 1970) or unequal in size like in the weighted up-down procedure (WUD) (Kaernbach 1991). The difference between these is the rule that controls the stimulus level (e.g., when to change the level) for example in regular up-down after each wrong or right answer the stimulus level is increased and decreased respectively. While in TUD wrong answer results in increase of the stimulus level but only either after two or three right answers stimulus level will be decreased. These procedures results in estimation of slightly different point on the psychometric function, 70.7% or 79% correct respectively. The WUD employ different step sizes to increase or decrease the stimulus level and is designed to estimate the 75% correct. Though all procedure are in use, traditional definition of threshold refer to 75% positive response and therefore this study employed WUD as will be further discussed below.

Adaptive procedures enjoy widespread use in psychophysics because they are designed to avoid trials with an inefficient placement of comparison values, that is, values that are either too small or too large, a thing that could happen with the non-adaptive predetermined procedure. Alternatively though not frequently, it could happen that the predetermined values are totally out of the subject's scale and additional pre-testing is needed in order to place the stimulus levels correct (for a review see Levitt, 1970; Treutwein, 1995; Simpson, 1988). Thus, in contrast to non-adaptive procedures, adaptive procedures are designated to concentrate the levels around the presume threshold and therefore to rapidly extract relevant information from a psychometric function that underlies discrimination performance without weaken accuracy or waste time (Emerson, 1984; García-Pérez & Alacalá-Quintana, 2005; Leek, 2001).

Comparisons of adaptive and non-adaptive procedures have sometimes been performed with computer simulations (e.g., Alacalá-Quintana & García-Pérez, 2005, García-Pérez & Alacalá-Quintana, 2005; Simpson, 1988; Watson & Fitzhugh, 1990) with conflicting results as will be discussed next paragraph. However, surprisingly little behavioral work has been directed to the question whether adaptive and non-adaptive procedures differ in estimating DL.[1] Empirical work is

[1]There are some studies, however, that employed a single fixed stimulus level within a block of trials, which can be regarded as a special case of the method of constant stimuli. These studies indicate lower discrimination performance with fixed-level than with adaptive procedures (for a review see, Leek, 2001).

important to validate the knowledge that had been reached by computer simulations. Such simulations require several assumptions that may be violated in practice. For example, simulations assume that the underlying psychometric function is unaffected by perceptual learning, or that the perceptual outcome on a certain trial is independent from the outcome of the preceding trials (see Leek, 2001, p. 1288).

A few empirical studies provide some clues on this issue (Brand & Hohmann, 2002; Dai, 1995; Hesse, 1986). According to these studies as well as to simulated studies, an efficient method would require relatively few trials to achieve a certain level of accuracy for estimating discrimination performance. Unfortunately, these studies provide a rather inconsistent picture of results. For example, Brand and Hohmann (2002) reported that adaptive procedures are more efficient than non-adaptive ones, Hesse (1986) that non-adaptive procedures are more efficient, and Dai (1995) that both procedures are equally effective. The same pattern of inconsistency is reflected with simulated data. While Simpson (1988) concluded that non-adaptive procedures are as efficient as the adaptive ones, Watson and Fitzhugh (1990) reported that adaptive procedures are much more effective then the non adaptive ones. Alacalá-Quintana & García-Pérez (2005) employed either fixed-length procedures or non-fixed and found them to have as little as neglected difference.

In conclusion then, and perhaps somewhat surprisingly, decisive empirical work on comparing adaptive and non-adaptive procedures is still lacking. Comparison of adaptive and non-adaptive procedure is typically concentrated in plotting the standard deviation and mean of the distribution of estimates as a function of number of trials. Hence, information on whether the two procedures in fact give the same DL estimates within fixed and equal number of trials is especially lacking. Therefore, the second goal of the present study was to address this issue. Specifically, I employed test-retest reliability methodology to assess temporal stability of the DL estimates using the both adaptive and the non-adaptive procedure. This study employed the weighted up-down procedure developed by Kaernbach (1991) for two reasons. First, Kaernbach's procedure provides an especially simple tracking algorithm that is easy to implement in an experiment. This may explain its increasing popularity in psychophysical research. Secondly, experimental work with this procedure has been often performed on duration discrimination (e.g., Grondin, 1993; Grondin, Ivry, Franz,

Perreault, & Metthe, 1996; Rammsayer, 1992; Rammsayer & Ulrich, 2005; Ulrich et al., 2006). In this domain of research, usually several experimental conditions are run within a single experiment and each condition requires the estimation of DL. Therefore, it is important to know for future research whether this approach provides sufficiently stable estimates of DL.

1.8 Overview and objectives of the current study

To sum up, this study had two major goals. The first was to assess whether DL estimates from the 2AFC task and the reminder task are of the same magnitude and the second was to assess whether the estimates yielded by the two main procedures of data collecting in psychophysics, namely adaptive and non-adaptive procedures are again equivalent. In order to evaluate potential differences between these two tasks and the two psychophysical procedures, each task was combined with an adaptive and non-adaptive procedure. In order to reduce the burden for the subjects, it was decided to decompose the complete factorial design, i.e., Task (reminder vs. 2AFC) × Procedure (adaptive vs. non-adaptive) × Order of Blocks (first vs. second), into two feasible designs. In Experiment 1, I excluded the 2AFC task from the complete design. In Experiment 2, I only excluded the non-adaptive version of the reminder task from the complete design. Thus, Experiment 1 together with Experiment 2 still enables a comparison between the 2AFC and the reminder task. Experiments 1 and 2 did not reveal any effect of procedure (adaptive or non-adaptive), rather a discrepancy between the two tasks (2AFC vs. reminder).

Therefore, Experiments 3 - 5 were designed to test specific hypotheses that emerged from the results of Experiments 1 and 2, and intend to point out the reason for that discrepancy. Experiment 5 investigate whether this discrepancy generalize to a different modality, and Experiments 7-8 are designated to further investigate whether these results also are generalized to non-temporal stimuli. After apparently identifying the reason for the observed discrepancy, that is to say, the presentation order of the standard and the comparison stimuli, mathematical models of three discrimination strategies are suggested and discussed as the underlie base of discrimination behaviour and its implication on the threshold estimation. An especially promising model that may account for the observed results is suggested. This model

assumes a use of internal standard as a base for comparison and most important it also suggests the manner of how this standard is created.

2 Experiments employing temporal stimuli

The following Experiments 1-6 employ temporal stimuli. In all tasks subjects are to discriminate the durations of auditory or visual stimuli. The threshold for duration discrimination is widely investigated using both tasks, as discussed in the introduction. Therefore, it is appropriate to use duration discrimination to compare the two tasks.

2.1 Experiment 1: Reminder task: adaptive vs. non-adaptive procedure

Subjects performed the reminder version of the duration discrimination task using filled auditory intervals. On each trial, a standard tone of 500 msec preceded a variable comparison tone, which could be shorter, equal to, or longer than the standard. At the end of each trial, the subject had to indicate which of the two stimuli was longer, i.e. the first or the second one. Each subject performed four consecutive blocks of trials. In the two adaptive blocks, the trial-to-trial changes of the comparison duration were governed by Kaernbach's (1991) weighted up-and-down procedure. There were two interleaved trial runs. One run estimated the 25% level of the psychometric function and the other run it's 75% level, which is referred to as the *doublet* procedure (see Leek, 2001). In the two remaining non-adaptive blocks, the duration of the comparison was sampled on each trial from a fixed set of pre-specified durations that were symmetrically arranged above and below the standard duration of 500 msec. Following previous suggestions (García-Pérez & Alacalá-Quintana, 2005; Hall, 1981; Leek, Hanna, & Marshall, 1992), I used a maximum likelihood analysis to estimate DL (and also the PSE) from the collected data in each block for each subject. Test-retest reliabilities were computed by correlating the DL estimates from the first and second block of each procedure. In a second set of analyses, I employed the Spearman-Kärber technique (Miller & Ulrich, 2001) to summarize each participant's psychometric function. In contrast to the maximum likelihood approach to estimation of DL, this technique does not require any assumption about the shape of the underlying psychometric function for assessing the slope of an observed psychometric function (see Miller & Ulrich, 2001, for a comparison of these techniques).

2.1.1 Method

Subjects. A group of thirty volunteers 24 female and 6 male students of the University of Tübingen (mean age ± SD: 22.7 ± 4.1 years) participated in a single experimental session that lasted approximately 60 min. They had normal hearing and were naïve about the purpose of the study.

Apparatus and stimuli. A PC controlled the presentation of the stimuli and the recording of the participants' responses. Auditory stimuli used for all tasks were temporal intervals of white noise that were generated by a SoundBlaster-compatible soundcard, and were presented binaurally via headphones (Philips SBC Hp 200) at an intensity of 85 dB SPL.

General procedure. The time course of a single trial was identical for the adaptive and the non-adaptive procedure. A trial started with the presentation of the 500-msec standard interval. 1,000 msec after the offset of the standard, the variable comparison interval was presented. At the end of the trial, subjects were asked to decide which of the two intervals was longer, the first or the second one. Subjects were not informed about the existence of a standard. They pressed the left-shift key of a computer keyboard when they judged the first interval as longer and the right-shift key when they judged the second interval longer. The experimenter emphasized accuracy over speed. Two seconds after the response, the next trial started with the presentation of the standard. The experiment consisted of four blocks each containing 110 trials. Previous research has indicated that at least 100 trials are required for satisfactory estimation of DL (see Leek et al., 1992). Two blocks involved the adaptive procedure and the remaining two blocks the non-adaptive procedure. There were six possible orders of the four blocks (i.e. AANN, ANNA, ANAN, NAAN, NANA, NNAA, with N=non-adaptive, A=adaptive) and these six orders were counterbalanced across the 30 subjects. To make sure that the subject understood the general task, a practice block was presented at the beginning of the experimental session (but not before each block). In order not to give advantage to one task over the other, the practice block consisted of 20 trials with a standard interval of 400 msec, in contrast to the experiment that engaged standards of 500 msec. In any case, no information was given on the duration of the standard.

Adaptive blocks. Durations of the comparison followed the weighted up-down rule (Kaernbach, 1991). Two step sizes were employed, 10 and 30 msec (in ratio of 1:3), that were kept constant throughout the experiment as in previous work (e.g., Rammsayer, 1992; Ulrich et al., 2006). Two separate interleaved runs (55 trials each) of the comparison duration were employed. One run targeted at the 75% level of performance and the second run at 25%, since these two target levels are the most common ones for estimating the DL. The starting value for the 75%-run was set at 600 msec, i.e. well above the standard duration. Whenever the comparison stimuli was judged as longer, its duration was decreased by 10 msec, and increased by 30 msec when it was judged as shorter. An exactly opposite procedure was employed for the 25%-run. That is, whenever the comparison was judged as shorter than the standard, it was increased by 10 msec and decreased by 30 msec when judged as longer. The starting value of the 25%-run was set well below the standard at 400 msec. The two independent interleaved runs, employed in the same block, made it impossible for a subject to predict the next signal level to be presented. Figure 2.1 provides an illustration of these two runs.

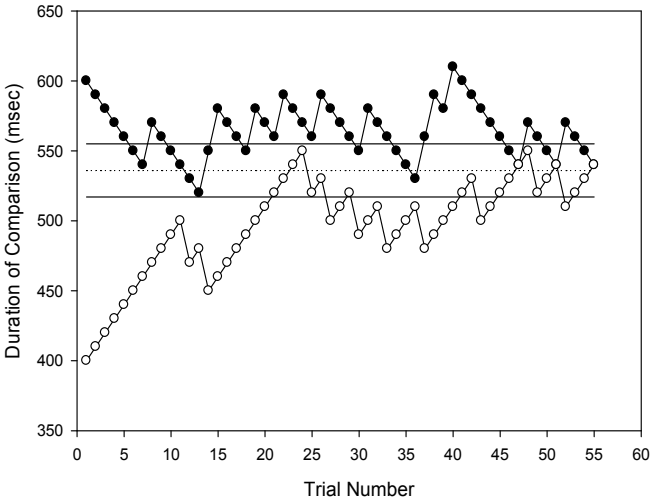

Figure 2.1: One block of the adaptive procedure consisted of two independent interleaved runs (55 trials each). This figure depicts the data for one block of a real subject. The graph shows the duration of the comparison stimuli (msec) on each trial as the experiment progressed. Stimulus levels followed an adaptive rule as described in the text corresponding to the test person's responses. 75%-run (filled circle) and 25%-run (open circles). The dashed line indicates the estimated PSE ($x_{.50}$), the lower solid line the estimated $x_{.25}$ and the upper solid line the estimated $x_{.75}$.

Non-adaptive blocks. The design was identical to the one of the adaptive blocks, with the only exception that the levels of the comparison stimuli were pre-selected and constant throughout a block (i.e. method of constant stimuli). Eleven levels were selected ranging from 400 to 600 msec in constant steps of 20 msec, so that five levels were below the standard and five levels were above it. The 11th level was equal to the standard, i.e. equal to 500 msec. Each level was presented to the subject ten times in random order, resulting in a total of 110 trials.

Estimation of DL. The technique for estimating DL was identical for adaptive and non-adaptive blocks and followed previous research (e.g., García-Pérez & Alacalá-Quintana, 2005; Leek et al., 1992). All 110 trials of a single block were always used to generate the full psychometric function. A logistic psychometric function

$$\Psi(x) = \frac{1}{1+\exp\left[-(x-a)/b\right]} \qquad (2.1)$$

was used to compute the maximum likelihood estimates of DL and the point of subjective equality (PSE), where x denotes the length of the comparison duration. The PSE is equal to a, and DL is equal to $b \cdot \log(.75/.25)$ (see Bush, 1967) . This function specifies the probability $\Psi(x)$ of the response "comparison longer than standard" for each level x of the comparison. The function $\Psi(x)$ ranges from 0 for extremely brief comparison intervals to 1 for extremely long comparison durations (Figure 2.2). In this and the following experiments, obtained estimates were checked by plotting the data together with the corresponding psychometric function in order to see whether the fit is reasonable.

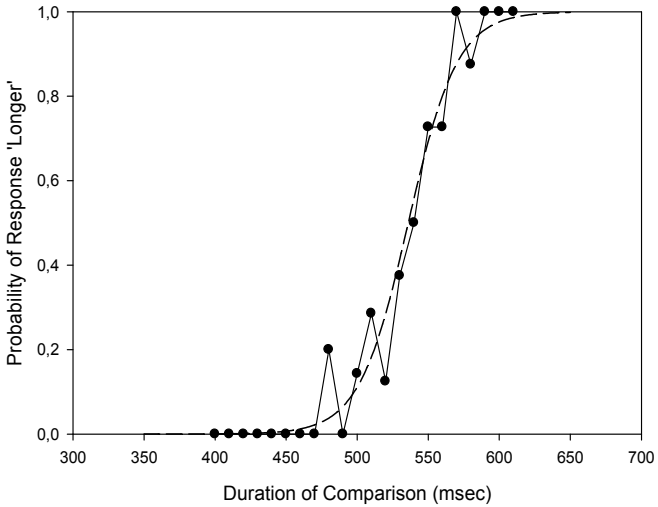

Figure 2.2: Psychometric logistic function fitted to the data depicted in Figure 2.1. The data themselves are represented by filled circles. This function has a PSE of 536 msec and a DL of 19 msec.

2.1.2 Results and Discussion

A separate two-way ANOVA with factors Procedure (adaptive vs. non-adaptive) and Block (first vs. second) was performed for DL and PSE. Figure 3 depicts the result for DL. Overall mean DL (± SE) was 32.2 ± 1.2 msec. DL estimates did not differ significantly between the two procedures, $F(1, 29)=2.43$, $p=.13$, $MSE=54.5$, $\eta^2= 0.08$; mean DL was 31.1 ± 1.9 msec and 33.2 ± 1.7 msec for the adaptive and non-adaptive procedure, respectively. This result suggests that both procedures give about the same estimates of DL. Although, performance slightly improved with practice, the main effect of Block did not reach statistical significance, $F(1,29)=2.73$, $p=.11$, $MSE=80.0$, $\eta^2= 0.09$. The mean DL was 33.5 ± 1.67 and 30.8 ± 1.93 msec for the first and second block, respectively. The interaction of both factors was not significant, $F(1,29)=0.02$, $p=.88$, $MSE=35.2$, $\eta^2= 0.00$. The obtained mean Weber fractions were 6.6% for the non-adaptive and 6.2% for the adaptive

procedure. Almost identical fractions have been reported for duration discrimination with auditory stimuli in previous studies (Getty, 1975; Grondin, 1993; Rammsayer & Ulrich, 2005). The present results support the idea that the way of collecting the data has very little effect on the DL estimates. The results depicted in figure 2.5 together with results for experiment 2.

A similar analysis of the PSE yielded again no significant main effect of procedure, $F(1,29)=0.2$, $p=.97$, $MSE=252.7$, $\eta^2= 0.01$. The overall mean PSE was 502.2 ± 1.7 msec and virtually identical for both procedures, that is, 502.1 and 502.3 msec for the non-adaptive and adaptive procedure, respectively. Neither the main effect of Block, $F(1,29)=3.5$, $p=.07$, $MSE=246.5$, $\eta^2= 0.11$, nor the interaction of both factors, $F(1,29)=0.2$, $p=0.681$, $MSE=174.2$, $\eta^2= 0.01$, became statistically significant.

In a second set of analyses, the nonparametric Spearman-Kärber technique was used to summarize each participant's psychometric function. This technique treats a psychometric function as a cumulative probability function (e.g., Finney, 1952; Trevan, 1927) and summarizes it in terms of its moments (i.e. mean, standard deviation, and higher moments). I used the computer program PMETRIC (Miller & Ulrich, 2004) to compute the mean, the standard deviation (SD), and the skewness of each participant's psychometric function. The mean, like the PSE, measures the central tendency of a psychometric function and the SD, like the DL, measures its steepness and thus a participant's discrimination performance.[1] I also employed a measure of skewness to assess whether the observed psychometric functions would deviate meaningfully from symmetry. As before, each summary statistics was submitted to a separate ANOVA. First, the overall mean was 502 ± 1.63 msec and thus virtually identical to the overall PSE. The average mean tended to be slightly larger in the first compared to the second block, $F(1,29)=3.1$, $p=.088$, $MSE=231.0$, $\eta^2= 0.10$, (i.e. 504.7 vs. 499.9 msec). No other effects were significant, $F<1$. Second, the overall SD was 46.5 ± 1.2 msec. The SD decreased somewhat, yet significantly, with practice indicating a slightly improved discrimination performance in the second

[1] Note that the DL of a logistic psychometric function is related to the SD of this function as following: SD=1.65·DL. Therefore, the overall mean DL of 32.2 msec is associated with an overall SD of 53.1 msec.

half of the experiment, $F(1,29)=5.6$, $p=.025$, $MSE=81$, $\eta^2= 0.16$, (i.e. 48.4 ± 1.5 vs. 44.5 ± 1.7 msec). There was no significant main effect of procedure, $F(1,29)=2.8$, $p=.106$, $MSE=57.0$, $\eta^2= 0.09$, nor a significant interaction, $F(1,29)=2.8$, $p=.106$, $MSE=57.0$, $\eta^2= 0.09$. Finally, the coefficient of skewness (i.e. third central moment divided by SD; see Evans, Hastings, & Peacock, 2000) was computed for each single psychometric function. The average coefficient was 0.238 and was significantly larger than zero, $F(1,29)=36.6$, $p<.001$, $MSE=0.186$, $\eta^2= 0.56$. Although this value indicates a small positive skewness, the observed functions were virtually symmetrical.[1] There were no further significant effects on this coefficient.

I also assessed whether the parametric and non-parametric estimates are correlated as one should expect if both index the same concept. In agreement with this expectation, the mean obtained from the Spearman-Kärber technique was highly correlated across subjects with the parametric PSE estimate, that is, the product moment correlation ranged from 0.95 to 0.99 across the different conditions. In addition, the SD of the Spearman-Kärber technique and the DL estimate were also highly correlated (0.973 to 0.997).

In summary then, the data of Experiment 1 show that the adaptive and the non-adaptive procedure produce virtually identical DL and PSE results. The same conclusion applies to the summary statistics from the Spearman-Kärber technique. In addition, the estimates of both analyses agree surprisingly well, though the underlying assumptions and the computational steps of both techniques differ greatly. (The results on test-retest reliability will be presented and discussed together with the ones of Experiment 2).

[1] For comparison, the coefficient of skewness of a 30-step gamma distribution is equal to 0.365 although this distribution nearly resembles a normal distribution according to the central limit theorem. It is therefore justified to use a symmetrical psychometric function (such as the logistic distribution) for probit analysis for the present data. In fact, and not surprisingly, the goodness of fit for the logistic psychometric function was satisfactory in all cases.

2.2 Experiment 2: The two procedures for the 2AFC task adaptive procedure for reminder task.

This experiment is similar to Experiment 1, except that the DL is now estimated by means of the 2AFC task. I also included the adaptive version of the reminder task in the design of Experiment 2 to enable a direct comparison between the 2AFC and the reminder task. DL was estimated by a standard parametric approach and again by the non-parameteric Spearman-Kärber technique (Ulrich & Miller, 2004).

2.2.1 Method

Subjects. A new group of thirty volunteers 21 females and 9 males (mean age: 27.8 ± 6.3 years) were recruited for this experiment. All had normal hearing and were naïve about the purpose of the study. Two subjects were replaced due to DL estimates that were two standard deviations above the mean of the entire sample.

Apparatus, stimuli, and procedure. The apparatus and the auditory stimuli were identical to those employed in Experiment 1. On a single trial of the 2AFC task, the 500-msec standard occurred either first or second. In contrast to Experiment 1, the comparison was always longer than the standard and the presentation order of the standard and comparison varied randomly from trial to trial. Subjects were asked to indicate the longer interval. An experimental session comprised six blocks: two adaptive 2AFC blocks (A), two non-adaptive 2AFC blocks (B), and two reminder adaptive blocks (C). There were six possible orders of the six blocks (i.e. ABCABC, ACBACB, BACBAC, BCABCA, CABCAB, and CBACBA) and these orders were counterbalanced across the 30 subjects.

Adaptive 2AFC blocks. In the adaptive 2AFC condition, the comparison duration followed the same adaptive rule as in Experiment 1, with the same step sizes of 10 and 30 msec converging to a stimulus level with 75% correct responses. The initial duration of the comparison was 600 msec. In contrast to Experiment 1, only one run of 100 trials was employed. Whenever the comparison duration reached the standard duration of 500 msec, the comparison duration was changed to its previous level for the next trial, in order to keep the comparison duration longer than the standard duration. Figure 2.3 provides an example of such a run. In order to estimate DL, the data of such a run were first

ordered according to the obtained levels of the comparison from short to long durations. Then the corresponding percentage of correct responses was computed for each level. As in Experiment 1, I adopted the standard definition of the threshold as the value yielding 75% correct performance in the 2AFC task. I again employed a standard parametric technique for estimating this threshold (cf., Ulrich & Miller, 2004). Specifically, the following 2AFC logistic function was fitted to the observed data points

$$\Psi(x) = 0.5 + \frac{0.5}{1 + \exp\left[-(x-a)/b\right]} \qquad (2.2)$$

in order to estimate the parameters *a* and *b*. Here *x* represents the difference between the comparison and the standard duration, and $\Psi(x)$ denotes the probability of a correct response at difference *x* (cf. Ulrich & Miller, 2004). In contrast to Equation 1, the parameter *a* represents now the threshold value (i.e. DL) instead of the PSE; that is $x_{.75} = a$, and $b>0$ a scale parameter (slope). Figure 2.4shows the fitted function for the data shown in Figure 2.3.

Non-adaptive 2AFC blocks. The estimation of DL was conducted in the same manner as for the 2AFC adaptive condition with the only exception that the comparison durations were pre-selected and kept constant throughout the experiment. Specifically, the following ten comparison durations were used: 515, 530, 545, 560, 575, 590, 605, 620, 635, and 650 msec. These durations were presented in random order across the trials in a single 2AFC block.

Reminder-adaptive blocks. This condition was identical to the one applied in Experiment 1

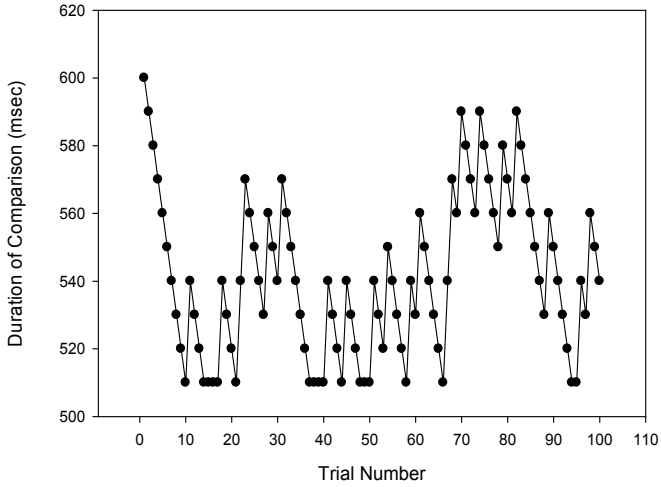

Figure 2.3: Illustration of the 2AFC adaptive task. Comparison duration is plotted as a function of trial number. Data were generated by a subject.

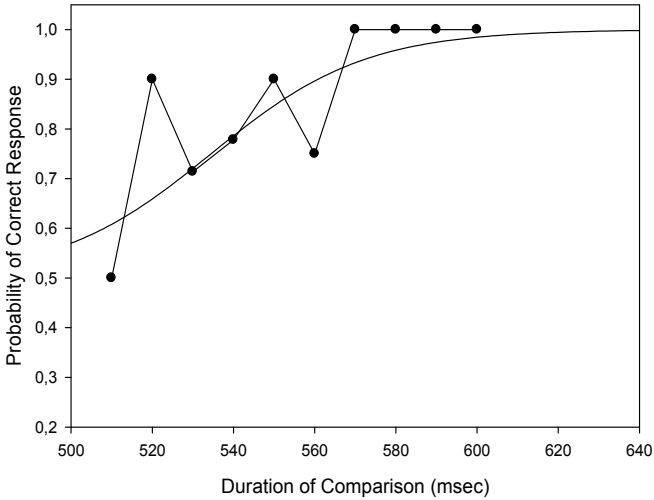

Figure 2.4: A typical psychometric function from the 2AFC task. The graph shows the percentage of correct responses as a function of comparison duration. Filled circles represent the actually visited comparisons duration (msec) and their corresponding probabilities of correct responses as generated by a real subject. Solid line represents the best fitted psychometric function. Notice that chance level is now equal to 0.5.

2.2.2 Results and Discussion

DL results. Figure 2.5 depicts the results of mean DL. As I hoped, the mean DL estimates from the reminder adaptive blocks were virtually identical to the ones of the previous experiment suggesting that the estimates were quite robust. A separate two-way ANOVA with factor Condition (2AFC adaptive vs. non-adaptive 2AFC vs. reminder adaptive) and Block (first vs. second) yielded a reliable main effect of Condition, $F(2,58)=32.7$, $p<.001$, $MSE=384$, $\eta^2= 0.53$. Mean DL was 59.3 ± 3.8, 56.5 ± 3.5, and 32.9 ± 1.5 msec for 2ACF-adaptive, 2AFC-non-adaptive, and reminder-adaptive blocks, respectively. A Scheffé-test revealed a significant DL difference between the reminder-adaptive procedure and each of the two 2AFC procedures, but no significant DL difference between the estimates of the two 2AFC procedures. I also employed the Spearman-Kärber technique to compute non-parametric DL estimates for the two 2AFC procedures (see Ulrich & Miller, 2004) and found almost the same results as those for the parametric estimates. In fact, the parametric and non-parametric DL estimates were again highly correlated (r=0.956-0.984). Rather surprisingly, the present results show that the DL estimates obtained with the 2AFC task were almost twice as large as those obtained with the reminder task. Consistent with Experiment 1, however, both the adaptive and the non-adaptive approach yielded virtually identical DL estimates in the 2AFC condition.[1]

[1] I re-ran the present experiment with another sample. In order to simplify the design, I omitted the reminder task, therefore making the design of this replication similar to the one of Experiment 1. As before there was no main effect of procedure, $F(1,23)=0.5$, $p=.467$, $MSE=183.8$, $\eta^2=0.01$, no main effect of block $F(1,23)=0.2$, $p=.630$, $MSE=252.4$, $\eta^2= 0.02$, and the interaction of both factors was again not significant $F(1,23)=1.0$, $p=0.320$ $MSE=334.5$, $\eta^2= 0.04$. Mean DL was 43.6 ± 5.0 and 45.6 ± 3.8 msec for the non-adaptive and adaptive procedure, respectively. Hence, even when omitting the reminder task from the design, the pattern of results remains the same.

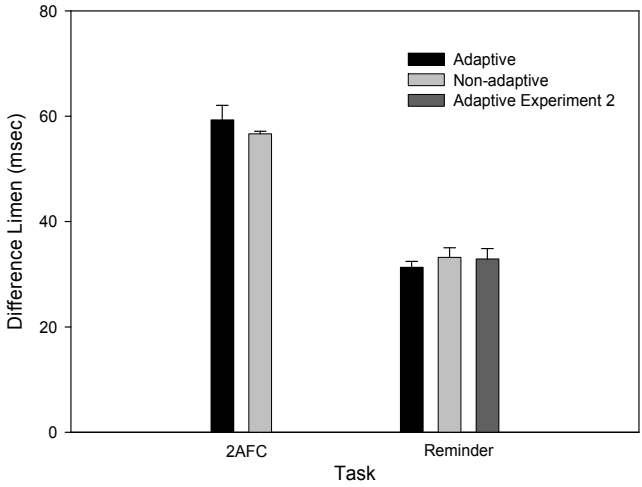

Figure 2.5: The graph includes the results of Experiments 1 and 2 and thus shows mean DL as a function of task and procedure. The x-axis shows the different procedures grouped according to task, while the y-axis shows DL (msec). Black bars represent the adaptive procedure, light grey bars represent the non-adaptive procedure, and the dark grey bar represents the additionally included control condition of Experiment 2.

Test–retest reliabilities. I used the data obtained in the two blocks of each task to compute the test-retest reliability of DL using the Pearson correlation coefficient (as performed by Linschoten et al., 2001). This was separately done for the 2AFC data in Experiment 2 and also for data of the reminder paradigm in Experiments 1 and 2. More specifically, I correlated for each single condition the parametric DL estimates that were obtained in the first and second block across all subjects. All tasks exhibited modestly strong but highly significant correlations (Table 2.1). Test-retest reliabilities of similar magnitude were reported by Linschoten et al. (2001) for a different discrimination task. Test-retest reliabilities were also computed for the non-parametric estimates and these results were virtually identical to the results of the parametric estimates.

Table 2.1 Test-Retest correlations of DL as a function of task and procedure separately for parametric estimates (probit analysis) and non-parametric estimates (Spearman-Kärber). The data are from Experiments 1 and 2.

	Task	
Procedure	2AFC	Reminder
Adaptive		
Parametric	0.75*	0.63*
Non-parametric	0.71*	0.62*
Non-adaptive		
Parametric	0.68*	0.70*
Non-parametric	0.76*	0.73*

*p< .001

2.3 Experiment 3. Short duration of comparison levels in 2AFC task

In the previous experiment, the DL estimates from the 2AFC task were almost twice as large as the ones from the reminder task. One may attribute this discrepancy to the different ranges of comparison durations that were employed by each task. In the reminder task, the comparison durations were shorter and longer than the standard duration. In the 2AFC task, however, the comparison durations were only longer than the standard duration. There is, however, evidence that the estimation of DL is sensitive to the distribution of the comparison durations relative to the standard duration (Grondin et al., 2001) indicating smaller DLs when the comparison durations are shorter than the standard duration and larger DLs when the comparisons are longer than the standard. Hence, one might assume that the 2AFC task yields larger DLs than the reminder task.[1] In order to test this hypothesis, only shorter comparison durations than the standard duration were employed in the 2AFC condition of this experiment. With this procedure, I expected smaller DLs in the 2AFC than in the reminder condition.

2.3.1 Method

Subjects. A new group of twenty volunteers, 12 females and 8 males (mean age: 30.3 ± 9.3 years) participated in this experiment. They had normal hearing and were naïve about the purpose of the study.

Apparatus, stimuli, and procedure. The apparatus and the auditory stimuli were identical to those employed in Experiments 1 and 2. A single session comprised one 2AFC block and one reminder block. The order of these two blocks was counterbalanced across subjects. The 2AFC task was identical to the non-adaptive version of this task applied in Experiment 2 with the exception of shorter durations than the standard only (350, 365, 380, 395, 410, 425, 440, 455, 470, 485). The reminder task was identical to the non-adaptive version of this task in Experiment 1. The method for determination of DL in the 2AFC task was analogous to the one of Experiment 2.

[1] I thank Jeff Miller for proposing this hypothesis.

2.3.2 Results and Discussion

Contrary to the prediction of the proposed hypothesis, mean DL was again significantly larger for the 2AFC (47.2 ± 3.7 msec) than for the reminder task (35.4 ± 2.6 msec), $t(19)=3.29$, $p=.002$, $\eta^2= 0.36$. A one-sided *t*-test for independent samples failed to reveal a significant difference between mean DL for the 2AFC task in this experiment and the corresponding mean DL of Experiment 2, $t(48)=1.36$, $p=.090$. In addition, also mean DL for the reminder task did not significantly differ from the corresponding one of Experiment 2, $t(48)=0.51$, $p=.308$. In summary, the observed discrepancy between the DL values obtained with the 2AFC and the reminder tasks in Experiment 2 cannot be attributed to the idea that the comparison durations were always longer than the standard duration. The present data rather suggest that the 2AFC task yields consistently larger DLs than the reminder task irrespective of whether the comparisons are shorter or longer than the standard duration.

2.4 Experiment 4. Random vs. fixed Interstimulus intervals

Another possible reason of the discrepant results might be the different temporal pattern of the two tasks. Note that in the reminder task, the duration of the first stimulus is always the same, whereas in the 2AFC task, the duration of the first stimulus varies from trial to trial. One may therefore assume that this temporal variability of the first interval increases the overall temporal uncertainty of when the second stimulus will be presented. It is well documented that an increase of temporal uncertainty impairs perceptual discrimination not only for non-temporal features (e.g., Rolke & Hofmann, 2007) but for temporal features as well (Grondin & Rammsayer, 2003). According to this explanation, the constant interstimulus interval (ISI) of 1,000 msec between the first and second interval would facilitate discrimination in the reminder but not in the 2AFC task. Thus, it is supposed that a random ISI in the reminder task might impair discrimination performance, and consequently yield DLs similar to those of 2AFC task with a constant ISI.

In fact, the observed discrepancy might also be explained within the framework of the entrainment model of temporal attention (Large & Jones, 1999). According to this hypothesis, the sequence of all temporal intervals before the presentation of the comparison forms an isochronous induction sequence in the reminder paradigm (Barnes & Jones, 2000). Note that this sequence includes the presentation of a 500-msec standard followed by a 1,000-msec break after the presentation of the standard. In other words, this sequence consists of two induction intervals (i.e. 500 and 1,000 msec) before the comparison is delivered and this may entrain an attending 2 Hz rhythm facilitating temporal processing of the comparison (cf. Barnes & Jones, 2000; McAuley & Jones, 2003). Such an induction process, however, would not be effective in the 2AFC task since the comparison can either occur in the first or second position. In this case, a stable induction sequence is not provided and thus the internal oscillator is expected to drift around, which hampers temporal discrimination (i.e., Ward 2003).

In order to evaluate these hypotheses, Experiment 5 combined each of the two tasks with a random and a fixed ISI. Although it will be unworkable to differentiate between these potential explanations it will

be possible to either acknowledge or decline them as a possible cause for the discrepancy.

2.4.1 Method

Subjects. A new group of twenty four volunteers, 16 females and 8 males (mean age: 21.0 ± 0.4 years), participated in this experiment. They had normal hearing and were naïve about the purpose of this study.

Apparatus, stimuli, and procedure. The apparatus and the auditory stimuli were identical to those employed in Experiments 1, 2, and 3. A single session comprised two 2AFC blocks and two reminder blocks. Each block combined one task with either a fixed ISI of 1,000 msec, or a random ISI that followed a normal distribution with mean of 500 msec and a standard deviation of 100 msec. In the random condition, a mean ISI of 500 rather than of 1,000 msec employed in order to omit exceedingly long ISIs. Order of these four blocks was counterbalanced across subjects. The 2AFC task was identical to the non-adaptive version of this task applied in Experiment 1 and the reminder task was identical to the non-adaptive version of this task in Experiments 2 and 3.

2.4.2 Results and Discussion

Figure 2.6 depicts mean DL. A two-way ANOVA with factors Task (reminder vs. 2AFC) and ISI (fixed vs. random) revealed again reliably larger DLs for the 2AFC (57.0 ± 4.1 msec) than for the remainder task (38.3 ± 2.7 msec), $F(1,23)=49.1$, $p<.001$, $MSE=170.4$, $\eta^2= 0.68$. As one might expect from research on temporal preparation (e.g., Rolke & Hofmann, 2007), discrimination performance was worse in the random (51.8 ± 10.6 msec) than in the fixed (43.5 ± 8.9 msec) ISI condition, $F(1, 23)=6.9$, $p=.015$, $MSE=242.6$, $\eta^2= 0.23$. The significant interaction of both factors indicates that performance on the 2AFC task worsened when the ISI varied from trial to trial rather than when kept constant across trials, $F(1, 23)=6.4$, $p=.019$, $MSE=121.8$, $\eta^2= 0.22$. In fact, separate *t*-tests indicated that this interaction effect emerged from the 2AFC task, $t(23)=2.98$, $p=.007$, $\eta^2=0.28$, and not from the reminder task, $t(23)=0.91$, $p=.37$.

The results of this experiment strengthen the notion that the 2AFC task produces generally enlarged DLs. In contrast to the expectations and somewhat surprisingly, however, the random ISI did not worsen discrimination performance in the reminder task. This finding rejects the

conjecture which assumes that the remainder task in Experiments 1 and 2 benefited from temporal certainty or from the entrainment of an internal oscillators by the preceding intervals. This finding, however, puts forward the idea that the superior discrimination performance in the remainder task stems from the continuous presentation of the standard in the first position.

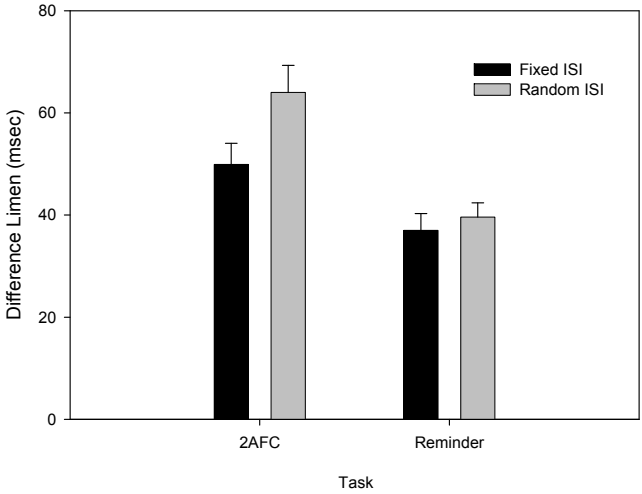

Figure 2.6: Mean DL as a function of task and ISI condition for Experiment 4.

2.5 Experiment 5: Effect of the position of the standard stimulus on discrimination performance in auditory modality

Experiment 5 examined this alternative explanation by manipulating the presentation order of the standard and the comparison in the reminder task. This order could be either standard in the first position or in the second position. If the order of these two stimuli matters, one should observe worse performance under the standard-second than under the standard-first condition.

As a matter of fact, experimental evidence suggests that the order of these two stimuli has an effect on both temporal (Marchman, 1969; McGavern, 1965; Rammsayer & Wittkowski, 1990; see Ulrich et al., 2006; Van Allen, Benton, & Gordon, 1966) and spatial (Nachmias, 2006) discriminations. Experiment 5 also included the 2AFC task as a control condition.

2.5.1 Method

Subjects. A new group of twenty-four volunteers participated in this experiment, 16 females and 8 males, (mean age: 21.4 ± 4.5 years). They had normal hearing and were naïve about the purpose of this study. One subject was replaced due to inability to comply with the task, that is, his DL was about three times larger than the mean of the entire sample.

Apparatus, stimuli, and procedure. These were identical to the previous experiments using the non-adaptive procedure. In the standard-second condition of the reminder task, however, the order of the standard and the comparison was reversed. Each subject performed three blocks. A single block of trials was used to measure DL in each condition (i.e. standard-first reminder, standard-second reminder, and 2AFC). The order of these three conditions was counterbalanced across subjects.

2.5.2 Results and Discussion

A one-way ANOVA revealed a significant difference between the three conditions $F(2,46)=13.9$, $p<.001$, $MSE=170.4$, $\eta^2= 0.38$. Mean DL was 43.3 ± 3.9 msec for the regular condition, 64.4 ± 6.2 msec for the irregular condition, and 68.6 ± 5.6 msec for the 2AFC condition. A

Scheffé test (α=.05, critical difference =12.9 msec) revealed a significant DL difference only between the standard-first and the two other conditions. This pattern of results supports the above idea that the order of the standard and the comparison in the remainder task has a strong effect on discrimination performance, and that the standard in the first position facilitates the discrimination process. Converging evidence for this *standard position effect* has been reported by Ulrich et al. (2006), although they did not manipulate standard position within a single experiment, and by Marchman (1969), McGavern (1965), Rammsayer and Wittkowski (1990), and Van Allen et al. (1966) who separated trials according to standard position after running the experiment. Interestingly, this finding has recently also reported for a spatial discrimination task (Nachmias, 2006). In fact, this effect may suggest an explanation for the performance difference between the regular remainder task and the 2AFC task.

2.6. Experiment 6: Effect of the position of the standard stimulus on performance in visual modality

All previous experiments revealed discrepant DLs between the 2AFC and the regular reminder tasks. All of them, however, employed auditory stimuli. An important question, therefore, is whether this discrepancy is restricted only to the auditory modality or whether it would generalize across modalities. In order to address this question, Experiment 5 was replicated with visual duration stimuli.

2.6.1 Method

Subjects. A new group of twenty-four volunteers, 18 females and 6 males (mean age: 25.7 ± 5.8 years), participated in this experiment. They had normal vision and were naïve about the purpose of the study.

Apparatus, stimuli, and procedure. This experiment was identical to Experiment 5 with the exception of the sensory modality. This experiment used the visual modality employing duration discrimination of light provided by a green light emitting diode (LED, diameter 0.48°, viewing distance 60 cm, luminance 48 cd/m^2). The LED was attached to the centre of the computer screen and the background of the screen was black. Standard duration was again 500 msec. The comparison durations were 300, 340, 380, 420, 460, 500, 540, 580, 620, 640, and 700 msec for the standard-first and the standard-second conditions of the reminder task, and 530, 560, 590, 620, 650, 680, 710, 740, 770, and 800 msec for the 2AFC task. These durations were selected in such a way that this task was about equally difficult as in the previous experiments. Subjects pressed the left-shift key when the first light appeared longer than the second one and they pressed the right shift key, when the second light appeared longer than the first one.

2.6.2 Results and Discussion

A one-way ANOVA revealed again a highly significant effect of condition, $F(2,46)=11.4$, $p<.0001$, $MSE=114.3$, $\eta^2= 0.33$. Mean DL was 77.1 ± 1.6 msec for the standard-first condition, 138.6±3.8 msec for the standard-second condition, and 146.2 ± 3.3 msec for the 2AFC condition. A Scheffé test ($\alpha=.05$, critical difference=41.2 msec) indicated a significant DL difference only between the standard-first and the two other conditions. These results clearly show that the discrepant

DL estimates between the 2AFC and the standard-first reminder task are not restricted to the auditory modality. Furthermore and consistent with the previous experiment, the position of the standard duration in the reminder task had again a strong effect on DL, being reliably larger when the standard appeared after the comparison.

3 Experiments employing non-temporal visual stimuli.

The previous experiments (1-6) exclusively employed duration discrimination tasks with auditory and also visual stimuli. Although it is quite clear that DL estimations yielded by the 2AFC task are up to double as large as the estimates yielded by the reminder task, this observation is currently limited to the temporal domain. The following two experiments are designed to test out whether these findings are generalizing across non-temporal tasks as well.

Specifically, random-dot pattern employed in the current experiment and line-discrimination task in experiment 8 that consider exhibiting high performance are chosen.

3.1 Experiment 7: Random-dot pattern discrimination

This experiment employs the 2AFC task and two versions of the reminder task. In the *regular* version, the standard precedes the comparison whereas in the *irregular* version the standard follows the comparison. The random-dot-pattern discrimination was similar to the one employed by Ross (2003). In brief, in each trial two visual random-dot patterns were successively presented and then the subject was asked to indicate which of the two patterns had more dots. If the results of the above experiments reported in paper of Lapid et al. (2008) generalize to the non-temporal domain, the presentation order of the standard and the comparison should again matter. Specifically, discrimination performance should be best in the regular reminder version and about equally worse in the two remaining tasks.

3.1.1 Method

Subjects. A new group of thirty volunteers, 23 females and 7 males (mean age: 27.1 ± 7.0 years) participated in this experiment. They had normal vision and were naïve about the purpose of the study.

Apparatus and stimuli. A PC controlled the presentation of the stimuli and the recording of the participants' responses. Visual stimuli used for all tasks were random patterns of black- filled circles on white background that were presented within an invisible rectangle region (300×500 pixels) at the centre of the screen. Each circle was 4 pixels in

diameter. These values (e.g., rectangle dimensions and circle diameter) permit large possibilities of positions in the presentation space, and facilitate a non-overlap presentation. Location of the circles was controlled by the program and was reselected in case of a circle overlap. Subjects sat approximately 60 cm from the screen and were instructed to look at the centre of the screen.

General procedure. The experiment comprised three blocks. Each block employed one of the following tasks: the 2AFC, the regular reminder, or the irregular reminder task. Subjects performed all three tasks in a counterbalanced order. In order to get familiar with the task, the subjects performed a short practice block with 20 trials at the beginning of the experiment. (This practice block employed the 2AFC task and a smaller number of dots than in the experimental blocks).

The time-course of a single trial was identical for all three tasks. On each trial, two successive displays were presented with a 1-second interstimulus interval. Each display was presented for 300 msec. One of the two displays was the standard and had a fixed number of 30 circles (spatial distribution of the circles was random in each presentation) while the other was the comparison stimulus. Subjects could respond within five seconds before a new trial started. A new trial started two seconds after a response. Subjects were instructed to indicate which of the two displays contained more circles, the first one or the second one. They pressed the left-shift key when they judged the first stimulus to contain more circles and the right-shift key when they judged otherwise. No feedback was provided.

2AFC block. In this block, the comparison levels were 32, 34, 36, 38, 40, 42, 44, 46, 48, or 50 circles. Each level was randomly presented ten times during a single block. In addition, the order of the presentation of the standard and the comparison stimulus varied randomly from trial to trial.

Reminder blocks. In these blocks, the comparison level was smaller, larger, or equal to the number of circles in the standard stimulus. Specifically, the comparison level was 15, 18, 21, 24, 27, 30, 33, 36, 39, 42, or 45 circles. Each value was randomly presented ten times.

In contrast to the 2AFC task, the position of the comparison was constant within a single block. In the regular reminder task, the standard stimulus was presented first, whereas in the irregular one, it was presented second.

Although the ranges of the stimuli are not identical they do match with the ranges that were use in the former experiments reported above and in Lapid et al. (2008). As reported previously the range of the stimuli did not affect the DL estimates when using adaptive vs. constant comparison stimuli, which by definition employ different values.

3.1.2 Results and Discussion

Like in the above experiments, a maximum-likelihood procedure was used for estimating the DL, which was obtained for each subject and for each task. Mean DL was 5.64 ± 0.57 circles for the 2AFC task, 3.93 ± 0.23 for the regular reminder task, and 4.99 ± 0.25 for the irregular reminder task. Weber fractions were 0.19, 0.13, and 0.17 for the 2AFC, the regular, and the irregular reminder task, respectively. Quite similar fractions have been reported in previous studies (e.g., Burgess & Barlow, 1983). A one-way analysis of variance with the within-subject factor Task (2AFC, regular, and irregular reminder task) revealed a statistically significant difference between the three tasks, $F(2,58)=6.8$, $p<.007$, $MSE=3.3$, $\eta^2= 0.19$. A Newman-Keuls test ($\alpha=0.05$, critical difference =0.94 circles) indicated a significant DL difference only between the regular reminder and the two other tasks; the DL estimated by the regular reminder task was about 44% smaller than the one estimated by the 2AFC task, and about 23% smaller than the one estimated by the irregular reminder task. A correlation analysis revealed a significant correlation coefficient of $r=0.34$, $z=1.83$, $p=.034$, between the DLs estimated by the regular reminder and the 2AFC task.

These results indicate a similar pattern of DL results that was obtained in the previous experiments employing a duration discrimination task (temporal stimuli).

3.2 Experiment 8: Line-Length discrimination

In order to strengthen the idea that the 2AFC task consistently yields a larger DL than the reminder task, subjects in Experiment 8 discriminated the length of two subsequently presented lines rather than the number of dots of two successive visual random-dot patterns. Subjects usually show a high performance in line-length discrimination with Weber fractions around 0.03 (e.g., Teghtsoonian, 1971). Therefore, it seems important to see whether the 2AFC task and the reminder task would again yield discrepant DLs when discriminative sensitivity is particularly high.

3.2.1 Method

Subjects. A new group of thirty volunteers, 16 females and 14 males (mean age: 27.4 ± 5.4 years) participated in this experiment. They had normal vision and were naïve about the purpose of the study.

Apparatus, stimuli, and procedure. This experiment employed the 2AFC task and the regular reminder task only. The apparatus and the time course of a single trial were identical to the previous experiment, except that the random dot pattern was replaced by a horizontal line that was displayed in the middle of a monitor screen (768 by 1024 pixels) in front of the subject. The size of a single pixel was 0.36 × 0.36 mm. The lines were presented in black on a white background. In order to force subjects to process line length and not positional line cues, the horizontal position of each line was randomly determined. Specifically, the midpoint of a line followed a uniform distribution. This distribution ranged from 20 pixels on the left to 20 pixels on the right of the middle of the screen. The standard consisted of an array of 300 pixels (107 mm). The length of the comparison stimuli in the 2AFC task were 303, 306, 309, 312, 315, 318, 321, 324, 327, or 330 pixels. In the reminder task, these levels were 280, 284, 288, 292, 296, 300, 304, 208, 312, 316, or 320 pixels. Subjects pressed the left-shift key, when the first line appeared longer than the second line. They pressed the right shift key, when the second line appeared longer than the first one.

3.2.2 Results and Discussion

Mean DL in the line-length discrimination task was 12.13±0.97 and 10.11±0.54 pixels for the 2AFC and the reminder task, respectively. The

corresponding Weber fractions are 0.04 and 0.03. A *t*-test detected a statistically significant difference between the two mean DLs, t=2.44, df=29, p=0.02 η^2= 0.17, that is, the DLs obtained by the reminder task were approximately 20% lower then those estimated by the 2AFC task. The DLs of both tasks were correlated significantly, r=0.54, z=3.14, p<.001. This experiment again indicates that the 2AFC tasks yield higher estimates of the DL then the reminder task in non-temporal, highly sensitive task.

4 General Discussion

The present work compared DL estimates produced by the 2AFC task and by the reminder task. Complementary comparison was made between the estimates produced by adaptive and non-adaptive procedures. The basic question was whether both tasks and both procedures yield DLs of similar magnitude and of about the same reliability. Therefore, each of the two tasks was combined with an adaptive and a non-adaptive procedure. Duration discrimination task was employed in Experiments 1-6 and non-temporal visual discrimination was employed in Experiments 7 and 8 to address these questions empirically.

Experiment 1 as well as Experiment 2 has not revealed any difference in the magnitude estimates produced by either way of data collecting procedures, adaptive or non-adaptive. This finding is true for the reminder task employed in Experiment 1 and for the 2AFC task employed in Experiment 2. The two procedures yielded virtually identical DL results. In addition, the test-retest reliabilities of the two procedures were also similar in magnitude, suggesting that both procedures yield equally stable DL estimates. This result indicate that an experiment conducted with an adaptive procedure that employ the same number of trials as an experiment conducted with non-adaptive procedure will results in the same magnitude of estimates, as to say, the same thresholds. This clearly supports the argument that the two procedures are equally efficient (e.g., Dai, 1995; Simpson, 1988). However, the current study does not rule out the possibility that with adaptive procedure the same threshold could be reached with smaller amount of trials. In case that the same threshold is continually estimated with smaller amount of trails, the adaptive procedure may be considered as more efficient because it wastes less time and trials of unnecessary, too small or too large stimuli level, which do not supply any further information on the threshold. However, as previously stated, the current study supplied information on the efficiency of the procedures within a specific number of trials.

In contrast to the lack of difference between the two procedures, a meaningful difference was found between the two tasks (for a summary, see Table 4.1). More specifically, the DL estimated by the 2AFC task turned out to be up to twice as large as the DL estimated by the reminder task in Experiment 2. Based on the combined results from Experiment 1

and 2, no more investigation was conducted on the issue of adaptive vs. non-adaptive procedure. Rather a more intriguing query arose, what was the reason of the discrepancy between the estimates produced by these two tasks. As discussed in the introduction, both the reminder task and the 2AFC task are widely used to measure discrimination performance of subjects. Each of the two tasks can provide an estimate of the difference limen (DL). Therefore, in practice, the reminder task and also the 2AFC task have been used to estimate DL. In fact, there is no a priori theoretical reason why DL estimates derived from the two tasks should systematically differ[1]. Furthermore, this difference cannot be attributed to potential biases in the way DL is estimated as can be confirmed by computer a simulation that was conducted to confirm this issue (see Appendix A). The following Experiments 3-5 where designed to reveal the cause for the disagreement between the two tasks. Since in the 2AFC task traditionally and in Experiment 2, the comparison stimuli that were used were only larger then the standard stimulus, Experiment 3 employed only smaller comparison stimuli. But again the 2AFC produced thresholds that are eloquently larger then those produced by the reminder task. Experiment 4 was designed to examine other two not mutually exclusive possible explanations of the discrepancy. One explanation claims that the uncertainty about the arrival time of the second stimulus in the 2AFC task impairs performance, whiles the other

[1] Theory of signal detection (SDT), however, might lead one to expect better discrimination performance in the 2AFC than in the reminder task. It must be stressed, however, that the reminder task as well as the 2AFC task that are employed in the domain of SDT research differ from the ones used in this study. In the SDT domain the stimulus sequences in each trial are either <S,C> or <C,S> in the 2AFC task, and either <S,S> or <S,C> in the reminder task, where S denotes the standard duration (e.g., 500 msec) and C the duration of the comparison (e.g., 600 msec). In this case, SDT predicts better performance (i.e. larger *d*') in the 2AFC task than in this type of reminder task employed in SDT studies (see Macmillan & Creelman, 2005). The reason for the higher task performance in the 2AFC than in the reminder task is the difference in stimulus magnitudes between the first and second stimulus. In the 2AFC task, this difference is either S-C (e.g., -100 msec) in <S,C> trials and C-S (e.g., 100 msec) in <C,S> trials. In the reminder task, however, this difference is either S-S (0 msec) in <S,S> trials and S-C (e.g., -100 msec) in <S,C> trials. Because the size of the two differences in the 2AFC task is twice as large as in the reminder task, SDT predicts better discrimination performance in the 2AFC task. In the present experiments, however, I did not use this type of reminder task that originated in the SDT domain. In this case, not only one comparison C but several comparisons stimuli are used to assess DL. In this kind of tasks, it is more appropriate to employ the classical psychophysical approach. In deed, as shown in the General Discussion, this approach predicts identical DLs

claims that the constant rhythm pattern in the reminder task induces an entrainment of the internal oscillator, and facilitates performance. In order to test those potential explanations a fixed vs. random ISI were employed with both tasks. However, results again revealed larger DL estimates produced by the 2AFC and that in contrast to expectation the random ISI had worsen performance in the 2AFC and not in the reminder task therefore both explanations were rejected. An alternative cause was tested. Of the main differences between the tasks is the order of the presentation of the stimuli. Therefore Experiment 5 manipulated the presentation order of the standard and comparison stimuli in the reminder task as well as employing the two tasks under investigation as a control condition. This experiment indeed elucidates the reason for this discrepancy. It turned out that the order of the presentation of the stimuli does matter and that presenting the standard in the first position facilitates performance. So far, only the auditory modality task was used therefore Experiment 6 employed the visual modality and replicated the results for Experiment 5. Experiments 7 and 8 employed non-temporal visual discrimination tasks, random-dot pattern discrimination and line-discrimination respectively, and revealed similar pattern of results indicating that the discrepancy is true for non-temporal as well as for temporal stimuli, meaning, that the 2AFC consistently produced larger DL estimation. The magnitude of the discrepancy found for non-temporal discrimination (approximately between 20 and 40 percent) is lesser then the one found for temporal discrimination (up to 100%). Nonetheless it is higher than the statistically significant discrepancy that was reported in Nachmias study (2006) which had a magnitude of approximately 6 %.

As emphasized in the introduction these discrepant DL estimates are of great concern because it produces a misleading picture when the results of several studies are compared without considering the task for estimating DL (e.g., Miller & McAuley 2005; Wright et al., 1997). For example in the domain of duration discrimination, several studies that examined the Weber fraction have employed one or the other task for estimating DL. According to our results and consistent with other research, estimates of the Weber fraction seem to be larger when employing a 2AFC (e.g., Grondin, 1993; Rammsayer & Lima, 1991; Wright et. al., 1997) than a reminder task (e.g., Getty, 1975; Rammsayer & Ulrich, 2005). Furthermore, a common issue in the field of time perception is to measure the validity of Weber law across different

stimuli length (Grondin, 2003; Killeen & Weiss, 1987). For example Treisman (1963) review several papers or data that are concerned with violation of Weber law. Specifically, Weber fraction was reported to have a minimum value that was linked to different durations, in these works. Moreover this lowest value ranged between 3% and 8%. That is in fact a disagreement of almost three folds between the minimums of the reported Weber fractions percentages, but there is no indication which methods were used to calculate those percentages.

The present study concentrated on the widely used tasks, 2AFC and reminder task. Nevertheless, in psychophysics the complexity is even worse since wide range of variations of those two tasks is used and compared. Therefore, the results of the present study strongly suggest that valid comparisons between studies can only be reached if the task for estimating the DL is taken into account. Finally, I also assessed quantification of discrimination performance by means of the Spearman-Kärber method. The present study reinforces previous results from computer simulations (Miller & Ulrich, 2001; Ulrich & Miller, 2004) showing high correlations between the estimates provided by traditional probit analysis and by the Spearman-Kärber method.

The rest of the discussion treats the discrepant results between these two tasks as factual and not stemming from calculation or methodological artefact. An attempt to explain it within the underlying strategies of the discrimination process is made.

Table 4.1 Observed mean DL (± standard error of the mean) in milliseconds (unless indicated differently) for each experiment.

	Reminder/ adaptive	Reminder/ non-adaptive	2AFC/ adaptive	2AFC/non-adaptive
Experiment 1	31.1 ± 1.9	33.2 ± 1.7		
Experiment 2	32.9 ± 1.5		59.3 ± 3.8	56.5 ± 3.5
Replication[a]				44.6 ± 3.5
Experiment 3[b]		35.4 ± 2.6		47.2 ± 3.7
Experiment 4[c]		38.2 ± 2.7		57.0 ± 4.1
Experiment 5[d]		43.3 ± 3.9[f]		68.6 ± 5.6
		64.4 ± 6.2[g]		
Experiment 6[e]		77 ± 1.6[f]		146 ± 3.25
		143 ± 3.8[g]		
Experiment 7[h]		3.93 ± 0.23[f]		5.64 ± 0.57
		4.99 ± 0.25[g]		
Experiment 8[i]		10.11 ± 0.54[f]		12.13 ± 0.97

Notes --- (a) This replication is reported in Footnote 6. (b) In this 2AFC task the comparision was always shorter than the standard. (c) Experiment 4 employed random ISIs. (d) Experiment 5 manipulated the position of the standard. (e) Experiment 6 used visual instead of auditory duration stimuli. (f) In this condition, the standard appeared in the first position. (g) Standard appeared in the second position. (h) Experiment 7 used non-temporal visual stimuli, random-dot pattern discrimination. (i) Experiment 8 employed a non-temporal, line-discrimination task

Strategies of discrimination

The law of comparative judgments -paired comparisons

Whenever two or more stimuli are presented to a subject and he/she is asked to make a decision such as to report which of them is the longest, it is said that the subject make a *Comparative judgment*. This term was suggested by Thurston (1927) in his work 'the *law of comparative judgment*'. This work was the foundation of methods which attempt to rate a sensation that is created in response to stimuli relative to each other (Gescheider, 1997; Guilford, 1954). However in its base, this work carries a theory about the discrimination process. Therefore

although the current study does not concern with scaled/ rated responses, the law of comparative judgment is the keystone to some assumptions that will be used in the mathematical models to follow and it does put forward one strategy of comparing two stimuli. All following expressions are termed by Thurston (1927).

When ever two stimuli A and B are presented to a subject each creates a response which is of neuronal, chemical, and psychological nature and generally termed *discriminal processes*. On occasions which stimuli A and B judged as different, for example one *seems* less or more then the other, it is likely to assume that the discriminal processes that correspond with each of them are different. Furthermore, even when the same stimuli A and B are repeatedly presented to the subject, the comparative judgment is not consistent, that is, on occasions A is judged as larger and vice versa in other occasions where B is judged as larger. The conclusion from this observation is that the discriminal processes corresponding with a given stimulus are not fixed. Nonetheless there is one discriminal process that corresponds more frequently with a specific stimulus. In fact the frequency of discriminal processes of a given stimuli form a normal distribution on the psychological scale. The standard deviation of each distribution is termed *discriminal dispersion*. In each occasion stimuli A and B are to be judged against each other. The difference between their discriminal processes is termed the *discriminal difference*. This difference as well varies on different occasions in accordance to the variability in the discriminal processes. Moreover this difference in turn is compared to some criterion, thus, whenever stimulus A is judged as more intense (louder, longer etc.) than B, the difference (A-B) is regarded as positive relative to that criterion, and in other occasions where B is judged as more intense then A, the difference (A-B) is regarded as negative relative to that criterion. When a pair of stimuli is repeatedly compared, a specific probability of 'larger' responses corresponds to either stimulus A or B. This strategy is largely the base of the signal detection theory (Creelman & Macmillan 1979).

Frame of reference and adaptation-levels

The two classical works that are shortly described in this section concluded a different mechanism than the one described above. The first work by Woodrow (1933) was designated to explain the *time-order error* (TOE). It is commonly observed that when a pair of stimuli is repeatedly presented to a subject, the percentage of the correct responses is dependent of the order in which the stimuli were presented. Thus, TOE is then defined as the difference between those percentages. For example in Woodrow's experiment on weight discrimination a comparison stimulus that was 3% larger than the standard was indicated heavier in 56% of the trials when it was presented first but 77 % of the trials when it was presented second.

The manipulation of the experiment was fixed standards (110 g and 200 g) versus varying standard (10 values between 110g to 200g in steps of 10 g) in each block of trials. Each of the standards had five comparison stimulus levels corresponding to a specific percentage of the standard (-3, +3, 0, +6 and +9). Varying the standard was made in order to prevent the subject to become 'set' (in Woodrow's words) as prepared, him/her self towards a familiar standard in either first or second place, which serve a base for comparison. As a small additional part of the work he found that the DL was higher with varying standard than with fixed standard. Based on those result Woodrow concluded that the percentage of correct responses can not be **solely** a function of the difference between the stimuli or the sensations that correspond with the two stimuli, even if this difference is the base of discrimination. In fact Woodrow pointed out that preceding data have an influence on discriminating process as an independent factor of the difference between them.

Woodrow assumed that the first stimulus of a pair attuned the subject to some level of expectation, also named adaptation (L). The L is assumed to approximately be an average of all preceding stimuli of the entire series given the series 'does not vary too much'. Further more he claimed that this L is mostly influenced by excitation level caused by the first stimuli (E1) presented in a pair. Specifically, the intensity of L immediately after E1 is approximately that of E1, however it could be slightly more or less. The level then sink back to the average level or some habitual level between the presentation of the first and the second stimuli in a pair, and the longer the time between them the lower L level is. Thus, the second stimulus is compared to L. Some points worth

noting (1) Woodrow has not manipulated the order of the presentation of standard and comparisons, in all blocks they were randomly presented. (2) the DL were an average of the DLs with both standards for fixed standard, vs. DL of the pooled data of all the varying standard series. That is to say, he did not compare for example the DL of 110g standard when presented alone or when presented as a part of varying series. (3) An interesting finding which was totally ignored is that the DL was approximately three folds higher when the standard was presented second than when the standard presented first.

The second work was presented by Helson (1947, 1948) and was conducted in manner of rating a range of stimuli (weights between 200 and 400 g in steps of 50 g) relative to two standard (90g or 900g). Responses were qualitative such as very very heavy, very heavy medium have, medium, medium light, very light and very very light, and later were translated to numerical scale from 90 to 10 respectively. As an example 400 g weights corresponded to scale of 59 which mean medium-heavy when judged relative to a standard of 900g. This is a paradoxical result since it is less than half as heavy than 900 g and would be expected to be judged as light. Helson used the term *adaptation-level*, which is described operationally as the stimulus evoking neutral or indifference response from a subject. Depending on the nature of the stimuli, the neutral stimulus is the one that creates the response 'doubtful' or 'equal' within the observer. Stimuli above this level will elicit one response (e.g., stimulus is large) and below a response in the opposite direction (e.g., stimulus is small). From his result Helson, as well as Woodrow, concluded that the judgments are not made with respect to the standard, but with respect to the adaptation-level (the neutral or medium point) regardless if the standard is explicitly given. The adaptation-level was then believed to be the 'pooled effect of all stimuli' that influences a subject from inside (neural processes) and outside (presented stimuli). Practically, only external stimuli that were presented to the subject were calculated, and in the current example, the adaptation level was calculated to be the value corresponding to 337 g when the standard was 900 g. If one considers that the comparisons are indeed made regarding to the adaptation level, this could explain why 400g was judged as medium heavy relative to 337g. It is worth noting that this is markedly different from the discrimination experiments in the current study, in which all the above

comparison stimuli would be judged in the same category, as 'lighter' relative to the standard of 900 g.

To summarize the above ideas both works suggest that there is no direct comparison between the given standard and a comparison in each trial, instead the second stimulus in a pair of stimuli is compared to an internal reference that is some combination of the previous stimuli.

Two discrimination strategies

The above strategies are commonly known today as the *paired-comparison strategy* (P-C) and the absolute identification (AI) strategy. In P-C the subject is comparing the sensory input of the second stimuli with a memory trace of the first stimuli and the difference between those is judged against a criterion. In AI however the second stimuli is judged against some internal criterion also termed internal standard. How this internal standard is set, still remains unclear. For example in the MSS subjects are classically believed to use the latter strategy since there is no explicit standard and only one stimulus is presented it each trial (Creelman & Macmillan, 1979; Lages & Treissman, 1998; Morgan et al., 2000; Vogels & Orban, 1986).

Vogels & Orban (1986) compared the prediction of performance for several designs of 2AFC task (different pairs) assuming the use of the different strategies. They came to a conclusion that a subject will benefit of using one or the other strategies in the different designs. However it must be emphasized that they assumed an independency of the stimulus representation between trials. Their results in fact based solely on the mean of the internal representations of the stimulus sequence in each trial, and they only parallelized their result **as if** subjects were using one or the other strategies. For example, they examined three 2AFC experimental designs: (1) in each trial subjects were presented with one of two pairs of stimuli: standard followed by standard + decrement or standard followed by standard + increment (SDSI) (2) in each trial subjects were presented with one of two following pairs of stimuli: standard followed by standard + increment or standard + increment followed by standard (SIIS) and (3) in each trial subjects were presented with one of the two following pairs of stimuli: standard follow by standard or standard follow by standard + increment (SSSI). They predicted equal performance for (SDSI) and (SIIS) when using the P-C rule and that this performance will be twice as good as using the P-C strategy with (SSSI). However using the AI strategy, the predicted DLs

become much lower for the SDSI design relative to the other two indicating better performance.

In fact, they imply that in order to optimize performance, subjects may change their discrimination strategy in accordance to the experimental design.

The next section provides predictions of discrimination performance in the 2AFC and the reminder task given the two discrimination strategies that may be used by a subject.

Mathematical models

In this section, a quantitative account is provided to explain why DL is larger in the 2AFC than in the reminder task and why there is a positional effect of the standard in the reminder task. Various quantitative models will be suggested for the underlying discrimination process and a comparison of the predictions of these models is made.

The following models of comparison behaviour are based on the assumption that both stimuli presented to the subject create a normally distributed perceived magnitude (dicriminal process) within the subject (Thurstone, 1927). The first model predicts performance in each task assuming subjects use the P-C strategy. It is shown that this assumption predicts identical DLs for the 2AFC and for the reminder task. The second model assumes that subjects use AI strategy for the reminder task and therefore use a stable internal standard, but uses the P-C strategy for the 2AFC task. This model predicts larger DLs in the 2AFC than in the reminder task but it cannot account for the position effect in the reminder task. The third model is in fact an elaboration of the second one. It assumes an internal standard for all tasks (AI strategy), both for the regular and the irregular reminder task, as well as for the 2AFC task. However, in addition it suggests a way of how this internal standard is created. This elaboration combines a model suggested by Morgan et al. (2000) and by Nachmias (2006). This model predicts larger DLs for the 2AFC than for the reminder task. Furthermore, it also implies a larger DL in the reminder task when the standard appears in second rather than in the first stimulus position.

Model 1: Trial-by-trial assessment of standard and comparison durations.

As was previously stated, this strategy proceeds from the assumption that subjects compute the difference between the standard

and the comparison in each trial. Their judgement is based on the size of this difference. I firstly apply this idea to the reminder task and then to the 2AFC task. Assume that each temporal interval in the reminder task– the standard interval t_S and the comparison interval t_C – generates a separate internal representation of the corresponding duration. Let X_S and X_C denote these internal representations, respectively. Just as in the theory of signal detection, it is assumed that these internal durations are noisy, that is, X_S and X_C are normally and independently distributed. To simplify the argument, let the expected mean of X_S be equal to $E[X_S] = t_S$, that is, the average perceived duration of the standard corresponds to its physical duration t_S. In addition, the variance of X_S is equal to $Var[X_S] = \sigma_S^2$. Likewise, the mean of X_C is equal to the physical duration t_C of the comparison, i.e. $E[X_C] = t_C$, and the variance associated with X_C is equal to $Var[X_C] = \sigma_C^2$. According to this model, the subject is assumed to judge the comparison larger than the standard, if $X_C > X_S$, and smaller than the standard, if $X_C < X_S$. (Note that the case $X_C = X_S$ can be ignored because the probability of its occurrence is equal to zero for continuous random variables).

On the basis of these assumptions, it is possible to derive the predicted psychometric function for the reminder task. Note that this function shows the conditional probability $P\{'C > S'| t_C\}$ that the comparison appears to be larger than the standard on the y-axis and the physical duration t_C of the comparison on the x-axis. According to the above assumptions, we therefore can write

$$
\begin{aligned}
P\{'C > S'|\, t_C\} &= P\{X_C > X_S \,|\, t_C\} \\
&= P\{X_C - X_S > 0 \,|\, t_C\} \\
&= 1 - P\{X_C - X_S \le 0 \,|\, t_C\}\,.
\end{aligned}
$$

The term $P\{X_C - X_S \le 0 \,|\, t_C\}$ denotes the probability that the difference $X_C - X_S$ is less or equal to zero. Note that this difference is normally distributed with mean $E[X_C - X_S] = t_C - t_S$ and variance $Var[X_C - X_S] = \sigma_C^2 + \sigma_S^2$. Employing z-transformation, the last expression can be rewritten as

$$
\begin{aligned}
P\{'C > S'|\, t_C\} &= 1 - \Phi\left[\frac{0 - E[X_C - X_S]}{\sqrt{Var[X_C - X_S]}}\right], \\
&= 1 - \Phi\left[\frac{0 - (t_C - t_S)}{\sqrt{\sigma_C^2 + \sigma_S^2}}\right],
\end{aligned}
$$

where $\Phi[z]$ denotes the cumulative distribution function of a standard normal random variable. Using the relation $\Phi[z] = 1 - \Phi[-z]$, the above expression may be rewritten more compactly as

$$
P\{'C > S'|\, t_C\} = \Phi\left[\frac{t_C - t_S}{\sqrt{\sigma_C^2 + \sigma_S^2}}\right] \text{ for } 0 < t_C < \infty. \tag{4.1}
$$

As example, assume $t_S = 500$, $\sigma_S = 50$, and $\sigma_C = 50$ msec. In this case, the psychometric function expressed by Equation 4.1 predicts a PSE and a DL of 500.0 and 47.4 msec, respectively. In general, the predicted PSE is equal to $PSE = t_S$ and the predicted DL equal to

$DL = \frac{t_C(.75) - t_C(.25)}{2} = 0.67 \cdot \sqrt{\sigma_C^2 + \sigma_S^2}$, where $t_C(.75)$ and $t_C(.25)$ are the 75% and 25% percentiles of the predicted psychometric function.

According to model 1, the judgemental process in the 2AFC task is the same as the one in the reminder task. It is straightforward to show that this strategy predicts identical DLs for the 2AFC and reminder task. Note that the psychometric function in the 2AFC task plots the probability of a correct response PC against t_C, for $t_C \geq t_S$. If $\langle CS \rangle$ and $\langle SC \rangle$ indicate the two presentation orders of the standard and the comparison, the probability of a correct response is computed as

$$\begin{aligned} PC(t_C) = & \; P(Correct \cap \langle CS \rangle \mid t_C) + P(Correct \cap \langle SC \rangle \mid t_C) \\ = & \; P(Correct \mid \langle CS \rangle \mid t_C) \cdot P(\langle CS \rangle \mid t_C) + P(Correct \mid \langle SC \rangle \mid t_C) \cdot P(\langle SC \rangle \mid t_C) \\ = & \; P(X_C > X_S \mid t_C) \cdot P(\langle CS \rangle \mid t_C) + P(X_C > X_S \mid t_C) \cdot P(\langle SC \rangle \mid t_C) \\ = & \; P(X_C > X_S \mid t_C) \\ = & \; \Phi\left[\frac{t_C - t_S}{\sqrt{\sigma_C^2 + \sigma_S^2}} \right] \text{ for } t_S \leq t_C < \infty. \end{aligned} \tag{4.2}$$

Note that this functions starts at 0.5 and approaches 1.0 as t_C increases. In fact, Equation 4.2 corresponds to the upper half of Equation 4.1 and consequently this strategy predicts identical DL for the 2AFC and the reminder task, which is, $DL = 0.67 \cdot \sqrt{\sigma_C^2 + \sigma_S^2}$.

Since the data obtained in this study consistently show larger DL for the 2AFC task than with the reminder task then the idea that subjects use the P-C strategy for both task is rejected. In other words, the difference in performance can not be accounted for if we assume that subjects use the P-C strategy for both tasks.

Model 2: Internal standard in the reminder task only.

As already mentioned above, the AI strategy holds that subjects ignore the external standard and decide whether the comparison is smaller or larger than an internal standard, say I_S, in memory. As suggested by Morgan et al. (2000), such an internal standard could be rapidly encoded during the initial testing phase. Like in the previous model, the response '$C > S$' is given whenever the perceived duration X_C of the comparison is larger than I_S, otherwise the subject will respond with '$S > C$'. If the standard is optimally calibrated, the mean of this internal standard would correspond to $E[I_S] = t_S$. Finally, let the variance that is associated with I_S be equal to $Var[I_S] = \sigma_I^2$. Thus, the predicted psychometric function is identical to Equation 4.1, except that σ_S has to be replaced by σ_I. Hence the predicted DL is $DL = 0.67 \cdot \sqrt{\sigma_C^2 + \sigma_I^2}$. In contrast to the reminder task, model 2 assumes that subjects cannot ignore the standard, because they are generally uncertain which of the two stimulus positions contains the standard. Hence, the judgemental process for the 2AFC task is assumed to be identical to the one of the 2AFC task that was assumed in model 1 that is, they must use the P-C strategy.

Model 2 can explain the discrepant results between the 2AFC and the reminder task, if we assume that the internal representation of the standard is stable. Specifically a smaller DL is predicted for the reminder than for the 2AFC task, if $\sigma_I < \sigma_S$. This assumption is supported by recent (e.g., Morgan et al. 2000; Nachmias, 2006; Viemeister, 1970) and classical (e.g., Woodworth & Schlosberg, 1954) psychophysical studies. These studies have revealed that discrimination performance does usually not worsen when the standard is omitted in the reminder task. For example, Viemeister (1970) employed a single-stimulus task. His subjects were asked to rate the intensity of an auditory stimulus on 4-point rating scale. In one condition, an intensity cue preceded the stimulus and in another condition, there was no cue. Interestingly, the discrimination performance did not differ between the two conditions. Based on this finding, Viemeister concluded that subjects ignore the cue but compare the intensity of the stimulus against a stored reference. Furthermore, in a hyperacuity study, Morgan et al. (2000) employed the reminder and the single-stimulus task to assess

performance using line separation discrimination. Discrimination performance was virtually identical in both tasks. In line with Viemeister (1970), this result let these authors to conclude that subjects employ an internal representation of the standard to judge the size of the comparison also when the standard is explicit like in the reminder task. They also conducted a simulation that indicates that such an internal representation is built up by sampling stimulus information over as many as 20 trials.

More recently, Nachmias (2006) assessed discrimination for simple visual patterns with the reminder task and with the single-stimulus task, which omits a preceding standard. He reported almost identical discrimination performance for the two tasks. In addition, he also has found that the position of the standard matters. Like in the current study, when the standard was presented in the first position, performance was better than when the standard occurred in the second position. A similar positional effect has been observed in some timing studies (Marchman, 1969; McGavern, 1965; Rammsayer & Wittkowski, 1990; Van Allen et al., 1966; Ulrich et al. 2006). In these studies, subjects also tended to give more correct responses for pairs of a fixed standard and a variable comparison interval, when the presentation order was standard – comparison than when it was comparison – standard.

To summarize, it is possible to explain the differences results obtain with the reminder task and 2AFC task if we assume that subjects in the reminder task use the AI strategy because they can create stable presentation of the standard. However assuming they can not create this stable representation they use the P-C strategy and performance is impaired. If we assume that the certainty of the position of the standard in the reminder task enables the subject to create this stable representation and ignore the external standard, why is the performance so reduced in the irregular reminder task where the position of the standard is also certain?

Model 3: Internal standard is employed in all tasks.

The present version of model 2 cannot account for this positional effect of the standard on discrimination performance. Consequently, model 3 is suggested to elaborate model 2 by incorporating a recent theoretical idea of Nachmias (2006) on the nature of the internal standard. Accordingly, subjects generate a virtual standard I_S that

combines information A from previous trials with information X_1 of the first stimulus in the current trial. It is assumed that subjects keep a moving average of the internal stimulus representations that precede the current trial; this moving average is denoted by A. For example, I_S might be a weighted combination of A and X_1, that is, $I_S = g \cdot A + (1-g) \cdot X_1$ with $0 < g < 1$. Like in model 2, subjects are assumed to judge the second stimulus longer than the first one, if the internal representation X_2 of the second interval is larger than the updated value of I_S. In order to evaluate the prediction of Nachmias' model, I implemented this in a computer simulation. In agreement with the data of Experiments 5-8, the model predicts a larger DL when the comparison appeared in the first than in the second stimulus position (Appendix B provides a more formal analysis of this prediction). Most importantly, this mechanism can also be applied to the 2AFC task. In agreement with the results of the present experiments, the simulations clearly indicate a larger DL for the 2AFC than for the standard reminder task.[1] In addition, the simulations reveal that this prediction holds whether the standard in the 2AFC task is always smaller than the comparison or always larger than the comparison. Thus, this model provides a promising research perspective for future modelling of discrimination behaviour. Specifically, it suggests a plausible mechanism of how the internal criterion I_S may evolve.

The above model can in fact account for the enhanced performance in the regular reminder task relative to both the 2AFC and the irregular

[1] These simulations assumed that the internal representations of the standard and the comparison (i.e. X_S and X_C) are noisy, that is, normally distributed with a mean equal to the physical duration of the stimuli and a standard deviation of 50 msec. Furthermore, the standard and comparison durations were identical to those used in the experiments. The moving average A was computed across 20 trials that preceded the current trial. We varied the weight g in separate simulations from 0 to 1. For example, for the reminder task with $g = 0.3$, DL was around 40 and around 58 msec when the standard was at the first and the second position, respectively. Increasing g, enhanced predicted performance when the standard was in the first position, but worsened performance when the standard was in the second position. The DL for the 2AFC task was approximately 54 msec, irrespective of whether the standard duration was always smaller or larger than the comparison durations.

reminder task. Taking into account the results from the current study, it seems reasonable to assume that subjects indeed use the AI strategy in all the examined tasks, providing the internal standard really is created as suggested.

5 Summary and conclusion

In summary, then, both the 2AFC and the reminder task are widely in use in psychophysics. It is commonly believed that estimates produced by these two tasks are equivalent. The present results, however, do not support this notion, neither for duration discrimination task nor for visual non-temporal discrimination tasks. DL estimates obtained with the 2AFC task were reliably larger than those obtained from the reminder task. The finding that a standard position effect can also occur in spatial tasks (Nachmias, 2006), supplies further support that although the general discrepancy between the tasks is higher for temporal tasks, as found in the current study, nonetheless, it exists for non-temporal tasks as well. In fact, the proposed models do not differentiate between temporal and non-temporal information processing.

In this study predictions of performance were given for the different tasks assuming subjects are using one of two conflicting strategies the P-C and the AI. A novel model was provided (i.e. model 3) that is based on ideas by Nachmias (2006) and by Morgan et al. (2000). In this model the differences in DL estimates between the tasks can be explained if in both tasks subjects use the AI strategy. Although the idea of using an internal criteria as a base for comparison is not new, this model however specifies a way in which this criteria is created and further more it supply predictions that specifically relate to the position of the standard and the comparison stimuli. In the classical work mentioned above which led to idea of an internal criterion, the position of the stimuli was never manipulated and no predictions were included towards a possible effect of such manipulation. Rather, it was believed that a wide range of standards would have an effect relative to a fixed standard, and that the criterion is somewhere at the mid point of all the preceding stimuli or the presented standards. As well, no attempt was made to compare the different discrimination strategies regarding the different methods, although it was already known on the 50's that threshold measurements depend on procedure (Creelman & Macmillan 1979).

The current study put forward the idea that the difference between the tasks lay within the underlying discrimination mechanism that is used by the subject. It is concluded that for both tasks the subject use the AI decision rule. This underlying mechanism implies that discrimination performance is sensitive to the position of the standard. Specifically,

discrimination performance worsens whenever the standard stimulus is presented in the second stimulus position within a given trial. The model might well account for other findings reported in the psychophysical literature. For example, for the finding that discrimination performance increases when the standard is repeatedly presented before a judgment is required (e.g., Drake & Botte, 1993; Ivry & Hazeltine, 1995; Schulze, 1989). Elaborations of the model might possibly also account for the important finding that discrimination performance is strongly affected by the number comparison intervals (Miller & McAuley, 2005). The model might also explain the finding that performance is better when the standard duration is fixed rather than varied across trials (i.e., roving standard). It may also account for the worsened performance when the range of standard durations is increased (Miller & McAuley, 2005). Future research is necessary to evaluate the prediction of this novel model in these related domains of temporal and spatial-.information processing. Finally, this study strongly indicates that a valuable and reliable comparison between various works can only be made if one takes into account the method in which those studies where conducted.

Zusammenfassung

Die vorliegende Studie beschäftigt sich mit der Fragestellung, ob die Zweifachwahlreaktionaufgabe die gleiche Unterschiedsschwelle (Differenzlimen) schätzt wie die so genannte „Reminder-Aufgabe“, die ursprünglich auch als Methode der konstanten Reize bezeichnet wurde. In einer Serie von sechs Experimenten sollten die Probanden jeweils zwei Zeitintervalle diskriminieren. In den Experimenten 1 bis 5 wurden jeweils auditive und in Experiment 6 visuelle Reize verwendet. In den Experimenten 1 und 2 wurde jede der zwei Methoden mit einer adaptiven und einer non-adaptiven Prozedur der Schwellenmessung kombiniert. In Experiment 3 wurde die Verteilung der Vergleichsstufen variiert, während in Experiment 4 zufällige Interstimulusintervalle verwendet wurden. In den Experimenten 5 und 6 wurde der Einfluss der Präsentationsreihenfolge von Standard- und Vergleichsreiz untersucht. Die Ergebnisse zeigen zum einen, dass sowohl die adaptive als auch die non-adaptive Prozedur die gleichen Schätzungen für die Unterschiedsschwelle ergeben, zum anderen jedoch, dass die Zweifachwahlreaktionsaufgabe konsistent größere Unterschiedsschwellen schätzt als die „Reminder-Aufgabe“. Zusätzlich nimmt die Unterschiedsschwelle zu, wenn der Standardreiz an zweiter und nicht an erster Reizposition präsentiert wird. Die Experimente 7 und 8 prüften, ob diese Ergebnisse nur für zeitliche Reize gelten oder sich auf nicht-zeitliche Reize generalisieren lassen. In Experiment 7 wurden zufällige Punktmuster und in Experiment 8 eine Aufgabe zur Längendiskrimination verwendet. Die Ergebnisse dieser Experimente bestätigen, dass sich die Diskrepanz zwischen den beiden Aufgaben auf die Diskrimination nicht-zeitlicher visueller Information übertragen lässt. Es wird daher angenommen, dass die Probanden statt dem tatsächlich dargebotenen Standardreiz einen internalen Standardreiz als Referenz für ihr Urteil benutzen.

References

Ahmed, I., Lewis, T.L., Ellemberg, D. & Maurer, D. (2004). Discrimination of speed in 5-year-olds and adults: Are children up to speed? *Vision Research*, *45*, 2129-2135.

Alcalá-Quintana, R. &. García-Pérez, M. A. (2005). Stopping rules in Bayesian adaptive threshold estimation. *Spatial vision, 18,* 347-374.

Allchorne, A.J., Broom, D.C. & Woolf, C.J. (2005). Detection of cold pain, cold allodynia and cold hyperalgesia in freely behaving rats. *Molecular Pain*, 36*,* 1-9.

Barnes, R. & Jones, M.R. (2000). Expectancy, attention, and time. Cognitive Psychology, 41, 254-311.

Bausenhart, K. M., Rolke, B. & Ulrich, R. (2007). Knowing when to hear aids what to hear. *The Quarterly Journal of Experimental Psychology*, *60*, 1610-1615.

Berens, M.S. & Pastore, R.E. (2005). Contextual relative temporal duration judgment: An investigation of sequence interruptions. *Perception & Psychophysics*, 67, 102-119.

Birnbaum, M. H. (1994). Psychophysics. *In Encyclopedia of Human Behavior*, *3*, 641-650. Academic Press.

Bode, D., & Carhart, R. (1973). Measurements of articulation functions using adaptive test procedures. *IEEE, Transactions on Audio and Electroacoustics, AU-21,* 196-201.

Boring, E.G. (1917). "A chart of the psychometric function" *American Journal of Psychology*, *28*, 465-. 470

Brand, T. & Hohmann, V. (2002). An adaptive procedure for categorical loudness scaling. *Journal of the Acoustical Society of America, 112,* 1597-1604.

Bush, R.R. (1967). Estimation and evaluation. . In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology. Vol. 1* (2nd ed., pp. 429-469). New York: Wiley.

Buss, E., Hall, J.W., Grose, J.H. & Dev, M.B. (2001). A comparison of threshold estimation methods in children 6-11 years of age. *Journal of the Acoustical Society of America*, *109*, 727-731.

Burgess, A., & Barlow, H.B. (1983). The precision of numerosity discrimination in array of random dots. *Vision Research, 23*, 811-820.

Cao, K.L & Wang, L.E. (2006). Current status and correlated issues on cochlear implantation. *Chinese Medical Journal*, *119*, 971-973.

Corso, J.F. (1963). A theoretico-historical review of the threshold concept. *Psychological Bulletin, 60*, 356-370

Cavanagh, P. & Anstis, S. (1991). The contribution of color to motion in normal and color-deficient observers. *Vision research, 31*, 2109-48.

Creelman, C.D. & Macmillan, N.A. (1979). Auditory phase and frequency discrimination: a comparison of nine procedures. *Journal of Experimental psychology: human perception and performance*, *5*, 146-156.

Dai, H. (1995). On measuring psychometric functions: A comparison of the constant-stimulus and adaptive up-down methods. *Journal of the Acoustical Society of America, 98*, 3135-3139.

Dawis, S. M. (1979). Light adaptation in cone photoreceptors – Occurrence and significance of unitary adaptive strength. *Biological Cybernetics, 34*, 35-41.

Donaldson, G.S., Viemeister, N.F. & Nelson, D.A. (1997). Psychometric functions and temporal integration in electric hearing. *The Journal of the Acoustical Society of America, 101*, 3706-3721.

Drake, C. & Botte, M.-C. (1993). Tempo sensitivity in auditory sequences: evidence for a multiple-look model. *Perception & Psychophysics, 54,* 277-286.

Dunn, W. (2001). The sensations of everyday life: empirical, theoretical, and pragmatic considerations. *American Journal of Occupational Therapy,* 55, 608-620.

Ekman, G. (1959). Weber's law and related functions. *Journal of Psychology, 47*, 343-352.

Emerson, P.L. (1984). Observations on a maximum likelihood method of sequential
threshold estimation and a simplified approximation. *Perception & Psychophysics,* 36, 199-203.

Evans, M., Hastings, N. & Peacock, B. (2000). *Statistical Distributions*, 3rd. Ed., John Wiley and Sons.

Farell, B. & Pelli, D. G. (1999) Psychophysical methods, or how to measure a threshold and why. In R. H. S. Carpenter & J. G. Robson (Eds.), Vision Research: A Practical Guide to Laboratory Methods, Oxford University Press, New York.

Finney, D.J. (1952). *Probit analysis,* 2nd Ed. Cambridge University Press, Cambridge.

Foster, D. H. & Bischof, W. F. (1987). , Bootstrap variance estimators for the parameters of small-sample sensory-performance functions. *Biological Cybernetics, 57*, 341–347.

Fronius, M., Cirina, L., Cordey, A. & Ohrloff, C. (2005). Visual improvement during psychophysical training in an adult amblyopic eye following visual loss in the contralateral eye. *Graefe's archive to clinical and experimental ophthalmology, 243*, 278-280.

García-Pérez, M. A. & Alcalá-Quintana, R. (2005). Sampling plans for fitting the psychometric function. *Spanish Journal of Psychology, 8*, 256-289.

Gelfand, S A. (1990). Hearing: An introduction to psychological and physiological acoustics. 2nd edition. New York and Basel: Marcel Dekker, Inc.

Gescheider, G.A. (1997). *Psychophysics: The fundamentals* (3rd ed). Hillsdale, NJ: Erlbaum.

Getty, D.J. (1975). Discrimination of short temporal intervals: A comparison of two models. *Perception & Psychophysics, 18*, 1-8.

Graham, C. H. (1950). Behavior, perception and the psychophysical methods. *Psychological Review*, 57, 108-120.

Green, D.M. & Swets, J.A. (1966). *Signal detection theory and psychophysics.* New York: John Wiley.

Grondin, S. (1993). Duration discrimination of empty and filled intervals marked by auditory and visual signals. *Perception & Psychophysics, 54*, 383-394.

Grondin, S. (2001). Discriminating time intervals presented in sequences marked by visual signals. *Perception & Psychophysics, 63*, 1214-1228.

Grondin, S. (2003). Studying psychological time with Weber's law. In R. Buccheri, M. Saniga & M. Stuckey (Eds.), The nature of time: Geometry, physics and perception (p. 33-41). Dordretch, Ne: Kluwer.

Grondin, S., Ivry, R. B., Franz, E., Perreault, L. & Metthe, L. (1996). Markers' influence on the duration discrimination of intermodal intervals. *Perception & Psychophysics, 58*, 424–433.

Grondin, S., Meilleur-Wells, G., Ouellette, C. & Macar, F. (1998). Sensory effects on judgments of short-time intervals. *Psychological Research, 61*, 261-268.

Grondin, S., Ouellet, B. & Roussel, M. E. (2001). About optimal timing and stability of Weber fraction for duration discrimination. *Acoustical Science and Technology, 22,* 370-372.

Grondin, S. & Rammsayer, T. (2003). Variable foreperiods and temporal discrimination. *The Quarterly Journal of Experimental Psychology, 56A*, 731-765.

Guilford, J.P. (1954). *Psychometric methods*. New York: McGrew-Hill.

Hall, J.L. (1981). Hybrid adaptive procedure for estimation of psychometric functions. *Journal of the acoustical society of America, 69,* 1763-1769.

Helson, H. (1947). Adaptation-level as frame of reference for prediction of psychological data. *American Journal of Psychology, 60*, 1–29.

Helson, H. (1948). Adaptation-level as a basis for a quantitative theory of frames of reference. *Psychological review, 55*, 297-313.

Hesse, A. (1986). Comparison of several psychophysical procedures with respect to threshold estimates, reproducibility and efficiency. *Acustica, 59,* 263-273.

Hill, N.J. (2001). Testing Hypotheses about Psychometric Functions - an investigation of some confidence interval methods, their validity, and their use in the evaluation of optimal sampling strategies. D. Phil. thesis, University of Oxford, UK.

Ivry, R., & Hazeltine, R.E. (1995). The perception and production of temporal intervals across a range of durations: Evidence for a common timing mechanism. *Journal of Experimental Psychology: Human Perception and Performance, 21,* 1-12.

Jones, M.R & McAuley J.D.(2005). Time judgments in global temporal contexts. *Perception & Psychophysics ,67,* 398-417.

Kaernbach, C. (1991). Simple adaptive testing with the weighted up-down method. *Perception & Psychophysics, 49,* 227-229.

Karmarkar, U. R. & Buonomano, D. V. (2003). Temporal specificity of perceptual learning in an auditory discrimination task. *Learning and Memory, 10,* 141-147.

Killeen, P.R. & Weiss, N.A. (1987). Optimal timing and the Weber function. *Psychological Review, 94*, 455–468.

Klein, S. (2001), Measuring, estimating, and understanding the psychometric function. *Perception and Psychophysics, 63*, 1421–1455.

Klein, S. A. & Macmillan, N. A. (Eds.) (2001). Psychometric functions and adaptive methods [Special issue]. *Perception & Psychophysics*, *63,(8)*.

Kling, J.W. & Riggs, L.A. (1971). Woodward & Schlosberg's Experimental psychology (3rd ed.). New York: Holt, Rinehart and Winston pp. 1281.

Lages, M. & Treisman, M. (1998). Spatial frequency discrimination: Long-term memory or criterion setting? *Vision Research, 38*, 557-572.

Lapid, E., Ulrich, R. & Rammsayer, T. (2008). On estimating the difference limen in duration discrimination tasks: A comparison of the 2AFC and the reminder task**.** *Perception & Psychophysics, 70,* 291-305.

Large, E.W. & Jones, M.R. (1999). The dynamics of attending: How people track time-varying events. *Psychological Review, 106*, 119-159.

Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Perception & Psychophysics, 63,* 1279-1292.

Leek, M.R., Hanna, T.E. & Marshall, L. (1992). Estimation of psychometric functions from adaptive tracking procedures. *Perception & Psychophysics, 51,* 247-256.

Levitt, H. (1970). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America, 49,* 467-477.

Linschoten, M.R., Harvey, L.O. jr., Eller, P.M. & Jafek, B.W. (2001). Fast and accurate measurement of taste and smell thresholds using a maximum-likelihood adaptive staircase procedure. *Perception & Psychophysics, 63,* 1330-1347.

Luce, R.D. (1963). A threshold theory for simple detection experiments. *Psychological Review, 70*, 61–79.

Luce, R. D. & Galanter, E. (1967). Discrimination. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology. Vol. 1* (2nd ed., pp. 191-244). New York: Wiley.

McMillan, N.A. & Creelman, C.D., 1991. Detection Theory: A User's Guide. Cambridge University Press, Cambridge, UK.

Macmillan, N. A.& Creelman, C. D. (2005). Detection Theory: A User's Guide (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Marchman, J. N. (1969). Discrimination of brief temporal durations. *Psychological Records, 19*, 83-92.

Marvit, P., Florentine, M. & Buus, S. (2003). A comparison of psychophysical procedures for level-discrimination thresholds. *The Journal of the Acoustical Society of America*, *113*, 3348-3361.

McAuley, J. D. & Jones, M. R. (2003). Modeling effects of rhythmic context on perceived duration: A comparison of interval and entrainment approaches to short-interval timing. *Journal of Experimental Psychology: Human Perception and Performance, 29,* 1102-1125.

McAuley, D. &. Kidd, G. (1998). Effects of deviations from temporal expectations on tempo discrimination of isochronous tone sequences. *Journal of Experimental Psychology: Human Perception and Performance, 24***,** 1786–1800.

McGavern, M. (1965). Memory of brief auditory durations in comparison discriminations. *Psychological Records, 15*, 249-260.

McKee, S.P., Klein S.A. & Teller, D.Y. (1985). Statistical properties of forced-choice psychometric functions: Implication of probit analysis. *Perception & Psychophysics, 37,* 286-298.

Meese, T.S. (1995). Using the standard staircase to measure the point of subjective equality - a guide based on computer- simulations. Perception & Psychophysics, 57, 267-281.

Miller, N.S & McAuley, J.D. (2005). Tempo sensitivity isochronous tone sequences: The multiple-look model revisited. *Perception & Psychophysics, 67,* 1150-1160.

Miller, J. & Ulrich, R. (2001). On the analysis of psychometric functions: The Spearman-Kärber method. *Perception & Psychophysics, 63,* 1399-1420.

Miller, J. & Ulrich, R. (2004). A computer program for Spearman-Karber and probit analysis of psychometric function data. *Behavior Research Methods, Instruments, & Computers, 36*, 11-16.

Morgan, M. J., Watamaniuk, S. N. & McKee, S. P. (2000). The use of an implicit standard for measuring discrimination thresholds. *Vision Research, 40,* 2341-2349.

Nachmias, J. (2006). The role of virtual standards in visual discrimination. *Vision Research, 46,* 2456-2464.

Nagarajan, S.S., Blake, D.T., Wright, B. A., Byl, N. & Merzenich, M. M. (1998). Practice-Related Improvements in Somatosensory Interval Discrimination Are Temporally Specific But Generalize across Skin Location, Hemisphere, and Modality. *The Journ**al** of Neuroscience, 18*, 1559-1570.

N'Diaye, K., Ragot, R., Garnero, L. & Pouthas, V. (2004). What is common to brain activity evoked by the perception of visual and auditory filled durations? A study with MEG and EEG co-recordings. *Cognitive Brain Research*, *21*, 250–268.

O'Regan, J. K. & Humbert, R. (1989). Estimating psychometric functions in forced choice situations: Significant biases found in threshold and slope estimations when small samples are used. *Perception & Psychophysics*, *46*, 434-442.

Rammsayer, T. (1992). An experimental comparison of the weighted up-down method and the transformed up-down method. *Bulletin of the Psychometric Society, 30,* 425-427.

Rammsayer, T. H. & Lima, S. D. (1991). Duration discrimination of filled and empty auditory intervals: Cognitive and perceptual factors. *Perception & Psychophysics, 50,* 565-574.

Rammsayer, T. & Ulrich, R. (2005). No evidence for qualitative difference in the processing of short and long temporal intervals. *Acta Psychologica, 120,* 141-171.

Rammsayer, T. H., & Wittkowski, K. M. (1990). Zeitfehler und Positionseffekt des Standardreizes bei der Diskrimination kurzer Zeitdauern [Time-order error and position effect of the standard stimulus in the discrimination of short durations]. *Archiv für Psychologie, 142*, 81-89.

Rolke, B. & Hofmann, P. (2007). Temporal uncertainty degrades perceptual processing. *Psychonomic Bulletin and Review,14*, 522-526.

Ross, J. (2003). Visual discrimination of number without counting. *Perception, 32,* 867-870.

Schulze, H.H. (1989). The perception of temporal deviations in isochronic patterns. *Perception & Psychophysics, 45,* 291-296.

Simpson, T.L. (1995). A comparison of 6 methods to estimate thresholds from psychometric functions. *Behavior Research Methods Instruments & Computers, 27*, 459-469.

Simpson, W.A. (1988). The method of constant stimuli is efficient. *Perception & Psychophysics , 44,* 433-436.

Smith, R.A. (1971). Studies of temporal frequency adaptation in visual contrast sensitivity. Journal of Physiology, 216, 531–552

Stellmack, M.A., Viemeister, N. F.& Byrne, A.J. (2004). Monaural and interaural intensity discrimination: Level effects and the "binaural advantage". *Journal of the Acoustical Society of America*, *116*, 1149-1159.

Strasburger, H .(2001).Converting between measures of slope of the psychometric function. *Perception & Psychophysics, 63*, 1348-1355.

Teghtsoonian R. (1971). On the exponents in Stevens' Law and the constant in Ekman's Law. *Psychological Review, 78*, 71-81.

Thompson, J.G., Schiffman, R. & Bobko, D.J. (1976). The discrimination of brief intervals. *Acta Psychologica, 40,* 489-493.

Thurstone, L.L. (1927). A law of comparative judgement. *Psychological Review*, *34*, 273-286.

Treisman, M. (1963). Temporal discrimination and the indifference interval: Implication for a model of the 'internal clock'. *Psychological Monograph, 77*, 1-31.

Treutwein, B. (1995). Adaptive psychophysical procedures. *Vision Research, 35,* 2503-2522.

Treutwein B., & Strasburger, H. (1999) Fitting the psychometric function. *Perception & Psychophysics, 61*, 87–106.

Trevan, J.W. (1927). The error of determination of toxicity. *Proceedings of the Royal Society of London: Series B 101*, 483-514.

Tse, P. U., Intriligator, J., Rivest, J. & Cavanagh, P. (2004). Attention and the subjective expansion of time. *Perception & Psychophysics, 66,* 1171-1189.

Ulrich R. & Miller, J. (2004). Threshold estimation in two-alternative forced-choice (2AFC) tasks: The Spearman-Kärber method. *Perception & Psychophysics, 66,* 517-533.

Ulrich, R., Nitschke, J. & Rammsayer, T. (2006). Crossmodal temporal discrimination: Assessing the predictions of a general pacemaker-counter model. *Perception & Psychophysics, 68,* 1140-1152.

Van Allen, M. W., Benton, A. L. & Gordon, M. C. (1966). Temporal discrimination in brain-damaged patients. *Neuropsychologia, 4*, 159-167.

Van Oeffelen M. P. & Voss P. G. (1982) Configurational Effects on the Enumeration of Dots. *Memory & Cognition, 10*, 396-404.

Viemeister, H.F. (1970). Intensity discrimination: performance in three paradigms. *Perception & Psychophysics, 8,* 417-419.

Vogels, R. & Orban, G.A. (1986). Decision processes in visual discrimination of line orientation. *Journal of Experimental Psychology: Human Perception and Performance, 12*, 115-32.

Watson, B.A. & Fitzhugh, A. (1990). The method of constant stimuli is inefficient. *Perception & Psychophysics, 47,* 87-91.

Ward, L.M. (2003). Synchronous neural oscillations and cognitive processes. *Trends in Cognitive Science, 7*, 553–559.

Wichman, F.A. & Hill, N.J. (2001). The psychometric function: I. Fitting, sampling and goodness-of-fit. *Perception & Psychophysics, 63*, 1293-1313

Woodrow, H. (1933). Weight discrimination with a varying standard. *American Journal of Psychology, 45*, 391–416.

Woodworth, R.S. & Schlosberg, H. (1954) Experimental Psychology, 3rd ed. London: Methuen

Wright, B.A., Buonomano, D.V., Mahncke, H.W. & Merzenich, M.M. (1997). Learning and generalization of auditory temporal-interval discrimination in humans. *Journal of Neuroscience, 17*, 3956-3963.

Wright BA, Sabin AT. (2007). Perceptual learning: How much daily training is enough? Experimental Brain Research, 180 (4), 727-736

Witew, I.B, Behler, G.K. & Vorländer, M. (2005). About just noticeable differences for aspects of spatial impressions in concert halls. *Acoustics Science & Technology, 26*, 185-191.

Zeng, F.G & Shannon, R. V. (1999). Psychophysical laws revealed by electric hearing. Auditory and Vestibular Systems. *Neuroreport, 10*, 1931-1935.

Appendix A: Monte Carlo Simulation

The purpose of the following simulations was to rule out the possibility that the discrepancy of the DL estimates between the 2AFC and the reminder tasks is due to the estimation procedure applied. Four programs, each matching exactly a single combination of task (2AFC vs. reminder) and procedure (adaptive vs. non-adaptive), were written to simulate the responses of a virtual subject according to psychometric functions (1) and (2). These data were used to estimate the DL with the routines that were also used to estimate the DL for the real subjects in the present experiments. All other aspects (e.g., stimulus levels, number of trials) were exactly matched to the real experiments. The parameters for the underlying psychometric function were DL=50 msec, for both tasks, and PSE=500 msec for the reminder task. Each program simulated 10,000 virtual subjects. Table A1 summarizes the results of these simulations. Specifically, it shows the overall mean DL and the standard deviation of the estimates. The last column shows the percentage of cases in which the estimation program did not converge. Generally, the procedures did satisfactorily recover the expected value of DL=50 msec. The results of the simulations do not indicate that the estimation procedure is a possible source of the discrepant DL from the 2AFC and the reminder tasks. The second parameter that was estimated from the simulation results was the PSE. For the adaptive as well as for the non-adaptive procedure the targeted value of PSE=500 msec was obtained and, thus, the simulation results suggest no bias in estimating the PSE (Table A2).

Table A1 DL simulation results. Mean estimated DL and the standard deviation of the DL estimates as a function of task and procedure.

Task/Procedure	DL (msec)	SD (msec)	Percentage of Non-Convergence
Reminder/adaptive	43.8	8.7	0
Reminder/non-adaptive	50.5	10.5	0
2AFC/adaptive	50.1	9.6	0.02
2AFC/non-adaptive	48.9	13.8	0.02

Table A2 PSE simulation results for the reminder paradigm. Mean estimated PSE and the standard deviation of the PSE estimates as a function of procedure.

Task/Procedure	PSE (msec)	SD (msec)
Reminder/adaptive	499.95	12.76
Reminder/non-adaptive	500.06	10.63

Appendix B: Moving Average Model and the Positional Effect of the Standard

This appendix proofs that the moving average model predicts a smaller DL when the standard is in the first than in the second stimulus position. In order to simplify things, the focus will be on the reminder task in this appendix and only address the standard position effect. (A formal analysis of the 2AFC task is complex, and thus, beyond the purpose of this study.) Note that the second interval is perceived longer than the first one, if the event $X_2 > I$ occurs, that is, when the internal representation X_2 of the second stimulus is larger than the internal standard I. Thus, the probability of the response $S_2 > S_1$ (i.e. second interval appears longer than the first interval) is given by

$$\begin{aligned} P\{'S_2 > S_1'|\, t_C\} &= P\{X_2 > I \mid t_C\} \\ &= 1 - P\{X_2 - I \le 0 \mid t_C\} \\ &= 1 - P\{X_2 - g \cdot A - (1-g) \cdot X_1 \le 0 \mid t_C\} \\ &= 1 - \Phi\left[\frac{0 - E\left[X_2 - g \cdot A - (1-g) \cdot X_1\right]}{\sigma_*}\right] \\ &= \Phi\left[\frac{E\left[X_2 - g \cdot A - (1-g) \cdot X_1\right]}{\sigma_*}\right]. \end{aligned} \quad \text{(B1)}$$

Note that $E\left[X_2 - g \cdot A - (1-g) \cdot X_1\right]$ and σ_* denotes the mean and the standard deviation of the random variable $D = X_2 - g \cdot A - (1-g) \cdot X_1$, respectively. The variable D represents the internal difference between the second stimulus and the internal standard. It can be shown that the standard deviation

σ_* does not depend on the order of the standard and the comparison, although, the mean of D does depend on this order.

Specifically, if the standard is in the first position, the expected value of D is

$$\begin{aligned} E\left[D \mid \langle S, C \rangle\right] &= E\left[X_2 - g \cdot A - (1-g) \cdot X_1\right] \\ &= E\left[X_2\right] - g \cdot E\left[A\right] - (1-g) \cdot E\left[X_1\right] \\ &= t_c - g \cdot t_s - (1-g) \cdot t_s \\ &= t_c - t_s . \end{aligned} \tag{B2}$$

Inserting (B2) into (B1) yields the psychometric function for the case that the standard occurs always in the first position,

$$P\{'S_2 > S_1' \mid \langle S, C \rangle, t_C\} = \Phi\left[\frac{t_c - t_s}{\sigma_*}\right]. \tag{B3}$$

The DL associated with (B3) is equal to $0.67 \cdot \sigma_*$.

A similar reasoning shows that when the standard occurs in the second position, the corresponding psychometric function is given by

$$\begin{aligned} P\{'S_2 > S_1' \mid \langle C, S \rangle, t_C\} &= \Phi\left[\frac{(1-g) \cdot (t_s - t_c)}{\sigma_*}\right] \\ &= \Phi\left[\frac{t_s - t_c}{\frac{\sigma_*}{1-g}}\right] . \end{aligned} \tag{B4}$$

Therefore, the DL associated with (B4) must be equal to $0.67 \cdot \sigma_* /(1-g)$. Note that this DL must be larger than the DL associated with (B3), which completes the proof. This prediction follows from the fact that mean D is smaller when the standard is presented in the second than in the first stimulus position. This reduction in mean D diminishes the perceptible difference between the standard and the comparison, and as a result, lowers discrimination performance.