The Sequence of Standard and Target

in Pairwise Magnitude Comparisons

Inauguraldissertation

zur

Erlangung des Doktorgrades

der Humanwissenschaftlichen Fakultät

der Universität zu Köln

nach der Promotionsordnung vom 18.12.2018

vorgelegt von

Victoria Marie Striewe

aus Paderborn

Köln, Juni 2021

Diese Dissertation wurde von der Humanwissenschaftlichen Fakultät der Universität zu Köln im März 2022 angenommen.

Erstgutachterin: Prof.'in Dr. Birgit Träuble
Zweitgutachterin: Prof.'in Dr. Hilde Haider

Acknowledgements

I want to thank the Center of Social Economic Behavior of the University of Cologne (C-SEB) that funded a major part of the empirical part of this dissertation *The Sequence of Standard and Target in Pairwise Comparisons*. C-SEB provided two Junior-Research Grants – the first in November 2017 titled *SNARC Effect in Price Perception* to finance the first four experiments. and the second in May 2018 titled *Sequence Effects of Target and Standard in Price Comparison,* for another five experiments. Without the financial support of C-SEB the experiments that are the basis of this dissertation could not have been realized.

The hypotheses and experimental designs lined out in this dissertation were created in cooperation with Prof. Dr. Sascha Topolinski who kindly dedicated his scientific expertise and experience to this project. Without his groundbreaking feedback and ideas the project would not have been possible. Furthermore, Prof. Topolinski made his lab and research infrastructure available for the studies of the project.

I want to thank the Social Cognition Center Cologne (SoCCCo) for being a school for excellent scientific work and a rich network of knowledge during my time as a doctoral student.

Foremost, I want to thank Prof. Dr. Jutta Stahl for empowering me as a woman to engage in cognitive research in order to raise female expertise in cognitive science and to harvest the output of my hard work. Prof. Stahl, you are a great role model for idealistic women that fight patriarchal structures. Besides being a supporter in scientific work, dealing with setbacks and challenges you were an outstanding, strong believer in me and my work. Without your support this dissertation would not have been possible.

Another special thank you goes to Prof. Dr. Birgit Träuble for supervising my dissertation and empowering me no less than Prof. Stahl. Prof. Träuble did an outstanding job

in building me up and taking my fear of failure and self-doubts as it is unquestionable that the greatest challenge of this work was the believe in me as a person who deserves to be honored for her work.

Thank you, Carola, for being my scientific sparring partner! And for your support that goes far beyond good turns in friendship. Your advice and feedback grounded me and were a calming guidance through my own thoughts. I will never forget this!

Summary

The present research introduces the effect of the presentation order of target and standard in paired magnitude comparisons on comparison performance. So far, this effect has been overlooked by most of the domains of psychological research on comparative thinking. The *standard-target-sequence-effect* (STSE) was demonstrated in eight out of eleven experiments (*N* = 1,018) presented in the work at hand. Participants repetitively performed simple magnitude comparisons of two objects (e.g. one digit numbers or geometric shapes) in various economic and social contexts. Results revealed a stable performance advantage (in terms of speed and accuracy) for trials in which the standard stimulus was encountered before the to be judged target stimulus. In three experiments the STSE could not be observed, most likely because of the relative spatial and temporal positions of stimuli. The diverse findings and experimental set ups are discussed as well as the underlying mechanism, the interaction of the STSE with the SNARC effect for numerical comparisons (Dehaene, Dupoux & Mehler, 1990; Dehaene, Bossini & Giraux, 1993; Fisher, Castel, Dodd & Pratt, 2003) and the ascending order advantage in magnitude judgement tasks (Turconi, Campbell & Seron, 2006; Müller & Schwarz, 2008; Schroeder, Nuerk & Plewnia, 2017). The effect of the order of target and standard on comparison processes had been mentioned in signal detection and stimuli discrimination tasks in psychophysics (so called Type B Effect, e.g. Dijas & Ulrich, 2014), while social and cognitive psychologists' research on judgements of similarity and contrast have provided inconsistent results for the influence of the sequence of standard and target on the comparison process (e.g. Tversky, 1978; Agostinelli, Sherman, Fazio & Hearst, 1986). Researchers on symbolic pairwise comparisons did not report such an effect at all. The research on the STSE outlined in the work at hand contributes to an interdisciplinary understanding of order effects of target and standard as well as to the debate on the origins of order effects in general and on the basic principles of comparative thinking.

Zusammenfassung

In der hier dargestellten Forschungsarbeit wird der Effekt der Präsentationsreihenfolge von Zielreiz (Target) und Vergleichsreiz (Standard) bei gepaarten Größenvergleichen auf die Performanz (gemessen in Reaktionszeit und Fehlerrate) vorgestellt. Bislang wurde ein solcher Effekt in den meisten Bereichen der psychologischen Forschung zum vergleichenden Denken übersehen. Der *Standard-Target-Sequence-Effekt* (STSE) wurde in der vorliegenden Arbeit in acht von elf Experimenten (*N* = 1.018) nachgewiesen. Die Teilnehmer:innen führten wiederholt einfache Größenvergleiche von zwei Objekten (z.B. einstellige Zahlen oder geometrische Formen) in verschiedenen ökonomischen und sozialen Kontexten durch. Die Ergebnisse zeigten einen stabilen Performanzvorteil für Durchgänge, bei denen der Standard vor dem Target präsentiert wurde. In drei Experimenten zeigte sich der STSE nicht, wahrscheinlich aufgrund der relativen räumlichen und zeitlichen Position der Stimuli. Die verschiedenen Befunde und Versuchsaufbauten werden diskutiert, ebenso wie der zugrundeliegende Mechanismus und die Interaktion des STSE mit dem SNARC-Effekt bei numerischen Vergleichen (Dehaene, Dupoux & Mehler, 1990; Dehane et al., 1993; Fisher, Castel, Dodd & Pratt, 2003) und dem *ascedning order advantag* in Aufgaben zu numerischen Vergleichen (Turconi, Campbell & Seron, 2006; Müller & Schwarz, 2008; Schroeder, Nuerk & Plewnia, 2017). Der Effekt der Reihenfolge von Target und Standard auf Vergleichsprozesse wurde in *Signal Detection*- und Stimulusdiskriminationsaufgaben in der Psychophysik erwähnt (sog. Typ-B-Effekt, z.B. Dijas & Ulrich, 2014), während die Forschung der Sozial- und Kognitionspsychologie zu Ähnlichkeits- und Kontrasturteilen inkonsistente Ergebnisse bezüglich des Einfluss der Reihenfolge von Standard und Target auf den Vergleichsprozess lieferte (z.B. Tversky, 1978; Agostinelli, Sherman, Fazio & Hearst, 1986). Forscher zu symbolischen paarweisen Vergleichen haben einen solchen Effekt bisher nicht berichtet. Die in der vorliegenden Arbeit dargestellte Forschung zum STSE leistet einen

Beitrag zum interdisziplinären Verständnis von Effekten der Anordnung von Target und Standard sowie zur Debatte über die ihnen zugrundeliegenden Mechanismen sowie über die Grundprinzipien des vergleichenden Denkens.

Table of Contents

List of Tables and Figures

Tables

Figures

1. Theoretical Background

Comparative thinking is part of our earliest cognitive development. It is the key mechanism of orientation in a complex environment and of the efficiency of human information processing. Without comparisons we would have to asses every single stimulus in every single situation when encountering it for the first time. Mental overload and time extensive processing would be the inevitable consequences. Comparisons the goal of resource preservation as they limit the range of information that has to be considered to evaluate or judge a given object (Mussweiler, 2009). Comparative information processing enables us to focus on a subset of knowledge that is immediately relevant. One could say, that comparative thinking as the basis of categorization processes and efficient information processing is a key native ability that has maintained survival since early human development (Leth-Steensen & Marley, 2000) – comparisons enable us to decide if an encountered subject is a predator or a friend, if one can react on an encountered object in a routine based way or whether it requires elaborate new behaviors. When we encounter an object (subject) for the first time, we encode its single features and based on this, perform contrast- and assimilation-decisions (Tversky, 1977; Tversky & Gati, 1978), sort it into preexisting accessible cognitive categories in which we store the acquired information and from which we can easily retrieve it when necessary in the future. This comparison- and categorization-based processing enables us to make sense of the occurrence of objects, subjects or a subject's motivation. And, with every new experience, step by step, these processes pave the way to a bigger picture of objects, the world and the self and to an energy saving processing of information. In terms of efficiency we categorize objects and subjects in good and evil, tall and small, beautiful and ugly, in a foremost preconscious automatic way to make fast decisions. From an evolutionary perspective especially magnitude information had to be assessed very quickly, because the relative judgement of an object's (subject's) size or spatial or hierarchical level on which it stands has been of behavioral relevance and surely maintained survival. For psychological research on comparative thinking this implies that the study of magnitude comparisons should unfold basic principles of comparative thinking in general. This assumption is evident from an extensive body of research on magnitude comparisons and from the common practice in psychological research to quantify ratings, judgements and any kind of measured property.

To study how humans form mental categories and define categories' borders, psychologists sample observations of behavior and decisions while test subjects engage in comparisons. While the first steps of this psychological research arose from the study of human experience of physical sensory input, today, psychological research is foremost interested in the experience of and reaction to rather complex stimuli – stimuli that range from objects or subjects, that consist of several, partly ambiguous, information, to the measurement of interindividual differences in preference and choice.

In the past 50 years of psychological research scientists shed light on the universal mechanisms, basic processes and neuronal circuits that are involved in comparative thinking and that these processes are widely shared between diverse comparison domains (e.g. Moyer, 1973; Link & Heath, 1975; Dehaene, 1989; Mussweiler, 2003). A well-studied example is the *distance effect* in comparative judgment that could be found in comparisons of psychophysical- (Moyer & Landauer, 1967), numerical- (Dehaene, Dupoux & Mehler, 1990), symbolic- (Moyer & Bayer, 1976) and social- (e.g. Festinger, 1954; Mullen & Hu, 2010) categorization performance: With decreasing distance on a critical dimension the discrimination of two objects gets more difficult.

Still there are discontinuities of comparative thinking that have been studied in only one or a few psychological domains of comparative thinking, like the effects of stimulus presentation order found in psychophysics. And furthermore, the basic unconscious patterns of attentional resource allocation as well as contrast- and assimilation processes shared by all human beings remain partly undisclosed.

This chapter provides an overview of the so far investigated principles and biases of comparative thinking that partly have been introduced as crossover effects from one research domain to another (like the distance effect) and discontinuities that currently are evident only for a limited range of stimulus material. The chapter stresses the relevance of the experimental paradigm of *pairwise comparisons* for experimental psychology and discovers the parallels of experimental set ups in diverse lines of psychological research. Focusing on magnitude comparisons it introduces experimental set ups, effects and models of psychophysicists, as well as the research on symbolic magnitude comparisons, similarity judgments, change detection and preferences ratings. As well, it discloses the gap of research on the effects of the sequence of target and standard in pairwise comparisons which the empirical part of this work deals with.

1.1 Pairwise Comparisons in Psychophysics

Psychophysics was the first psychological discipline that investigated relative magnitude judgements. Its evolvement went hand in hand with the rise of psychometric approaches that attempted to measure and quantify human experience of physical stimuli varying in intensity with numerical scales. Linguistically magnitude expressions were used already before, to talk about sensations of varying intensity, but with numbers these sensations could be measured discretely and objectifiable. The laws of the relation of human experience and physical stimuli, first stated by experimental psychologists in the in second half of the 19th century, are accepted approaches to explain human relative thinking and decision making until today. Pairwise comparisons, first employed by psychometricians to create scales for human experience and behavior, form the basis of contemporary experimental psychology. But even after 150 years of research on the relative encoding of magnitude information of physical stimuli, the mechanisms underlying the observed systematical discontinuities of comparative judgments remain unresolved.

1.1.1 The Measurement of the Subjective Experience of Physical Magnitudes

Studying human experience of physical stimuli raises the question, at which threshold on a continuum of a certain perceptible characteristic a stimulus is categorized to belong to one category or another. Whether a stimulus is perceived as larger, heavier or more intense than the stimulus encountered just before depends on the *just noticeable difference* (JND). The term JND is interchangeably used with *difference limen* (DL), whilst recent publications use the term DL most frequently. For instance, in psychophysical experiments of weightlifting the DL of the subjective experience of weight increase in relation to actual weight increase is measured by gradually adding weight to a to be judged stimulus in a pairwise presentation. The amount of added weight that is necessary to let the participant perceive an increase of weight, is defined as the DL. This procedure is similar in all psychometric experiments independent of the to be measured modality. The amount of added weight (light, decibel etc.) that is heavy (bright, loud etc.) enough to exceed the threshold of JND differs between modalities and is affected by situational environmental factors, like the order of the stimuli and their relative intensity. Experimental psychologists began to study these situational environmental factors to clarify the laws of subjective experience and decision making (e.g. Fechner, 1860; Cattel, 1902; Michel & Helson, 1954; Audley & Wallis, 1962; Banks & Root, 1979).

The Weber-Fechner-Law. The founding fathers of psychophysical research formulated *laws* of the human experience of physical stimuli that have remained the key assumption of every psychophysical study today. In the 1830s, Ernst Heinrich Weber (Weber,

1851) discovered that the subjective discriminability of two stimuli is affected by the relative intensity of these stimuli. The more intense two stimuli were, the greater the difference had to be to be just noticeable (DL). Further investigating this effect Gustav Fechner reformulated the *Weber law* and named it *Weber-Fechner-Law* in 1860 (Fechner, 1860). He stated that the experienced intensity of a stimulus equaled the logarithmic function of the stimulus' intensity (also referred to as *psychometric function*).

Stevens' Power Function. In the middle of the 20th century Stanley S. Stevens (Stevens, 1957) stated a *power law* for the relation of actual physical magnitude (in terms of brightness, weight or loudness etc.) increase and experienced magnitude increase. In his experiments, he deviated from the standard procedure of testing the DL and let participants compare two stimuli updated from trial to trial in paired presentation. The participants had to perform ratio estimations instead of absolute judgements. With this approach he attempted to account for a decrease of sensitivity due to perceptual adaption that he stated to be inevitable in the previously accepted method of the stepwise forced choice to measure DL. His results pointed at a power function of the relation of experienced intensity and physical intensity with a special exponent for every perceptual modality. Psychophysics still debate on whether Fechner's Law or Stevens' Law is more valid (Krueger, 1989; Dehaene, 2003).

In actual psychophysical studies – besides the DL – reaction times (RT) in discrimination tasks are measured to describe participants' performance in paired comparisons as they reflect the difficulty of the task over all participants (e.g. Link & Heath, 1975), thus allowing to draw conclusions on discrimination sensitivity.

1.1.2 Classical Experimental Setups for Relative Judgments of Physical Magnitudes

Today, the most common experimental set up in psychophysical studies of relative judgment is the *two-alternative forced choice* (2AFC) paradigm. Mostly, participants are instructed to select the stimulus of a pair that fits a predefined decision criterium best. The to be compared stimulus pairs are constructed or selected by the experimenter in advance of the experiment. To control for unintentional effects and measurement errors, stimuli have to be identical in all characteristics, despite of the characteristic in question. The critical characteristic is the only one that varies between the stimuli. For instance, when two tones must be compared regarding their duration, their loudness and frequency are held constant throughout the trials of the experiment (e.g. Hellström & Rammsayer, 2015; Hellström, Patching & Rammsayer, 2020). In psychophysical experiments the two stimuli that are to compare within a trial are presented successively, as one can only focus attention on one stimulus at a time. All other sensory input in an experimental session is reduced to a minimum of potential distractors whenever they are not part of a systematically installed and controlled factor.

Participants are instructed to respond to each experimental trial by choosing one of two optional answers – for instance, ‘first larger’ versus ‘second larger’. To analyze response patterns over participants, the experimenter distinguishes between a standard stimulus and a target stimulus in each trial. In some publications the standard stimulus is referred to as the *referent* and the target sometimes is labeled as *test stimulus*, *study stimulus* or the *comparison*.

In psychophysics most common are *classification tasks*, indicated by an instruction to decide whether a target is the same or different from a standard stimulus. Here, the distinguishability of target and standard is necessary for the participant to perform the task and for the experimenter who computes the DL with the collected data.

Depending on the research question and on the independent variables that had been hypothesized to influence the DL measurement, sometimes *selection tasks* instead of classification tasks are performed. Selection tasks are appropriate when the influence of the presentation order of stimuli on the discriminative sensibility is in the focus of the investigators' interest. In this kind of task, the instructions indicate to 'choose' one of the stimuli that subjectively fits the outlined decision criterium. In this case, which stimulus is the target and which is the standard is not a necessary information and is untransparent for the participants. Only the experimenter knows which stimulus is the standard and which is the target of a trial to be able to calculate the DL. During the experiment, the experimenter or a computer, programmed according to the research question, controls the sequential presentation of the stimuli.

In the beginning, starting with Weber and Fechner, and for a long period in psychophysical research, the standard was the stimulus that was held constant throughout all the trials of an experiment. This modus operandi caused learning effects and systematical response behavior on the side of participants and was time consuming for the experimenters. So, according to Stevens' findings (Stevens, 1957) researchers felt compelled to vary the expression of the standard stimulus between the test trials depending on the discrimination response in the trial before (e.g. Dyjas, Bausenhart & Ulrich, 2012). In contrast to the old measurement when both stimuli of a trial, the standard and the target, had been selected from a predefined list of stimuli, in modern adaptive measurements, stimuli are generated according to previously programmed algorithms with the help of a computer. The experimental computer can adjust the stimuli online in direct reaction on the participant's response behavior (Leek, 2001; Rammsayer, 2012).

1.1.3 Discontinuities in the Judgment of Physical Magnitude

The relativity of human perception and human judgement that Weber and Fechner investigated is an example for the proneness to implicit biases and heuristics of human information processing. Today, psychologists broadly explain these processing shortcuts as a consequence of the omnipresent principle of the human mind to work efficiently and goal directed. It has become a standard experimental approach to observe and record behavior and its systematic *errors* to conclude on underlying mechanisms in the human mind. The most prominent discontinuities examined in psychophysical studies are the *law of comparative judgement*, the distance effect, the *semantic congruity effect*, the *time-order-error* and the *standard-position effect* (or *Type B effect*).

Law of Comparative Judgement. One of the first very influential findings in the study of relative human experience was the *law of comparative judgments* observed by Thurstone in 1927. Thurstone reported that the relative judgement of the same two stimuli varies between points of measurement within the same participant. Thurstone’s observation led to today’s common practice of repeated measurement – the aggregation of several points of measurement into one mean value – and led to a fundamental questioning of absolute categories in human cognition and experience.

Time Order Error. Fechner (1860) discovered that one stimulus in a sequence of two identical stimuli tends to be overestimated, thus being judged as greater (heavier, longer etc.) than the other. He named this effect *time order error* (TOE). When the first stimulus was overestimated compared to the second stimulus, Fechner called it a positive TOE, when the second stimulus was overestimated, he called it a negative TOE. The common definition of the positive TOE is a rating of higher magnitude (weight, duration etc.) for the first stimulus compared to the second stimulus although they have the same magnitude (weight, duration etc.). The negative TOE is defined as a rating of higher magnitude (weight, duration etc.) of the second stimulus compared to the first stimulus of actually equal magnitude (weight, duration etc.). Psychophysicists later referred to the TOE as *Type A effect.* The Type A effect is mostly observed in studies that measure the *point of subjective equality* (PSE) with undirected equality judgment tasks (Dyjas & Ulrich, 2014). The positive Type A effect is more frequently reported than the negative Type A effect. Most attempts to explain the Type A effect assume it to be a perceptual or cognitive effect, not a bias at the judgment level of reaction (Hellström, 1985; Hellström et al., 2020) but the debate on the effect's origin remains unresolved until today.

Standard Position Effect. Early psychophysical studies found that the psychometric function differs depending on the presentation order of the standard and the target. The DL increased and the psychometric function was steeper, indicating decreased discrimination sensitivity when the target stimulus preceded, instead of following, the standard stimulus. This effect was first reported by Martin and Müller in 1890 for weightlifting. A *standard position effect* (SPE), psychophysicists refer to as *Type B effect*, just like the Type A effect, can be positive or negative. A positive Type B effect indicates an increased discrimination sensitivity for the order target precedes standard, a negative Type B effect indicates an increased discrimination sensitivity for the order standard preceding target. The negative Type B effect is found much more frequently than the positive Type B effect and was reported for different stimuli (e.g. duration discrimination, weightlifting and visual shapes; for an overview see Dyjas et al., 2012), for various tasks (comparative judgement tasks as well as for equality judgement; Dyjas & Ulrich, 2014) and for various outcomes, as well as for a lower DL and a lower PSE (Dyjas & Ulrich, 2014). The common explanation for the effect is that the target has a greater impact on the comparison outcome than the standard.

The difference between Type A and Type B effect is, that the Type A effect refers to the stimulus order in an undirected equality judgement task, affecting the PSE, and the Type B effect refers to the stimulus order in a directed 2AFC task, affecting the DL. The observation of a SPE requires the experimental variation of the position of standard and target. However, the common experimental procedure due to technical reasons for a long time was to present a constant standard before a varying target, discovering only the TOE (Dyjas &

Ulrich, 2014). Current studies that employ roving standards made it possible to disentangle both effects.

Distance Effect and Congruity Effects. Further effects of environmental factors have been reported to affect RT in discrimination tasks. In an experiment by Cattell, published in 1902, participants' RT in the discrimination of the brightness of two subsequently presented cards increased as the difference in brightness of the two cards decreased. Cattel concluded that psychophysical stimuli are easier to discriminate with increasing distance on the measured dimension. This finding could be replicated for other modalities (e.g., Woodworth & Schlosberg, 1954; Moyer, 1973).

Furthermore, a linguistic effect was demonstrated in Cattel's experiments in the RT patterns of the discrimination tasks. When instructions indicated to 'choose the brighter' of two stimuli, participants reacted faster (and more likely correct) on two relatively bright stimuli, when the instruction indicated to 'choose the darker' stimulus, participants performed faster (and more likely correct) when the two to be compared stimuli were relatively dark. This effect was frequently replicated after Cattel's publication with varying stimulus material. Audley and Wallis (1964), in their study on the perception of differences in illuminance, called this effect the *crossover effect*. In 1979, Banks and Roots introduced the term *semantic congruity effect* (SCE) in a publication on the discrimination of tones regarding loudness. They hypothesized the effect to originate from the closeness of the semantic code of the instruction (e.g. 'chose larger') and the code for the perceived stimulus (e.g. louder).

1.1.4 Models of the Discontinuities of Comparisons of Physical Magnitudes

Until recently, the TOE and the SPE mainly were reported as side effects of of psychometric measurements. For a long time, researchers overlooked effects of stimulus presentation order due to aggregating data across presentation orders, underestimating their role as confounds of the measurement of DL and PSE. Whereas, in recent studies, since the second half of the 2010s (Hellström & Rammsayer, 2004; Dyjas et al., 2012; Dyjas & Ulrich, 2014; Hellström & Rammsayer, 2015; Hellström et al., 2020), the origins and underlying mechanisms of stimulus order effects have been of key interest. While in early psychophysical studies the standard stimulus was held constant in physical characteristics, stimulus onset, duration and sequential position to avoid biases like the TOE and SPE, studies that investigate the effects' underlying mechanisms systematically vary these four properties and integrate them as independent variables in statistical analyses. The aim of the models introduced below, is to explain human comparative information processing behind DL, PSE and its discontinuities due to stimulus order.

Simple difference model. The first model of comparisons of sensory input was stated by Thurstone in 1927. He theorized that comparisons were performed corresponding to a *simple difference model* (Thurstone, 1927) stating that outcomes of comparisons were completely determined by the difference of the to be compared physical stimuli. Furthermore, Thurstone assumed that physical stimuli would not be mentally represented as discrete expressions like the experimental input, but that the mental representations rather varied on *psychological continua*. With this theory, Thurstone attempted to explain the law of comparative judgement and the finding that humans could not make absolute judgements of physical magnitudes.

Internal Reference Models. Based on Thurstone's theory, researchers in the 1950s concluded that the human adaptive sensory systems formed internal representations of stimuli rather than absolute encodings, and that these internal representations, rather than the sensory input itself, were compared to other stimuli to judge their relative magnitude (e.g. Michels & Helson, 1954). Since then, the approach of internal reference points has been very influential in many fields of psychological research on pairwise comparisons. These reference points have been described as internal unconscious mnemonical anchors in judgment processes (e.g. I/E model by Marsh, 1986; Dyjas et al., 2012; Jamieson & Petrusic, 1975). For instance, Marks (1972) and Jamieson & Petrusic (1975) hypothesized that the distance effect and the SCE were caused by decreased ease of discrimination with increasing distance from these internal anchors (further examples for the reference point models in section 1.2.3).

Psychophysicists’ investigations on the stability of the effects of stimulus order under various conditions promise further insight into the formation of internal reference points and refine according models.

In 1954, Michels and Helson stated a model in order to explain the TOE effect. The *Michels-Helson-model* (MH) claimed that in pairwise comparisons of physical stimuli the participants did not compare the second presented stimulus directly to the first presented stimulus, but to a weighted compound of the first presented stimulus and to the serie’s adaption level (that is “[…] a weighted geometric mean of previously experienced stimuli with weights according to their degree of recency”, Hellström et al. 2020, p. 3197).

In 2012, Dyjas, Bausenhart and Ulrich reformulated the MH-model. Their *Internal Reference Model* (IRM) assumes that the second stimulus of a to be compared pair was not compared with the first stimulus but with an *internal reference* (IR). The IRM, in contrast to what was stated by the MH model, is updated in a dynamical process, where the IR in the current trial is a weighted mean of magnitudes of the first stimulus in the current pair and the IR of the previous trial.

To test this hypothesis of a trial by trial updating of the IR and to rule out systematical response behavior as an explanation for stimulus order effects, Dyjas et al. (2012) set up experiments on duration discrimination with a variation of the sequence of target and standard between the test trials1. Participants performed either a selection task (comparative judgement) or a classification task (equality judgement). In both tasks, the standard stimuli were hold constant throughout the experiment with a duration of 500 ms while the duration of the target stimuli varied between 400 ms and 600 ms in steps of 20 ms from trial to trial. The sequential order of target and standard was varied between blocks of several trials within participants. Although participants could not identify which stimulus of a pair was the target and which was standard, a strong negative Type B effect showed up. The DL was twice as high in ‘standard preceding target’-trials compared to ‘target preceding standard’-trials. The authors found shorter RT (only in the equality judgment condition) and better discrimination sensitivity (in both conditions) for the sequence of standard preceding target. The results were interpreted as proof for the IRM. The authors admitted, that the IRM was not sufficient to account for positive Type B effects as this effect had not been found in their study. To further investigate these findings, the same experiment was conducted again employing a visual duration discrimination task. The same data patterns were found. This data suggested that the Type B effect was not only observable in classical 2AFC tasks but also in equality judgment and across modalities and therefore should be considered as a “general phenomenon” (Dyjas & Ulrich, 2014, p. 1132).

Sensation Weighting Model. The IRM proposes that only the first of two to be compared stimuli is an integrated weight of previously encountered target stimuli. Hence, it is not sufficient to explain the findings of positive Type B effects or the absence of a negative Type B effect, as reported in recent studies (Hellstöm & Rammsayer, 2004; Bausenhart, Dyjas & Ulrich 2015; Hellström & Rammsayer, 2015; Ellinghaus, Ulrich & Bausenhart,

[1] Designs of studies that share essential aspects with or differ in essential aspects from the studies presented in the empirical part of this thesis are described in detail. In Chapters 2 and 4, the empirical part will be described and discussed in reference to the papers described in Chapter 1.

2018; Hellström et al., 2020). In these studies, the negative Type B effect did not show up in trials with very short interstimulus intervals (ISI, 300 ms). For very short standard stimulus durations (50 ms) a positive Type B effect was reported.

The most promising model to account for these findings is the *sensation weighting model* (SW) by Hellström (1979, 1985, 2000, 2003) that integrates the MH model and the IRM. The SW assumes that in preparation of the comparison of two stimuli a weighted compound was calculated for each of the stimuli, not only for the first. Each of these compounds combined the subjective magnitudes of a stimulus and of its *reference level* (ReL; Hellström et al., 2020, p. 3198).

Whereas in studies focusing on the IRM, e.g. Dyjas et al. (2012), an invariant standard was used while the target varied very diversely, in the studies to support the SW model various, even very short, standard duration were employed (*adaptive staircase method*) and the duration of ISI was varied as well as whether the ISI were filled or not filled (e.g. Hellström & Rammsayer, 2015: Hellström et al., 2020).

Until recently, the adaptive staircase method, introduced by Rammsayer (2012), is the method that allows for the most flexible testing of DL with the 2AFC paradigm. This is especially necessary if one wants to prevent a systematical overestimation of one of the stimuli in a pairwise testing due to its expression. The adaptive staircase method enables the researchers to measure the effect of the standard’s position on discrimination sensitivity, disentangled from sensory adaption or memory effects. In the experiments of Hellström et al. (2020) participants performed an according task on a computer in a “sound-attenuated and dimly lit room”, initiating the task and responding to the trials by pressing one of two designated keys on the keyboard labeled “first interval longer” and “second interval longer” (Hellström et al. 2020, p. 3200). Participants were told to perform as correctly as possible, without receiving feedback on the correctness of their performance in single trials. Each participant had to perform four test blocks. In the first block, in each trial the standard preceded the target, starting with a standard higher than the target in the first trial. In the second block, in each trial the standard preceded the target but starting with a lower standard. In the third block, in each trial the target preceded the standard, starting with a higher target. The fourth block presented targets before standards, starting with a lower target in the first trial. Each block consisted of sixty-four trials. In each block the standard duration was held constant, which was not transparent to the participant, while the target duration varied according to a weighted up down method. After the experiment, the participants were asked whether they had been aware of different presentation orders, of a constant standard and a variable comparison interval, no case reported awareness.

While previous studies testing the IRM failed to find positive Type A effects, in these studies a positive Type B effect for very short durations (50 ms) and short ISI (less than 300 ms) was observed. Both findings indicated that the effect's origin was not in a systematic response behavior but rather in a perceptual or cognitive mechanism. Hellström and colleagues (2020) concluded that a negative Type B effect that occurred in a study with high stimulus variability argued in favor of the SW, as the model would not limit the value of the weights of the first presented stimulus and the second presented stimulus. With regard to this aspect, the SW differed from the IRM, that stated that the ratio of the two stimulus weights to be below a value of 1 (s1/s2 < 1) in any case (Hellström, 1979, Hellström, 2003).

Dyjas et al. (2012) already had discussed whether the observed Type B effect was a mnemonic effect: The memory trace of the first stimulus faded, therefore, an accurate relative judgement was easier in 'standard preceding target'-trials. The weighted difference model argues that the sensory system compensates for this effect by differential weighting of the stimuli depending on their sequential position. This assumption stands in line with the *distinct* *timing hypothesis* (e.g. Rammsayer, 2008). It states that, for the discrimination of very small durations, a sensory mechanism is employed while the discrimination and comparison of rather long durations (defined by Rammsayer, 2008, as longer than 300 ms) required a cognitive mechanism. In combination with the finding that the negative Type B effect has only been observed in discrimination tasks of rather long durations and that it reverses for small durations, one could conclude on a cognitive effect related to the time comparison system that employs cognitive strategies.

A general constraint of the psychophysical measurement of discrimination sensitivity is, that a complete randomization of the position and expression of the standard stimulus cannot be realized as the instruction comparative of a classification task has to be directed. A standard that varies in sequential position and in magnitude would not be distinguishable from the experimental target neither by the experimenter nor by the participant – in such a setting, a DL, as traditionally defined, could not be measured. A completely variable 2AFC paradigm for psychophysical DL measurement has not been realized yet. In the debate on the Type B effect being an effect of systematical response behavior, happening on the encoding stage or being a cognitive effect, most researchers conclude on a cognitive mechanism as the effects origin resulting from the 2AFC tasks. This is especially supported by the findings, that the Type Be effect depended on peculiarities of the experimental design (Hellstöm & Rammsayer, 2004; Bausenhart, Dyjas & Ulrich 2015; Hellström & Rammsayer, 2015; Ellinghaus, Ulrich & Bausenhart, 2018; Hellström et al., 2020).

1.2 Pairwise Comparisons of Symbolic Magnitudes

Relative magnitude, as one of the most elementary information of human cognition, can be extracted from nearly every kind of stimuli that are perceived by human beings. As a behavioral correlate, linguistic, gestural or symbolic magnitude expressions have been used by humans to describe and communicate the intensity of a physical stimulus, the volume or height of objects and subjects, the numerosity of items and, more recently in human civilizational evolution, the relative positions of persons and groups on social dimensions such as power or attractiveness (e.g. 'she has a lot of power'; 'he is highly attractive').

Psychological researchers in the middle of the 20th century began to use the settings of psychometric studies to measure and quantify human experience of abstract concepts in higher level processing. Pairwise comparisons, most specifically the 2AFC paradigm explained in section 1.1.2, have been adopted by researchers to investigate the principles and mechanisms of magnitude comparisons. Also, the quantification of the experience of physical magnitudes was adopted in other fields of psychological measurement. Today, it is common practice to scale ratings on continua with linguistic anchors, e.g. from 'small' to 'large', from 'less' to 'a lot', from 'little' to 'very much', and sometimes figurative with a horizontal line with a movable slider – all expressions of magnitude. Thurstone (1929) was the first to report the attempt to quantify attitudes that are relevant for social psychological behavior on scales. He used Arabic numbers as symbols for the relative specification on a predefined attitude dimension with a minimum and a maximum. In today's psychological research, magnitude serves as an appropriate general dimension to mark off sensations and abstract ratings.

1.2.1 The Measurement of the Subjective Experience of Symbolic Magnitudes

Psychologists basically distinguish between two types of magnitude information – analogue magnitude information, like physical percepts that are processed directly and symbolic (indirect) magnitude information that have to be encoded and plotted into discrete categories on a continuum (Dehaene, 1989). The advantage of using symbolic stimuli in contrast to physical stimuli to investigate the cognitive compound of the discontinuities in relative magnitude ratings (e.g. the Weber fraction, the distance effect etc.) is, that symbolic stimuli, especially numbers, are to a lesser degree ambiguous (Dehaene, 1989). Therefore, the processing and comparison of numerical stimuli is assumed to require less conscious cognitive operating. The ambiguity of a stimulus' magnitude depends on how well it has been learned in advance to the comparative judgement task and/or on whether there are familiar standards to compare the to be judged stimulus with. In other words, the knowledge of the magnitude information of a symbolic stimulus depends on previously acquired knowledge, the frequency of stimulus encounter and preexisting standards for the same category of stimuli. The efficiency of participants' performance in such judgement tasks is affected by the accessibility of the comparison's standard, as well (Mussweiler, 2003). Besides the frequency of previous encounters, the accessibility of the standard is a result of its relative saliency (Tversky & Gati, 1978). These factors – accessibility, saliency, familiarity – influence the outcomes of interest of psychological studies on relative magnitude assessment of symbolic objects. In contrast, psychophysical magnitude assessments are to a greater degree affected by environmental factors than by memory aspects. Exceptions are the TOE and the SPE, actually categorized as memory effects (memorized reference points according to previous trials influence actual judgements) as they are moderated by ISI (see section 1.1.4).

In comparative judgements of symbolic stimuli, discontinuities in comparison performance similar to those in the relative experience of physical stimuli were found. The previously mentioned distance effect and the SCE have been studied extensively for a broad range of symbolic stimuli (see 1.1.3 and 1.2.3), whereas effects of stimulus order on symbolic comparisons had been overlooked for a long time (see section 1.3.3). Only rarely researchers drew the conclusion that the mechanisms underlying the discontinuities of the experience of and reaction on symbolic stimuli could be the same or even identical mechanisms of the effects of psychophysics (e.g. Banks, 1977; Dehaene 1989; Dehaene, 2003; Petzschner, Glasauer & Stephan, 2015). Scientists of different fields scarcely referred to each other;

hence, each discipline of psychology – psychophysics, cognitive experimental psychology, and social psychology – theorized their own domain specific explanations. So far, this lack of interdisciplinary communication might have been due to the different dependent variables of interest: While psychophysicists are mainly interested in DL, experimental psychologists examined RT and error rates and social psychologists studied liking and attraction, to name only a few examples.

During the past 20 years the intersection approach of *social cognition* research began to observe and interpret performance variables such as error rates and RT in studies testing social psychological hypotheses. For example, *processing fluency,* as a dependent variable, can be derived from RT of liking ratings and is positively correlated with positive affect and positive ratings (e.g. Jacoby & Dallas, 1981; Reber, Winkielmann & Schwarz, 1998; Newell & Shanks, 2007). Performance outcomes (namely RT and error rates) in speeded decision-making tasks (tasks in that the participants are instructed to indicate a response as fast as possible) can be used to derive higher level, partly unconscious, cognitions and have the advantage not to be affected by social desirability or conscious control in general and therefore grant insight into preconscious judgements, behavioral tendencies and experiences.

With regard to an interdisciplinary approach of the underlying mechanisms of discontinuities of pairwise comparisons there is still a lack in the reported research.

If symbolic stimuli as well as the information about their magnitude on various dimensions are learned, it would be interesting to find out how this magnitude information is represented in the human mind, where it is stored in memory, and where this information necessary to make a decision resulting in a behavioral response is integrated (Chen, Lu & Holyoak, 2014). The following section reviews pioneering research on symbolic pairwise magnitude comparisons.

1.2.2 Classical Experimental Designs for Relative Judgments of Symbolic Magnitudes

To gain insight in how comparisons of symbolic magnitude are mentally performed, cognitive scientists adopted the experimental paradigms used in psychophysical research (see section 1.1.2). Taking the law of comparative judgment into account, repeated measurement – instructing participants to perform a bunch of trials of an identical task and aggregating the outcomes to mean values – is common practice. Experimentally reducing and controlling the fast and unconscious process of comparative judgement enables cognitive researchers to conclude on its facilitating and inhibiting factors. This is especially relevant for the comparative judgement of symbolic magnitudes, that are processed indirectly involving higher-level cognitions and memory aspects.

Instructions. To measure the basic processes of mental comparisons that are free from conscious systematic behavior of the participants, usually, instructions indicate to perform the task as fast and as accurate as possible (Liesefeld & Janczyk, 2018). To explore the locus of the discontinuities in participants' response behavior researchers may vary the position of the instructions as this allows them to detect the influence of task specific stimulus encoding and the influence of task specific memory of stimuli. Instructions provided before the presentation of stimuli influence the encoding process of the stimuli (Banks & Flora, 1977) in terms of attention allocation, partly consciously, partly unconsciously and relative assessment according to the information provided by the instructions. Whereas instructions placed after the presentation of the stimuli unveils memory effects on task performance (Petrusic, Shaki & Leth-Steensen, 2008). Therefore, comparing experimental conditions that vary in their instructions' position unveils if the hypothesized effect happens on an encoding stage of stimulus processing or at the response selection stage of the comparison of the stimuli.

Dependent variables. Participants' performance of a comparison task in terms of speed and accuracy reflects the ease of the task and the extent to which conscious processing and decision making are involved in task execution. One performance related outcome of key interest of the study of symbolic magnitude comparison tasks is RT, that is measured for trials in which participants gave correct responses. Therefore, the erroneous trials usually are excluded from data analyses and RT patterns are compared for different characteristics of the stimulus input or of the stimulus' presentation format. In some cases, when error rate and RT are not positively correlated, or the experimental manipulation is assumed to affect the RT and error rates differently, error rates are analyzed and interpreted independently from the RT (e.g. Müller & Schwarz 2008; Ben Nathan, Shaki, Salti, & Algom, 2009). Separate analyses of error rates and RT may indicate, depending on the task, different memory processes, systematic response behaviors or speed-accuracy-trade-offs (Liesefeld & Janczyk, 2018).

By analyzing RT patterns researchers aim to conclude how the stimuli are represented in memory and selected by attentional processes. Recently, researchers have interpreted RT in terms of accessibility of the to be compared stimuli and in terms of working memory load during task performance (van Dijck and Fias, 2011).

In contrast to psychophysical studies, in symbolic magnitude comparisons the employed stimuli symbolize learned or mentally constructed concepts that are perceived and represented as discrete, not as continuous. Therefore, not the DL but the difficulty of contrast and assimilation decisions defining category borders is in the focus of interest of this research.

Stimuli. In studies that employ symbolic magnitudes different from numbers or numerosity (e.g. dot clouds), like the ferocity of animals (Banks & Flora, 1976), temperature (Petrusic et al., 2011) etc., the relation of the to be compared items, objects or subjects is part of semantic memory (Banks & Flora, 1976) or has to be learned episodically in advance of the experimental comparison task (e.g. Jou, Escamilla, Torres, Ortiz & Salasar, 2018).

An example for an experimental training phase for just learned object lists is the *linear syllologic reasoning task* (Leth-Stensen & Marley 2000). Performing this task, the participants had to learn order relations of the test stimuli, presented pairwise, similar to the following test phase. A training phase was installed between encoding and test phase. It comprised repeated comparative judgements identical to the task of the test phase. During training, participants got feedback on their responses after each trial, whereas in the test phase no feedback was given.

Classification versus Selection. The instructions of a 2AFC magnitude comparison task can indicate a classification (for example ‘Is the target stimulus larger (smaller) than the standard’ or ‘Is the target stimulus same as (different from) the standard?’), also called *directed* comparison task, or a selection task (for example ‘choose the larger (smaller) item’), also called *undirected* comparison task. A special case of an undirected classification task of number comparisons is a parity judgement task (‘Is the number odd or even?’, e.g. Dehaene et al., 1993; van Dijk & Fias, 2011) that is especially appropriate to investigate whether task irrelevant magnitude information affects response patterns (see section 1.3.2).

Which of these types of tasks is used depends on the research question respectively on the mechanisms assumed to drive the hypothesized effect. Usually, in classification tasks the differentiation between standard and target is made in the instruction. Hence, it is transparent to the experimenter and to the participant which one of the two stimuli is the (mostly throughout the experiment) stable standard and which is the to be judged target stimulus. Just like in psychophysical studies, the test of the robustness of found effects and to distinguish them from situational and individual error variance, implies to replicate the findings with roving standards (e.g. Ben Nathan et al., 2009) and different spatial and temporal arrangements of target and standard and with variation of the ISI (Hellström et al. 2020, Kaan, 2005; Müller & Schwarz, 2008). Mostly the two items in a paired comparison experiment with symbolic stimuli are presented simultaneously in a horizontal line except for the studies that investigate moderating effects of spatial arrangement (e.g. the SNARC effect, see section 1.3.2), spatial cues (attSNARC effect, see 1.3.2) or latencies of stimulus presentation (e.g. Schwarz & Stein, 1998; Müller & Schwarz, 2008) to draw conclusions on underlying mechanisms of the discontinuities of comparison performance.

An effect of the employed type of task reveals whether found (not found) discontinuities in responses are biases of the comparison process or rather effects of absolute stimulus processing. To exemplify, the distance effect appears in classification and selection task, whereas the SCE and the *size effect* (performance advantage for larger stimuli, Moyer & Landauer, 1967; Henik & Tzelgov, 1982) only was reported in selection tasks (Dehaene, 1989). This indicates that the distance effect (compare Cattel, 1902) is produced by the relative encoding of the standard and the target and that the SCE and the size effect are rather related to the processing manner of absolute magnitudes.

1.2.3 Effects and Models of Comparative Judgements of Symbolic Magnitudes

Five discontinuities in the performance of pairwise symbolic magnitude comparisons have been frequently reported and hypothesized to be provoked by contextual factors (like range of tested stimuli, their presentation and the comparative instructions) in interaction with human attention- and memory functions: The SCE, the distance effect, the *serial position effect*, the size effect and the *end effect* (Ebbinghaus, 1913; Krajcsi & Kojouharova, 2017).

As mentioned in section 1.1.3, the SCE and the distance effect had been evident in psychophysical studies. However, the vast majority of literature reporting and discussing these effects focused on comparative judgements of symbolic magnitudes (Leth-Steensen & Marley, 2000). These studies aimed to clarify how people represent and process relational magnitude information, as well as how it is stored and retrieved from memory (Leth-Steensen & Marley, 2000, p. 63). Only a few authors aimed to clarify the extent of the equality between the comparison processes of symbolic magnitudes and those of comparisons of physical magnitudes. Whereas, the question posed by both research lines is concerns the origins of the discontinuities of pairwise comparisons. Both lines of research aim to unveil whether the comparison takes place at the encoding (perceptual) or response (cognitive) level and if the underlying mechanisms of the above mentioned effects found in psychophysics, symbolic magnitudes, non-symbolic magnitudes (e.g. numerosity of dots) and numbers are the same or if one needs to define special models for each of these domains.

Distance, Semantic Congruity and Serial Position. The publication of Banks, Fuji and Kyra-Stuart in 1976 was one of the first to report a distance effect in the comparisons of numerical digits. When shown a pair of digits and asked to select the larger of the two, participants made their choice more quickly as the numerical difference between the digits increased. In 1977, Banks and Flora replicated these findings for pictures (drawings of items) and words (names of items). In their experiments, the discrimination speed was higher for stimulus pairs that differed more from one another on the dimension in question (e.g. door – airplane) than relatively closer pairs (e.g. butterfly – ant). Additionally, they reported an SCE: The item of a pair that corresponded to the comparative presented in the instruction was identified quicker and more likely accurate. Banks and Flora concluded that both, symbolic and figurative magnitude information, were “ […] processed in terms of linguistic codes rather than mental images” (p. 1). They referred to this finding as an explanation for both found effects “[…] The stimuli and the instructions are represented as discrete codes, and […] processing proceeds until one and only one of the stimulus codes is the same as the code for the instructions." (p. 1). The authors hypothesized that the RT advantage for items that are semantically "closer" to the concept of the comparative instruction was caused by the overlap of close concepts in semantic memory. They named their approach the *Semantic Coding Theory*. The theory states three stages of the comparison process of symbolic magnitudes – encoding, choice and response. First, an analogue coding would be performed: The instruction comparative on the one hand and the magnitude information of the to be classified items on the other hand would be transformed into binary codes. The subsequent choice level would be divided into two substages. First, there would be a discrimination decision – are the two to be compared items different from each other? Then, the stimuli would be matched with the comparative dimension of the instruction and the stimulus identical to the instruction code would be identified. According to this theory, the distance effect would occur on the discrimination substage, as more distant stimuli were easier to discriminate. The SCE, however, would happen on the matching stage – the stimulus that matched the comparative instruction code best was identified fastest. And also, stimulus pairs that were semantically closer to the code of the instruction comparative, the starting point of the matching process, were processed faster.

Banks and Flora (1977) found shorter RT and higher accuracy for the selection task of pictures compared to the selection performance of words (pictures and words symbolizing the same objects). This RT difference appeared in addition to the distance effect and the SCE in both stimulus classes. In summary, one could conclude that the same processes of magnitude comparisons operated for different kind of stimuli. However, the more direct the magnitude information could be extracted from the stimuli, or to be more precise, the closer the stimuli were to the actual sensory input, the easier the comparison could be performed. Following the interpretation of the authors, even though the pictures in Banks and Flora's experiments did not convey direct sensory input, they could be processed more directly than words because words would have a higher degree of symbolization. The extraction and decoding of magnitude information from linguistic concepts would take more time and was more prone to errors.

Shoben, Cech, Schwanenfluegel and Sailor (1989) concluded that a serial position effect in symbolic magnitude comparisons could be explained by the semantic code model as well and in some cases, it would even be difficult to distinguish from SCE and distance effects. Effects of serial position had most frequently been observed in memory tasks with symbolic stimuli when participants had to learn and recall single items from lists. In these tasks, a *primacy effect*, an increased memory performance for items that stand in initial positions, as well as a *recency effect*, an increased memory performance for items on end positions have been reported (e.g. Deese & Kaufmann, 1957). According to this, in paired comparison tasks *bowed* serial position effects have been observed (Shoben et al. 1989a). This label describes the finding that the discrimination performance is increased for pairs of extreme magnitude compared to pairs of intermediate magnitude. Some authors referred to this finding as the end effect (e.g. Leth-Steensen & Marley, 2000; see below). In their experiments, Shoben et al. (1989a) let participants perform paired selection tasks. In a repeated measurement design, participants were exposed to words describing objects in paired presentation. Right before each stimuli pair, either the word ‘larger’ or ‘smaller’ was displayed, indicating the decision criterium. Besides the bowed serial position effect, the authors found an SCE and a distance effect. The authors argued that the semantic coding model accounted for all three effects. According to the model, RT reflected the time required to code and match the stimuli’s magnitude information and the magnitude dimension of the instructions. In the logic of semantic coding, the bowed serial position effect would represent the fact that extreme magnitudes were easier to discriminate than intermediate magnitudes, because they were semantically closer to the semantic code of the decision criterium.

Therefore, stimuli of intermediate magnitude required more time to be encoded and matched correctly.

Shoben et al. (1989a) lined out that the bowed serial position effect unlike the distance effect and congruity effects only showed up in certain tasks. For example, it had not been reported for numerical stimuli until then. Shoben et al. (1989a) hypothesized that the “degree of arbitrariness” (Shoben et al., 1989, p. 273) of the stimuli determined the likelihood of the effect’s occurrence. They further provided an explanation about the locus of the effect in 2AFC tasks with symbolic stimuli. While judging the size of real-world objects would require long term memory, the judgement of items from an order learned during the experiment (or a learned limited range of real-world objects) usually would require short term memory. The comparison of numbers and other “overlearned” orders (e.g. Nuerk & Schroeder, 2017) would rather require working memory capacities. This supported the notion of Shoben et al. (1989a) that bowed serial position effects were context (in terms of stimulus range) dependent. Other researchers employed the same explanation for the less frequently reported end effect (e.g. Leeth- Steensen & Marley, 2000). The so-called end effect describes the finding that discrimination performance is increased for pairs that involve single items from one of the ends of a prelearned list (Leeth- Steensen & Marley, 2000). It could also be explained by the semantic coding approach as it is quite similar and sometimes hard to distinguish from the bowed serial position effect.

As semantic coding could explain many discontinuities in pairwise forced choice tasks of symbolic stimuli, it appeared to be very promising but the account had the serious limitation on pairwise comparisons of symbolic stimuli, whereas the SCE and the distance had also been reported for physical stimuli. Magnitude information of physical stimuli rather is processed directly and does not have to be transferred into linguistic codes to be compared.

Therefore, the semantic coding theory had to be rejected as a broad explanatory approach for RT discontinuities in magnitude comparisons.

As a concurring explanation for the SCE, Banks and Flora (1977) also tested the approach of the *expectancy hypothesis* (Marschark & Paivio, 1979; Shoben, Sailor & Wang, 1989). From an expectancy point of view, the instruction comparative, given before the to be compared stimuli, serves as kind of a prime to what has to be detected in the following stimuli pair (Marschark & Paivio, 1979). According to this, Banks and Flora investigated whether the position of the instruction comparative affected the congruity effect. If the congruency effect showed up exclusively in an experimental setting where the comparative instructions were presented before the test stimuli, the role of an expectancy created by the instructions would have to be stressed; if the congruity effect occurred no matter the instruction was presented before or after the test stimuli, the semantic coding theory would be supported. The finding that the SCE occurred in both conditions led Banks and Flora to the conclusion that the expectancy approach had to be rejected. Petrusic et al. (2008) argued that an expectancy leading to a faster detection and classification of the semantic code could also be built up when the test stimuli are presented before the instruction. According to this view, expectancy formed by the relative size of the encoded stimuli being matched with the subsequentially presented instruction comparative could influence RT in terms of semantic coding, too. In both presentation orders corresponding response codes of instructions and stimuli were facilitated.

‘Holistic’ versus ‘symbolic’ Models. In 1990, Dehaene, Dupoux and Mehler clustered the models that attempted to account for the above-mentioned effects into `symbolic models` and opposed them to ‘holistic models’. According to their definition, symbolic models referred to linguistic priming as the key mechanism causing SCE and distance effects (Holyoak, 1978), while holistic models stressed the relevance of mental random walks between encoded stimuli and decision criteria, and the relevance of reference points, that had already been mentioned as promising aspects of psychophysical models of comparative judgements (see section 2.1.4).

The first, very influential, holistic approach for discontinuities in symbolic pairwise comparisons also stressed the role of the instructions of comparative judgement tasks – the *reference point account*. Its most prominent and most cited advocate is the cognitive psychologist Keith Holyoak, who replicated the finding of the distance effect in numerical paired comparisons in 1978 across instruction positions. Relative magnitude judgments were made faster when stimuli were numerically closer to a reference point introduced in the instructions. This reference point could be either implicit, when the word ‘large’ (‘small’) set the reference point on the end of the rating continuum with the larger (smaller) expressions, or explicit, when the instructions introduced a certain number as reference point. He concluded that internal reference points on a continuum of numerical magnitude served as internal standards to which the actually perceived target stimuli were compared, to be able to judge the targets’ sizes – just like psychophysical researchers had previously theorized for psychophysical magnitude judgement tasks (see section 1.1.3). Holyoak hypothesized the internal reference point and the underlying mechanism to be located in working memory (Holyoak, 1978; Chen et al.,2014). His approach broadly corresponds with the assumption of Jamieson & Petrusic (1975) for distance effects in psychophysical comparisons (see section 1.1.3) and with assumptions of theorists of numerical cognition (e.g. Dehaene, 1989; see section 1.3). For numerical comparisons, Dehaene (1989) concluded that the compression of the slopes of the RT of one-digit number comparisons could be explained by two reference points at the ends of the magnitude continuum. Holyoak and Dehaene referred to the random walk model by Buckley and Gillman from 1974, who already had identified the implicit use of reference points in paired comparisons of digits and dot patterns to be the underlying mechanism. Buckley and Gillman (1974) showed that the compression of the RT slope of digit comparisons equaled the RT slope of comparisons of dot patterns and theorized that this supported the hypothesis of analogue coding of both types of quantity information (find accounts for analogue coding of numbers in detail in section 1.3). The model of Buckley and Gillman assumed that the internal representation of a number was a random variable fitting the Weber-Fechner-Law, namely the logarithmic compression of the actual magnitude of the number. In a random walk from one internal representation to the other, the magnitude of the difference between the two was computed and evaluated as a “subjective ratio” (Jamieson & Petrusic, 1975, cited by Holyoak, 1978, p. 236).

Serial Position Based Distinctiveness. Holistic as well as symbolic models came to the conclusion that the SCE was difficult to distinguish from the distance effect, the end effect and the bowed serial position effect. The interpretation of the compression of the RT curve’s slope to the ends of the stimuli range appeared to depend on the class of tested stimuli – symbolic, physical or numerical until it became evident that the four mentioned effects represented behavioral manifestations of one single underlying effect. After three decades of theorizing and testing theories against one another, recently some reviews have been published attempting to get to the bottom of the underlying effect (e.g. Jou, 2010; Jou, Escamilla, Torres, Ortiz & Salasar, 2018; Jou, Matos, Martines, Sierre, Guzman & Hut, 2020).

The most recent review of the semantic coding model, the expectancy theory and the reference point account, was published by Jou et al. in 2020. Within their publication they discussed and defended the latest theory accounting for the effects mentioned above: The *serial position based distinctiveness account* (SPBD) stated by Jou et al. in 2018. Jou et al. (2018) had replicated the SCE in comparative judgments of previously learned symbolic stimuli. In their experiments, participants had to recall and compare the heights of persons learned as ordered information in advance to a speeded comparison test phase. According to Jou et al. (2018), in episodic memory tasks, the SCE actually was a serial position effect within a pair of items. The authors based their assumption on the observations that the effects of symbolic magnitude comparisons were rather effects of learned orders than domain specific codings, and that participants' discrimination performance in a pairwise comparison depended on the *serial distinctiveness* of each stimulus. Jou et al. (2018) had discarded the assumption of the causal role of the anchors (or reference points), set from trial to trial by instruction comparatives, and theorized that these eventual primes were just a "coincidental factor" (Jou et al., 2020, p. 226). They were able to show that the SCE also occurred when no task instruction had been given. To simplify, Jou et al. theorized that the serial distinctiveness of an item increased the closer its magnitude information came to the end of an introduced, overlearned continuum crucial to the measurement. According to the authors' approach, the speed and accuracy of the response to a target increased with its serial distinctiveness.

Computer Models for Comparisons of Symbolic Magnitude. As another attempt to explore the processes that underlie the mentioned discontinuities of symbolic comparisons, cognitive researchers at the beginning of the 21st century started to program computer models to simulate mental comparisons (Leth-Steensen & Marley, 2000; Page, Izquierdo, Saal, Codnia & El Hasi, 2004; Chen et al., 2014). The most successful models to account for all of the above mentioned effects in pairwise comparisons integrate contextual- , mnemonical and attentional factors.

Leth-Steensen and Marley (2000) and Page et al. (2004) integrated the common robust effects in the reaction patterns of paired comparisons in 'recursive models' that combined learning-, representational-, comparative-, and decisional processes. According to their models, all these processes contributed to the performance of comparative judgement tasks and were ran over and over again until the outcome of a mental 'searching' process fitted an implicit or explicit criterium or reference point. This assumption corresponds to the previously mentioned random walk model by Buckley and Gillman (1974). The RT reflected how often this ‚cycles' have to be run. The comparison of RT of different experimental conditions is the most diagnostic behavioral marker for the differentiation of the processing stages of comparisons.

To explain how RT reflect the underlying processes of pairwise magnitude comparisons, Link and Heath referred to the *Relative Judgment Theory* in 1975. This theory assumes that every comparison contained a subjective referent (or standard) which was based upon a person's experience. They claim that this subjective referent was comparable to the adaption level introduced by the model of Michel and Helson, mentioned in section 2.1. Link and Heath claimed that test stimuli, in a paired comparison, served as 'probes' that were tested against the referent established in advance. The duration of this comparison depended on the given task. In tasks in which the referent varied from trial to trial, naturally, there was no improvement of RT over the trials. In general, according to their theory, the RT heavily depended on the frequency of previous exposures to the referent stimulus, especially encountered during the task. This crucially decreased the RT in a certain trial. Therefore, the best practice to measure comparison processes should be a variation of both stimuli of repeated paired comparisons from trial to trial. Only in this way could one be sure that the trial wise installed standard stimulus would serve as the actual reference point in the respective trial.

In their recursive model, Page et al. (2004) theorized two sequential steps to be run repeatedly until the decision criterion was met – the *exponential accumulation rate* and the *accumulation stopping conditions*. "These features are directly associated with the distance effect and the congruity effect. The end effect results neither from the dynamical behavior nor from the stopping conditions, but it will be accounted for by a plausible selection of the encoded stimuli.” (p. 197). They named the process consisting of these two aspects the *recurrent accrual process* and proposed that the internal numerical representation in a number comparison task depended on the way in which these stimuli were learned. This corresponded to the SPBD account that stressed the role of learned orders.

The computer model defined by Chen et al. in 2014 highlighted the role of memory functions in paired comparisons in the tradition of the reference point model stated by Holyoak (1978). It assigns a crucial role to working memory operations instead of to semantic-memory representations of previously learned magnitude information or other features of the range of stimuli that is employed in a certain task. The authors claimed that the magnitude distributions needed by the participants to perform the comparison tasks were formed in working memory. These could be influenced by contextual factors, like the range of the stimuli and the polarity of the comparative instructions (Chen et al., 2014, p. 27). Chen et al. (2014) argued that, just like for the comparison of physical stimuli, for a quasi-perceptual dimension as size the pre-storage in long term memory of the magnitude information was not decisive for the task performance. As physical magnitudes cannot be stored in long term memory precisely, but repeated pairwise comparisons produces distance, end and congruity effects in performance outcomes. They formulated the *BARTlet* model (a further development of the *Bayesian Analogy with Relational Transformations Model*) that simulates how magnitudes can be composed in working memory based on previous learning. BARTlet has been the only model so far that accounts for all the previously mentioned discontinuities of symbolic magnitude comparisons. It demonstrates that due to the limited capacities of working memory some ratings are attributable to the selective attentional focus in a task. In the case of an SCE and a distance effect attention was guided by the reference points given with the word used to introduce the comparative instructions (e.g. ‘indicate which one is

LARGER') and therefore, the speed of the reaction on a stimulus close to that referent was enhanced. The model could also account for framing effects in social psychological *framing* experiments (Tversky & Kahneman, 1981, cited in Chen et al. 2014), where response patterns depending on the markedness of scalar adjectives (like 'tall' or 'large') were observed.

1.3 Special Case: Numbers

Numerical cognition is a special case of symbolic magnitude processing. Different from research that investigated comparative ratings on symbolic magnitude scales, like attractiveness, intelligence, liking etc., digits as symbols for numerosity do not depend on subjective rating and, different from physical input, are unambiguous as rating errors cannot be attributed to peripheral factors, like noise or masking (Dehaene, 1989). Therefore, numbers, as stimuli in psychological experiments on comparative processing, combine the advantages of symbolic and physical magnitudes – unnoisy perception and direct processing. On that ground, the study of numerical magnitude comparisons is promising to unveil basic comparison mechanisms shared by both fields and maybe even by every kind of comparisons (Moyer & Dumais, 1987).

1.3.1 The Mental Number Line

Moyer and Landauer in 1967 were the first who found a distance effect and a size effect in pairwise comparisons of single-digit numbers. Since these effects had only been known from discrimination tasks in psychophysical studies, the authors concluded, that one-digit numbers were rather processed like analogous physical stimuli. Plotting the discrimination accuracy of the tested participants, Moyer and Landauer (1967) found a curve that fitted the Weber fraction – the discrimination accuracy could be modelled by the logarithmic function of the numerical value of the to be judged stimuli. This was the first step into a research domain Stanislas Dehaene (1989) named „the psychophysics of numbers".

Stanislas Dehaene is one of the most influential and prominent researchers in the field of numerical cognition. In his early studies in the 1980s, he investigated the origins of the distance effect in symbolic and numerical comparisons observing reaction time patterns in paired comparison tasks. He found that participants reacted to one-digit numbers very fast and accurately, in contrast to more digit numbers. He concluded that one-digit numbers must be quite un-noisy and unambiguous to human perception - like a direct physical percept (Dehaene, Dupoux & Mehler, 1990). An influential publication of Dehaene et al. (1990) dealt with the role of decades versus ones in number comparisons and the question whether two-digit numbers were processed holistically or rather stepwise (encoding and retrieving digit per digit) to judge their relative magnitude. This was tested comparing the results of three experiments, with different ranges of test stimuli. In every experiment, the standard had been held constant throughout the trials in the numerical center of the range of targets. Dehaene et al. (1990) instructed their participants to judge the relative magnitude of numerical targets compared to a fixed numerical standard, that appeared only once in the beginning of the experiment. A trial consisted of a target number appearing in the middle of a computer screen. Participants were instructed to indicate as fast and as correct as possible whether the target numbers were numerically larger or smaller than the standard (a more detailed description of the experimental setting is lined out in section 1.3.2, 'the SNARC effect'). In their experiments the authors found a distance effect in the RT of participants' decisions and that it was mediated by minimal changes in the magnitude of the standard stimulus (e.g. standard '65' in Experiment 1, standard '55' in Experiment 2) although the standard in every experiment was the numerically center of the range of target numbers. For example, participants were faster reacting on the target 99 than on 11 when the standard was 55. The authors concluded that this would indicate a compression of the continuum of numerical magnitudes to the end of the larger expressions, here higher numbers. And, according to this, the internal representation of numerical magnitude would obey Fechner's law instead of being linear. Theorists of numerical cognition like Stanislas Dehaene (1989) and William P. Banks (1977) already had theorized, shortly before Dehaene and colleagues' experiments in 1990, that numerical symbols stand for categories on a magnitude continuum. According to these publications, Dehaene et al. (1990) concluded that the distance effect was an effect of the digit's representation, not an effect of its magnitude. The authors suggested, comparable to what had been stated by the semantic coding approach for symbolic stimuli in general, that during a number comparison "first, the digital code of numbers had to be converted into an internal magnitude code on an analogical number line" (Dehaene et al., 1990; p. 638). Afterwards, the analogical comparison could be performed, without access to the digital appearance of the numbers. "Finally, in the last stage of the comparison, the analogical comparison algorithm triggers a response buffer to make one of two discrete responses, namely, larger or smaller" (Dehaene et al., 1990; p. 638). Later other researchers replicated these findings (e.g. Gallistel & Gelman, 2000).

In 1992, Dehaene introduced the metaphor of the *mental number line* (MNL), illustrating that one-digit numbers were mentally represented on an analogue number line with smaller numbers on the left and larger numbers on the right (Dehaene, 1992). This notion has been held by a broad scientific community until today (for a review see Leibovich et al., 2017).

Several studies found the distance effect and the size effect in numerical comparisons in animals (e.g. Schwarz & Stein, 1998) stressing the evolutionary foundation of number processing. Gallistel and Gelman (1992) underlined the role of the evolutionarily important capacity of preverbal computation and its figural correlate of finger counting as evidence for an explicit categorical verbal (in terms of verbal rehearsal) computation system and as a support for Dehaene's MNL metaphor.

Based on the idea of the MNL, Dehaene developed the *triple-code model of numerical cognition* (Dehaene, 1992). This model postulates that during the processing of magnitude information there were three different representation modi of numerical size present: a verbal representation-, a symbolic- (Arabic numeral) and a magnitude- code. The codes were, to a different extent, needed to perform various mathematical operations and successively develop during childhood. Dehaene assumed that humans had an innate sense for small quantities. Therefore, during the individual development of visuospatial working memory in early childhood, attentional control, spatial orientation and also the idea of volume and quantity in spatial relations would build up simultaneously.

Recently, the close link between the mental representation of space and numbers has been scientifically proven (e.g. van Dijk & Fias, 2011). This highlights the role of visuospatial imagery during mathematical reasoning and corresponds to the reports of many people on the imagination of a mental number line, comparable to an abacus, during mathematical tasks of adding and subtracting numbers (Dehaene, 1997).

After three decades of research on the MNL, the concept recently has been criticized. Evidence from neuroimaging studies suggests that overlapping brain areas are activated during symbolic and non-symbolic number processing (for a review see Krajsci, Lengyel & Kojohouharova, 2018). In the studies of Krajsci et al. (2018) dot clouds were used as non-symbolic stimuli. It could be shown that even though the sensitivity in judgement tasks of symbolic and non-symbolic number comparisons could differ largely, both classes of stimuli would be discriminated following the Weber fraction. Hence, the researchers concluded that, in both cases, the magnitude discrimination performance was based on the ratio of standard and target, not on their absolute magnitude encoding. According to this, Krajsci, et al. (2018) theorized that humans rather had a ‘sense for magnitudes’ than a ‘sense for numbers’.

Due to the contradicting findings of numerical magnitude processing more research is needed to support, neglect or extend the MNL-concept. Apart from this, results from other magnitude comparisons could provide further conclusions on the representation of symbolic magnitudes in general.

1.3.2 Grounded Cognition

Another line of research in numerical pairwise comparisons identified discontinuities, first observed by chance than systematically studied, that resulted from an association of the systematic manipulation of spatiality and temporality of the to be compared stimuli and the spatial position of the response devices. Prominent and well-studied effects of this domain are the *spatial-numerical association of response codes* (SNARC; Dehaene et al., 1993) effect and the *Spatial Temporal Association of Response Codes* (STEARC; Ishihara, Keller, Rossetti & Prinz, 2008). These effects have been seen as examples for the *grounded cognition* approach that came up in the end of the 1990s. The term, interchangeably used with *embodied cognition*, was especially coined by the cognitive psychologist Lawrence Barsalou, who claimed in 1999 that human higher cognition was rooted in *perceptual symbol systems* (Barsalou, 1999). According to his theory, information about objects was stored and retrieved as it was perceived. Also abstract concepts were processed in a modal analogical fashion, just how they were encountered in the first place, namely perceptually. The assumed reason for this fashion of information processing is that humans processed external information in relation to their body. According to Barsalou, body related information, in an evolutionary sense, was the most relevant aspect of object related information. He promoted the assumption that the perceptual system would not only be an auxiliary system that transmitted information to higher order processing systems but, just like the proprioceptive and the introspective information systems, would be an equally relevant part of human abstract cognition. He stated that abstract concepts would not only exist in the linguistic system, but also in bodily structures. These were activated on a low level every time, concepts that had been bodily experienced once were encountered or activated again.

The assumptions of the mental number line and the analogical coding of symbols to compare their magnitude paved the way for the approach of grounded cognition as they already connected higher level cognition about abstract concepts with basic dimensions of human perception.

The SNARC Effect. Dehaene et al. (1990), in their paper on the distance effect in number comparisons reported a *response side effect* as a side finding (later referred to as the SNARC effect). Half of the participants was instructed to indicate their response by pressing a right-side key with their right hand for ‘target lager than standard’ and a left-side key with their left hand for ‘target smaller than standard’. The other half of participants was instructed to react on a smaller target with the right hand on the right-side key and on a larger target with the left hand on the left side key. Instructions said to respond as fast and as accurate as possible. The numerical standard of each experiment was presented only once at the beginning of the test session. The experimental trials consisted of a target stimulus (a two-digit number) appearing in the middle of a computer screen for two seconds. The participants had to indicate their response within an interval of four seconds. If no response had been given within this interval, the reaction was counted as an erroneous trial. Two seconds after the target from the previous trial had disappeared, the next target appeared on the screen. They found that participants’ reactions on targets larger than the standard were faster when they were instructed to indicate larger targets with the right hand and when smaller targets had to be responded to with the left hand. Dehaene et al. (1990) also found this response side effect within a French sample of Arabic native speakers and concluded that the effect was not caused by language specific reading direction. Later, this intercultural robustness of the response side effect had to be rejected as it could not be replicated for subjects from cultures of non-western reading direction (Dehaene et al., 1993; Gevers & Lammertyn, 2005).

Building on the finding of the response side effect, Dehaene et al. (1993) published a set of nine experiments, devoted to the clarification of the embodied representation of the MNL. They referred to the effect as the SNARC effect for the first time. In contrast to the experimental design of Dehaene et al. (1990), the authors let the participants judge the parity of target numbers. For this type of judgement task, the magnitude and therefore analogue coding as it had been ascribed to numerical comparisons by Dehaene, is irrelevant. Between the experiments they used different presentation modes of numbers – some tasks employed one- and two-digit Arabic numerals, others employed French number words. In each experiment, half of the participants was instructed to react with the right hand on a key on their right (left) side to an even number (number word) and with their left hand on a key on their left (right) side to an odd number (number word), whereas the other half of participants was instructed vice versa with crossed hands. Throughout their experiments the authors found robust SNARC effects: They observed that when participants' hands were crossed (the right side key had to be pressed by the left hand and the left side key had to be pressed with the right had) the response code 'right side for larger numbers' and 'left side for smaller numbers' could be replicated – participants made faster, correct responses with their left hand on the right side key responding to larger numbers and with their right hand on the left side key responding to a smaller target number. Dehaene et al. (1993) concluded that magnitude information was automatically activated in numerical judgement tasks even if the magnitude information was irrelevant for the task and that this was an embodied effect.

The parity judgment task has become the common experimental set up to study the SNARC effect, because its instructions do not prime the mental number line whereas instructions with some kind of comparative magnitude instruction do (compare the common instructions to test the SCE, see section 1.2.3). Later, the SNARC effect could also be found in classification and selection tasks pairwise comparison set ups mentioned in section 1.2.2 (Shaki & Petrusic, 2005; Ben Nathan, Shaki, Salto & Algom, 2009; Shaki, Petrusic & Leth - Steensen, 2012).

In 2003, Fischer, Castel, Dodd, and Pratt found that, according to SNARC, spatial shifts of attention could be induced by mere presentation of numbers. In their experiments, participants were significantly faster detecting targets on the right (left) side of the screen when in advance a relatively large (small) number, that had been introduced as irrelevant for the task, had been displayed. This effect was called *attentional SNARC effect* (attSNARC). Also, several studies reported a vertical SNARC effect (e.g. Petrusic, Lucas & Leth-Steensen, 2011; Ito & Hatta, 2004): Numerically smaller (larger) numbers are associated with bottom (top) positions. This challenged the assumption of the underlying MNL as causal explanation for the SNARC and brought up new research questions.

Synesthesia. As an extension of the SNARC, synesthesia effects were explored in the second decade of the 21st century: the interaction of temporal, spatial and numerical (or generally magnitudinal) information on human decision making and behavior (for a review see Winter, Marghetis & Matlock, 2015). Herewith the approach of spatially grounded numerical cognition (e.g. SNARC) and stroop like experiments that had revealed the size congruency effect (the detection performance advantage for numerals that match in font size and numerical size; e.g. Henik & Tzelgov, 1982; Winter et al., 2015) was addressed and expanded.

Kaan (2005) and Müller & Schwarz (2008) investigated the question if there was a *temporal number line* besides the well-studied spatial mental number line. In their studies they presented two numbers with a small stimulus onset asynchrony (SOA) and instructed the participants to indicate, after both numbers had disappeared, which one had been larger (smaller). Both studies found, besides a SNARC effect (in terms of a response side effect) and an SCE, a performance advantage for the larger (smaller) of two stimuli appearing as the second (first) stimulus of a trial. Comparison performance was measured in terms of RT and error rates. Kaan’s and Müller & Schwarz’ findings corresponded to the results of earlier studies (e.g. Sekuler, Tyann & Levinson, 1973), in which participants performed stimulus onset judgements and found significantly shorter RT and lower error rates when relatively large (small) stimuli were presented at the second (first) position of temporally asynchronous presented stimulus pairs. The authors of these studies concluded that there was a processing advantage for temporally ascending digit orders evidencing a mental temporal number line.

The finding of the STEARC (Spatial Temporal Association of Response Codes) by Ishihara et al. (2008) – left side responses were facilitated for stimuli of relatively early onset and right-side responses were facilitated for late onset stimuli (not observed for vertical alignment of responses) – supported the suggestion of a general *Spatial Quantity Association of Response Codes* (SQUARC) hypothesized by Walsh in 2003. Walsh set up the *A Theory of Magnitude (*ATOM) approach to account for the frequently reported conceptual association of temporal, spatial and numerical magnitude information. He theorized that the same neural circuits were active when a person processed temporal, spatial and numerical information and concluded that there was one domain general magnitude system. According to Walsh, this explained why humans across cultures used magnitude related linguistic expressions to talk about the relativity of time as well as about volume or height. In their review Winter, et al. (2015) cited studies on macaques that evidenced, as interpreted by the authors, that the mapping of time and space in humans was biologically rooted (Merritt et al. 2012, as cited in Winter et al., 2015). From an evolutionary perspective, in early human development one general system for magnitude information could have meant an advantage for rudimental survival related behaviors like grasping, squeezing and punching. Lesion studies with animals, as well as clinical studies with subjects suffering from various neurological disorders, reported domain general impairments of magnitude assessments. The bilateral intra-parietal sulcus (IPS), the temporal-parietal sulcus (TPS) and the prefrontal cortex were simultaneously involved in tasks where magnitude, time and space judgement had to be performed (Buetei & Walsh, 2009, as cited in Winter et al., 2015). Newborn humans also seem to be able to associate spatial extent, temporal duration, and approximate numerical magnitude (Winter et al., 2015, p. 210).

There is another account of grounded cognition that locates the connection of time, space and number on higher cognitive levels. The *Conceptual Metaphor Theory* (CMT) hypothesizes that we cognitively map abstract concepts in the physically concrete dimension of space to *figure out* or *make a picture* about relative information on other continua in relation to each other and to the own body (e.g. Lakoff, 1993; Lakoff, 2008). This approach has been supported by the behavioral studies of grounded cognition theorists that frequently referred to linguistic metaphors: In many languages magnitude expressions were used to talk about temporal and spatial expansion (e.g. Winter et al 2015). Winter et al. (2015) reviewed and discussed ATOM and the CMT and advocated to integrate both accounts. They concluded, that ATOM „focuses on interactions between low level magnitudes [while] the CMT focusses on higher level reasoning and language understanding“ (p. 210).

The representational overlap of time, space and number develops in humans associated with the acquisition of language, that includes the learning of linguistic expressions talking about time in an early stage of child development. Which metaphors are used to talk about time is culture dependent (e.g. while English speakers talk about time in terms of length, Greek speakers use amount metaphors (e.g. de Hevia & Spelke, 2010, cited in Winter et al., 2015). The metaphor of the mental number line also implied a movement along a path when performing arithmetic operations and therefore was compatible with ATOM as well as with CMT (Winter et al., 2015). According to Winter et al. (2015), the metaphorical acquisition of the concept of time served as the foundation for the development of ATOM, a more general magnitude system. The authors argued that it needed both systems, the phylogenetic explanation stated by ATOM and the ontogenetical explanation of CMT, to account for all lower-level and higher-level cognitions about time, space and magnitude and that both approaches were rooted in biological determinations.

An integration of ATOM and CMT would presume that „[the] evolutionarily older magnitude system in parietal cortex posited by ATOM might be subject to neural reuse or recycling as a result of culture and experience (Anderson, 2010; Dehaene & Cohen, 2007), shaped throughout ontogeny by cultural artifacts and practices including language and writing to produce more directional, asymmetric mappings.“ (Winter et al., 2015, p. 219).

Polarity correspondence. Recently a general embodied mapping of binary poles of the concrete dimension of the response device (e.g. left and right) and the abstract dimension of the stimuli (e.g. large and small) in experimental psychological studies has been discussed as *polarity correspondence* (Proctor & Cho, 2006). Proctor and Cho in 2006 introduced the polarity correspondence approach to account for the observed flexibility of participants in 2AFC experiments to map a perceptual and a conceptual dimension on each other, namely the lateralized response devices and the two categories of stimuli verbally introduced in the instructions (e.g. large and small). It is especially relevant for an experiment where the perceptual input has to be classified onto a binary dimension and the corresponding answer has to be given by reaction on an also binary dimension. Proctor and Cho stated that the mere structural similarity was sufficient to cause RT advantages and a conceptual correspondence of both dimensions was not necessary (Proctor & Cho, 2006, p. 416). The Simon effect (Simon & Rudell. 1967), referring to as a vertical-horizontal congruency of response device and lateralized stimulus presentation, is a prominent example for this phenomenon (Proctor & Cho, 2003). The observation underlying the Simon effect is, that responses are facilitated when the response device and the to react on stimulus are spatially close to each other although their spatial position is irrelevant for the task. Proctor and Cho (2003) found a performance advantage (in RT and error rates) for stimuli presented at the upper-left side of a screen when the response devise deviated a bit to the left from the center of the screen's position as well as for an eccentric response device deviating to the right and upper-right side stimuli.

Proctor and Cho (2006) assumed that a lateralized reaction was confounded by the tendency to react with the device or hand side that was closest to the to be rated stimulus, leading to a systematical effect of the response device position on participants' response behavior. The earlier mentioned expectancy theory and the semantic coding model already had stressed the role of the semantic anchors of the instructions in producing the SCE (Santigo & Lakens, 2014; Jou et al., 2020). Until now, the SCE could not be clearly disassociated from polarity correspondence effects (Santigao & Lakens, 2014). Santiago and Lakens (2014) could show, that the SNARC effect was not an effect of polarity correspondence, because the effect was not modulated by the response device's eccentricity.

1.3.3 Ascending Order of Magnitudes

Comparable to the SPBD Account by Jou et al. (2018) and the BARTlet model (Chen et al., 2014) that stressed the role of learned order information for the occurrence of discontinuities in pairwise comparisons of symbolic stimuli, researchers studying the origins and underlying mechanisms of the SNARC effect discovered associations of order information and spatial position as an alternative explanation for the data patterns that had let to the assumption of the mental number line.

Within the last 15 years, more and more data has been published that relativized the space-number-association by varying the experimental set up. For example, in an examination of the SNARC effect, Ben Nathan, Shaki, Salto and Algom (2009) compared two within participant conditions of a directed comparison task (it was transparent to the participant which of the two to be compared stimuli was the standard and which was the target). In one condition they used a fixed standard between the trials, in the other condition they used a “roving standard” that changed from trial to trial. Targets and standards comprised the range from 1 to 9. Target and standard stimuli were presented simultaneously on the screen in a vertical arrangement. The standard was easy to distinguish from the target, as it always appeared on the top position, 500 ms earlier than the target and printed in a little bit smaller font. According to common practice in computer-based experiments, participants were instructed to indicate their response in each trial by pressing one of two lateralized keys. After half of the trials the assignment of the keys was reversed so that the lateralization was counterbalanced within participants to test for the SNARC effect. Participants were instructed to answer as fast and as accurate as possible. Just like Dehaene et al. (1993) had reported – however, not having aroused scientific interest so far, Ben Nathan et al. (2009) found that the SNARC effect’s occurrence was range depended. “When a given number was larger than the momentary standard [in the roving standard condition], the responses were faster by the right-hand key. However, when the same number was smaller than the momentary standard, the responses were faster by the left-hand key” (p. 581). Ben Nathan and colleagues concluded that, due to the association of numbers and magnitudes the SNARC effect looked like a numerical effect at first sight, but it was rather an association of order and space because the relative size of the target stimulus compared to the standard produced a SNARC-like response side effect not the absolute magnitude of the target.

This finding corresponded to the results reported by Gevers, Reynvoet and Fias (2003) and Schroeder, Nuerk and Plewnia (2017) of a SNARC effect in pairwise ordered position comparison of weekdays, months and letters in the alphabet. These stimuli ranges could be denoted as “overlearned” orders like it is the case of numbers (Schroeder et al., 2017). Again, like in the parity judgement task in Dehaene et al. (1993) the SNARC effect showed up in Gevers et al. (2003) and Schroeder et al. (2017) although the order information was irrelevant for the task.

Parallel to the research on the origins of the SNARC effect, another research line highlighted the role of order information in quantity processing in general. Van Opstal, Gevers, de Moor and Verguts (2008) found the distance effect not only for magnitude information but also for order information (letters of the alphabet). Turconi, Campbell and Seron (2006) reported a performance advantage for ascending numerical orders compared to descending numerical orders (one digit numbers) in paired numerical quantity judgments as well as in paired numerical order judgments. The participants performed speeded undirected selection tasks and were instructed to indicate their response after the stimuli of a trial had disappeared. Turconi et al. (2006) also reported that in the quantity judgement task, the distance effect was more pronounced in pairs of descending numerical order than in pairs of ascending numerical order. In the order judgement task, they observed a mediation effect of order on the distance effect – while they reported the standard distance effect for descending number pairs (stimuli were presented in a horizontal arrangement), for ascending numerical order they found a better performance for consecutive number pairs than for numerically more distant pairs. The authors concluded that the found modulation of the distance effect by numerical order in quantity judgements and its moderating effect in order judgements pointed at different mechanisms underlying the processing of descending orders and ascending orders. At least, quantity as well as order judgments of number pairs would employ magnitude and order judgment processes. They stated a performance advantage for consecutive ascending number pairs over rather distant ascending ordered number pairs due to the fluency of successive ascending numerical orders as a possible explanation for the reversed distance effect in order judgements.

Turconi et al. argued that the reason for their findings was the strong association of orders to our mental number line and that orders fitting the ascending order from left to right of the mental number line had a processing advantage. According to the latest findings of order effects for learned orders beyond numbers (as mentioned above e.g. Schroeder et al.,2017; Gevers et al., 2003), Turconi et al.'s causal interpretation could also be reversed – numerical comparisons that fit the mental number line had an processing advantage because ascending orders in general had an processing advantage.

Working memory account. Recently researchers that claimed the roots of the discontinuities of symbolic magnitude comparisons to lay in working memory (Van Dijck & Fias, 2011; Chen et al., 2014; Deng, Chen, Zhu & Li, 2017) attempted to explain the SNARC effect with the modus operandi of this memory system. Van Dijck and Fias (2011) tested the hypothesis of the SNARC effect being a mnemorical effect instead of an automatic intrinsic spatial frame as the ATOM and the CMT suggested. In their experiments, they increased the working memory load of the participants that had to perform a parity judgements task of number pairs (like in the standard SNARC effect assessment, e.g. Dehaene et al., 1993). In their experiments, participants were instructed to respond only to stimuli that had been part of a previously encountered pool of numbers. The SNARC effect could not be found. Similar to the reasoning against accounts of semantic long-term memory theorists for the SCE and distance effect (see section 1.2.3, e.g. Holyoak, 1978, Jou et al., 2018; Jou et al., 2020), Van Dijck and Fias (2011) argued that the association of space and numerical magnitude observed in previous studies was more flexible than overlearned orders or phylogenetically established associations could explain. They hypothesized the previously observed SNARC effect rather resulted from actual relevant orders that were represented in working memory while performing the crucial decision task. According to this assumptions, the SNARC effect would happen on the response level of the comparison process rather than on the encoding level, as spatial mapping theories suggested. This had already been claimed by Keus and Schwarz (2005) who reported the SNARC effect in an experiment where the response to a parity judgement task had to be given vocally instead of manually as in the standard SNARC paradigm.

According to the SPBD account introduced by Jou and colleagues (2018), van Dijck and Fias argued that the serial position of the items in a range relevant for the actual to be performed task was coded in working memory. Hence, positional coding often had produced a data pattern that had been interpreted as spatial mapping (e.g. by Dehane et al., 1993) leading to the postulation of the SNARC effect.

Recently a broad preregistered replication attempt of Fischer et al. (2003) by Colling et al. (2020) on 1105 participants in 17 different labs failed. In the study several potential moderators of the failed replication of the attSNARC effect were tested, like finger-counting habits, reading and writing direction, handedness, and mathematics fluency and mathematics anxiety (Colling et al., 2020, p. 15). Colling et al. (2020) only found significant correlations for the length of the ISI and the association of numbers and space. The only condition in that a SNARC effect showed up had a latency of 500 ms between two to be compared stimuli. This result stressed the working memory account and let the authors conclude that the attSNARC could not be a key argument for the association of numbers and space. The contradicting interpretations in favor of the SNARC effect were addressed by Colling et al. (2020) as well. They pointed out, that the task employed by Dehaene et al. (1993) had been inappropriate to conclude on spatial mapping because the reported response side effects had been range depended – for instance, reactions on the digit 5 in a set from 4 to 9 were faster with the left hand and in a set from 0 to 5 faster with the right hand (Colling et al, 2020, p. 2). According to Colling et al., a response side effect in the RT patterns of absolute magnitude judgements would have been necessary to conclude on SNARC.

1.4 Effects of the Sequence of Target and Standard in Paired Comparisons

In directed pairwise comparison tasks a standard stimulus and a target stimulus are experimentally defined and installed. The purpose of this setting is to reduce the human decision process that we engage in frequently, in a quick and unconscious manner, on its basic elements not to be confounded by processes of visual search, target detection and standard selection. This is necessary to carefully explore the influence of the target's and standard's features and their interaction with the presentation mode (e.g. spatial and temporal arrangement, linguistic introduction) on the performance of the comparison. A factor that exclusively has been investigated on a systematic basis in psychophysics (SPE, see section 1.1), not in studies on symbolic magnitude comparisons, is the sequential order of standard and target.

1.4.1 Effects of the Direction of Comparisons

In cognitive and social psychological research, two prominent lines of research investigated the influence of systematically installed standards on target ratings. On the one hand, there is the *anchoring* paradigm, investigating the influence of referent information presented in advance of a target on the rating (Tversky & Kahnemann, 1974; Biernat & Manis 1994; Biernat, 2005). In the priming paradigm the anchor stimulus is referred to as a standard but it is usually not explicitly installed as a stimulus the target should be compared to but rather as a side information. On the other hand, there is the research on the influence of the direction of comparisons on relative judgements of (the magnitude of) targets. The binary variable of direction of a paired comparison distinguishes between *upward comparisons*, cases in which the target has a lesser expression of the characteristic in question than the standard, and *downward comparisons*, cases in which the target has a higher expression (e.g Suls & Wheeler, 2012). The direction of the comparison could be shown to influence the outcomes of comparisons and therefore the categorization process. These outcomes respectively affect ratings of the compared objects and subjects. A well-studied example is the effect on self-esteem when comparing oneself with superior (minor) others on a certain dimension (Festinger, 1954; Aspinwall & Taylor,1993; Collins, 1996).

In both research lines, anchoring and direction of comparison, the implicitly presumed starting point of a comparison is the target. The attribute *upward* and *downward* refers to a perspective from the target's position to the standard. Most studies implicitly refer to an of the standard to be more common and the most accessible stimulus associated with the target (Mussweiler, 2003). Hence, to become a standard, a stimulus had to be known better than the target (Tversky & Gati, 1978). As Link and Heath (1975) argued the standard to be the "yardstick of a comparison" (Link & Heath, 1975, p. 3). Implicitly in the history of encountering the to be compared stimuli, the standard must have been encountered before and more frequently to the target.

In experimental research on pairwise comparisons target and standard were not per se identified and distinguished by their temporal occurrence but by the positional coding introduced in the task's instructions.

1.4.2 Asymmetry in Judgements of Similarity and Difference

Amos Tversky was the first psychological researcher who found that the informativity and the outcome of a comparison vary with its direction (Tversky, 1977), to be precise: with the sequence of occurrence of standard and target. Tversky's publications in the late 1970s dealt with the basic mechanisms of categorization, an essential organizing principle of human information processing and retrieval. With the *feature matching model* of comparison Tversky claimed that, when humans engage in relative judgments, they needed to recall previous knowledge about a familiar similar object from memory to match its features with the features of a target that is, at the moment of the comparison, in the focus of attention. Into which category we sort the target, depends on the task, respectively in the real world on accessibility, availability and applicability of category cues (Higgins, 1996). These three factors for standard selection were specified in more recent models for social comparisons, the selective accessibility model (SAM) by Mussweiler (2003) and the shiftig-standard model by Biernat, Manis and Nelson (1991). In studies according to these models, the effects of the direction of comparisons, were measured with the paired comparison paradigm employing predefined standards, respectively reference points.

Feature matching model. The feature matching model theorizes that humans performed similarity judgements by matching the features of the referent retrieved from memory and the features of a newly encountered target. In experiments testing the model, the features of target and standard are usually provided or shall explicitly be reported by the participants (Tversky and Gati, 1978). Tversky theorized in 1977, that the direction of a comparison (standard preceding target versus target preceding standard) influenced the degree of perceived similarity of two objects moderated by the objects' saliency (unique, better accessible features). Furthermore, he assumed that judgements of similarity would be facilitated whenever the order of a less salient target preceding a more salient standard would be presented. According to Tversky, for judgments of difference the opposite case would be favored: A more salient standard should precede a less salient target. Tversky had denoted this moderating effect of order of target and standard the *asymmetry of judgements of similarity.*

In their studies, Tversky and Gati (1978) used the names of countries as stimuli. In their first experiment, participants were randomly assigned to two groups, a ‘similarity judgement group’ and a ‘difference judgement group’. They were instructed to select pairs of countries from a list of pairs according to the pairs’ degree of similarity (difference). The similarity group had to select the most similar pairs, the difference group had to select the most different pairs. Randomly sorted after a prerating, there were prominent pairs of countries (e.g. USA, U.D.S.S.R) and non-prominent pairs (e.g. Paraguay, Ecuador). Tversky and Gati found that in both groups the prominent pairs were more frequently selected, slightly more often in the similarity judgement group though. The authors interpreted the data as supporting Tversky’s (1977) *focusing hypothesis* that people attended more to common features in similarity judgements than in differences judgments, where they attended more to differing features.

Moreover, Tversky and Gati attempted to test the asymmetry of similarity and difference judgements in pairs of mixed saliency and constituted another experiment. According to Tversky’s previously proposed feature matching model, the authors hypothesized for pairs of countries with varying saliency that the perceived similarity of a country was higher when the less salient country was presented before the more salient one and the vice versa effect for difference judgements. The most popular example from their studies, that has frequently been used to illustrate asymmetry in similarity judgements, is the confrontation of the two sentences “Red China is like North Korea” versus “North Korea is like Red China” (Tversky & Gati, 1978, p. 84). The first sentence in which the more prominent, salient country was mentioned first led to a lower similarity judgement than the second sentence, in which the less prominent one stood at the starting point of the comparative sentence. Transferring Tversky’s theory upon difference judgements, the opposite asymmetry rule was reported to be valid. When participants were presented with the two reformulated sentences: ‘Red China is different from North Korea’ versus ‘North Korea is different from Red China’ the difference between the two counties was rated larger for the first sentence (Tversky & Gati, 1978). To facilitate difference judgements the standard had to precede the target as it had more unique features and therefore, participants found the differences between target and standard more efficiently.

Interviewing the participants of Tversky’s experiments on the preference of one of the two orderings of standard and target, participants indicated to prefer the more fluent, easier to perform sentences that followed the syntactic rule of subject (target) preceding object (standard). At this point of scientific evidence, a linguistic effect cannot not be discarded to be responsible for the found asymmetries. One could assume that the cognitive operation of comparing a target to a standard is a preferred ordering and that language adapted to this order in communicational evolution – the syntactic order of comparative sentences would fit the mental operation of pairwise comparison. Further studies on the asymmetries of comparisons without linguistic confounds could support this assumption (see Chapter 2).

1.4.3 Asymmetry in the Detection of Addition and Deletion

Comparable to the experimental set up of studies that tested the origins of the semantic congruency effect (e.g. Banks & Flora, 1977, see 1.2.3), in 1986, Agostinelli, Sherman, Fazio and Hearst varied the position of the comparative instruction in a change detection task to affect the relative saliency of the to be compared stimuli. They found that the detection and identification of change in two subsequentially presented stimuli was affected by the awareness (no awareness) of the task’s instruction during the encoding of the stimuli. In line with Tversky’s feature matching model, Agostinelli et al. found that the accuracy of judgements, whether a feature had been added or deleted from one stimulus to another, was differently affected by the relative position of target and standard.

In their first experiment, Agostinelli et al. (1986) tested how stimuli were compared when the task instruction were given after the presentation of the stimuli. Unaware of the task, participants were exposed to sixty ‘study sildes’ showing hand drawings of unambiguous everyday items (e.g. car, shoe). The task was to examine the sequentially presented slides. Afterwards, the main task’s instructions were provided. Then the experimenter handed out a booklet of eighteen ‘test slides’, a collection of items participants had seen on the previously presented study slides. For each test slide, participants were instructed to indicate whether the item on the test slides (hypothesized standards) differed from the item in the study slides (hypothesized targets) and when the answer was yes, participants should further indicate whether an aspect of the drawing had been deleted or added in the test slide compared to the study slide. 61% of the changes had been detected correctly, additions were easier to detect than deletions. From the gathered data, the authors concluded that the during task performance perceptually present test slides had served as the standard of the comparison. They further theorized that the first presented study slides (the targets) had been encoded holistically, while the test slides could have been reexamined in a feature-based manner. After encoding, the targets’ features were plotted mentally on the encoded representation of the study slides to detect differences.

In a second experiment, Agostinelli et al. (1986) attempted to reverse the directionality of the comparison and therefore presented the full instructions before the encoding phase of the study slides. Apart from this, the experiment was implemented identically to the first. The authors hypothesized for this experiment that the study slide should be encoded in a feature-based manner, as the comparative instructions had been given in advance. The features of the study slides (standards) were plotted directly on test slides (targets) during encoding to detect changes. Agostinelli et al. observed, as expected from Tversky’s feature matching model, the deletions to be detected more easily and the change detection performance to be better in general. The overall detection accuracy in the second experiment was 85%, deletions were easier to detect than additions. Hence, Agostinelli et al. (1986) could replicate the findings of Tversky and Gati (1978) for difference judgements: in a change detection task performance was better with a presentation order standard preceding target.

1.4.4 The Order of Target and Standard in Preference Judgements

In 1989, Houston, Sherman and Baker attempted to apply the by the feature matching model hypothesized influence of the relative salience of multiple feature stimuli on preference judgements. Adopting Agostinelli et al.'s (1986) experimental set up, Houston et al. attempted to replicate the finding that the instructions' position determines which stimulus of a pair becomes the target and which one becomes the standard. Houston et al. further hypothesized that pairwise presented objects (subjects) described by lists of attributes should vary in ratings of likeability due to unique features weighted differently depending on whether they were target or standard.

Houston et al. (1989) conducted a series of experiments varying the unique features of targets and standards and manipulated their relative position through a variation in the position of the preference rating instructions (Experiment 1: between the two stimuli; Experiment 2: before both stimuli). The characteristics on the lists describing the objects (subjects) had been prerated and combined to result in several equivalent object (subject) prescriptions with an equal amount of good (bad) features and one unique bad (good) feature. Within each of the three experiments, there were two experimental conditions between participants. One group was provided with descriptions of pairs with the same amount of equal good and unique bad features, the second group received descriptions of pairs that had the same amount of equal bad and unique good features. The preference judgements were supposed to vary depending on the interaction of relative position and the valence of the unique features, as, according to the feature matching model, the features of the target were expected to be the starting point of the comparison and therefore weighted more in the preference judgement than the features of the standard.

According to the findings of Agostinelli et al. (1986), Houston et al. had theorized that which stimulus becomes target and which becomes standard should vary due the position of the instructions between the experiments. In Experiment 1, the standard preceding target condition, the authors assumed that unique good features in targets (second stimulus) should lead to a preference of targets even when the standard (first stimulus) would have an equal number of good features; and that unique bad features in both stimuli should lead to a preference of standard, even with an equal amount of bad features of the target. According to the feature matching model by Tversky (1977), vice versa results were expected for the target preceding standard condition (Experiment 2) regarding the position of the stimuli: It was predicted, that participants, on average, would prefer the standard (second stimulus) more often than the target, when the prescriptions of both objects had unique bad features. In the ‘ unique good feature’ condition, the target should be preferred more often. Contrary to the authors’ predictions the patterns of preference judgements across all conditions indicated that the object presented second had become the target of the comparisons performed by the participants. Hence, Agostinelli et al.’s findings that the position of target and standard presented sequentially could be switched by the placement of the instructions could not be replicated.

To test whether the order of standard preceding target was experimentally induced, Houston et al. conducted a third experiment. Because an alternative explanation for participants using the second object as the target of comparison was that it was present in the moment when the preferences choice had to be made. To attend to this possibility, in the third experiment, again, the instructions were provided before the encoding of the first booklet like in Experiment 2. After having examined both description booklets, participants had to perform an unrelated filler puzzle for 5 minutes to disturb the memory trace of the stimulus presented before. After the puzzle, participants were shown a copy of the description of either the first or the second object they had encoded before the puzzle and were instructed to compare this reinstated object to the other previously encoded object, and to make a preference choice between them. Thus, participants had present a copy of the description of one object when they made their preference choices, while the other object had to be retrieved from memory. Again, the same result pattern revealed, indicating that the object presented second became the target as in the previous two experiments.

Houston et al. discussed possible specific mechanisms for relative preference choices, in contrast to judgements of similarity, to account for the failure of the replication of Agostinelli et al.'s findings. They assumed the crucial difference between the tasks was, that the similarity- and difference judgments were to be made more elaborate employing a mental resorting of target and standard, whereas preference judgements were mad in a more spontaneous way being performed right when the second stimulus appeared – making the second stimulus to the inevitable target of the feature matching process.

1.5 Summary: Stimulus Order Effects in 2AFC Tasks

Researchers that experimentally measure the discontinuities of pairwise discrimination and decision making in many domains of psychological research employ the pairwise comparison paradigm adopted from psychophysical DL measurement. The latest models to explain the distance effect, the end effect (or serial position effect) and the SCE are related to the logarithmic compression of DL slopes to the ends of the continuum stated by the Weber-Fechner-Law (Krajcsi & Kojouharova, 2017; Krajcsi et al., 2018) while working memory accounts are the most recent approaches to account for distance effects in symbolic pairwise comparisons and for the stimulus order effects in Psychophysics. It is difficult to bridge the findings of psychophysics and symbolic comparison theorists, because the crucial difference between physical and symbolic magnitudes is, that symbols can be stored as discrete information in memory whereas physical stimuli are processed directly and cannot be judged absolutely. The study of numerical stimuli is fruitful to further clarify the mechanisms shared by comparisons of physical and symbolic stimuli, not only to further explore the relevance of learned or implicit orders but also to shed light on a spot so far overlooked by symbolic comparison researchers: the effect of the sequential order of target and standard. This approach is supposed to be aiding attempts of psychophysicists to model the comparisons processes that underly the Type B effect and the TOE. And it will provide new ideas about which mechanisms drive the comparison and judgment of continuous stimuli (physical percepts) that are shared with the assessment of discrete stimuli.

By now, effects of the sequential order of target and standard have been investigated for discrimination performance in pairwise comparisons of physical stimuli and complex multi-feature objects but have been overlooked so far for pairwise comparisons of symbolic stimuli. While the role of serial position in learned orders of symbolic stimuli on discrimination accuracy and speed of symbolic have been studied broadly to investigate the underlying mechanisms of the distance effect, the SNARC effect and the SCE, the relative position of standard and target has not been questioned in this line of research (see sections 1.2 and 1.3). Even in psychophysics, the elaborate investigation of the origins of the Type B effect and the TOE, only started in 2009 (Ulrich & Vorberg, 2009).

One possible reason for the omission – that effects of the relative temporal occurrence of target and standard were not being analyzed in the studies on pairwise comparisons of symbolic magnitudes – might be the implicit definition of the comparison’s standard per se having a higher familiarity and being encountered earlier than the to be judged target. The

Relative Judgment Theory (Link & Heath, 1975) defines the target in a comparison as the newly encountered stimulus while the standard being the stimulus being encountered in advance. This immanent assumption might have hindered researchers to question the sequence of occurrence of standard and target in the field and in the laboratory, although it might enlighten basic principles of relative judgments.

According to the focusing hypothesis and the feature matching model by Tversky the target was always the starting point of a paired comparison, its features were plotted onto the standard and this led to a rating of higher similarity compared to a case when the standard would be the comparison's starting point. For difference judgements, according to Tversky, this relation was reversed. To be more sensitive to differences between two stimuli in a directed comparison the standard had to stand in the syntactic position of the subject (in a linguistic sense). Tversky's findings and interpretations partly correspond to the findings of Diyas & Ulrich in 2014. Psychophysical studies investigating difference judgments to measure discrimination sensitivity (DL), a lower DL (equates higher sensitivity) was found for 'standard preceding target' trials (negative Type B effect) which corresponds to Tversky's findings. For equality judgements, Dyjas and Ulrich reported PSE lower than the standard in 'standard preceding target'-trials (negative TOE = underestimation of the first stimulus) and in 'target preceding standard' trials the PSE was higher than the standard (positive TOE = underestimation of the second stimulus). As a lower PSE can be translated as more likely/earlier judgement of equality, which could be interpreted as a higher sensitivity to equality. Regarding this point, Tversky and psychophysicists came to divergent findings. The decisive difference between Dyjas & Ulrich's equality judgements and Tversky's judgements of similarity is, that equality judgement tasks in psychophysics are undirected (selection task) while Tversky's task was directed and a standard and a target were defined by the syntactic order of the comparative sentence (see section 1.4.2). Tversky's asymmetry effect, as replicated by Agostinelli et al. (1986), therefore can be seen as corresponding to the positive Type B effect for similarity judgments and to a negative Type B effect for difference judgements. Tversky & Gati (1979) and Agostinelli et al. (1986) found, that when the target was encountered first, the similarity of target and standard seemed to be greater. When the standard was encountered first the difference seemed to be greater, in other words the judges were more sensitive to the differences of the stimuli.

The recently favored explanation for the Type B effect to occur, is that the target has a greater impact on the comparison outcome that the standard (e.g. Hellström et al., 2020, see section 1.1.4) which also corresponds to the focusing hypothesis of Tversky. Taking the recent findings of psychophysicists and the research on asymmetry in similarity judgements together, the processing advantage for the sequence of 'standard preceding target' has its origin in an accessibility advantage for the target when it is presented in the recent position of the to be compared pair and the comparative instructions are known in advance. The selective accessibility hypothesis could ideally be studied with roving standards, varying instruction position and varying working memory load in repeated measurement settings of 2AFC tasks. This has so far not been possible for studies on pairwise comparisons for physical stimuli (as lined out in section 1.1.4). The investigation of the performance advantage for the recency of the target in pairwise comparisons employing numerical stimuli, combining the advantages of symbolic and physical stimuli (see section 1.3), could reveal further insights into origins and boundary conditions of the effect.

Exploring the origins of the SNARC effect (see section 2.3), researchers recently came to the conclusion that momentary relevant order information that are hold and compared in working memory are responsible for the SNARC (see Ben Nathan et al., 2009; Fischer & Shaki, 2016). In favor of this approach, the idea of the mental number line as the underlying long term memory concept had to be discarded (Colling et al., 2020). The relevance of serial order coding opposed to the spatial coding approach has been be proven frequently over the past few years (e.g. Page et al., 2004; van Opstal et al., 2008; Chen et al., 2014). The ascending order advantage appears to be a domain general principle of information organization that revealed its performance advantage on the response level in 2AFC tasks (e.g. Van Dijck & Fias, 2011). Regarding the mentioned order effects (see section 1.3.3) it remains an open question whether the crucial role of order information is only evident for overlearned orders (like number, month or weekdays), or, as the SPBD account (Jou et al. 2018; 2020) and the recursive computer models for comparisons of symbolic and numerical stimuli suggest, orders held in working memory are responsible not only for the prominent discontinuities of symbolic pairwise comparisons but also for the attSNARC effect. As the effect could not be replicated recently (Colling et al., 2020) it remains suggestable that the previous reports of attSNARC effects are rather generated by actual relevant range dependent order information.

According to all the research lines presented in Chapter 1, that highlight the importance of the processing of order information for the explanation of the discontinuities of pairwise magnitude comparisons, it appears to be necessary to combine the knowledge and scientific expertise of all the mentioned domains to conclude on the unifying underlying mechanisms of order and sequence effects.

2. Empirical Part

The empirical work followed an inductive approach. Hence, the hypotheses were successively adapted to the results of the conducted experiments. The initial attempt was to generate evidence for a SNARC-like performance advantage in paired price comparisons employing a 2AFC paradigm. Based on the existing research on the association of magnitude and space and the according performance advantages (see section 2.3, e.g. Shaki et al. 2012), the first hypothesis was formulated.

The SNARC effect of numerical cognition affects the perception of prices presented in a comparative manner. Bargains that are presented in a SNARC-fitting way can be processed, as well as reacted on, faster and more accurate. (H_1)

Results of Experiment 1 pointed at a so far overlooked effect of the sequence of standard and target in number comparison that had been overlooked in previous research. The focus of interest in the work at hand switched from the SNARC effect to this new aspect of pairwise comparisons and was the beginning of a series of 10 experiments examining the effect of the order of standard and target. The following hypothesis was the first hypothesis of this new research line.

In a comparison task of two prices of different size, participants' performance, measured in RT and error rates, is enhanced when the experimental temporal occurrence of the previous sales price (standard) and the actual sales price (target) fits the real market place temporal occurrence of standard price preceding target price. (H_2)

In the course of the exploratory empirical work a third hypothesis was generated to test the found sequence effect in spatial orders.

In a comparison task of two prices of different size, participants' performance, measured in RT and error rates, is enhanced when the spatial arrangement of the two prices fit their real *market place temporal occurrence, namely standards being presented at left or upper positions, targets being presented at right or bottom positions.* (*H_3*)

A fourth hypothesis was generated to test for an alternative explanation of the found effect of the sequential order of target and standard in paired comparisons of prices against the temporal order of standard and target.

In a comparison task of two prices of different size participants' performance, measured in reaction times and error rates, is enhanced when the standard is presented before the target, even when the theoretical temporal occurrence of the stimuli was 'target occurring before standard'. 'Prospective bargains' are identified faster and more often correct when future sales prices (standard) being presented temporally before actual sales prices (target). (*H_4*)

To test the found *standard-target-sequence-effect* (STSE) with numerical stimuli in diverse contextual frames, another hypothesis was formulated.

In directed pairwise comparisons of numbers, participants' performance, measured in reaction times and error rates, is enhanced when the standard temporally precedes the target. (*H_5*)

Another hypothesis was formulated based on the previous findings. The STSE should be extended on another magnitude dimension – namely volume of geometrical figures.

In directed pairwise comparisons of the volume of quadrats, participants' performance, measured in reaction times and error rates, is enhanced when the standard temporally precedes the target. (*H_6*)

Experiment 11 was a rather exploratory experiment testing for the underlying mechanism of the STSE by varying the instructions position (compare Banks & Flora, 1977; Holyoak, 1978; Agostinelli et al 1986). An effect of the instruction's temporal position would reveal whether the effect happened on an encoding or on a response stage of comparison performance.

2.1 Experiment 1: SNARC in Price Perception

The aim of the initial study was to test for a performance advantage of visual detection and behavioral reaction on bargains presented in a SNARC fitting spatial arrangement (see H_1), higher prices standing in upper positions and lower prices standing in bottom position. The vertical SNARC effect (Ito & Hatta, 2004; Petrusic et al., 2011) has so far been demonstrated for the association of response side, spatial position and numerical value but not in a setting with roving standards and roving position of the standard.

Recently, researchers in cognitive science studied the influence of the task-irrelevant relative spatial position of items on judgements of their relative value (Fischer et al. 2003, Gevers, Lammertyn, Notebeart & Verguts, 2006) as well as on relational judgments of persons on dimensions associated with magnitude or various qualities (e.g. Meier & Dionne, 2007). However, these findings have not been adopted on the relative assessment of prices so far.

In times of increasingly competitive marketing strategies, the study of price presentation formats that fit the principles of human visual perception and the characteristics of intuitive decision making is of enormous relevance. An everyday life example for intuitive decision making, based on pairwise comparisons, is the assessment of bargains, usually presented in the format of an actual sales price placed near by a previous higher sales price. The spatial arrangement of actual and previous price varies between brands, marketplaces and brochures – it seems to be arbitrarily selected.

In Experiment 1 the influence of the spatial arrangement of prices on participants' performance in a comparative judgement task was assessed. Speed and accuracy of participants’ detection reaction on bargains when presented in a vertical price arrangement were measured (compare Ben Nathan et al., 2009). A *bargain-judgement-paradigm* was developed to combine a classic SNARC effect task with a realistic setting of number comparisons.

2.1.1 Open Science Statement and Power Analysis

The preregistration of Experiment 1 as well as materials and data are available online in the Open Science Framework (https://osf.io/w9k3m). A 2 by 2 within-participant design was chosen, resulting from the experimental combination of two factors: SNARC (compatible vs. incompatible price arrangement) and Bargain (yes = previous sales price higher than actual sales price; no = smaller previous sales price than actual sales price). An a priori power analysis was performed using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) to calculate the required sample size for a preset power of 80% and a conservatively estimated medium to large effect size of $f = .30$ for the difference between SNARC compatible and incompatible trials in this within-subjects design. Power analysis resulted in a required sample size of $N = 90$.

2.1.2 Method

Subjects. 97 students participated in a 20 minute test battery in a laboratory at the campus of the University of Cologne, compensated with course credit or 3 Euro. Demographic data for one of the participants was missing. 66 participants identified as women and 30 as men. Their mean age was 24 years.

Material. The number of test trials was arbitrarily set to 100. To compare 4 conditions resulting from the combination of the two binary factors within participants, 25 photos of real-world drug store products (monochrome bottles and tubes typically encountered as industrial cases for lotions and shower gel) were sampled. Labels and any kind of writing on the products were erased. The photos were digitally brought to the same size (ranging from 163 x 445 px to 201 x 467 px) and combined with two prices standing underneath the product photo. Each pictorial stimulus consisted of a vertical alignment of a product photo on top and two prices beneath. Every participant that completed the task encountered the same product photo four times with four different price arrangements. The four price arrangements resulted from the permutation of the two binary factors SNARC (compatible vs. incompatible) and bargain (yes vs. no). Every within participant condition was characterized by the special arrangement of the prices. Figures 1 to 4 display exemplary stimuli for each condition. The stimuli appeared in full screen during the experiment. The whole frame (the photo and the two prices against a white background) had the size of 1280 x 720 px resulting in actual 45 x 25 cm on a 60 cm screen. The prices were printed in red (font: *Calibri*, size: 36) with one of them being crossed by two lines with the same thickness as the font.

All prices were presented in Euro and consisted of a single digit ranging from 1 to 9 followed by a comma and two decimal places. The higher of the two prices within a stimulus slide was accompanied with two times 9 behind the comma, the lower price was followed by two times 0 after the comma. This should support the SNARC effect to show up in the participants' response behavior. To avoid a distance effect, the difference between the two prices was held constant on 1,99 throughout the trials.

Procedure. In the laboratory with a capacity of four parallel testings, participants were seated in front of a table with a distance of 60 cm to the computer screen. The keyboards, that the participants should use to indicate their responses, were fixed on the tables in the middle between participant and screen to prevent a Simon Effect in the response behavior.

At the beginning of the experiment, participants were instructed to decide in every trial whether the displayed offer indicated a bargain or not. They were instructed to just look at the prices to make a decision, and that whether they would really buy the product was irrelevant.

The CTRL-keys were introduced as the two keys to indicate the responses. To prevent an effect of polarity correspondence overall cases, half of the participants were instructed to react on a bargain pressing the left CTRL-key and to react on a ‘fake bargain’ pressing the right CTRL-key. The other half was instructed to indicate their decision with vice versa key assignments. Instructions said to respond as fast and as accurate as possible.

By pressing the space key the participants got to the next instruction page. They were instructed to lay their index fingers on the two CTRL-keys of the keyboard in front of them and to press the space key again whenever they felt ready to start the task.

The offers appeared on the screen for a maximum of 5 seconds before the next trial appeared. RT was measured from the onset of the stimulus to the participants’ response. If, within this interval, no response had been given, the trial was recorded as an erroneous trial.

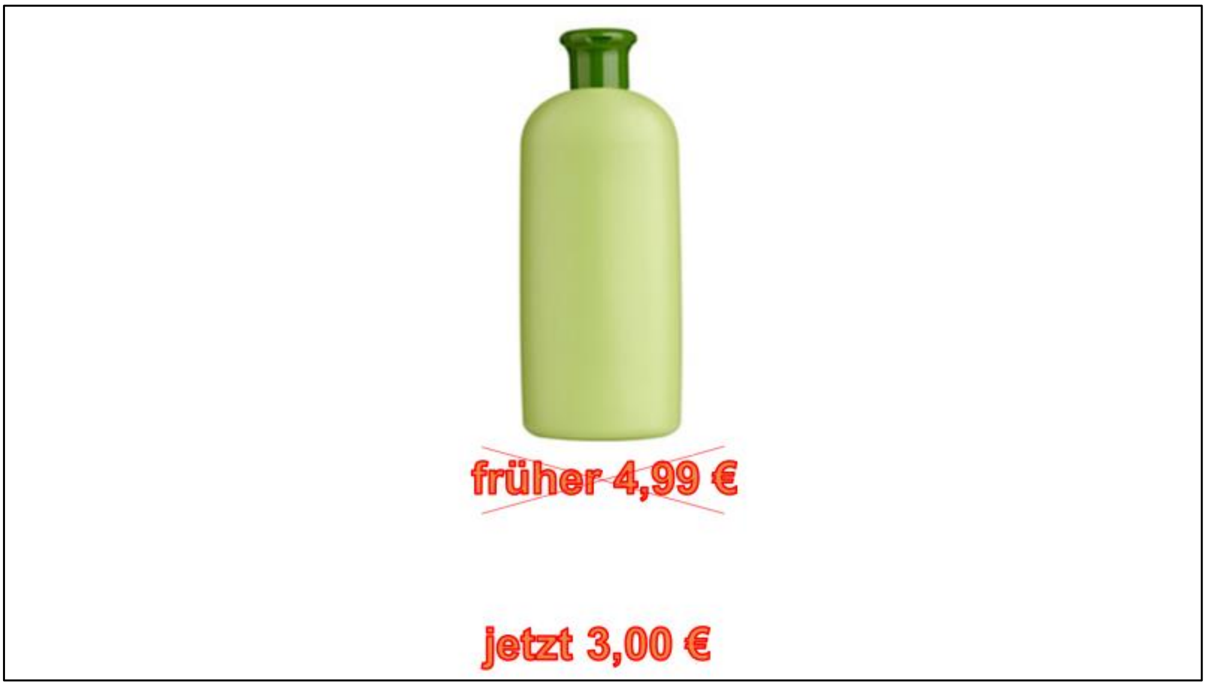

Figure 1. Example for a stimulus of Condition 1 in Experiment 1.
A bargain (higher previous price than actual sales price) is presented in a SNARC compatible manner (higher price above smaller price). “Früher“ is the German word for “previously”, ”jetzt“ is the German word for “now“.

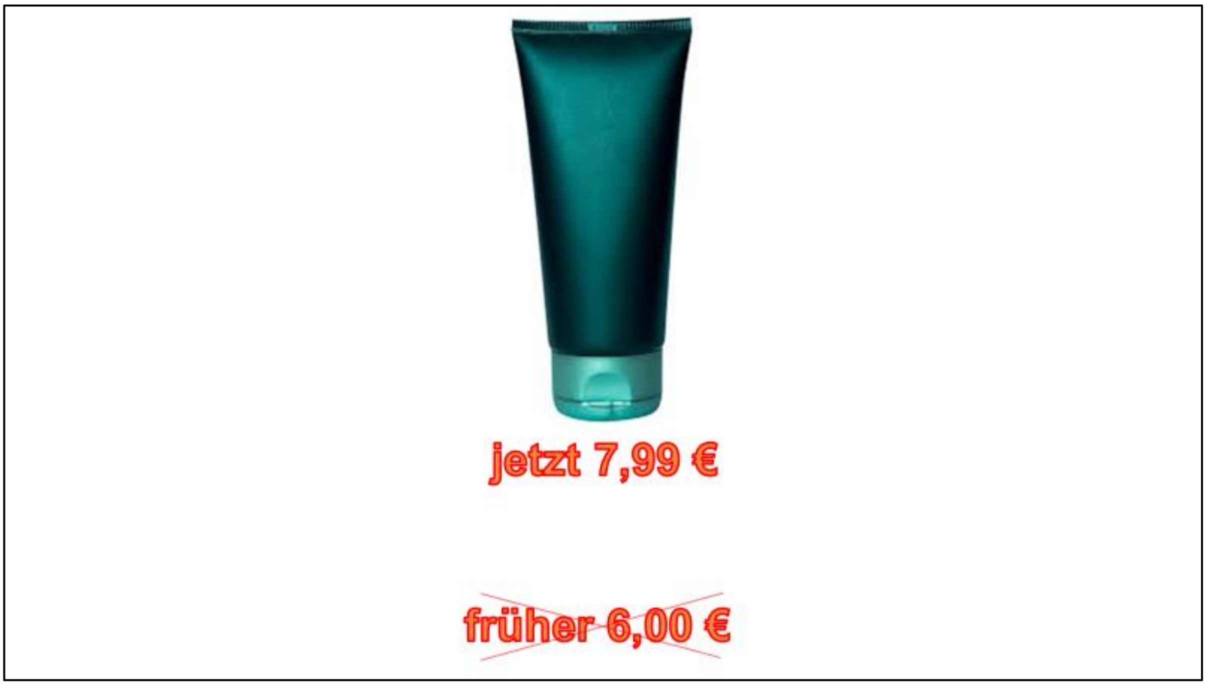

Figure 2. Example for a stimulus of Condition 2 in Experiment 1.
A fake bargain (lower previous price than actual sales price) is presented in a SNARC compatible manner higher numbers above smaller numbers).

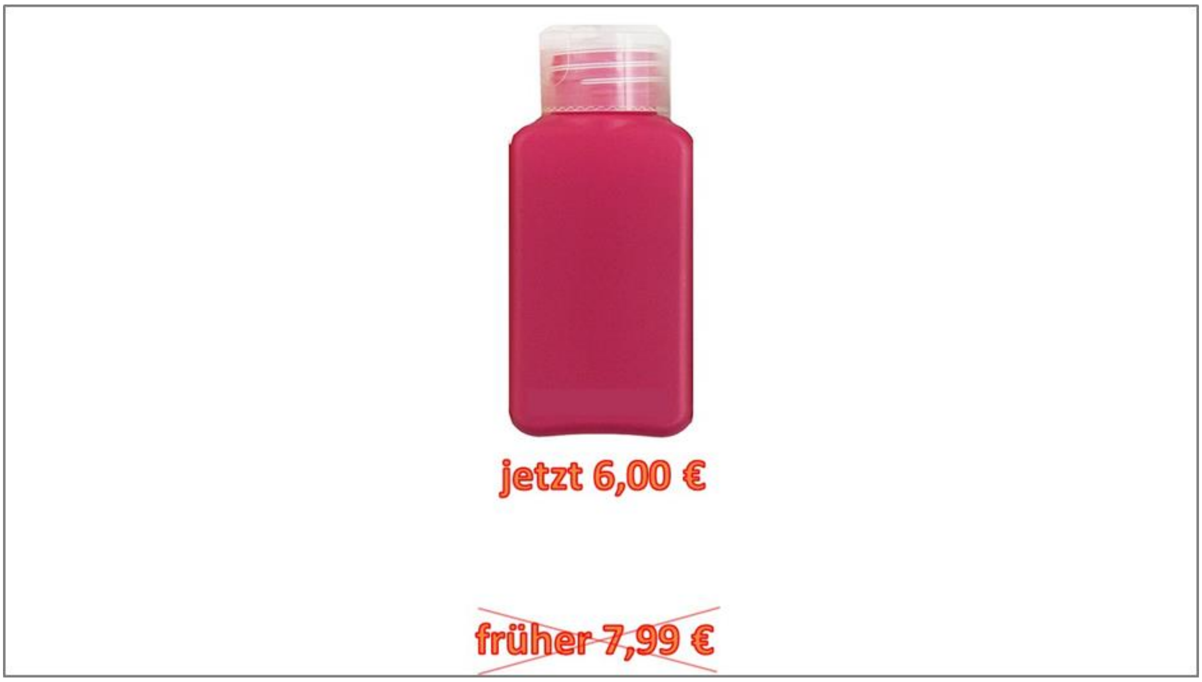

Figure 3. Example for a stimulus of Condition 3 in Experiment 1.
A bargain (higher previous price than actual sales price) is presented in a SNARC incompatible manner (smaller price above higher price).

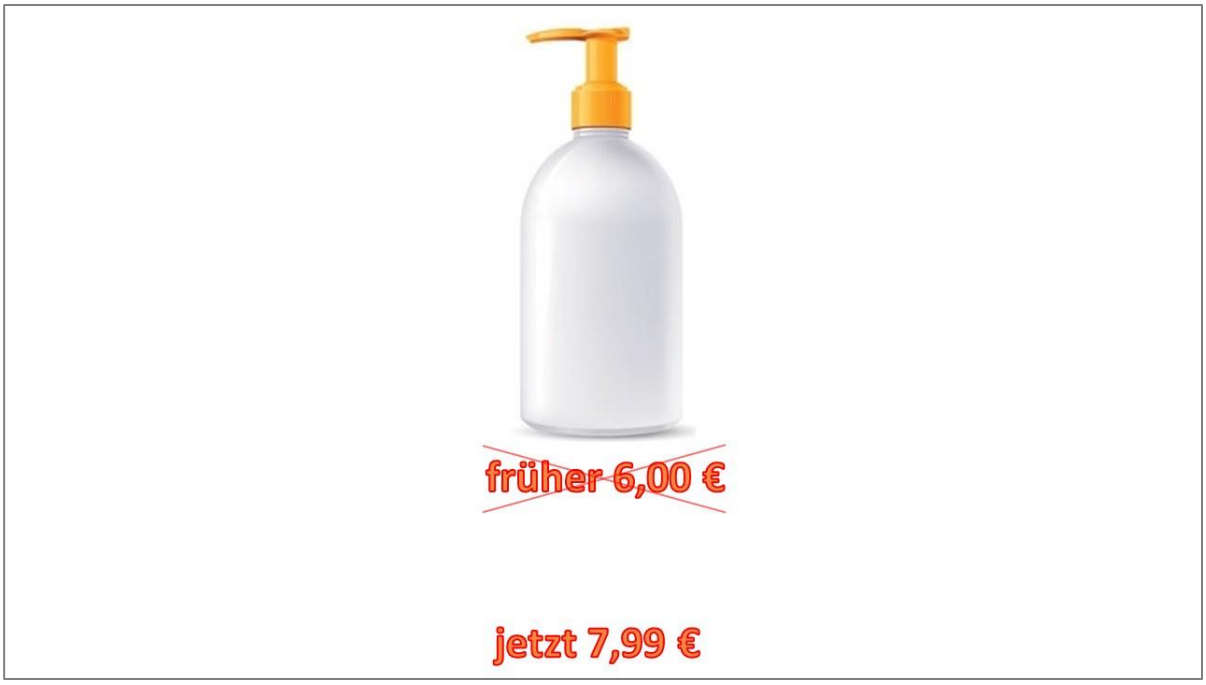

Figure 4. Example for a stimulus of Condition 4 in Experiment 1.
A fake bargain (lower previous price than actual sales price) is presented in a SNARC incompatible manner (smaller price above higher price).

2.1.3 Results

The data of 14 subjects fell below the minimal criterion of 70% correct and therefore was excluded from further analyses, so that RT and error rates were statistically analyzed with the remaining sample of 83 cases. Noteworthy 6% of the participants showed an error rate of over 85%.

Furthermore, trials with an RT that deviated more than 2.5 times the standard deviation from the mean were discarded from further analyses. This concerned 2 % of all trials over all cases. The mean response time of the trials that were included in the data analysis was *M* = 1,383.44 ms (*SD* = 704.78 ms).

Error rates. 8.4 % of all trials included in the data analysis were erroneous. Condition means are displayed in Table 1 and visualized in the bar chart in Figure 5. A 2 by 2 rmANOVA (SNARC x Bargain) was run. SNARC, in terms of SNARC fitting price presentation – namely larger price on top position smaller price on bottom position – showed no significant influence on the error rates, *F*(1,82) = .26, *p* = .613. Bargain showed a highly significant effect of medium size, *F*(1,82) = 18.54, *p* < .001, η_p^2 = .18. A significant interaction of SNARC and Bargain, *F*(1,82) = 66.52, *p* < .001, η_p^2 = .45, was observed, which had the largest effect size in all analyses performed on this data. Inspecting the descriptives and the interaction diagram in Figure 6, error rates were the lowest in the condition where the presentation format neither fitted the SNARC nor the format of a true bargain. Fake bargains could be detected more often correctly.

Table 1

Condition Means of Error Rates (Exp. 1)

	Bargain			No Bargain		
SNARC	*n*	*M* (*SD*)	95% CI	*n*	*M* (*SD*)	95% CI
Compatible	83	.095 (.13)	[.07, .12]	83	.08 (.06)	[-.04, .19]
Incompatible	83	.135 (.13)	[.11, .16]	83	.03 (.04)	[.02, .11]

Note. CI = confidence interval.

RT. For the analysis of RT, erroneous trials were discarded from the data. A Kolmogorov-Smirnov-Test (*p* < .001) revealed that the distribution of the RT did not correspond to the normal distribution. The distribution of RT was extremely right skewed. To meet the requirements of valid parametric tests the natural logarithm of RT [M_{lnRT} = 7.15 ln(ms); SD_{lnRT} = .50 ln(ms)] was used for further analyses.

Condition means are displayed in Table 2 and visualized the bar chart in Figure 7.

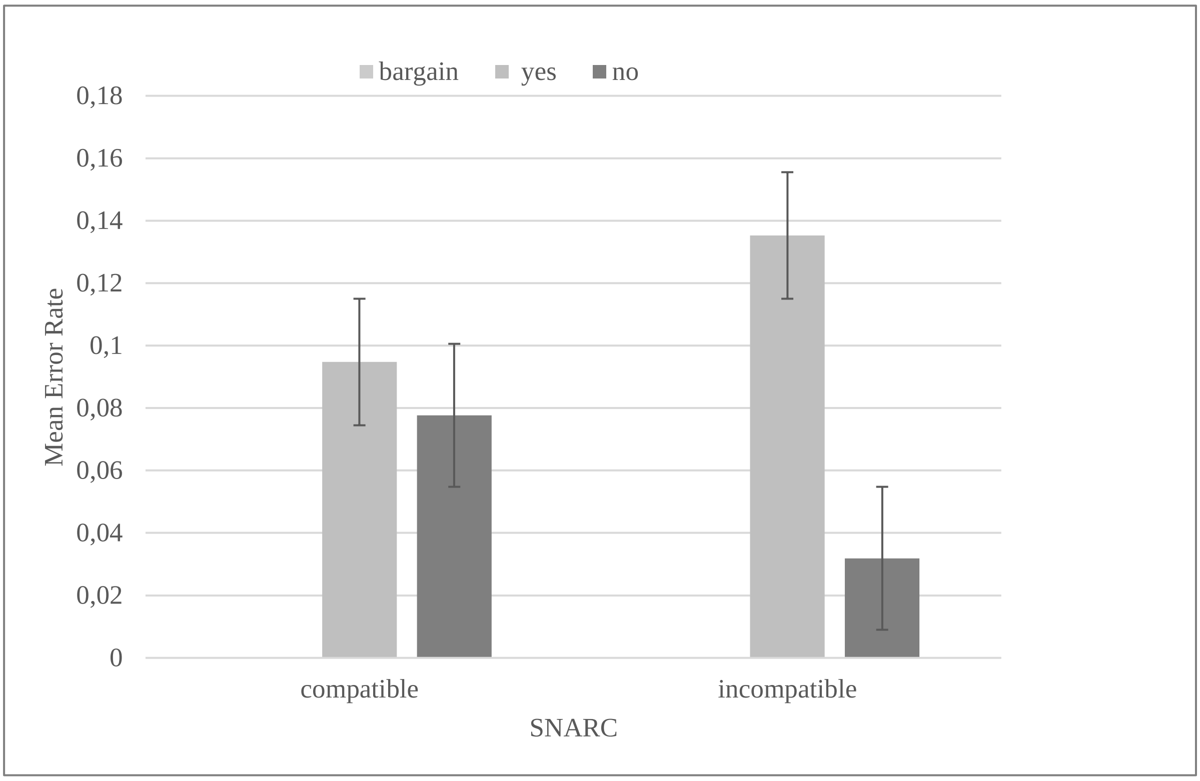

Figure 5. Mean error rates in Experiment 1 as a function of SNARC (compatible vs. incompatible price presentation) and Bargain (yes vs. no). Bars stand for discrete within-participant conditions in the experiment.

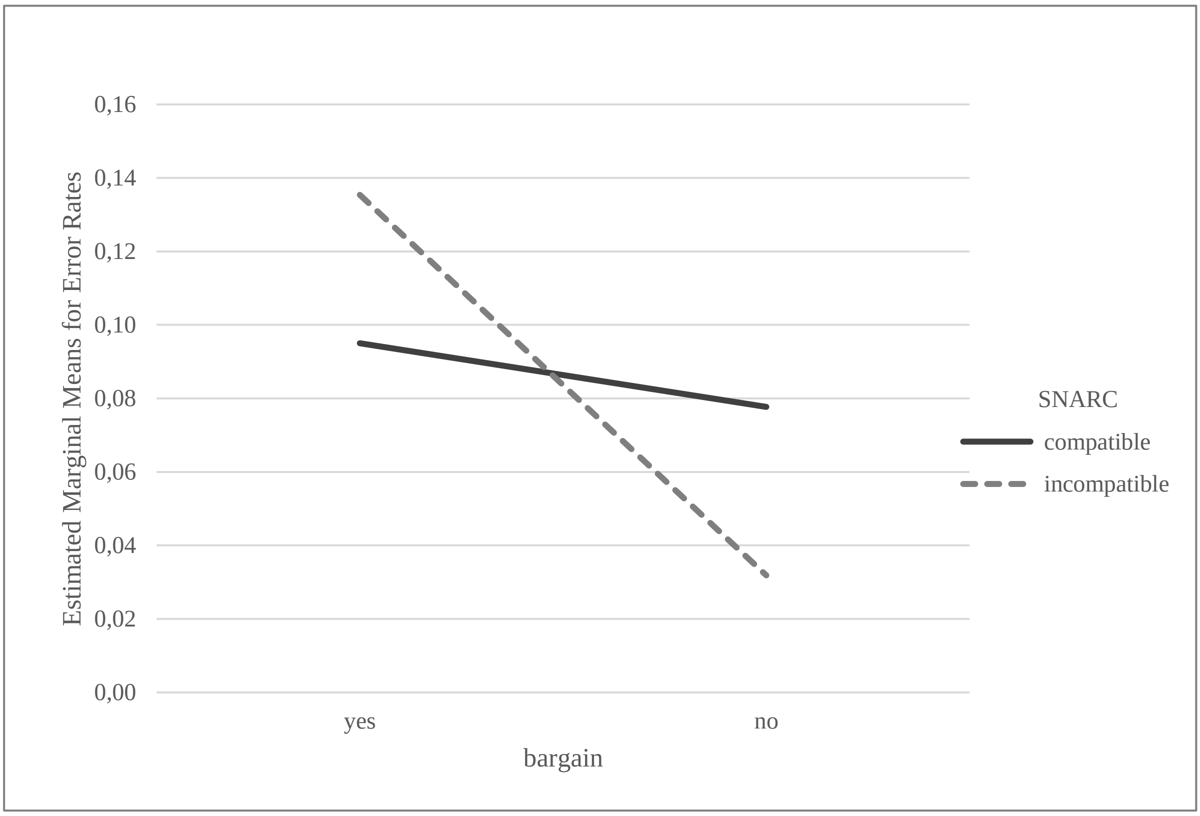

Figure 6. Interaction plot of error rates in Experiment 1.

Table 2

Condition Means of Reaction Times (Exp.1)

	Bargain			No Bargain		
SNARC	*n*	*M* (*SD*)	95% CI	*n*	*M* (*SD*)	95% CI
Compatible	83	1,313.58 (43.11)	[1,229.08; 1,398.08]	83	1,481.13 (46.14)	[1,390,70; 1,571.56]
Incompatible	83	1,355.95 (44.78)	[1,268.18; 1,443.71]	83	1,388.73 (41.88)	[1,306.65; 1,470.81]

Note. Reaction time in ms. CI = confidence interval.

A 2 by 2 rmANOVA (SNARC x Bargain) was calculated. SNARC, like in the analysis of error rates, did not reveal a significant effect, $F(1,82) = 2.47$, $p = .120$ and a highly significant effect of Bargain, $F(1,82) = 42.39$, $p < .001$, with a large effect size, $\eta_p^2 = .34$. Inspecting the descriptives in Table 2, displayed in Figure 7, reveals that participants responded faster to true bargains. There was a significant interaction effect of SNARC and Bargain, $F(1,82) = 14.39$, $p < .001$, as well with $\eta_p^2 = .15$, a large effect size. In the condition in that the participants showed the best performance (fastest correct reaction) on average, the to be judged prices indicated a true bargain and were presented in a SNARC compatible arrangement.

Post-hoc analyses. Taking a closer look at the bar chart in Figure 1, a general performance advantage in trials in that the previous price was presented above the actual sales price was found descriptively in the data pattern. A t-test for paired samples was performed to test for statistical significance. Trials in that the previous price was presented above the actual sales price were compared to trials in that the actual sales price was presented above the

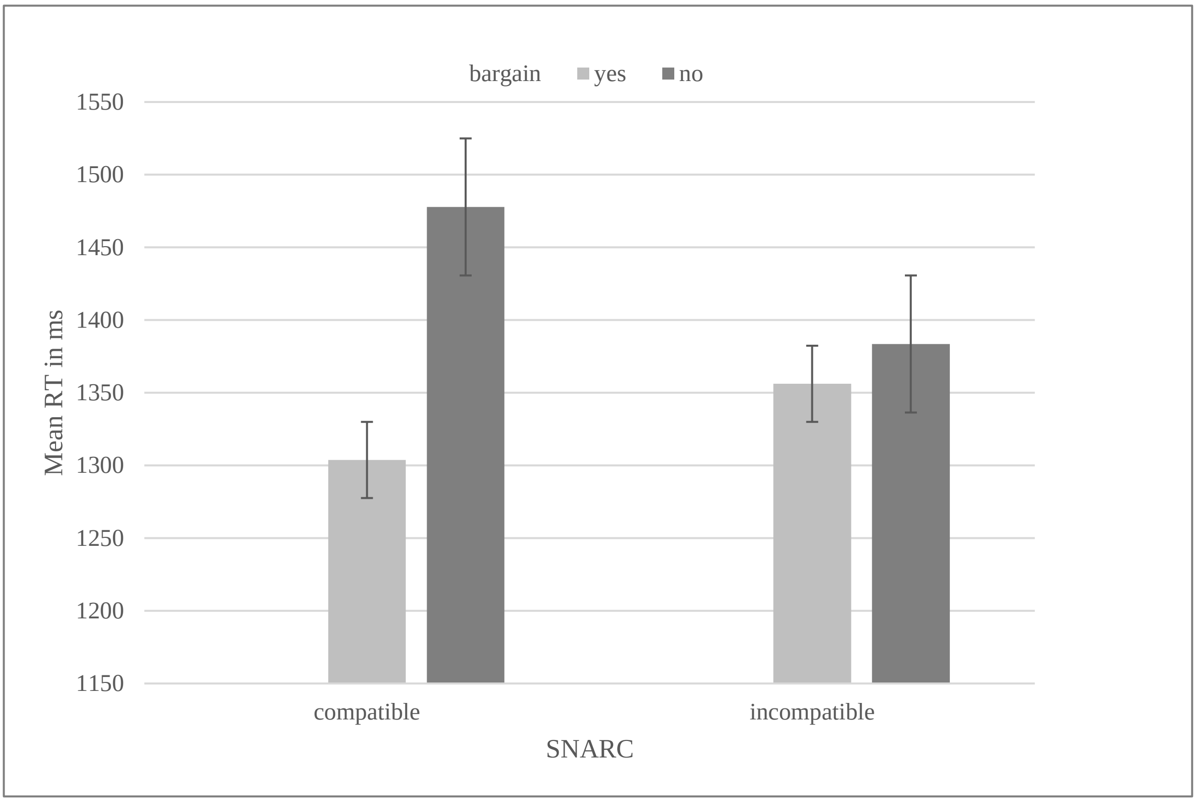

Figure 7. Mean RT in Experiment 1 as a function of SNARC (compatible vs. incompatible price presentation) and the factor bargain (yes vs. no). Bars stand for discrete within-participant conditions in the experiment.

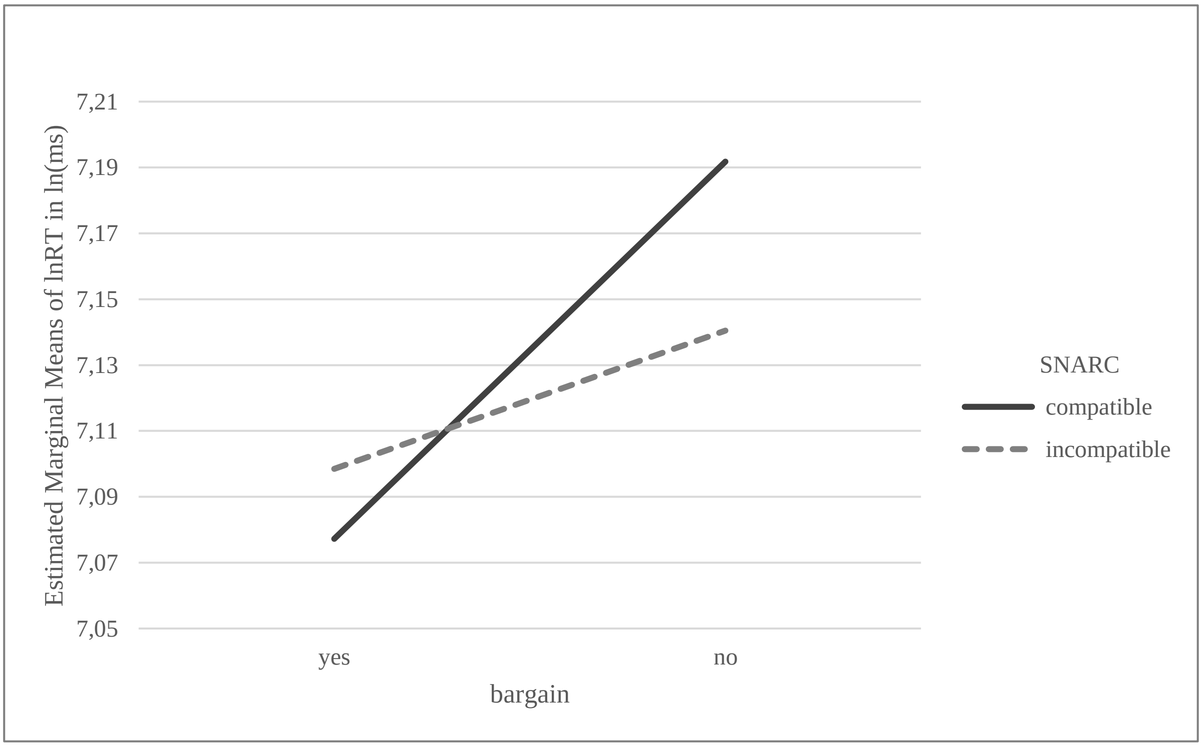

Figure 8. Interaction plot of RT in Experiment 1.

previous price. For error rates a significant advantage for the vertical order of previous price above sales price (*M* = .06; *SD* = .07), compared to the ‘actual price above previous price’-trials (*M* = .11; *SD* = .07) was found, t(82) = 8.16, p < .001. This effect had a large effect size of d_z = .90. The same effect was found for RT. Trials in which the previous price stood above the actual price (*M* = 1,351.15; *SD* = 369.8) showed faster responses than trials in that the actual price stood above the previous price (*M* = 1,418.54 ms; *SD* = 402.13 ms), t(82) = 3.79, p < .001. This effect had a medium effect size of d_z = .42.

2.1.4 Discussion

The a priori defined analyses of error rates and reaction times in the experiment described above revealed contradicting data patterns. Albeit, for both outcomes Bargain and the interaction of SNARC and Bargain showed significant effects, but in opposite directions. While the error rates were lowest (indicating a facilitated performance) in the ‘fake bargain’ and SNARC incompatible condition, the RT were smallest (indicating a facilitated performance) in the ‘true bargain’ and SNARC compatible condition. The large interaction effect in the error rates analysis (η_p^2 = .45) and the comparatively small interaction effect on the reaction times (η_p^2 = .15) seemed to be inverted. The two interaction plots, Figure 6 for the analysis for error rates and Figure 8 for the analysis of RT, seemed to be mirrored on the y-axis, which could indicate a speed-accuracy-trade-off.

H_1, supposing a performance advantage for bargains presented in a vertical SNARC fitting manner, could not be supported by the presented data. Although the higher price in each trial was accompanied with ‘,99’, having expected to support the SNARC effect in terms of the attention guiding effect stated by the attSNARC research. The only aspect that matched the previous predictions was the mean RT in the SNARC compatible true bargain condition.

Taking a closer look at the condition means revealed that participants' performance was affected systematically by the relative position of the previous and the actual sales price. In post-hoc analyses, t-tests revealed that the participants performance was enhanced when previous prices were presented above actual prices with medium to large effect sizes for both performance outcomes (RT: d_z = .42; error rates: d_z = .90).

2.2 Experiment 2: Retrospective Bargains

The results of Experiment 1 indicated that the performance of paired price comparisons, presented in a vertical arrangement, was not affected by a SNARC fitting presentation of higher prices standing in top positions and lower prices in bottom positions. The data pattern of the post hoc analyses of Experiment 1 revealed that the performance of paired price comparisons was enhanced when the standards price was presented in the top position. Following these results, a new focus of research was initiated. Accordingly, Experiment 2 was a spin-off of Experiment 1, attempting to replicate the findings of the post hoc analyses of Experiment 1. It was hypothesized that the temporal occurrence of a previous sales price preceding the occurrence of an actual sales price enhanced participants performance, opposed to vice versa temporal sequence of the prices. The bargain-judgement paradigm was used again, this time simplified by reducing the price stimuli employed in Experiment 1 to their essential numerical content.

2.2.1 Open Science Statement and Power Analysis

The preregistration of Experiment 2, as well as materials and data, are available online in the Open Science Framework at https://osf.io/64pa9. A 2 by 2 within-participant design was registered, resulting from the permutation of two factors with two factor levels each: Sequential order of prices (standard preceding target [ST] vs. target preceding standard [TS]) and Bargain (yes vs. no). The a priori power analysis with G*Power (Faul, Erdfelder, Lang, &

Buchner, 2007) for a preset power of 80% and an estimated medium to large effect size of *f* = .40 on the basis of the d_z = .90 for the effect on error rates and d_z = .42 for RT found in Experiment 1, resulted in a required sample size of *N* = 15. The study was arbitrarily overpowered with a preregistered *N* = 60.

2.2.2 Method

Subjects. 74 students participated in a 20 minutes test battery in a laboratory at the campuses of the Universities of Cologne and Würzburg. The sample consisted of 57 women and 17 men, with a mean age of 23 years.

Materials. The to be compared prices were displayed as one digit numbers without decimal places. Target prices as well as standard prices could range from 1 to 9 and were randomly paired. The number of trials was set to 100, 50 'standard preceding target'-trials and 50 'target preceding standard'-trials. Standard and target were marked by words giving temporal information about the prices' temporal occurrence in a fictional bargain comparison situation. The word *Gestern* (German for yesterday), indicating a former sales price, was presented left to the standard, the word *Heute* (German for today), indicating the actual sales price, was presented left to the target price. Figures 9 to 12 show examples for each of the four conditions resulting from orthogonal rotation of the two factors mentioned above, for readability reasons the size ratio of frame to number is modified in the figures. In the experiment, prices were printed in font size 26 (font: *Calibri*) on a 60 cm screen. The numerical expression of target and standard was randomized and neither controlled by the experimenter nor could it be anticipated by the participants. Due to the randomization of numbers, it was possible that the target and the standard had the same numerical expression. In these trials, hereafter referred to as *ties*, the correct response would be 'no bargain'.

Procedure. The procedure equaled Experiment 1, except for the target and standard stimuli appearing sequentially on the screen on the identical spatial position in the middle of the screen. Task instructions were given before the task started and were displayed until the space bar was pressed by the participant to start the task. The key assignment of the two CTRL-key was counterbalanced between participants to control for the SNARC effect in terms of an effect of response side (e.g. Dehaene et al., 1990; Dehaene et al., 1993).

A trial started with a fixation cross for 500 ms, followed by the first stimulus for 1000 ms and then another stimulus that stayed on screen until the participant had indicated their answer. RT was measured from the onset of the second stimulus to the participant's response. The next trial started immediately after the participant had responded.

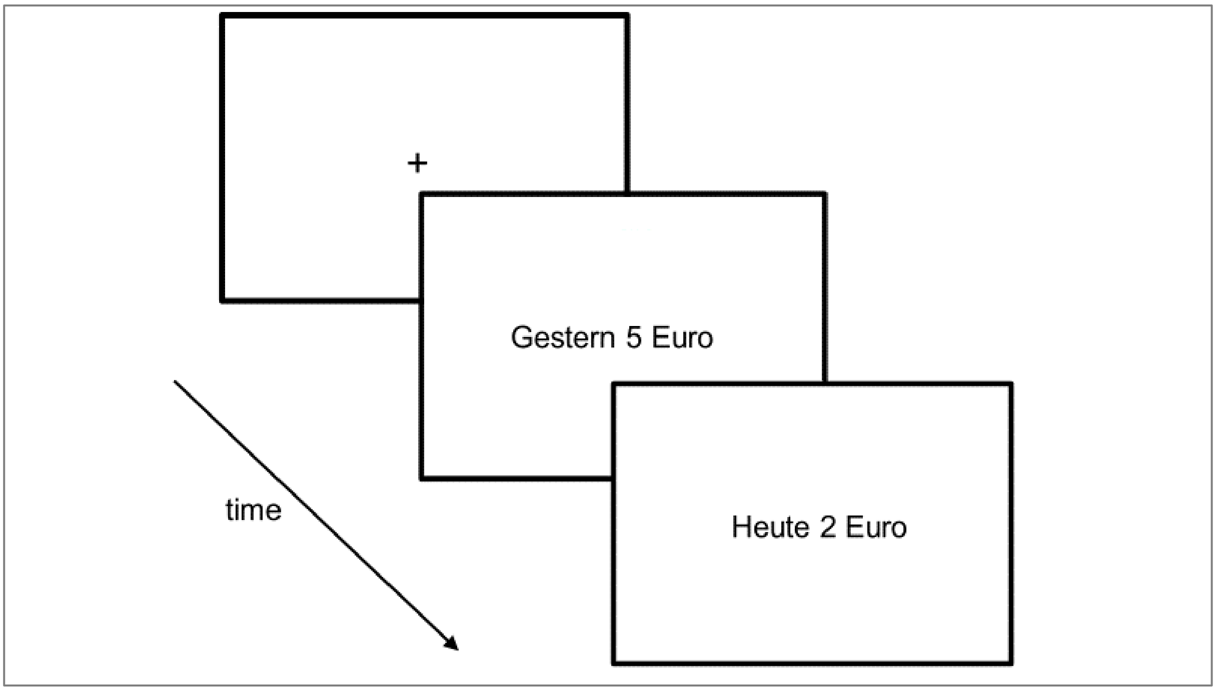

Figure 9. Example for a trial of Condition 1 in Experiment 2.
A bargain (higher previous price than actual sales price) presented in a 'standard preceding target' sequence.

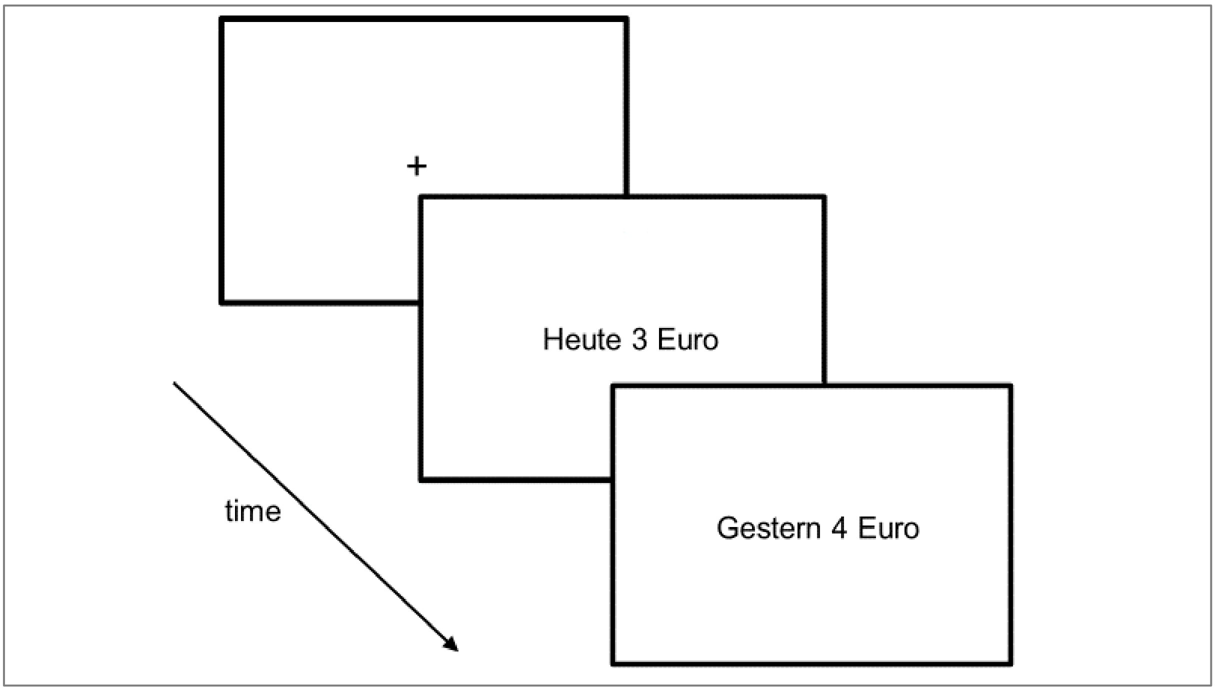

Figure 10. Example for a trial of Condition 2 in Experiment 2.
A bargain (higher previous price than actual sales price) presented in a 'target preceding standard' sequence.

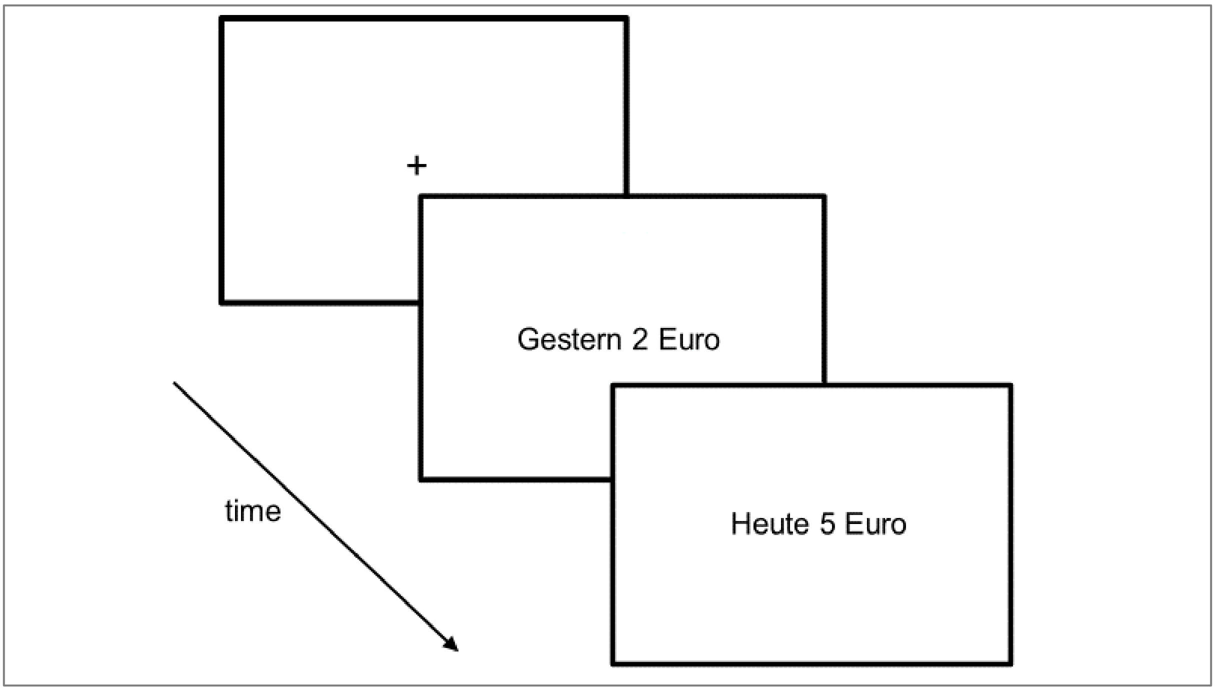

Figure 11. Example for a trial of Condition 3 in Experiment 2.
A fake bargain (higher previous price than actual sales price) presented in a 'standard preceding target' sequence.

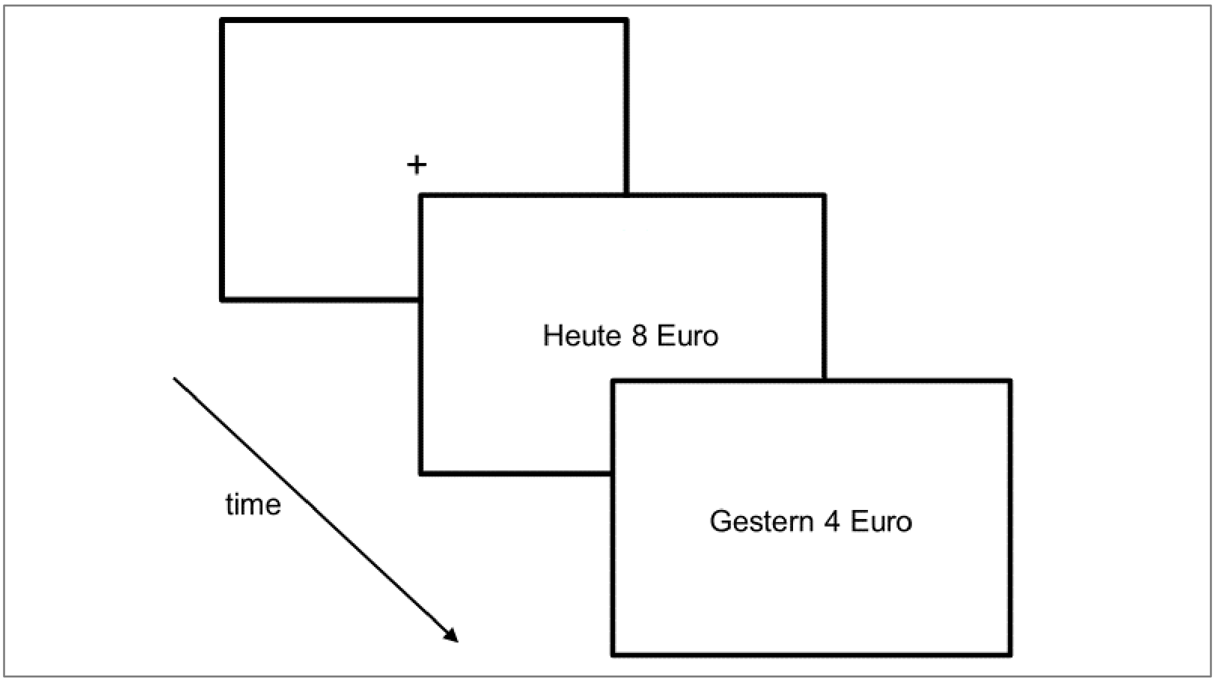

Figure 12. Example for a trial of Condition 4 in Experiment 2.
A fake bargain (higher previous price than actual sales price) presented in a 'target preceding standard' sequence.

2.2.3 Results

The data of 18 subjects fell below the inclusion criterion of 70% correct trials and therefore had to be excluded from data analyses. Seven of the deleted cases showed an error rate over 80%. RT and error rates were statistically analyzed with a remaining sample of 56 cases.

The next step of data cleaning was the deletion of trials with *response primes*. Response primes were defined as stimuli on first positions that had the expression of 1 or 9, as this would mean that the participant already could conclude on the correct response when the first stimulus appeared. One could argue that in this experiment, were ties could appear, participants could not be sure about the correct answer when encountering a response prime at the first position of a trial but as the likelihood of a tie to occur was 11 % the likelihood to with a response prime informed about the correct answer was still 89%.

Trials with an RT that deviated more than 2.5 times the standard deviation from mean RT were discarded from further analyses. This concerned 2 % of all trials over all 56 cases. The mean RT of the trials that were included in the data analysis was M = 1,217.46 ms (SD = 670.42 ms).

Error rates. Overall error rate was 8.8 %. Error rates were analyzed as preregistered in a 2 by 2 rmANOVA (Sequence x Bargain). Sequence, $F(1,55) = 40.29$, $p < .001$, $\eta_p^2 = .42$, showed a highly significant influence on the error rates with a large effect size. Bargain was significant as well and had a small effect size, $F(1,55) = 5.02$, $p = .029$, $\eta_p^2 = .08$, explaining less than 10% of the total variance. There was a highly significant interaction effect of Sequence and Bargain, $F(1,55) = 22.12$, $p < .001$, $\eta_p^2 = .29$.

To assess the influence of the ties, a 2 by 3 rmANOVA (Sequence x Bargain) with ‘tie’ as third factor level of Bargain was run. The sample size for this analysis was N = 53, as in two cases no tie had been presented. Again, Sequence showed a significant effect but this time with a slightly smaller effect size, $F(1,53) = 22.35$, $p < .001$, $\eta_p^2 = .29$. Bargain was marginally significant and compared to the 2 by 2 ANOVA reported above, showed a slightly increased effect size, $F(2,53) = 3.41$, $p = .04$, $\eta_p^2 = .11$. The interaction effect showed the largest effect size of all effects revealed by the analysis, $F(2,53) = 21.65$, $p < .001$, $\eta_p^2 = .45$. The interaction plot in Figure 13 illustrates the differences between the ST and TS trials in the ‘bargain’, ‘no bargain’ and ‘tie’ trials. For the ties the sequence effect turns: Error rates were lower ‘target preceding standard’ trials whereas the effects of the other two factor levels of Bargain were as predicted.

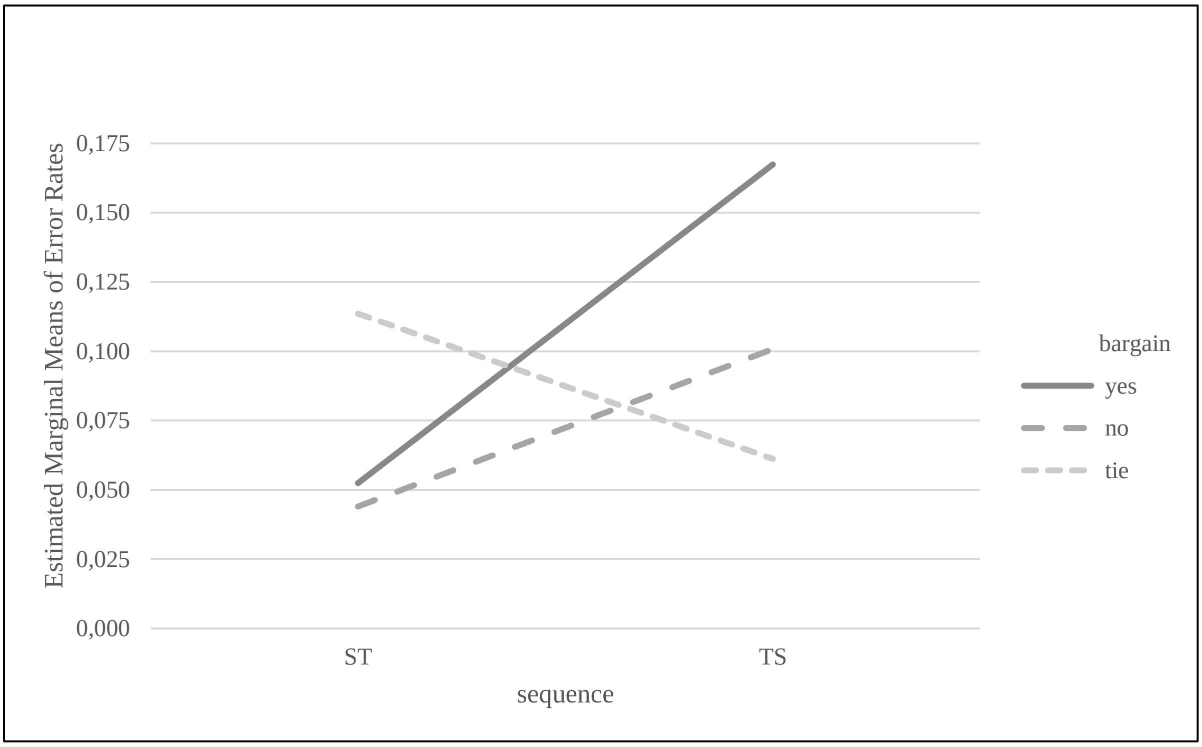

Figure 13. Interaction plot of a 2x3 rmANOVA of error rates in Experiment 2.

Excluding the ties from data analysis, the 2 by 2 rmANOVA (Sequence x Bargain) showed the same results as the 2 by 3 rmANOVA with ties, except for the effect sizes. The two main effects were higher, for Sequence, $F(1,55) = 43.58$, $p < .001$, $\eta_p^2 = .44$, for Bargain, $F(1,55) = 6.70$, $p = .012$, $\eta_p^2 = .11$, and the effect size of the interaction was smaller but still had a significant effect, $F(1,55) = 10.32$, $p = .002$, $\eta_p^2 = .16$. Simple effect analyses revealed that the interaction indicated a moderation of the differences between the bargain factor levels by sequence. The performance advantage for fake bargains, that could descriptively be found in both sequence conditions as displayed by the interaction plot in Figure 14, was only significant in the TS trials, $F(1,55) = 11.68$, $p = .001$, $r = .42$, not in the ST trials, $F(1,55) = 0.41$, $p = .527$. As the presentation format of stimuli in this condition fitted the temporal number line (smaller numbers preceding larger numbers, e.g. Müller & Schwarz, 2008), further analyses were run to test for sequential numerical order.

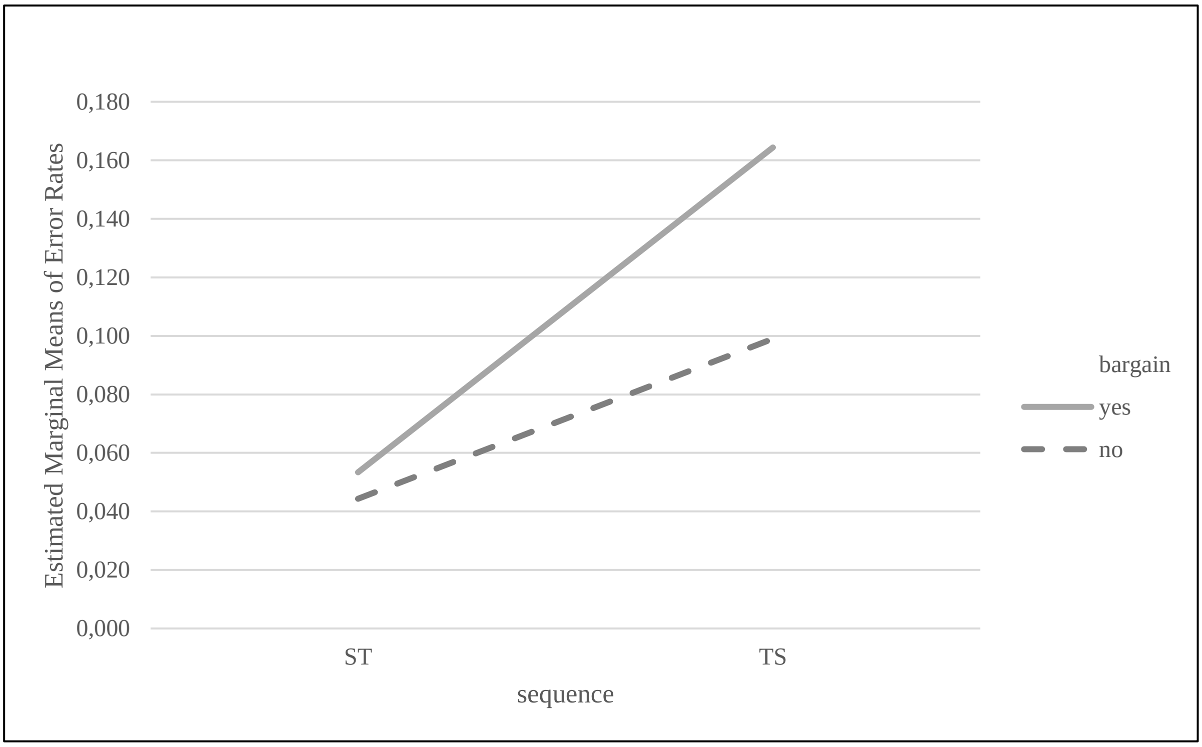

Figure 14. Interaction plot of a 2x2 rmANOVA of error rates in Experiment 2.

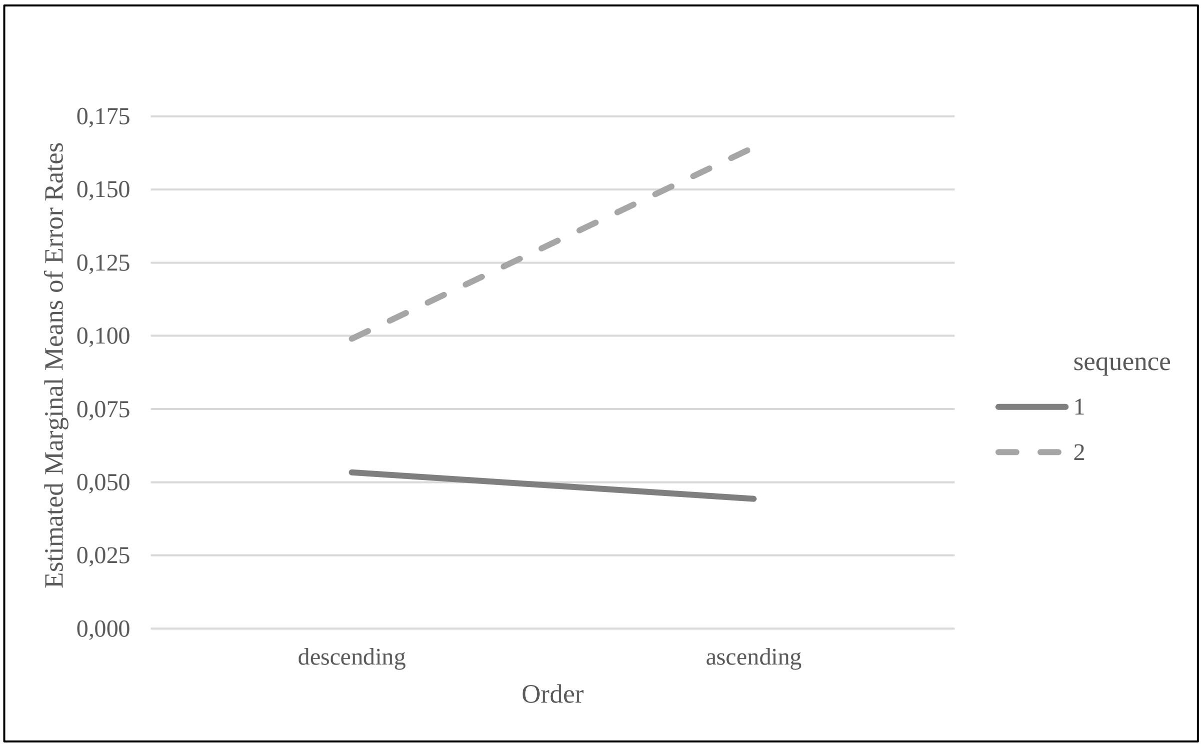

Figure 15. Interaction plot of a 2x2 rmANOVA (Sequence x Order) of error rates in Experiment 2.

2 by 2 rmANOVA (Sequence x Order) was run. The numerical order of the stimuli showed a significant medium effect, $F(1,55) = 10.32$, $p = .002$, $\eta_p^2 = .16$, with a performance advantage for descending numerical order (ascending order: $M = .11$, $SD = .08$; descending order: $M = .07$, $SD = .08$). Sequence had a significant and large effect, $F(1,55) = 43.58$, $p < .001$, $\eta_p^2 = .44$, of the predicted direction. The interaction of both factors was also significant, $F(1,55) = 6.70$, $p = .012$, $\eta_p^2 = .11$. Error rates were lowest in the condition that fitted the ascending numerical order and the sequence of standard preceding target, followed by the condition of ‘standard preceding target’ and descending numerical order. Within the TS trials the order effect turned. Sequence mediated the order effect, but only within the TS trials the order effect was significant, $F(1,55) = 11.68$, $p = .001$, $r = .42$. This was corresponding to the analysis including the bargain factor.

RT. For the analysis of RT the erroneous trials were discarded. For the parametric tests RT were transformed to their natural logarithm. The analysis of RT was performed in the same stepwise manner as the analyses of error rates. The 2 by 2 rmANOVA (Sequence x Bargain) revealed a similar effect pattern as the analysis of error rates. The sequence effect was highly significant and showed a large effect size, $F(1,55) = 267.32$, $p < .001$, $\eta_p^2 = .83$, while Bargain did not show a significant effect, $F(1,55) = 1.7$, $p = .20$. Again, the interaction of Bargain and Sequence showed a significant effect of large size, $F(1,55) = 29.52$, $p < .001$, $\eta_p^2 = .35$. As the analysis of error rates had revealed a contradicting data pattern of tie trials compared to the other bargain factor levels, a 2 by 3 rmANOVA (Sequnece x Bargain), with the additional factor level ‘tie’ for Bargain was run. The sequence effect showed a large effect size, $F(1,55) = 104.75$, $p < .001$, $\eta_p^2 = .67$, compared to a smaller but still large effect of Bargain, $F(2,55) = 12.90$, $p < .001$, $\eta_p^2 = .34$, and an equally small effect of the interaction of Sequence and Bargain, $F(2,52) = 11.54$, $p < .001$, $\eta_p^2 = .31$. Simple effect analyses revealed that, while the difference between ST and TS trials was significant in bargains, *F*(1,55) = 197.86, *p* < .001, *r* = .88, with a large effect size, and in the ‘no bargain’ condition, *F*(1,55) = 49.31, *p* < .001, *r* = .69, with a medium effect size, the difference between ST and TS was not significant in the tie-trials, *F*(1,55) = 1.602, *p* = .211.

Excluding the ties from the data analysis, the 2 by 2 rmANOVA (Sequence x Bargain) showed an increased effect size for sequence, *F*(1,55) = 246.07, *p* < .001, η^2 = .82, a little bit smaller than in the 2 by 2 rmANOVA with ties, a decreased effect size for bargain, *F*(1,55) = 11.05, *p* = .002, η_p^2 = .17, and a decreased size for the interaction effect, *F*(1,55) = 18.24, *p* < .001, η_p^2 = .25. The interaction plot of the 2 by 2 rmANOVA without ties (see Figure 16) reveals that the interaction effect of Bargain and Sequence revealed in the moderation of Bargain by Sequence. There was only a minimal difference between the condition means for ‘bargain’ and ‘no bargain’ that was not significant, *F*(1,55) = 0.03, *p* < .86. The simple effect analyses supported the observation that Bargain only showed an as predicted significant effect within the ST-trials, *F*(1,55) = 23.10, *p* < .001, *r* = .54.

To check for the influence of numerical order of the stimuli, a 2 by 2 rmANOVA (Sequence x Order) was run. The main effect for Sequence was highly significant with a very large effect size, *F*(1,55) = 246.07, *p* < .001, η_p^2 = .82. Order had a significant effect of large size as well, *F*(1,55) = 18.24, *p* < .001, η_p^2 = .25, and also the interaction showed a significant medium effect, *F*(1,55) = 11.05, *p* = .002, η_p^2= .17 . The post hoc simple effect analyses revealed that Order had a significant effect within the ST trials, *F*(1,55) = 23.10, *p* < .001, *r* = .54, while within the TS trial Order has no significant effect, *F*(1,55) = 0.03, *p* = .862.

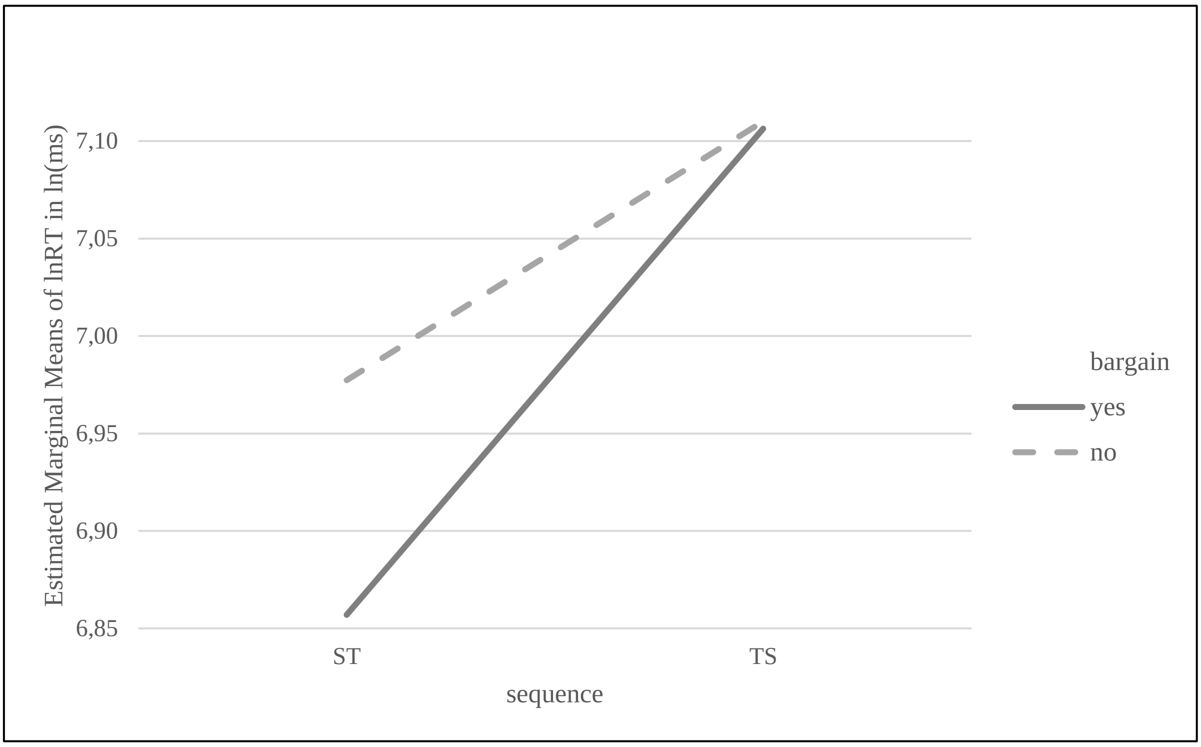

Figure 16. Interaction plot of a 2x2 rmANOVA of RT in Experiment 2.

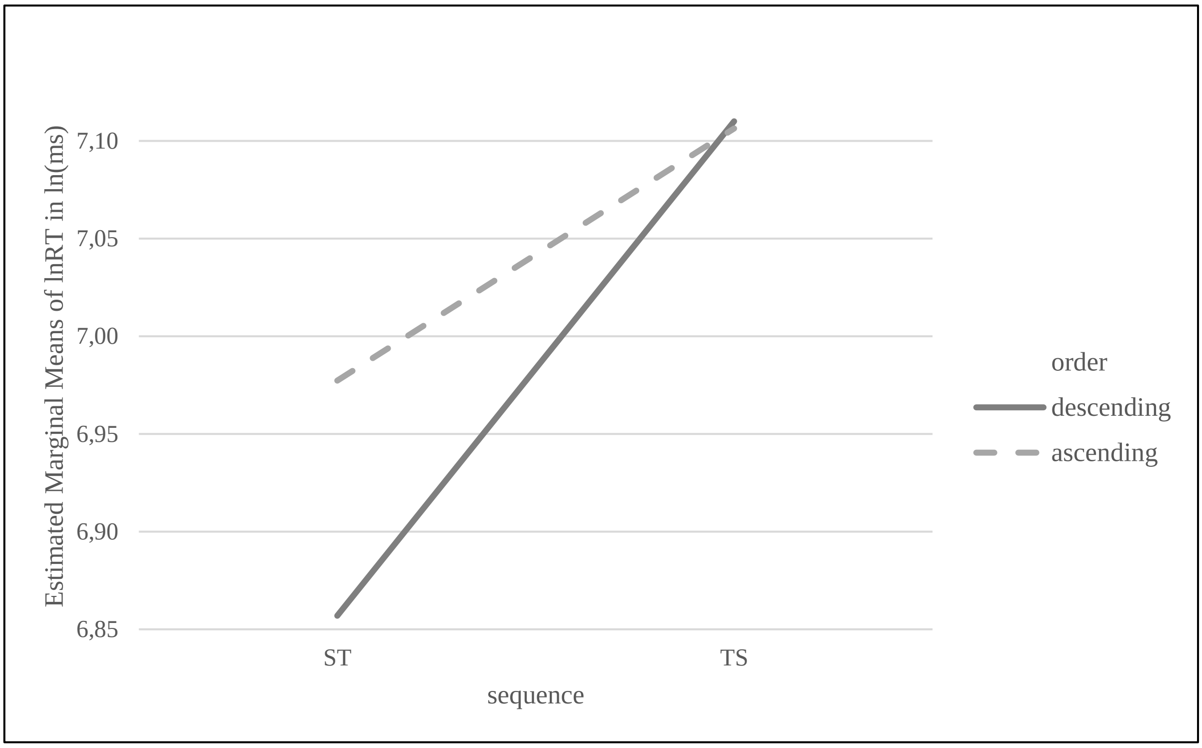

Figure 17. Interaction plot of a 2x2 rmANOVA (Sequnece x Order) of RT in Experiment 2.

2.2.4 Discussion

The results of Experiment 2 replicated the exploratory finding of Experiment 1, that the presentation order ‘standard preceding target’ has a performance advantage in terms of error rates and RT compared to ‘target preceding standard’. In this Experiment, the effect of presentation order was tested to check for the temporal number line (Müller & Schwarz (2008). Comparable to Experiment 1, the effect of presentation order of standard and target showed a higher effect size, $\eta_p^2 = .44$ for error rates and $\eta_p^2 = .83$ for RT, than the effect of numerical order of the stimuli. The interaction effect of Order and Sequence in the analysis without ties exactly equaled the bargain effect in both outcomes (for RT, $\eta_p^2 = 11$; for error rates, $\eta_p^2 = 17$), suggesting that the bargain effect could be explained completely by the interaction of the two factors Sequence and Order. The order effect had higher effect sizes than Bargain on both outcomes (for RT, $\eta_p^2 = 16$; for error rates, $\eta_p^2 = 25$) but cannot interpreted independently as the decision criterium for the participants was bargain detection.

The tie-trials had a performance advantage compared to trials in which target and standard had different numerical expressions. Two equal numbers are easy to recognize and would not have to be compared in working memory. Therefore, the reaction on two equal numbers in a trial could be performed immediately when the second number was displayed. Nevertheless, as the task was to indicate whether a pair of two subsequently presented numbers was a bargain or not, ties were unexpected and unusual in price comparisons in the real world. This explains why the RT varied depending on the sequence of target and standard also in tie trials. The RT for the ties in ST-trials were the slowest of all conditions (ST: *M* = 1002.00 ms, *SD* = 298.29 ms; TS: *M* = 1096.97 ms, *SD* = 467.29 ms) followed by bargains presented in the standard preceding target trials (*M* = 1086.97 ms, *SD* = 353.90 ms).

This could be interpreted as an indication for working memory process underlying the found *standard-target-sequence-effect* (STSE).

Before further conclusions can be drawn and research on the underlying mechanisms of the effect continues, it needs further replication and validation in diverse experimental designs.

2.3 Experiment 3: Horizontal Bargains

The spatial coding of the temporal sequence of standard and target was tested in two further experiments. In Experiment 3 the simultaneous horizontal presentation of previous and actual sales price was hypothesized to produce a spatial STSE grounded on the association of time and space (see *H_3)*. It was expected to find a performance advantage in bargain detection for previous sales price being presented in left positions and actual sales prices presented on right side positions.

2.3.1 Open Science Statement and Power Analysis

The preregistration, the materials and the data of Experiment 3 are available online in the Open Science Framework at https://osf.io/bk3ac. The study design and the corresponding power analysis equaled Experiment 2. For the effect size of the STSE found in Experiment 2, $\eta^2 = .42$ for error rates and $\eta^2 = .83$ for RT, the power analysis with G*Power (Faul, et al. 2007) was based on an effect size of f = .40 for the rmANOVA. The recommended sample size of *N* = 10 was arbitrarily overpowered and *N* = 50 was preregistered.

2.3.2 Method

Subjects. 66 students, 48 women, 17 men and one person of a diverse gender identification participated in a 20 minutes test battery in a laboratory at the campus the

University of Cologne for course credit or a compensation of 3 Euro. Their mean age was 23 years.

Materials and Procedure. Materials and procedure equaled the setup of Experiment 2, except the to be compared prices in Experiment 3. They were demonstrated simultaneously in a horizontal arrangement in the middle of the screen with no SOA. Again, the task was to compare one-digit prices and to decide whether the pair of prices indicated a bargain, namely a higher previous price than actual sales price. Prices only differed in their numerical size and in their prefix indicating their temporal occurrence, “yesterday” or “today”. Figure 18 to 21 show examples for trials of all four conditions of Experiment 3. Parallel to Experiment 2, due to the randomization of numbers, target and the standard could have the same numerical expression (ties).

Because in in Experiment 1 6% and in Experiment 2 9.5% of the participants had error rates above 80%, five training trials had to be performed by the participants after they had read the task’s instructions and before the relevant test trials began. In the training phase, participants received feedback when they had given an incorrect response in a trial. After they had performed the training trials, they were reminded of the instruction and told that in the following test phase they would not receive feedback.

Reaction time was measured from the onset of the stimuli in each trial.

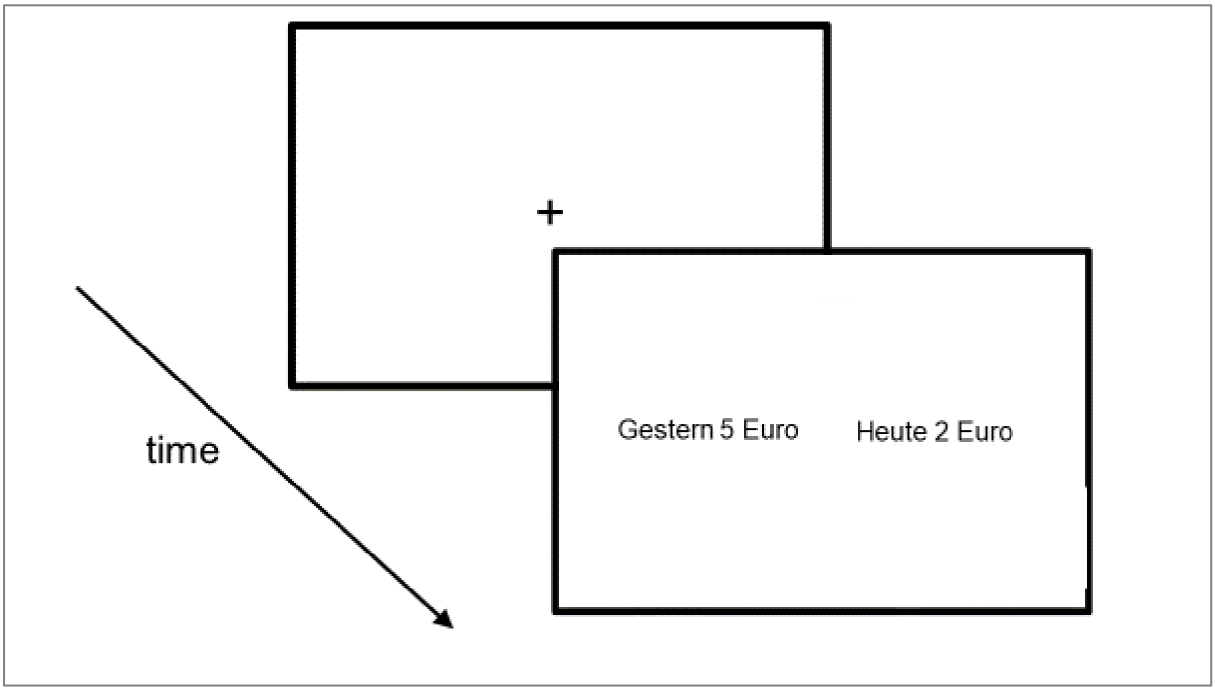

Figure 18. Example for a trial of Condition 1 in Experiment 3.
A bargain (higher previous price than actual sales price presented in a ‘standard preceding target’ sequence.

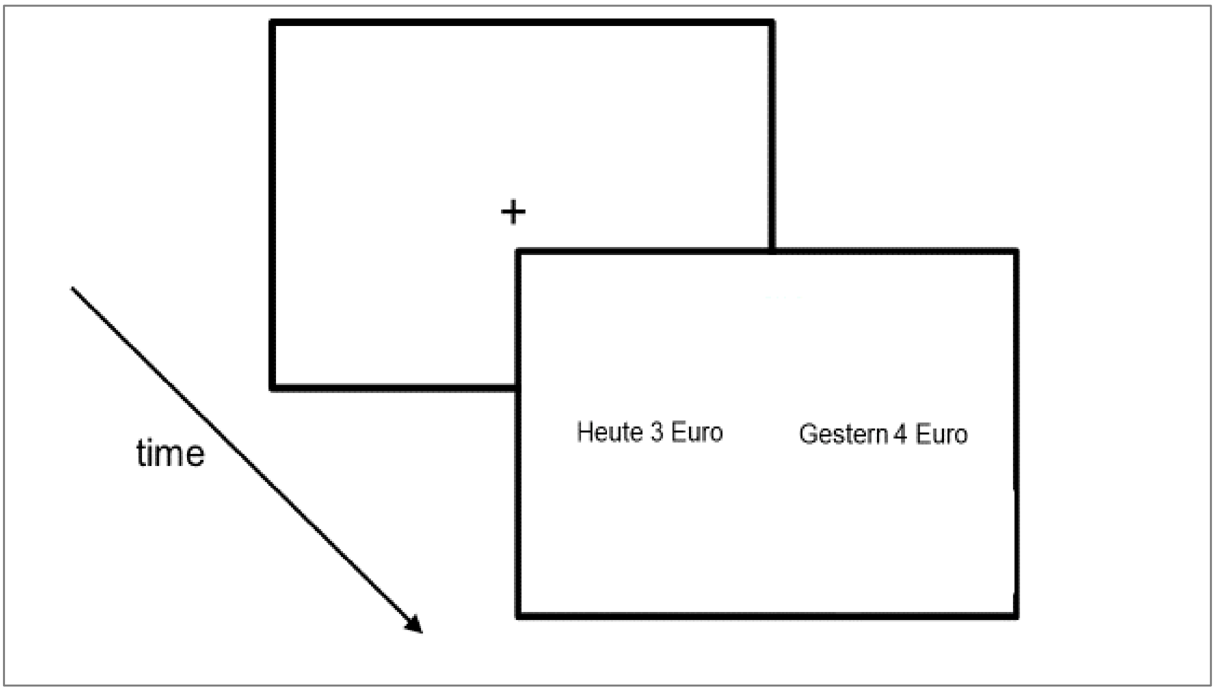

Figure 19. Example for a trial of Condition 2 in Experiment 3.
A bargain (higher previous price than actual sales price) presented in a ‘target preceding standard’ sequence.

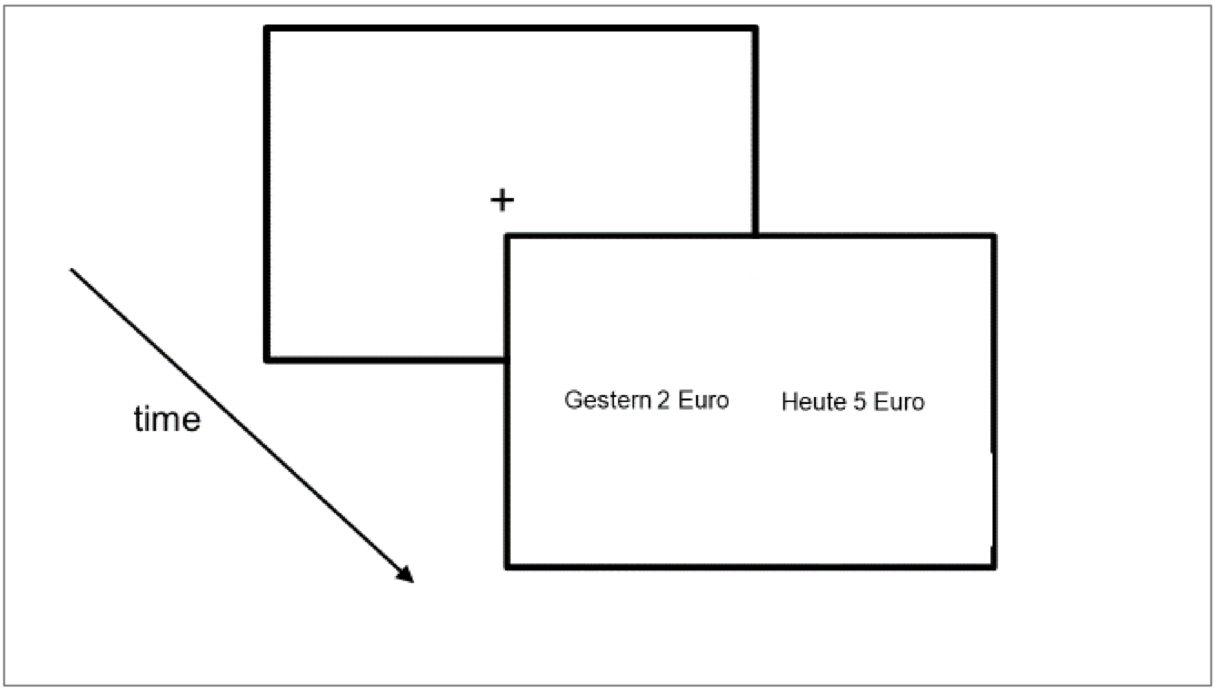

Figure 20. Example for a trial of Condition 3 in Experiment 3.
A fake bargain (higher previous price than actual sales price) presented in a ‘standard preceding target’ sequence.

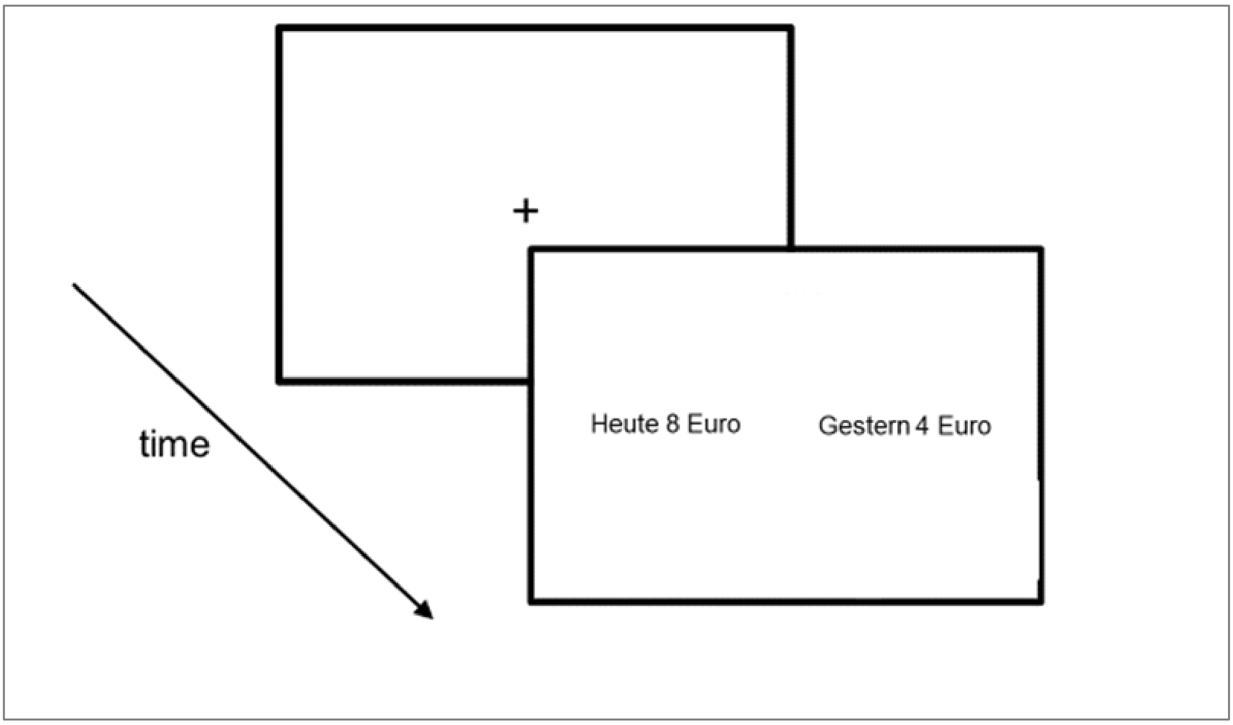

Figure 21. Example for a trial of Condition 4 in Experiment 3.
A fake bargain (higher previous price than actual sales price) presented in a ‘target preceding standard’ sequence.

2.3.3 Results

The data of 12 subjects fell below the inclusion criterion of 70% correct trials and therefore was excluded from data analyses. Reaction times and error rates were statistically analyzed with a remaining sample of 54 cases.

Trials with an RT that deviated more than 2.5 times the standard deviation from mean RT were discarded from further analyses. The mean response time of the trials that were included in the data analysis was *M* = 1,843.47 ms (*SD* = 907.95 ms).

Error rates. Overall error rate was 4.8 %. A 2 by 2 rmANOVA with Sequence (ST vs. TS) as one factor and Bargain (yes vs. no) as the second factor was run. Both factors had no significant main effect [sequence: *F*(1,54) = 0.11, *p* = .737; bargain: *F*(1,54) = 0.05, *p* = .829], the interaction of both factors was marginally significant and had a small effect size, *F*(1,54) = 3.51, *p* = .07, η_p^2 = .06. The 2 by 3 ANOVA with 'tie' as a third level of Bargain neither resulted in significant main effects [sequence: *F*(1,54) = 0.04, *p* = .843;

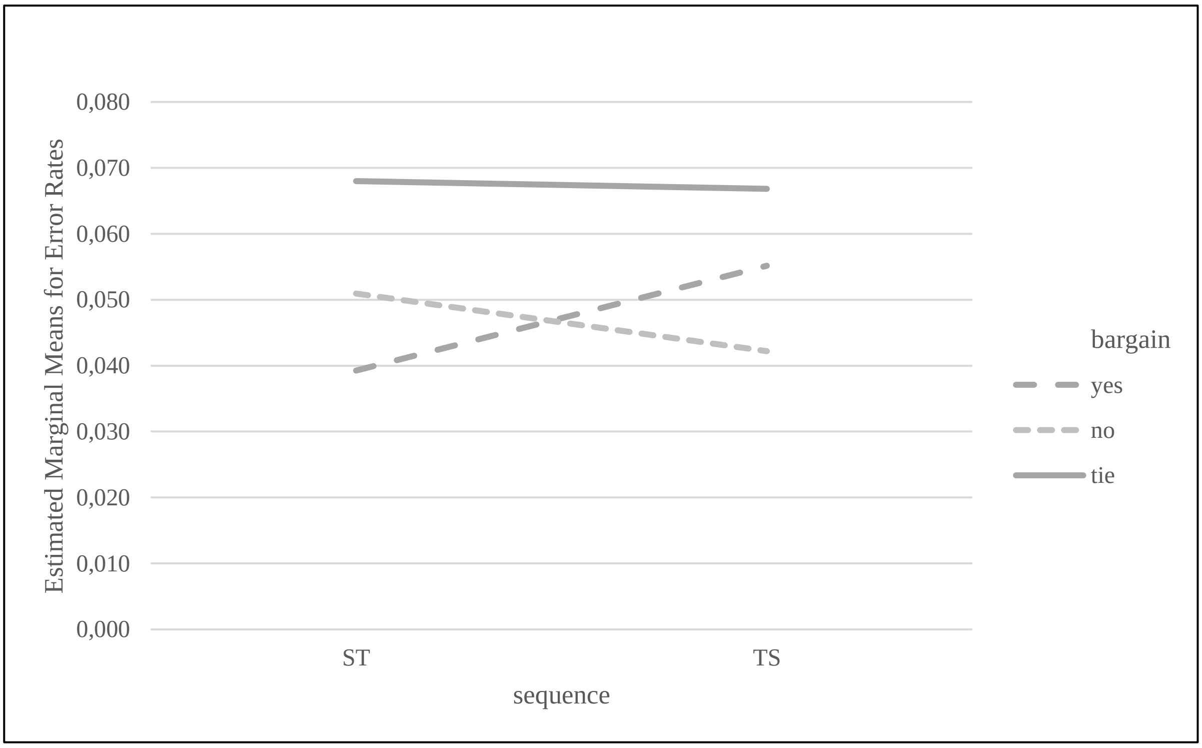

Figure 22. Interaction plot of the 2x3 rmANOVA of error rates in Experiment 3.

bargain: *F*(2,54) = 0.57, *p* = .567] nor a significant interaction of Sequence and Bargain, *F*(2,54) = 1.58, *p* = .215. The interaction plot in Figure 22 illustrate that the error rate of ties-trials was higher than on the other factor levels of Bargain. Again, a 2 by 2 rmANOVA (Sequnece x Bargain) was run excluding the tie trials. This analysis revealed the same results as the 2 by 2 rmANOVA for the data set with ties. Neither the single factors [sequence: *F*(1,54) = 0.13, *p* = .722; bargain: *F*(1,54) = 0.01, *p* = .920] nor the interaction of Bargain and Sequence, *F*(1,54) = 3.23, *p* = .078, showed significant effects on the error rates.

Another 2 by 2 rmANOVA (Sequence x Order) was run to test for the effect of numerical order of stimuli, revealing neither significant main effects, for Sequence, *F*(1,54) = 0.13, *p* = .722, for Order, *F*(1,54) = 3.23, *p* = .078, nor a significant interaction effect, *F*(1,54) = 0.01, *p* = .920. Descriptively the trials with descending numerical order from left to right were performed more often correct than trials with ascending numerical order.

RT. For the analysis of RT the erroneous trials were discarded. For the parametric tests RT were transformed to their natural logarithm. A 2 by 2 rmANOVA (Sequence x Bargain) was run. Sequence revealed to have a significant main effect, *F*(1,54) = 17.67, *p* < .001, η_p^2 = .25, as well as Bargain, *F*(1,54) = 8.13, *p* = .006, η_p^2 = .13, the interaction of both factors was not significant, *F*(1,54) = 0.69, *p* = .411. The 2 by 3 rmANOVA (Sequence x Bargain) with ‘ties’ as a third factor level of Bargain revealed highly significant main effects for Sequence, *F*(1,54) = 15.04, *p* < .001, η_p^2 = .22, with a large effect and for Bargain, *F*(2,54) = 78.55, *p* < .001, η_p^2 = .75 , with a large effect size as well. The interaction of Sequence and Bargain showed no significant effect, *F*(2,54) = 1.83, *p* = .17.

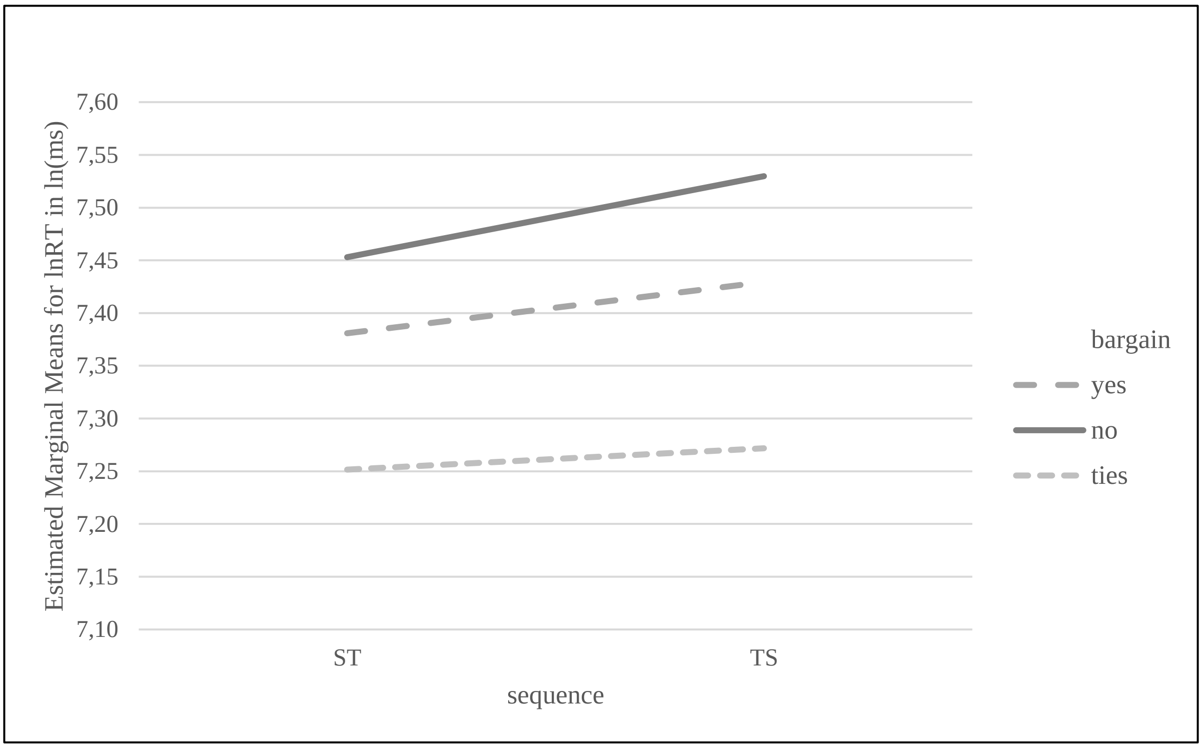

Figure 23. Interaction plot of a 2x3 rmANOVA of RT in Experiment 3.

Simple effects analyses revealed significant effects for Sequence on the first two factor levels of Bargain but not on the tie level, *F*(1,54) = 0.84, *p* = .363. To avoid an underestimation of the effect size of the STSE, a 2 by 2 rmANOVA (Sequence x Bargain) was run excluding the tie trials from the analysis. In this analysis Bargain had a large effect on the RT of participants, *F*(1,54) = 47.29, *p* < .001, η_p^2 = .47. Sequence had a highly significant main effect as well, *F*(1,54) = 19.07, *p* < .001, η_p^2 = .26. The interaction effect was not significant, *F*(1,54) = 2.10, *p* = .154. Trials in that the presentation fitted the STSE effect and the pair of prices was a bargain were performed the fastest of all conditions (*M* = 1,779.83 ms, *SD* = 571.85 ms). A 2 by 2 rmANOVA (Sequence x Order) should check for an effect of numerical order of the to be compared stimuli in a trial. The STSE was highly significant and had a large effect size, *F*(1,54) = 19.07, *p* < .001, η_p^2 = .26, Order had no significant effect, *F*(1,54) = 2.10, *p* = .154, the interaction of both factors was significant and had a large effect size, *F*(1,54) = 47.29, *p* < .001, η_p^2 = .47. The post hoc simple effects analyses revealed that the STSE only was significant in the trials of descending order, $F(1,54) = 58.18$, $p < .001$, $\eta_p^2 = .52$, but not in the trials of ascending order, $F(1,54) = 1.61$, $p < .210$. This aligned with the bargain effect reported previously.

2.3.4 Discussion

In Experiment 3, the STSE found in Experiment 1 and 2 could be only replicated in the analysis of RT. Mean error rate and RT over all conditions were lower than in Experiment 2. This goes hand in hand with no effects being found in the analysis of error rates. As both to be compared stimuli of a trial were visible during response, participants could check their judgment without having to remember one of the stimuli.

An assumption for the underlying mechanism of the STSE in RT is that longer RT for the TS trials compared to the ST trials indicated a process of attention allocation according to temporal or spatial position. When the to be compared stimuli were presented fitting the reading direction of standard preceding target (STSE), the response could be initiated immediately when the target was attended (or appeared as in Experiment 2). In TS trials the participant's attention had to go back to the target after the standard at the later attended spatial position, to make a correct magnitude judgment. To affirm this assumed mechanism, more empirical evidence for the STSE in divers experimental set ups and diverse stimuli is needed.

The experimental set up lined out in section 2.3.2 was similar to the setup of Turconi et al. (2006), who found an ascending order advantage for paired number comparisons (selection task) with horizontal stimulus arrangement. Due to the additional factor Bargain in the setup of the experiment presented here, and the according task to detect bargains, an ascending order advantage might have been suppressed. In the error rates even a descending order advantage was found. As the definition of a bargain per se was a temporally descending price, this might explain why for error rates, descriptively a general performance advantage for descending stimuli pairs from left to right was found.

2.4 Experiment 4: Vertical Bargains

In Experiment 4 the same hypothesis was tested as in Experiment 3, this time the simultaneous vertical presentation of previous and actual sales price was hypothesized to produce a spatial STSE grounded on the association of time and space (see *H_3)*. It was expected to find a performance advantage in bargain detection in terms of RT and error rates for previous price standing above actual sales price, as it had been reported in the results of Experiment 1 in for a more noisy price presentation format where standards had been intensely marked.

2.4.1 Open Science Statement and Power Analysis

The preregistration, the materials and the data of Experiment 4 are available online in the Open Science Framework at https://osf.io/eqp5b. The study design and the corresponding power analysis equaled Experiment 3. The sample size for Experiment 4 was arbitrarily overpowered and *N* = 50 was preregistered.

2.4.2 Method

Subjects. 55 students, 38 women, 8 men and 3 of diverse gender identification, participated in a 20 minutes test battery in laboratories at the campuses of the universities of Cologne and Würzburg. Demographic data of four participants were not collected due to a technical error. Participants received a course credit or 3 Euro for compensation. Their mean age was 22 years.

Materials and Procedure. Materials and procedure equaled the setup of Experiment 3 except that the stimuli described in section 2.3.2 were demonstrated simultaneously in a vertical arrangement in the middle of the screen with no SOA. Before the relevant test trials started, there were five training trials, in which the participants received feedback for false responses. Reaction time was measured from the onset of the stimuli in each trial.

2.4.3 Results

The data of 2 subjects fell below the inclusion criterion of 70% correct trials and therefore were excluded from data analyses. RT and error rates were statistically analyzed with the remaining sample of 53 cases.

Trials with an RT that deviated more than 2.5 times the standard deviation from the mean were discarded from further analyses. The mean RT of the trials that were included in the data analysis was *M* = 1,580.96 ms (*SD* = 756.04 ms).

Error rates. 4.5% of all trials were erroneous. The 2 by 2 ANOVA with the factors Sequence (ST vs. TS) and Bargain (yes vs. no) revealed no significant effects [Sequence: *F*(1,52) = 0.76, *p* = .388; Bargain: *F*(1,52) = 1.24, *p* = .270; interaction: *F*(1,52) = 0.29, *p* = .592]. Descriptively the ST trials had a lower error rate *M* = .042 (*SD* =.045) than the TS trials, *M* = .048 (*SD* =.060) and the 'bargain' trials had a lower error rate, *M* = .043 (*SD* =.20), than the 'no bargain' trials, *M* = .046 (*SD* =.21). The lowest mean error rate over all conditions could be found in the ST-bargain-condition, *M* = .042 (*SD* =.056).

In a 2 by 3 rmANOVA (Sequence x Bargain) with ties as the third factor level of Bargain, no significant effects were found. However, the error rate in the tie trials was higher over all Sequence factor levels, *M* = 0.087, compared to the other factor levels of Bargain, *M* = 0.039. Excluding the ties from the data analysis, the 2 by 2 rmANOVA (Sequence x Bargain) revealed no significant effects [Sequence: *F*(1,52) = 0.73, *p* = .400; Bargain: *F*(1,52) = 2.79, *p* = .101; interaction: *F*(1,52) = .17, *p* = .683].

To check for the effect of numerical order, a 2 by 2 ANOVA (Sequence x Order) was run. None of the main effects [Sequence: $F(1,52) = 0.73$, $p = .396$; Order: $F(1,52) = 0.17$, $p = .683$] and neither the interaction of both factors, $F(1,52) = 2.79$, $p = .101$, was significant.

RT. For the analysis of RT the erroneous trials were discarded**.** For the parametric tests RT were transformed to their natural logarithm. The same stepwise analyses as for error rates were run. The 2 by 2 rmANOVA (Sequence x Bargain) revealed no significant main effect of Sequence, $F(1,52) = 0.34$, $p = .559$, but a highly significant effect of Bargain, $F(1,52) = 43.33$, $p < .001$, $\eta_p^2 = .46$. The interaction of both factors was not significant, $F(1,52) = 3.32$, $p = .07$. The 2 by 3 rmANOVA (Sequence x Bargain) with ties as a third factor level of the factor Bargain revealed the same results. While the sequence effect was not significant, $F(1,52) = 1.79$, $p = .187$, Bargain showed a highly significant effect of a very large size $F(2,52) = 59.53$, $p < .001$, $\eta_p^2 = .54$. The interaction of Sequence and Bargain showed no significant effect, $F(2,54) = 0.55$, $p = .579$. The mean RT of the tie trials was lower over all conditions, $M = 7.13$ ln(ms), than compared to the other factor levels of Bargain, $M = 7.29$. The 2 by 2 rmANOVA (Sequence x Bargain) without ties revealed a highly significant bargain effect, $F(1,52) = 57.28$, $p < .001$, $\eta_p^2 = .524$, no significant sequence effect, $F(1,52) = 0.50$, $p = .485$, and no significant interaction effect, $F(1,52) = 0.01$, $p = .924$.

A 2 by 2 rmANOVA (Sequence x Order) revealed neither an STSE, $F(1,54) = .50$, $p = .485$, nor an effect of order, $F(1,54) = 0.01$, $p = .924$. The interaction of both effects was significant and had a large effect size, $F(1,54) = 57.28$, $p < .001$, $\eta_p^2 = .52$. The interaction of the factors Sequence and Order was accordant with the Bargain factor, as bargains or ‘fake bargains’ resulted from a combination of the sequences of target and standard and an ascending or descending numerical order.

2.4.4 Discussion

In a vertical arrangement of the bargain-judgement-paradigm, already employed in Experiment 2 and 3 to measure the STSE, the hypothesized effect did not show up in participants' performance, neither in RT nor in error rates. Neither was there an effect of numerical order.

Experiment 4 was a spin-off of Experiment 1 with noise-reduced stimuli. In Experiment 1, where standards had been intensely marked by two times 9 as decimal places and a red cross, an STSE had been found probably mediated by the markedness of the standard. The increased salience of the standard might have attracted participants' attention serving the standard's being the starting point of the comparison. In Experiment 4, there was no such markedness – target and standard were presented simultaneously and equally designed.

In the horizontal arrangement of stimuli in Experiment 3 the STSE showed up. In this case, the STSE might have been a language dependent effect in the German sample. Western languages have a reading direction von left to right, as to say an association of a proceeding direction from left to right (Guida, Megreya, Lavielle-Guida, Noël, Mathy, van Dijck & Abrahamse, 2018). Context- and culture- dependency had been reported for the SNARC effect in vertical arrangements as well (e.g. Ito & Hatta, 2004). A vertical SNARC effect in number comparison tasks, namely an enhanced performance when larger targets had to be responded to with an upper response key and smaller targets with a spatially lower response key, had only been reported in a setting where vertically presented pairs of numbers where framed as temperatures with a thermometer presented next to the stimuli (Petrusic et al. 2011), in an Asian sample (Ito & Hatta, 2004) and when the standard was intensely marked by earlier occurrence, different font and steady upper position (Ben Nathan et al. 2009).

2.5 Experiment 5: Prospective Bargains

To explore whether the found STSE was an effect of temporal occurrence of previous and actual sales price, comparable to a processing advantages for temporally ascending sequences (compare Schroeder et al. 2017), an experiment with prospective bargains was conducted. In this setup, the standard prices laid in the future and the sales prices in the presence. If the STSE was an effect of ascending temporal order of events, detection of bargains should be facilitated by the sequence of targets (actual sales prices) being presented before standards (future sales prices). In Experiment 5 H_4 (see section 4) was tested.

2.5.1 Open Science Statement and Power Analysis

The preregistration, the materials and the data of Experiment 5 are available online in the Open Science Framework at https://osf.io/v5grp. The study design and the corresponding power analysis equaled Experiments 3 and 4. The sample size for Experiment 5 was arbitrarily overpowered and N = 50 was preregistered.

2.5.2 Method

Subjects. 55 students, 42 women, 12 men and one of diverse gender identification, participated in a 20 minutes test battery in laboratories at the campuses of the universities of Cologne and Würzburg. Participants received a course credit or 3 Euro for compensation. Their mean age was 23 years.

Materials and Procedure. Materials and procedure equaled the setup of Experiment 2 except that the standard stimuli were not marked by the German word "Gestern" (English:yesterday) on their left side but with "Morgen*"* (English: tomorrow). A trial started with a fixation cross for 500 ms, followed by the first stimulus for 1000 ms and then the second stimulus that stayed on the screen until the participant had indicated their answer. RT was measured from the onset of the second stimulus.

2.5.3 Results

The data of 8 subjects fell below the inclusion criterion of 70% correct trials and therefore were excluded from data analyses. Reaction times and error rates were statistically analyzed with a remaining sample of 47 cases.

The next step of data cleaning was the deletion of trials with response primes, trials with 1 or 9 on the first position of a trial.

Trials with an RT that deviated more than 2.5 times the standard deviation from mean RT were discarded from further analyses. The mean response time of the trials that were included in data analyses was M = 1,132.46 ms (SD = 554.08 ms).

Error rates. Overall error rate was 5.8 %. A 2 by 2 rmANOVA (Sequence x Bargain) was run. Both factors had no significant main effect [Sequence: $F(1,46) = 2.69$, $p = .108$; Bargain: $F(1,46) = 0.62$, $p = .437$], but the interaction of both factors was significant, $F(1,46) = 14.29$, $p < .001$, and had a large effect size of $\eta_p^2 = .24$. The interaction plot in Figure 24 illustrates that in the 'bargain' conditions the 'standard preceding target' presentation order showed lower error rates on average than the 'target preceding standard' trials. In the 'no bargain' conditions the STSE turned. Here, the 'target preceding standard' trials showed less errors on average.

Checking for the influence of ties on these effects a 2 by 3 rmANOVA (Sequence x Bargain) with ties as third level of Bargain was run. This analysis revealed the same data pattern as the 2 by 2 rmANOVA (Sequence x Bargain) described above [Sequence: $F(1,46) = 0.01$, $p = .930$; Bargain: $F(2,46) = 1.08$, $p = .349$; interaction: $F(2,46) = 7.76$, $p = .001$, $\eta_p^2 = .28$]. The interaction plot in Figure 25 reveals that the performance advantage for ST was descriptively present in the bargain trials but turned into a performance advantage for the TS sequence in the tie trials. The 2 by 2 rmANOVA (Sequence x Bargain) in a data set

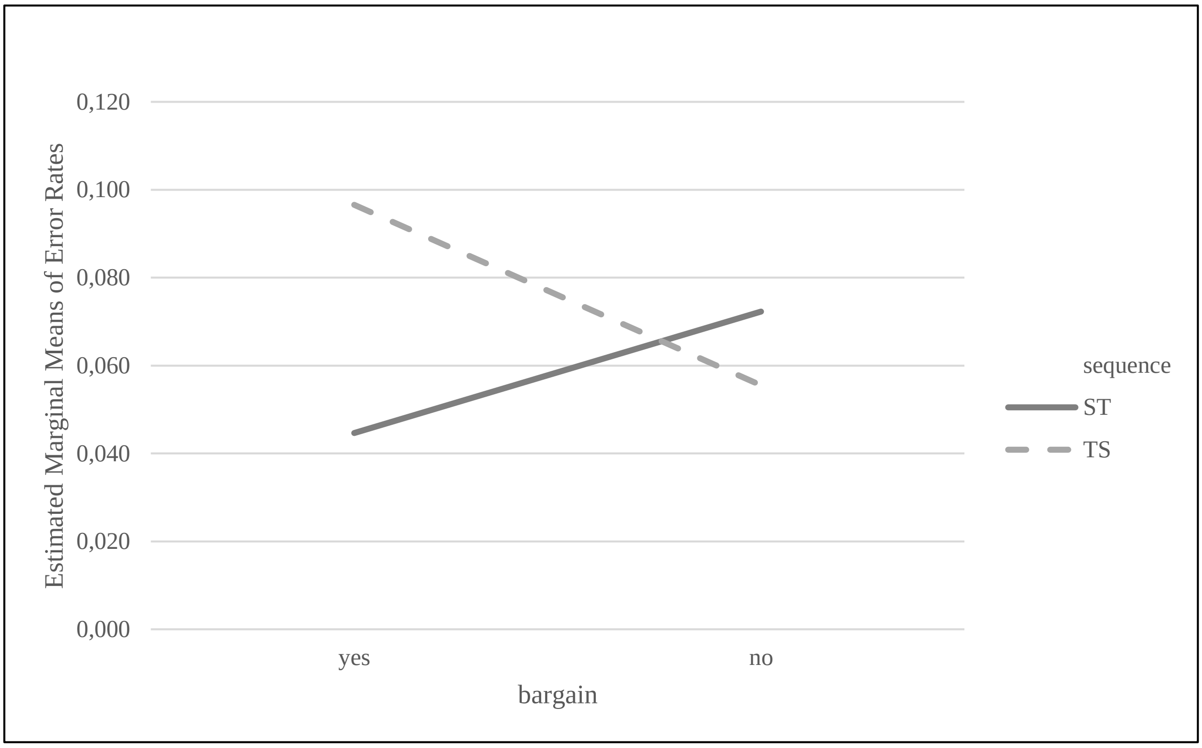

Figure 24. Interaction plot of a 2x2 rmANOVA of error rates in Experiment 2. Ties are included in the no bargain factor level.

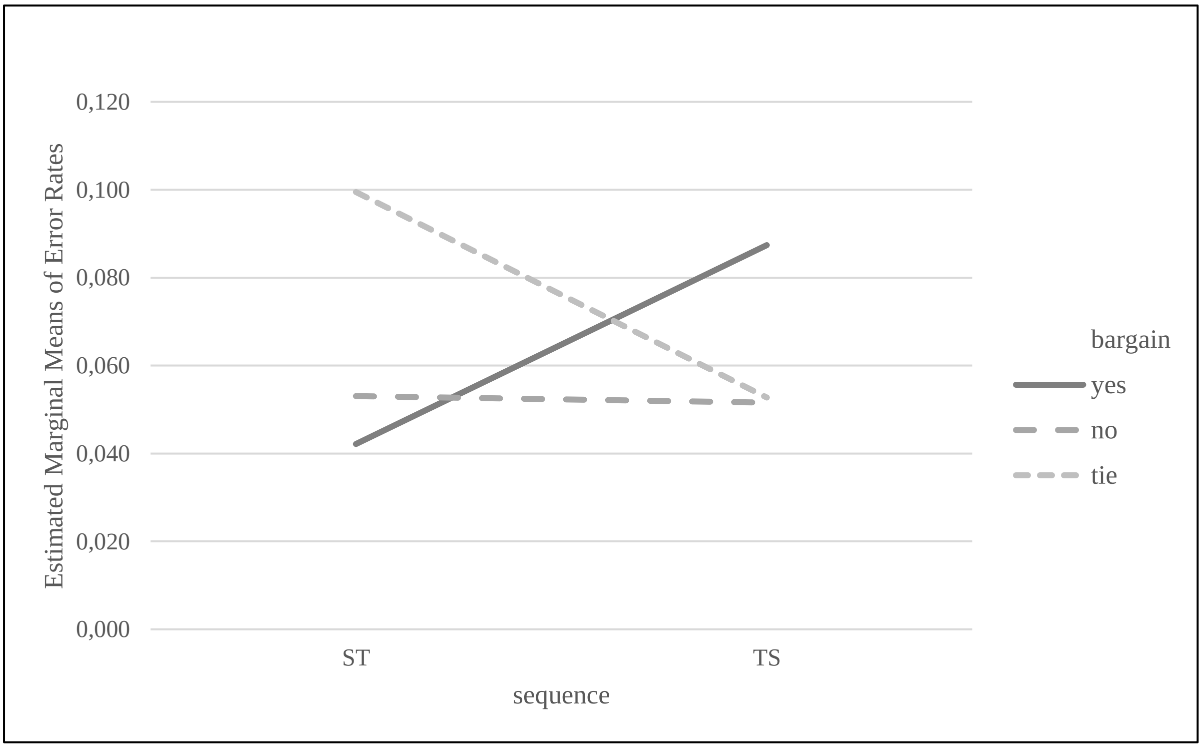

Figure 25. Interaction plot of a 2x3 rmANOVA of error rates in Experiment 5.

from that the ties had been excluded, Sequence had no significant effect, *F*(1,46) = 2.96, *p* = .092, η_p^2 = .06, the bargain effect neither, *F*(1,46) = 0.34, *p* = .565, the interaction effect was highly significant and had a medium size, *F*(1,46) = 13.31, *p* = .001, η_p^2 = .22. The interaction plot in Figure 26 illustrates that the interaction was caused by the mediating role of bargain for the sequence effect. In the bargain trials STSE was present and significant, *F*(1,46) = 8.98, *p* = .004, *r* = .40, while in the ‘no bargain’ trials there was an advantage for the TS trials, but the difference of means was not significant, *F*(1,46) = 0.84, *p* = .365.

To check for the influence of numerical order on the found data pattern, a 2 by 2 rmANOVA (Sequence x Order) was run. Sequence was just not significant, *F*(1,46) = 2.96, *p* = .092, η_p^2 = .06, comparable to the 2 by 2 rmANOVA (Sequence x Bargain). Order was highly significant, *F*(1,46) = 13.31, *p* = .001, η_p^2 = .22, meeting exactly the expressions of parameters of the interaction of Bargain and Sequence. Trials with numerically descending stimuli were performed more often correctly. The interaction of Sequence and Order was not

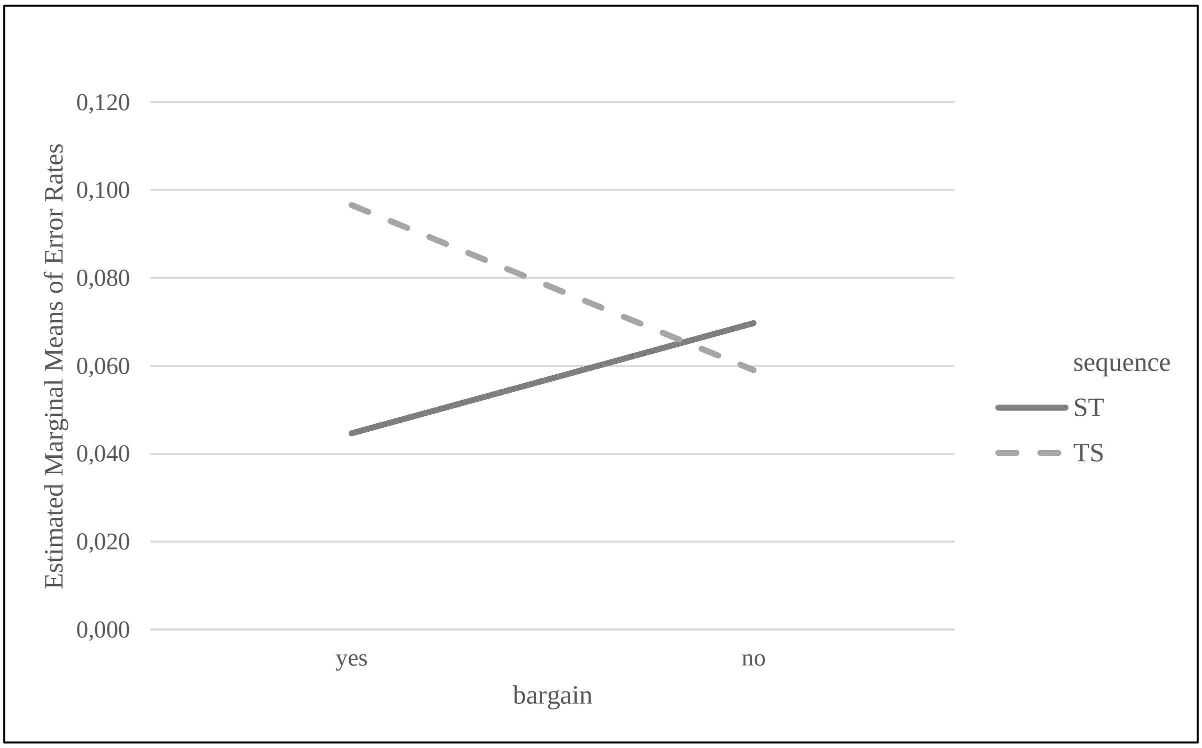

Figure 26. Interaction plot of a 2x2 rmANOVA of error rates without ties in Experiment 5.

significant, $F(1,46) = 0.34$, $p = .565$, meeting exactly the expressions of the parameters of the bargain effect in the 2 by 2 rmANOVA (Sequence x Bargain).

RT. For the analysis of RT the erroneous trials were discarded. For the parametric tests RT were transformed to their natural logarithm. The analysis of RT was performed in the same stepwise manner as the analyses of error rates. The 2 by 2 rmANOVA (Sequence x Bargain) revealed a highly significant effect for Sequence, $F(1,46) = 13.39$, $p = .001$, $\eta_p^2 = .22$, a highly significant effect for Bargain, $F(1,46) = 26.41$, $p < .001$, $\eta_p^2 = .37$, a no significant effect of the interaction of both factors, $F(1,46) = 2.43$, $p = .126$, $\eta_p^2 = .05$. The sequence effect and the bargain effect were in the predicted direction, as displayed in the interaction plot in Figure 27.

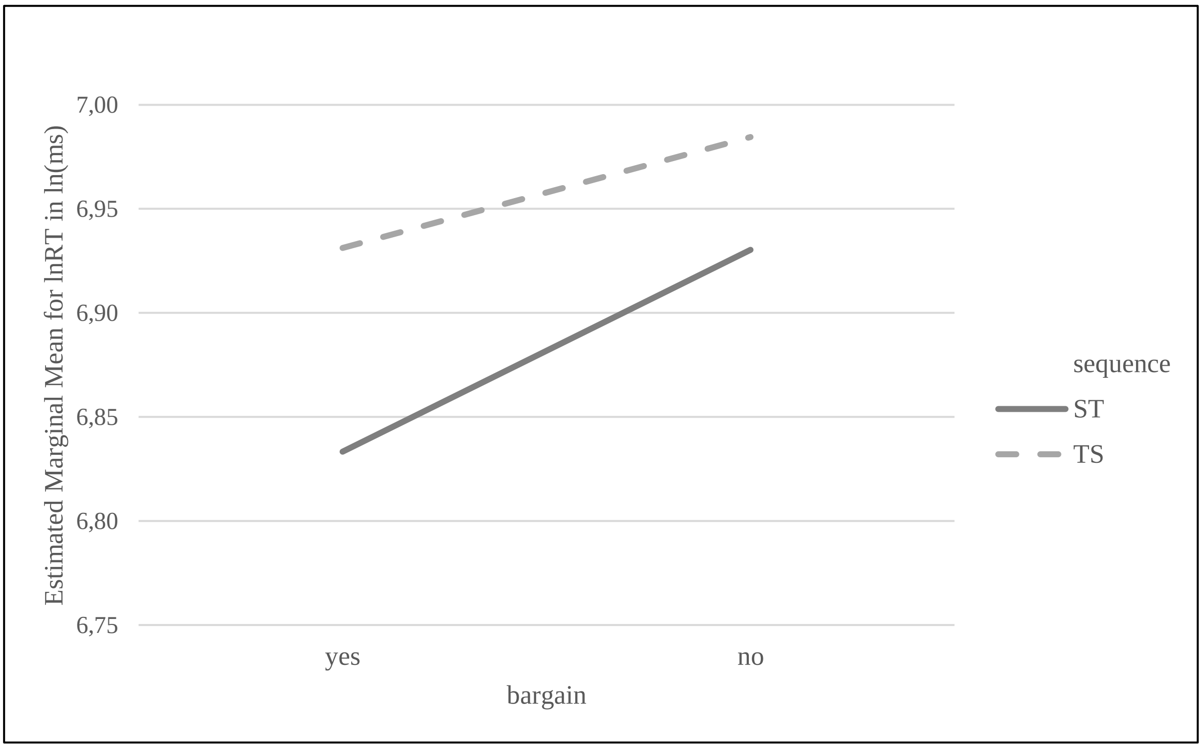

Figure 27. Interaction plot of a 2x2 rmANOVA of RT with ties in Experiment 5.

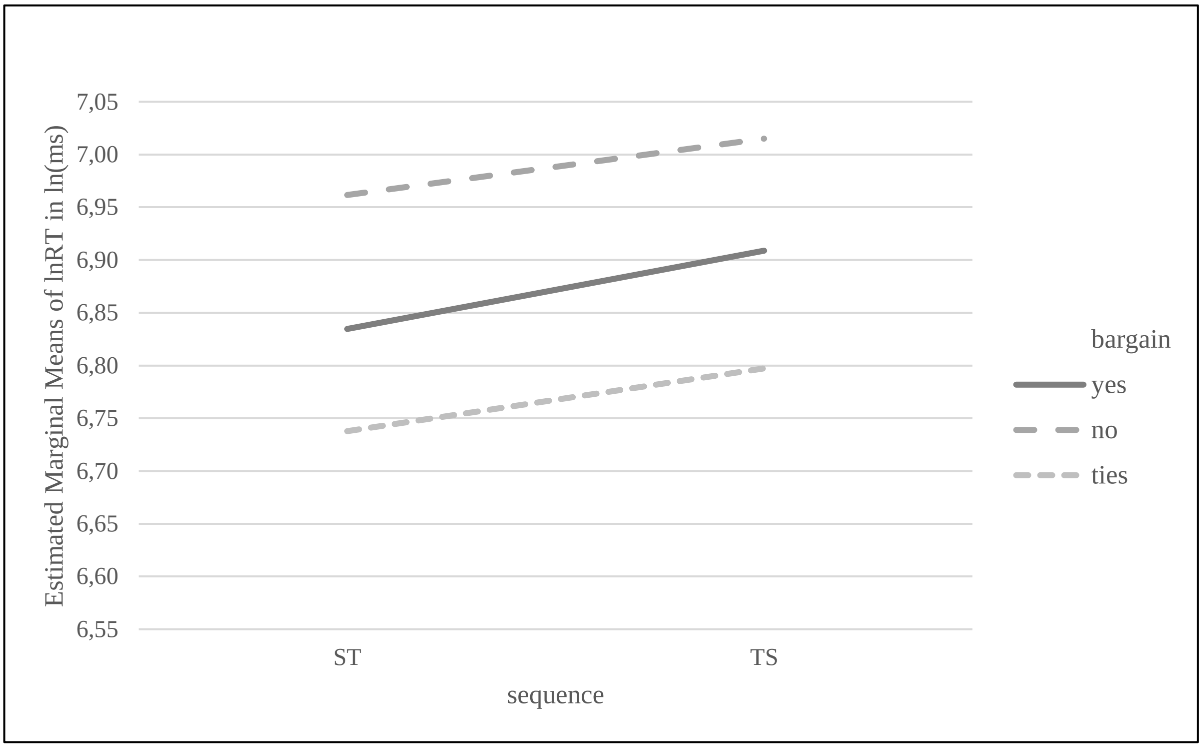

Figure 28. Interaction plot of a 2x3 rmANOVA of RT in Experiment 5.

To check for the influence of the ties a 2 by 3 rmANOVA (Sequence x Bargain) with 'ties' as third factor level of Bargain was run. Results equaled the 2 by 2 rmANOVA (Sequence x Bargain) with ties, except for a slightly increased effect size for the STSE, $F(1,46) = 13.29$, $p = .001$, $\eta_p^2 = .25$, and a highly significant effect of bargain, $F(2,46) = 58.61$, $p < .001$, $\eta_p^2 = .76$, as the ties had a much lower average RT than the other factor levels of Bargain. The differences of the bargain factor levels are displayed in the interaction plot in Figure 28.

The 2 by 2 rmANOVA (Sequence x Bargain) without ties revealed exactly the same result patterns as the analysis including ties, except for an increased effect size for the baragin effect [Sequence: $F(1,46) = 12.82$, $p = .001$, $\eta_p^2 = .22$; Bargain: $F(1,46) = 64.87$, $p < .001$, $\eta_p^2 = .56$; interaction: $F(1,46) = 1.56$, $p = .218$].

To check for the influence of numerical order on the found effects, a 2 by 2 rmANOVA (Sequence x Order) was run. The STSE found in the 2 by 2 rmANOVA (Sequence x Bargain) showed up with exactly the same parameters, $F(1,46) = 12.82$, $p = .001$, $\eta_p^2 = .22$, the effect of the interaction of Sequence and Order equaled the bargain effect found in the 2 by 2 rmANOVA (Sequence x Bargain), $F(1,46) = 64.87$, $p < .001$, $\eta_p^2 = .56$, and there was no main effect of order, $F(1,46) = 1.56$, $p = .218$.

2.5.4 Discussion

Error rates and RT revealed divergent result patterns in Experiment 5. For RT the same data patterns were reported as in previous experiments. The STSE and the Bargain effect were found as predicted, however, there was no effect of numerical order on the RT.

For error rates a local STSE effect in the bargain trials showed up, while no effect of sequence was found in the ‘no bargain’ trials. The effect of numerical order could explain this effect, as it was the only significant main effect in the analysis of error rates ($\eta_p^2 = .22$). In a task examining prospective bargains, this fitted a temporally ascending order from standard to target. The temporal framing of the task in Experiment 5 was opposite to the temporal framing of the experiments described previously, where no order effect had been reported. In the design of Experiment 5, Bargain and Order interacted but Sequence had no main effect on the error rates. One could conclude that the STSE was interfered by the temporal markers of standard and target in this experiment that contradicted a temporal order of standard preceding target. The standard (marked by the word tomorrow) was set in the future, while the target (marked by the word today) was temporally set in the present. In Experiments 1 to 4 it was the other way around, so the temporal markers of target and standard fitted the (according to the contextual framing of standard marked by the word yesterday and target marked by the word today) theoretical temporal occurrence of the two prices within a trial. Experiment 5 used a surreal setting, as participants were not used to be exposed to future sales prices as referents in bargain judgments in real marketplaces. So, bargain detection in this task was not as fluent as in the previously presented setting. It appears logical that the contradicting sequence of the temporal markers in a bargain detection task interfered with the STSE in error rats. The effects found in the analyses of error rates lead to the conclusion that an empirical examination of the STSE in a setting without Bargain as an extra factor of variance is needed to be able to disentangle the effects of time, sequence and numerical order.

2.6 Experiment 6: Numerical Comparisons

The purpose of Experiment 6 was to replicate the STSE found in Experiments 1, 2, 3, and 5 in pairwise comparisons of mere single-digit numbers. H5 was tested (see section 4).

2.6.1 Open Science Statement and Power Analysis

The preregistration of Experiment 6 as well as materials and data are available online in the Open Science Framework at https://osf.io/7z8cp. A 2 by 2 by 2 mixed model rmANOVA was preregistered, including Responseside as between factor and Sequence (standard preceding target [ST] vs. target preceding standard [TS]) and Order (descending vs. ascending) as within participant factors. The actual data analyses excluded the between factor as the data collection had been counterbalanced between participants for the relative size of the target (see section 2.6.2) to prevent it from causing systematic variance. The calculation of the preregistered mixed model can be found in the Appendix B1.1. The a priori power analysis was performed using G*Power (Faul et al., 2007) to calculate the required sample size for a t-test with an, according to the results of the previously conducted experiments, large effect size of d_z = .80 and a preset power of 80%. The required sample size to replicate the effect, *N* = 10, was arbitrarily overpowered due to an estimated high drop rate of the cases, that would not meet the inclusion criteria (see section 4.6.3), according to what had been reported in Experiment 1. Moreover, a larger sample size would be necessary to find smaller effects in a more difficult task. Therefore, N = 100 was preregistered.

2.6.2 Method

Subjects. 121 students, 91 women, 29 men, and one person of diverse gender identification, participated in a 30 minutes test battery in laboratories at the campuses of the universities of Cologne and Würzburg. As compensation they received 4 Euro or a course credit. Their mean age was 23.

Materials. Materials and procedure of Experiment 6 equaled Experiment 2, except for the reduction of the stimuli to simple single digit numbers. Target and standard were marked by printing them in different font. To one half of the participants the targets were introduced as being presented in a bold font, to the other half of the participants the standard was introduced to be presented in bold font. Targets and standards could range from 1 to 9. In this experiment, to avoid ties, the pairing of numbers was controlled so that every possible pair of target and standard was presented once in the 'standard preceding target' within participant condition and once in the 'target preceding standard' condition. Programming resulted in 9 times 8 trials per condition of the sequence factor. Every participant had to perform a total amount of 144 trials, 72 'ST'-trials that were of ascending numerical order and 72 'ST'-trials that were of descending numerical order. Trials of both within participant conditions were presented intermixed in randomized order.

Procedure. There were three between participant counterbalancing factors leading to a split of the sample into 6 groups through permutation (the instruction texts per group of Experiment 6 can be found in Appendix A). Half of the participants saw bold targets and thin standards (font: *Arial*, size 26), the other half saw bold standards and thin targets. Half of the participants were instructed to indicate whether the target number was smaller than the standard, the other half was instructed to indicate whether the target was larger than the standard. Half of the participants were instructed to use the right CTRL-keys indicating matching trials and to press the left CTRL-key to indicate a mismatch. The other half of the participants was instructed with vice versa key assignments. These counterbalancing factors were employed to control for effects of polarity correspondence (e.g. SCE) and for a SNARC effect in terms of a response side effect (e.g. Dehaene et al. 1990; Dehaene et al. 1993).

After the instructions had been displayed on the screen and the participant had indicated to have understood the task, five training trials had to be performed. In these training trials, feedback for false response was given. Before the relevant test trials started participants were told that no feedback on their performance was given during the test.

2.6.3 Results

26 subjects performed worse than the minimal criterion of 70% correct and therefore their data were excluded from further analyses. Among them were 9 cases with an error rate above 80%, indicating a confusion of the key assignments. Reaction times and error rates were statistically analyzed with a remaining sample of 95 cases.

Trials with an RT that deviated more than 2.5 times the standard deviation from the mean were discarded from further analyses. This concerned 4 % of all trials over all 96 cases four additional outlier trials were deleted as well. The mean RT of the trials that were included in the data analysis was *M* = 804.27 ms (*SD* = 367.58 ms).

A next step of data cleaning was the deletion of trials with response primes. Response primes were defined as trials that started with a 1 or a 9.

Error rates. 7.4% of all trials were erroneous. To check for the STSE on error rates, trials in that standard preceded target were compared with trials in that target preceded standard in a paired t-test. The STSE was significant having a small effect size (Cohen, 1988), *t*(94) = 2.84, *p* = .032, d_z = .22. Trials in that the standard preceded the target were performed more often correct, *M* = 0.07 (*SD* = 0.06), than trials in that the target preceded the standard, *M* = 0.08 (*SD* = 0.06).

To control for the effect of numerical order, a 2 by 2 rmANOVA (Sequence x Order) was run. Sequence, as in the t-test, showed a significant effect of a small size, *F*(1,94) = 4.86, *p* = .030, η_p^2 = .05, while Order did not affect the error rates significantly, *F*(1,94) = 0.01, *p* = .977. The interaction of both factors included in the analysis revealed to be significant, *F*(1,94) = 6.91, *p* = .010, η_p^2 = .07. The interaction plot in Figure 29 reveals that the STSE was especially pronounced and significant in the numerically descending trials, *F*(1,94) = 9.97, *p* = .002, *r* = .31, while it was very small and not significant in the numerically ascending trials, *F*(1,94) = 0.25, *p* = .621. The order effect mediated the STSE.

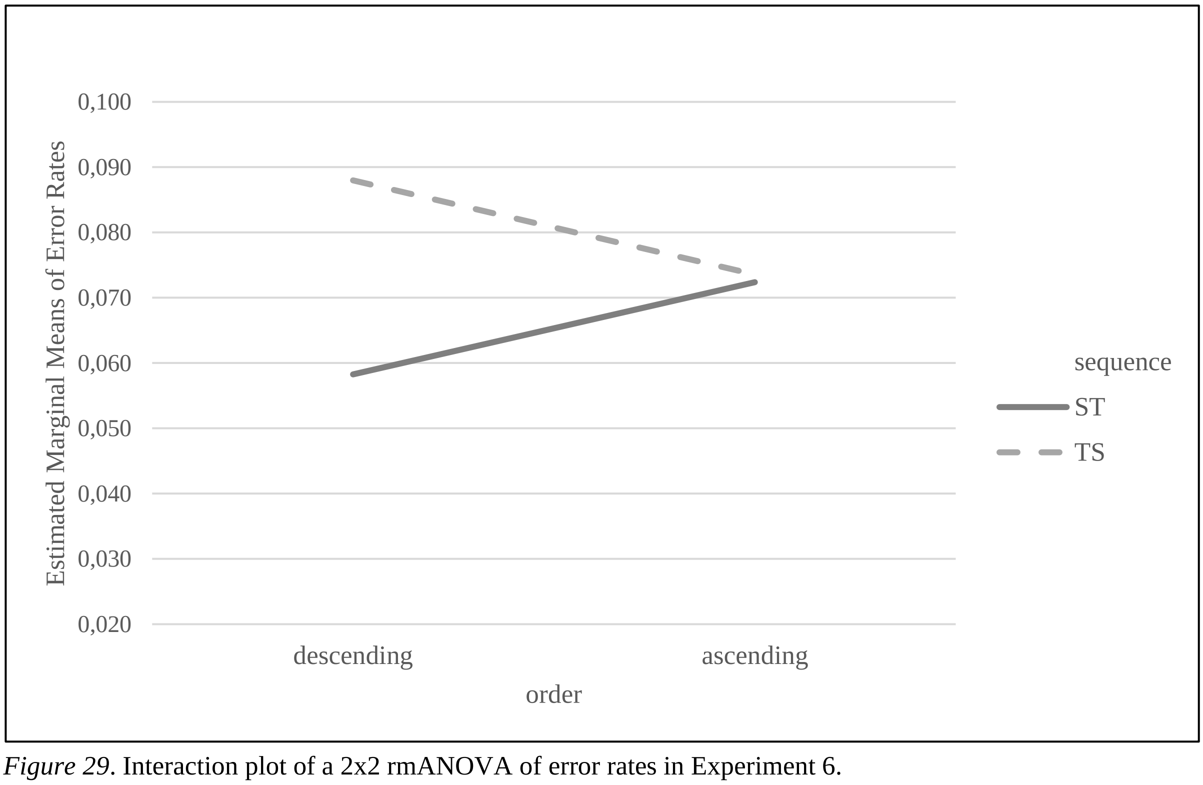

Figure 29. Interaction plot of a 2x2 rmANOVA of error rates in Experiment 6.

RT. For the analysis of RT the erroneous trials were discarded and the natural logarithmized RT scores were used. The paired t-test revealed a highly significant STSE of a medium effect size (Cohen, 1988), *t*(94) = 3.66, *p* < .001, *dz* = .38. Trials in that the standard preceded the target were performed faster, *M* = 800.48 ms (*SD* = 222.61 ms), than trials in that the target preceded the standard, *M* = 837.65 ms (*SD* = 221.85 ms).

A 2 by 2 rmANOVA (Sequence x Order) resulted in a highly significant STSE of medium size, *F*(1,94) = 14.04, *p* < .001, η_p^2 = .13, while order did not affect RT, *F*(1,94) = 2.67, *p* = .106. The interaction of both factors was not significant as well, *F*(1,94) = 1.05, *p* = .309. The interaction plot in Figure 30 illustrates that the mean RT of numerically descending trials were lower than the mean RT of numerically ascending trials.

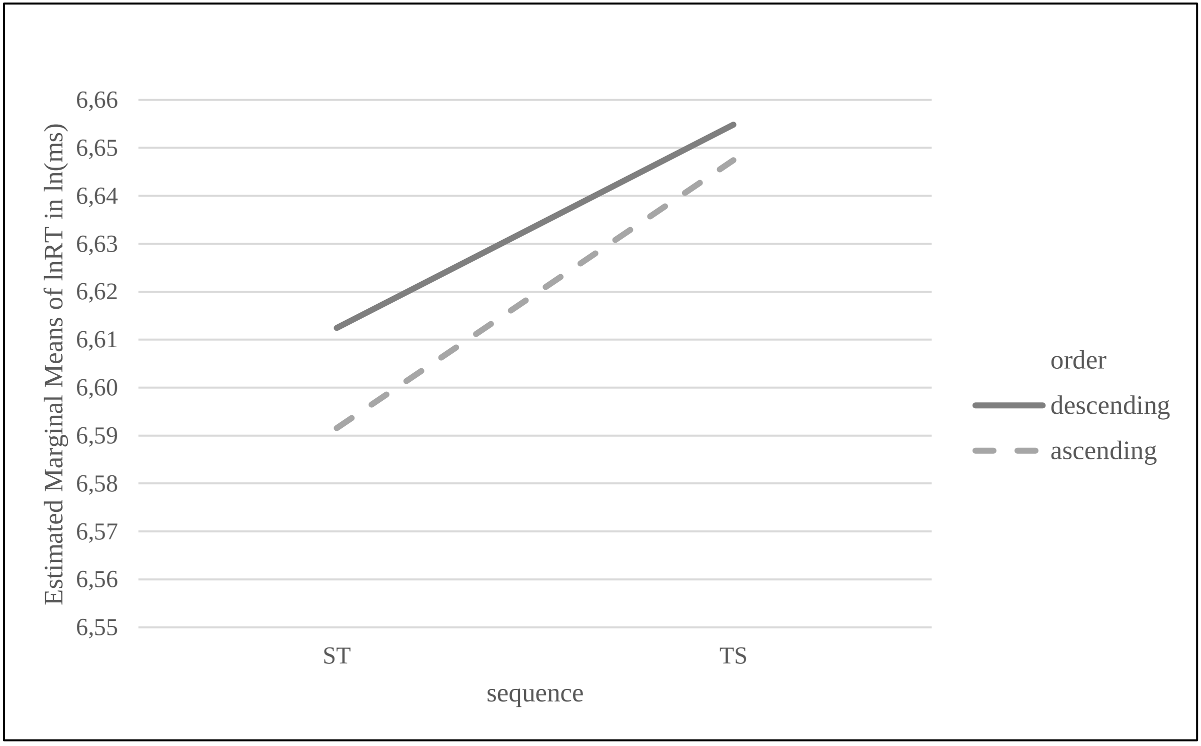

Figure 30. Interaction plot of a 2x2 rmANOVA of RT in Experiment 6.

2.6.4 Discussion

The test for the STSE in paired comparisons of mere single digit numbers revealed a performance advantage for the sequence of standard and target in RT, $d_z = .38$. A local STSE was found for error rates in ascending numerical trials. In numerically descending pairs the STSE was only found descriptively. There was a speed-accuracy-trade-off between the numerically ascending and descending trials. Descriptively ascending trials were performed faster but caused more errors.

The task employed in Experiment 6 revealed a higher overall error rate than the experiments employing the bargain-judgement-paradigm (see Experiment 1 to 5), although in Experiment 6, participants had to perform a training phase with error feedback. This indicated that the 2AFC task presented in this section was harder to perform and instructions were more difficult to remember than the task with higher practical relevance to identify bargains in paired price comparisons.

2.7 Experiment 7: Comparisons of Four Digit Numbers

The purpose of Experiment 7 was to replicate the findings of Experiment 6 for a 2AFC task of four-digit numbers. In this experiment H_5 was tested.

2.7.1 Open Science Statement and Power Analysis

The preregistration of Experiment 7 as well as materials and data are available online in the Open Science Framework at https://osf.io/8bx9z. The power analysis equaled Experiment 6, except for the estimated average effect size of $d_z = .30$, on the basis of $d_z = .22$ for error rates and $d_z = .38$ for RT, found in Experiment 6. The t-test required $N = 72$ to replicate the effect. Sample size was arbitrarily overpowered and N = 100 was preregistered.

2.7.2 Method

Subjects. 135 students, 112 women, 21 men, and 2 persons of diverse gender identification participated in a 30 minutes test battery in a laboratory at the campus of the University of Cologne. For compensation they received course credit or 4 Euro. Their mean age was 23.

Materials. The material equaled the material employed in Experiment 6 except for the stimuli being four-digit numbers instead of one digit numbers, ranging from 1000 to 9999. The number of trials in this experiment was arbitrarily set to 100. It was programmed to randomly pair four-digit numbers of the predefined range. Because of this programming, ties were possible (compare Experiments 2 to 5).

Procedure. The procedure of Experiment 7 equaled Experiment 6, employing the same counterbalancing between participant factors and instructions. Participants performed five training trials with feedback for false responses before the relevant test trials without performance feedback started. After half of the trials, participants took a mandatory break for one minute. They were advised to relax and close their eyes during the break. After the break the task instructions appeared on the screen again and participants were told to start the second half of the trials by pressing the space bar.

2.7.3 Results

24 subjects performed worse than the minimal criterion of 70% correct and therefore were excluded from further analyses. RT and error rates were statistically analyzed with a remaining sample of 111 cases.

Trials with an RT that deviated more than 2.5 times the standard deviation from the mean were discarded from further analyses. All trials under 300 ms were deleted as well, as stimuli in this experiment were a lot more complex to process and faster RT were assumed to indicate guessing rather than attending to the task. In total 1.9 % of all trials were discarded due to extreme RT. The mean RT of the trials that were included in the data analyses was *M* = 1,054.13 ms (*SD* = 577.66 ms). The next step of data cleaning was the deletion of trials with response primes. Response primes were defined as trials that started with a 10,11, 99 or 98. By chance, the randomized stimuli presentation resulted in no tie trials.

Error rates. 9.0 % of all trials were erroneous. To check for the STSE in error rates, a t-test comparing trials in that standards preceded targets and trials in that targets preceded standards was performed. The STSE did not show up, *t*(110) = 0.05; *p* = .962.

To control for the effect of numerical order within the trials, a 2 by 2 rmANOVA (Sequence x Order) was run. Sequence, as in the t-test, did not result in an significant effect, *F*(1,110) = 0.81, *p* = .371, while the numerical order affected the error rates significantly, *F*(1,110) = 6.72, *p* = .011, η_p^2 = .06. The interaction of Sequence and Order revealed to be significant, *F*(1,110) = 6.21, *p* = .014, η_p^2 = .05.

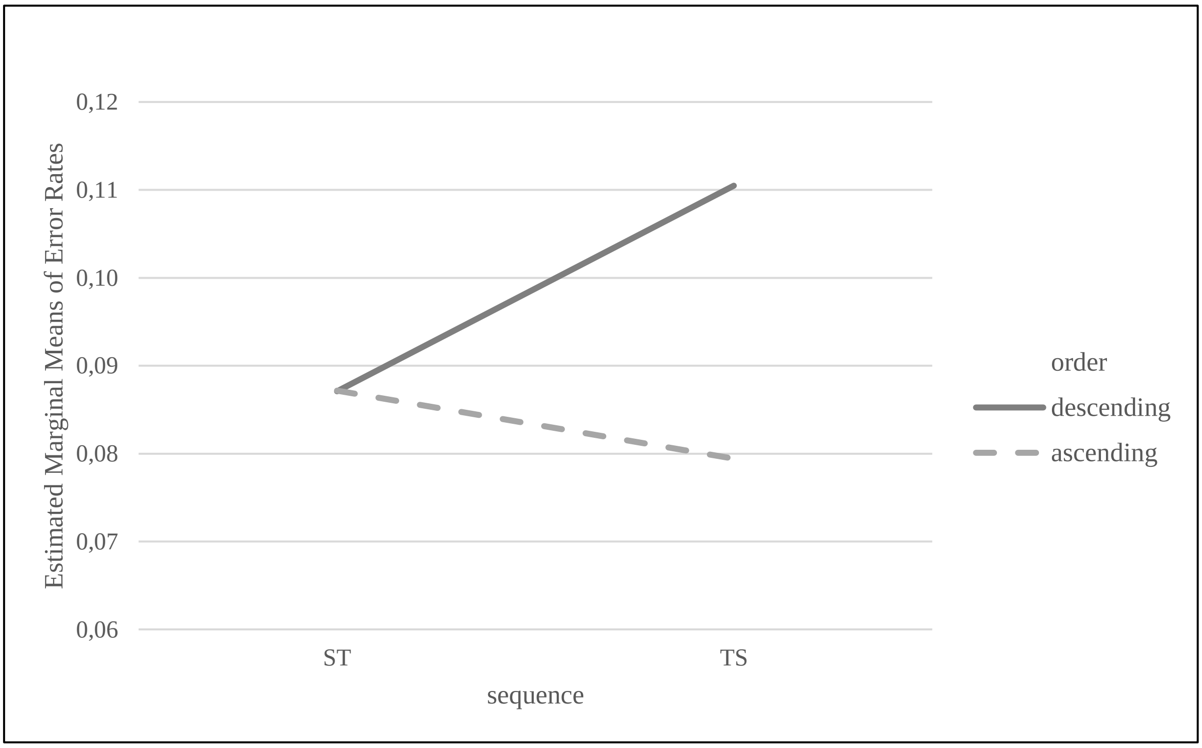

Figure 31. Interaction plot of a 2x2 rmANOVA of error rates in Experiment 7.

The interaction plot in Figure 31 illustrates that the as predicted STSE shows up in the numerically descending trials, $F(1,110) = 5.09$, $p = .026$, $r = .20$, but not in the numerically ascending trials, $F(1,110) = 0.49$, $p = .486$. The order effect was only significant in the TS trials, $F(1,110) = 13.08$, $p < .001$, $r = .33$, but not in the ST trials, $F(1,110) < 0.01$, $p = .993$.

RT. For the analysis of RT the erroneous trials were discarded and the natural logarithmized RT scores were used for the parametric tests. The paired t-test for the STSE, revealed a highly significant effect of a medium effect size (Cohen, 1988), $t(110) = 5.32$; $p < .001$; $d_z = .51$.

A 2 by 2 rmANOVA (Sequence x Order) resulted in highly significant STSE of medium size, $F(1,110) = 30.01$, $p < .001$, $\eta_p^2 = .22$, while Order did not significantly affect RT, $F(1,110) = 2.29$, $p = .133$. The interaction of both factors was not significant as well, $F(1,110) = 2.35$, $p = .128$.

2.7.4 Discussion

The results of Experiment 7 revealed an ascending order advantage (compare Turconi et al., 2006) in the error rates of a speeded 2AFC magnitude comparison task with four-digit numbers, with a small effect size, $\eta_p^2 = .05$. A local STSE within the numerically descending trials could be found as well, $r = .20$. For RT a global STSE of medium effect size, $d_z = .51$, showed up. Overall conditions, the error rate of 9% was higher than in previously described experiments, and overall RT was over 1,000 ms higher than in the speeded 2AFC task with one-digit numbers as stimuli. The higher complexity of the task could account for the longer RT as well as the higher error rate, requiring different comparison mechanisms or heuristics to perform the task as fast as possible. Furthermore, the working memory load in a comparison task with four-digit numbers was much higher than in a task with one-digit numbers. The task in Experiment 7 required more elaborate arithmetic processing than the task in Experiment 6.

The differences in the result patterns of Experiment 6 and 7 could reveal the crucial role of working memory function for the STSE. This assumption needs to be investigated with further research, for example employing a dual task paradigm (compare van Dijck & Fias, 2011).

2.8 Experiment 8: Comparisons of IQ Scores

Experiments 8 and 9 were conducted to transfer the STSE on real-world comparisons beyond marketplaces. In Experiment 8, testing H_5 (see section 4), numerical stimuli were framed as IQ scores of target and standard persons. Participants had to perform a speeded 2AFC task, just like in Experiments 2 to 7, deciding whether the target person, presented repeatedly in a sequential pair with a standard person, had a higher (lower) IQ score than the standard person.

2.8.1 Open Science Statement and Power Analysis

The preregistration of Experiment 8 as well as materials and data are available online in the Open Science Framework at https://osf.io/4ehzc. The data analysis presented in section 2.8.3 deviates from the 2 by 2 by 2 mixed model ANOVA that was preregistered for this experiment, as lined out in section 2.6.1 regarding the preregistration of Experiment 6. The a priori power analysis equaled Experiment 6 as well. For a preset power 80% and an estimated medium to high effect size of $f = .40$ on the basis of the $d_z = .90$ for the effect on error rates and $d_z = .42$ for RT found in Experiment 1, resulted in a required sample size of $N = 10$. The study was arbitrarily overpowered with a preregistered $N = 130$. Experiment 8 was even more overpowered than Experiment 6 due to a higher estimated error rate and latency for the applied context of Experiment 8 – participants had to judge persons regarding their IQ scores, not only numbers.

2.8.2 Method

Subjects. 152 students participated in a 20 minutes test battery in laboratories at the campuses of the universities of Cologne and Würzburg. The sample consisted of 114 women and 38 men, with a mean age of 24 years.

Materials. The design of the stimuli equaled Experiment 2 to 7 except for the contextual frame of the to be compared numbers being IQ scores of persons. IQ scores ranged from 100 to 140 and were randomly paired for the comparison trials. The number of trials was set to 100, 50 'standard preceding target'-trials and 50 'target preceding standard'-trials. Figure 32 displays an example for a trial in Experiment 8. It was possible, due to the randomization of IQ scores, that target and standard in a trial had the same numerical expression (ties).

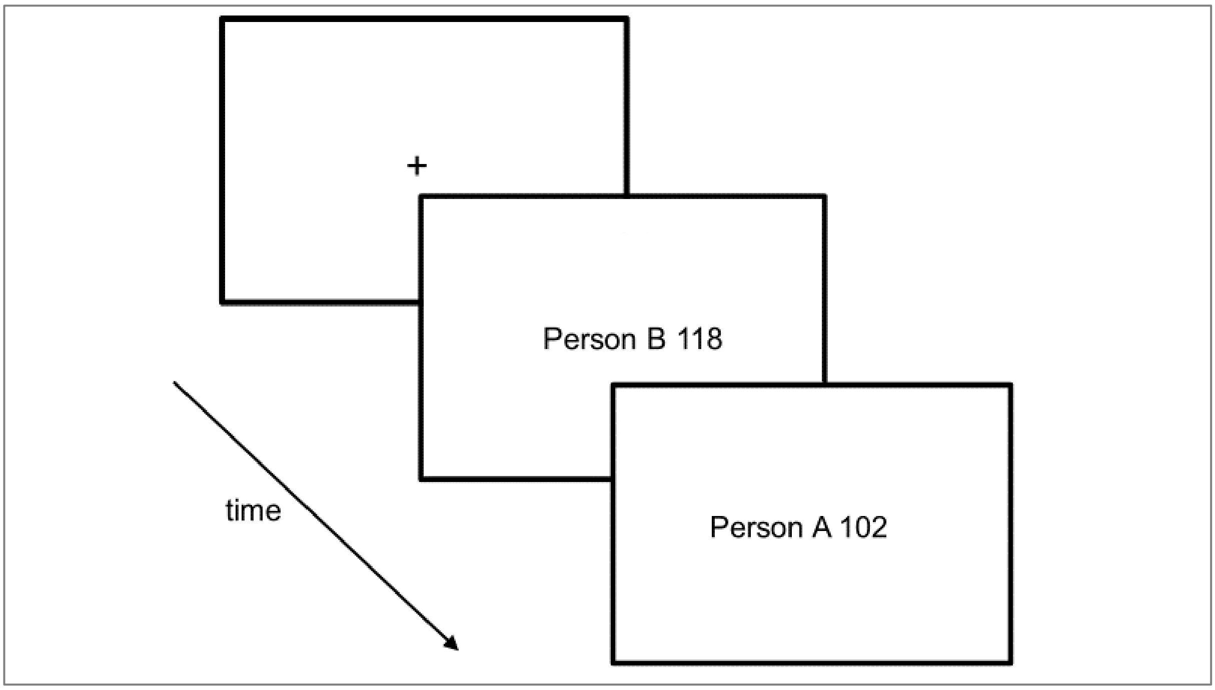

Figure 32. Example for a trial of Experiment 8. Letters A and B describing the persons served as markers for standard and target (see *Procedure*).

Procedure. The procedure equaled Experiment 2 to 7. In each trial the IQ scores of Person A and Person B were sequentially presented. Half of the participants were instructed to judge Person A's IQ score (target) compared to Person B's IQ score (standard), the other half was instructed to judge Person B's IQ score (target) to Person A's IQ score (standard, the instruction texts of Experiment 6 can be found in Appendix A). Half of the participants were instructed to indicate whether the target IQ score was smaller than the standard, the other half was instructed to indicate whether the target IQ score was larger than the standard. Half of the participants were instructed to use the right CTRL-key to indicate a trial that matched the instructions and to press the left CTRL-key to indicate a mismatch. The other half of participants was instructed with vice versa key assignments. These three counterbalancing factors were employed to control for effects of polarity correspondence (e.g. SCE or a response side effect, reported by Dehaene et al. 1990) and for a SNARC effect in terms of a response side effect (e.g. Dehaene et al. 1990; Dehaene et al. 1993).

After the instructions had been displayed on the screen and the participant had indicated to understand the task, five training trials had to be performed. In these training trials feedback for false response was given. Before the relevant test trials started participants were told that during the test no feedback on their performance was given.

2.8.3 Results

The data of 22 subjects performed worse than the minimal criterion of 70% correct and therefore was excluded from further analyses. RT and error rates were statistically analyzed with a remaining sample of 130 cases.

Trials with an RT that deviated more than 2.5 times the standard deviation from the mean were discarded from further analyses. This concerned 2.9 % of all trials over all 130 cases four additional outlier trials that undercut an RT of 300 ms were deleted as well. The mean RT of the trials that were included in the data analysis was *M* = 907.27 ms (*SD* = 395.52 ms).

2.5% of all trials included in the data analyses were ties and were not discarded because they had no impact on the results presented in the following (see Appendix B2).

Error rates. 8.5% of all trials were erroneous. To check for the STSE, error rates of 'standard preceding target' (ST) trials were compared to error rates in 'target preceding standards' (TS) trials in a paired t-test. The STSE was significant having a small effect size (Cohen, 1988), *t*(129) = 2.74; *p* = .007; d_z = .24.

To control for the effect of numerical order of the stimuli, a 2 by 2 rmANOVA (Sequence x Order) was run. Sequence, as in the t-test, showed a significant effect of a small size, *F*(1,129) = 6.40, *p* = .013, η_p^2 = .05, Order showed a significant effect as well, *F*(1,129) = 7.99, *p* = .005, η_p^2 = .06. The interaction of both factors had no significant effect, *F*(1,129) = 0.75, *p* = .387. The interaction plot in Figure 33 illustrates that for both main effects the predicted direction was met, ST trials were performed more often correct, *M* = .08 (*SD* = .07), than TS trials, *M* = .10 (*SD* = .08); trials of ascending numerical order were performed more often correct, *M* = .08 (*SD* = .06), than trials of descending numerical order, *M* = .09 (*SD* = .08). On average the best performance of participants in terms of error rates could be found in trials that matched the STSE and an ascending numerical order (*M* = .07, *SD* = .08).

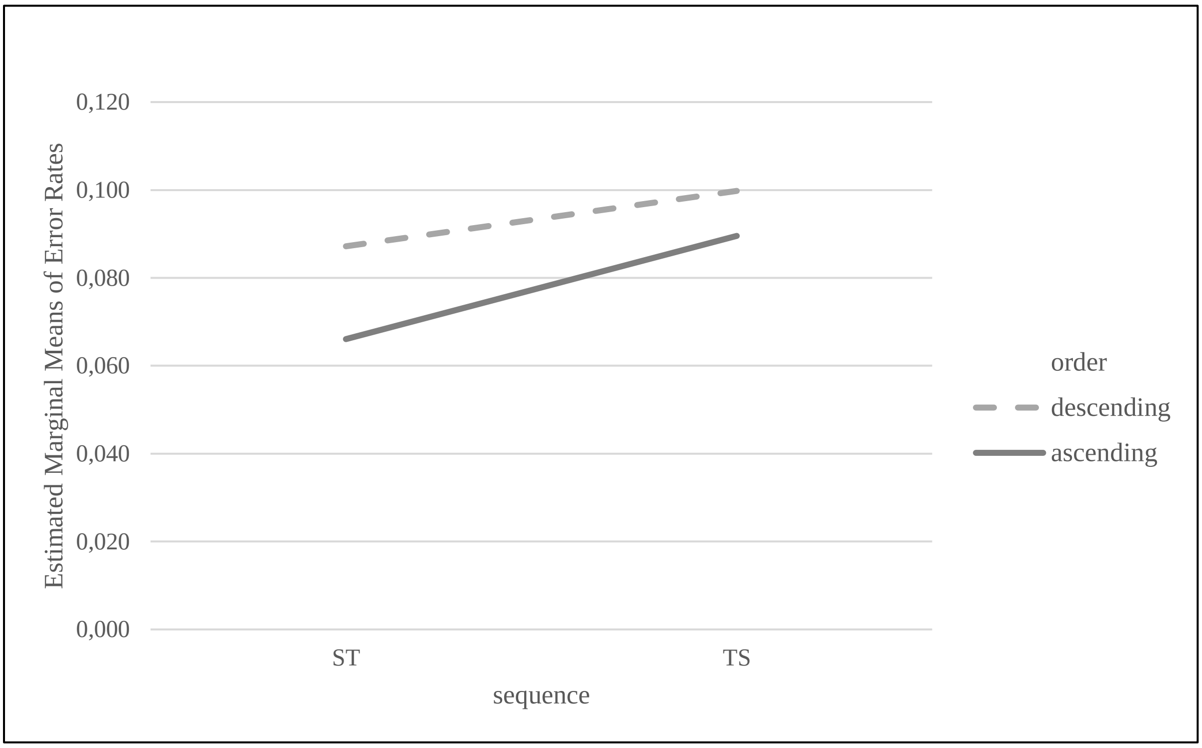

Figure 33. Interaction plot of a 2x2 rmANOVA of error rates in Experiment 8.

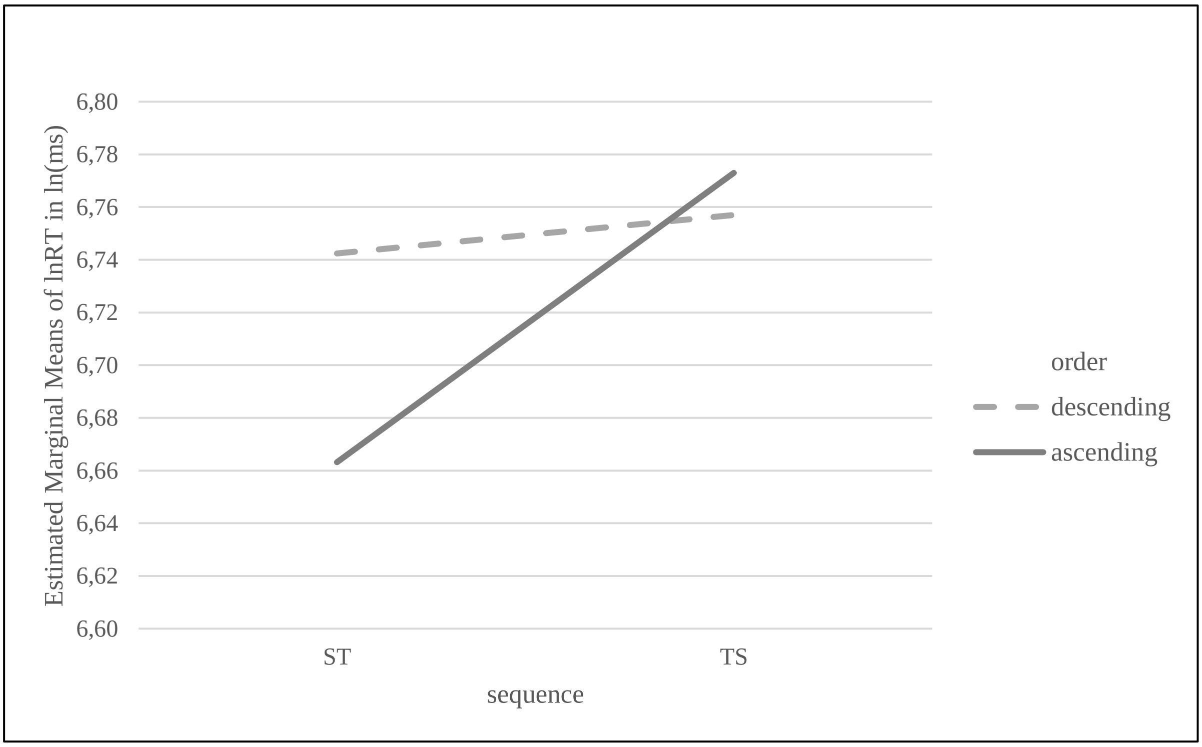

Figure 34. Interaction plot of a 2x2 rmANOVA of RT in Experiment 8.

RT. For the analysis of RT the erroneous trials were discarded and the natural logarithmized RT scores were used for the parametric tests. The paired t-test for the STSE revealed a highly significant difference of mean RT and a medium to large effect size, $t(129) = 6.26$; $p < .001$; $d_z = .55$.

The 2 by 2 rmANOVA (Sequence x Order) resulted in a highly significant STSE of medium size, $F(1,129) = 37.87$, $p < .001$, $\eta_p^2 = .23$, a highly significant effect of numerical order, $F(1,129) = 13.47$, $p < .001$, $\eta_p^2 = .10$, and a highly significant interaction of both factors, $F(1,129) = 22.83$, $p < .001$, $\eta_p^2 = .15$.

Simple effects analyses resulted in a highly significant difference for ST versus TS trials of ascending order, $F(1,129) = 55.00$, $p < .001$, $r = .55$, but a not significant difference within the trials of descending order, $F(1,129) = 1.16$, $p = .283$. Descriptively the performance advantage for ST trials compared to TS trials could be found in ascending as well as in descending trials, but the order effect turns from an ascending order advantage in the ST trials to a descending order advantage in the TS trials. This is illustrated in the interaction plot in Figure 34. The difference of the numerically ascending trials and the descending trials was only significant in the ST trials, $F(1,129) = 35.70$, $p < .001$, $r = .47$, but not in the TS trials, $F(1,129) = 1.49$, $p = .224$.

2.8.4 Discussion

The test of the STSE in pairwise comparisons of two persons' IQ scores revealed a small to medium STSE in participants' performance parameters. The effect was larger in RT, $d_z = .55$, than in error rates, $d_z = .24$. Also, the ascending order advantage (compare Turconi et al., 2006) within the trials was significant and had a small effect in the analysis of error rates, $\eta_p^2 = .06$, and a small effect in RT, $\eta_p^2 = .10$. But the order effect was mediated by the STSE in RT. For RT it turned into a descending order advantage for TS trials. The trials of ascending order and ST had the best performance parameters.

2.9 Experiment 9: Comparisons of Monthly Income

In Experiments 9, testing H_5 (see section 4), numerical stimuli were framed as monthly income of fictional persons. Participants had to perform a speeded 2AFC task, just like in Experiment 6, deciding whether the target person, presented repeatedly in a sequential pair with a standard person, had a higher (lower) income than the standard person.

2.9.1 Open Science Statement and Power Analysis

The preregistration of Experiment 9 as well as materials and data are available online in the Open Science Framework at https://osf.io/z6wa4. The data analysis described in section 2.9.3 deviates from the 2 by 2 by 2 ANOVA that was preregistered for this experiment just like the analyses of Experiments 6 and 8. The a priori power analysis was the same for Experiment 9 as for Experiment 6 to 8. For a preset power 80% and an estimated medium to large effect size of f = .40 on the basis of the d_z = .90 for the effect on error rates and d_z = .42 for RT found in Experiment 1, resulted in a required sample size of N = 10. The study was arbitrarily overpowered with a preregistered N = 130.

2.9.2 Method

Subjects. 177 students participated in a 20 minutes test battery in a laboratory at the pus of the university of Würzburg. The sample consisted of 102 women, 73 men and two with diverse gender. Their mean age was 26 years.

Materials. The design of the stimuli equaled Experiment 8 except for the contextual frame of the to be compared numbers being monthly incomes of persons instead of IQ scores. Monthly income ranged from 1000 Euro to 1900 Euro. Figure 35 shows an example for a trial of Experiment 9. Like in Experiment 6, to avoid ties, the pairing of monthly incomes was controlled. Hence, every possible pair of target and standard was presented once in the ‘standard preceding target’ within participant condition and once in the ‘target preceding standard' condition. Programming resulted in 9 times 8 trials per condition of the sequence factor. Every participant had to perform a total amount of 144 trials, 72 'standard preceded target' (ST) trials in ascending numerical order, 72 ST trials in descending numerical order, 72 'target preceded standards' (TS) trials in ascending numerical order and 72 TS trials in descending numerical order. Trials of both within participant conditions were presented intermixed in randomized order.

Procedure. The procedure of Experiment 9 equaled Experiment 8. In each trial the monthly income of Person A and Person B were sequentially presented. Half of the participants was instructed to judge Person A's monthly income (target) compared to Person B's monthly income (standard), the other half was instructed to judge Person B's monthly income (target) to Person A's monthly income (standard). Find the instruction texts in Appendix A. To prevent effects of polarity correspondence and SNARC in terms of a response side effect, two further between participant factors were employed. Half of the participants were instructed to indicate whether the target's income was smaller than the standard's, the other half was instructed to indicate whether the target's income was bigger than the standard. Half of the participants were instructed to use the right CTRL-key to indicate a trials that matched the instructions and to press the left CTRL-key to indicate a mismatch. The other half of participants was instructed with vice versa key assignments. After the instructions had been displayed on the screen and the participant had indicated to understand the task, five training trials had to be performed. In these training trials feedback for false response was given. Before the relevant test trials started, participants were told that they would not receive feedback on their performance during the test phase. After half of the trials, participants took a mandatory break of one minute to prevent fatigue. They were advised to relax and to close their eyes during the break. After one minute the task instructions appeared on the screen again and participants were told to start the second half of the trials by pressing the space bar.

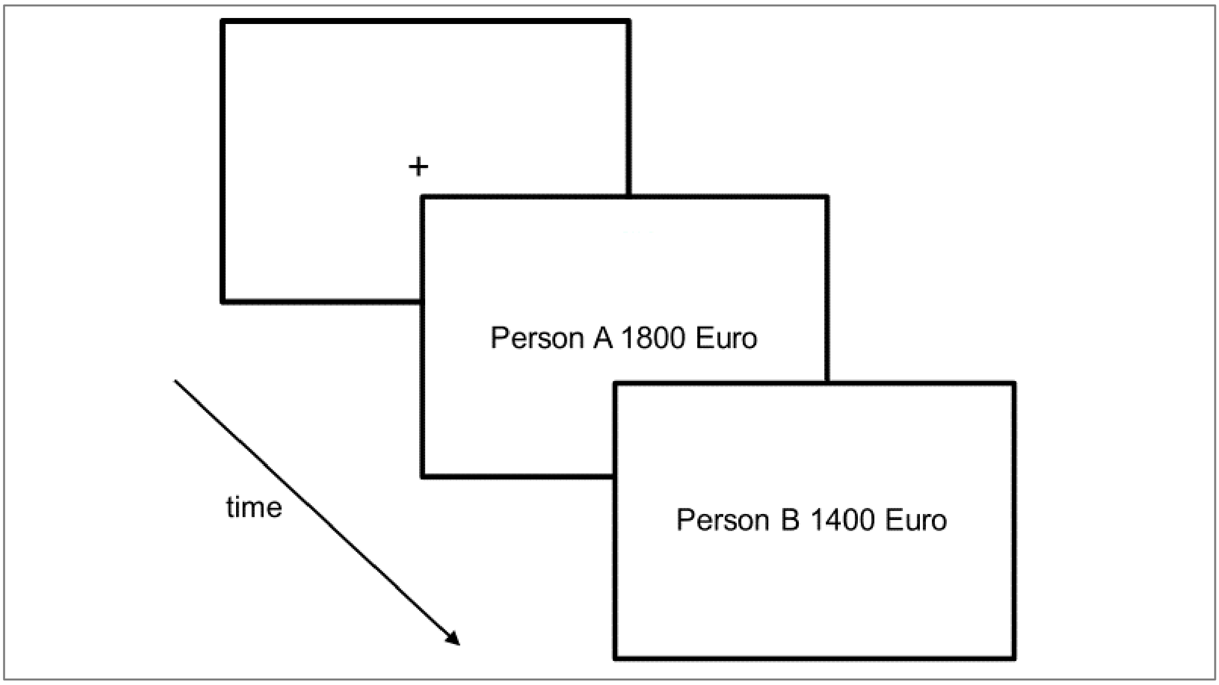

Figure 35. Example for a trial of Experiment 9. Letters A and B describing the persons served as markings for standard and target (see *Procedure*).

2.9.3 Results

27 subjects performed worse than the minimal criterion of 70% correct answers and therefore were excluded from further analyses. RT and error rates were statistically analyzed with a remaining sample of 150 cases.

Trials with an RT that deviated more than 2.5 times the standard deviation from the mean were discarded from further analyses. This concerned 3.2 % of all trials over all 150 cases. The mean response time of the trials that were included in the data analysis was *M* = 739.24 ms (*SD* = 280.66 ms).

The next step of data cleaning was the deletion of trials with response primes. Response primes were defined as trials that started with 1100 Euro or 1900 Euro, because participants could assume the right response after the first stimulus had appeared. When 1100 Euro appeared in the first stimulus of a trial, the second stimulus could only be larger, when 1900 Euro appeared in the first stimulus of a trial, the second stimulus could only be smaller.

Error rate. 8.5% erroneous trials. To check for the STSE in error rates of participant's performance, a paired t-test was performed to compare the mean error rates of ST trials and of TS trials. The STSE was significant with a small effect size (Cohen, 1988), $t(149) = 3.72$; $p < .001$; $d_z = .30$.

To control for the effect of numerical order, a 2 by 2 rmANOVA (Sequence x Order) was run. Sequence, as in the t-test, showed a significant effect of a small size, $F(1,150) = 13.89$, $p < .001$, $\eta_p^2 = .09$, numerical order showed a significant effect as well, $F(1,150) = 20.84$, $p < .001$, $\eta_p^2 = .12$. The interaction of both factors had no significant effect, $F(1,150) = 0.45$, $p = .502$. The interaction plot in Figure 36 reveals for both main effects an as expected direction, ST trials were performed more often correct than TS trials; trials of ascending numerical order were performed more often correct than trials of descending numerical order. On average the best performance of participants in terms of error rates could be found in trials that matched the STSE and had an ascending numerical order ($M = .07$, $SD = .08$).

RT. For analysis of RT the erroneous trials were discarded and the natural logarithmized RT scores were used. The paired t-test for the STSE revealed a highly significant difference of mean RT and a medium to lagre effect size, $t(149) = 6.79$; $p < .001$; $d_z = .55$. The 2 by 2 rmANOVA (Sequence x Order) resulted in highly significant STSE of large size, $F(1,150) = 48.38$, $p < .001$, $\eta_p^2 = .25$, a highly significant effect of numerical order, $F(1,150) = 27.43$, $p < .001$, $\eta_p^2 = .16$, and a highly significant interaction of both factors, $F(1,150) = 28.13$, $p < .001$, $\eta_p^2 = .16$.

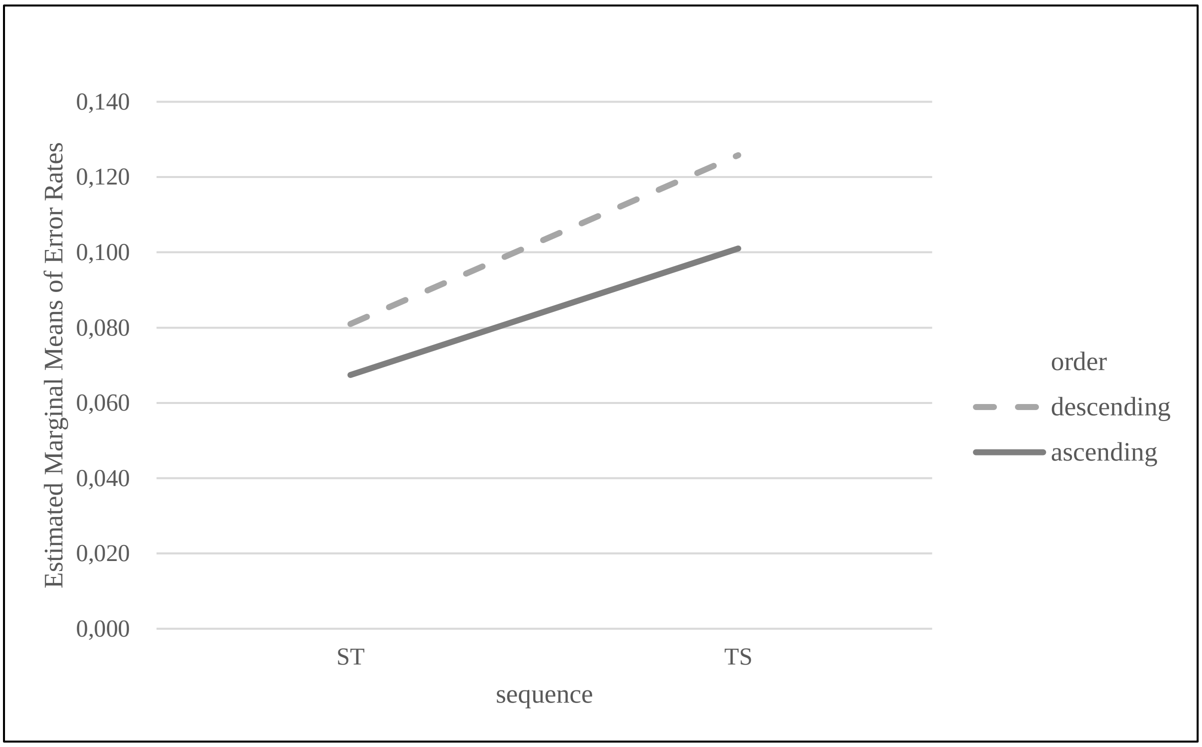

Figure 36. Interaction plot of a 2x2 rmANOVA of error rates in Experiment 9.

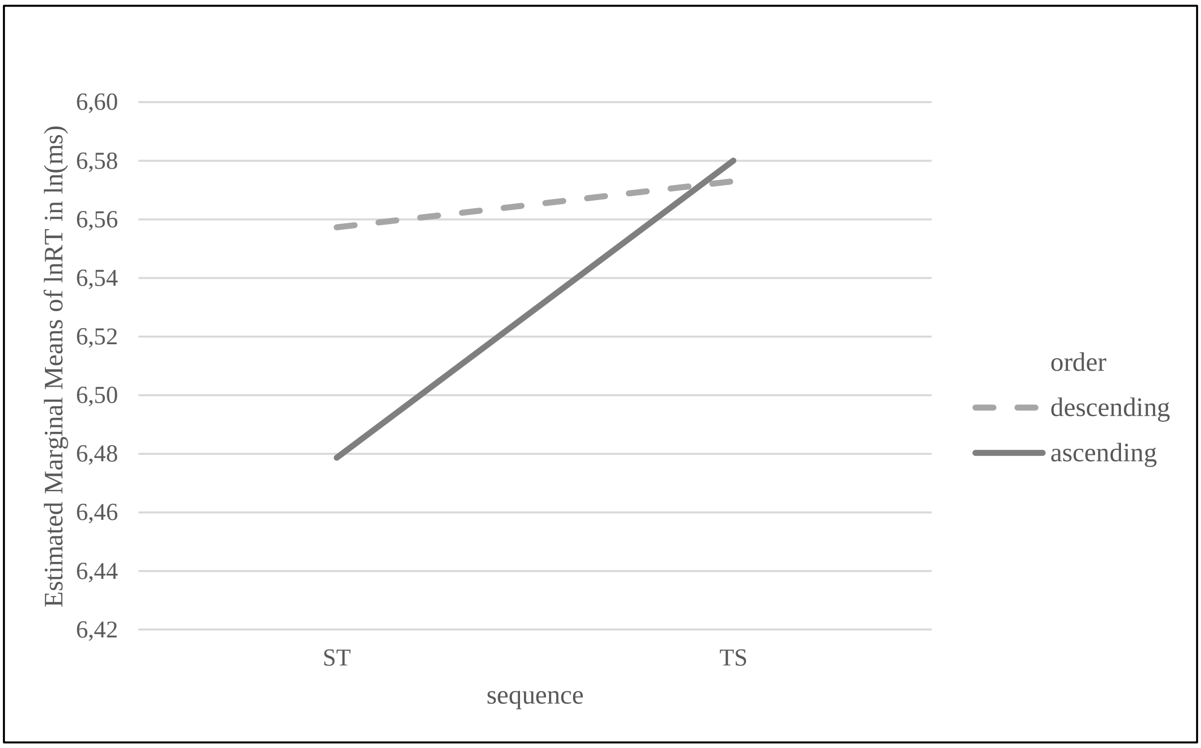

Figure 37. Interaction plot of a 2x2 rmANOVA of RT in Experiment 9.

Simple effects analysis resulted in a highly significant difference for ST versus TS trials within the trials of ascending numerical order, $F(1,147) = 71.59$, $p < .001$, $r = .82$, however, not within the trials of descending numerical order, $F(1,147) = 2.92$, $p = .168$. Descriptively the performance advantage for ST trials compared to TS trials could be found in both factor levels of Order. The order effect turns between the factor levels of Sequence, in ST trials there was a significant ascending order advantage, $F(1,147) = 52.70$, $p < .001$, $r = .51$, while in TS trials there was a not significant descending order advantage, $F(1,147) = 0.47$, $p = .500$ (see Figure 37).

2.9.4 Discussion

The data analyses of Experiment 9 revealed the same effects as Experiment 8. The performance advantage for standard-target-sequences compared to target-standard-sequences could be found for error rates and RT. The STSE had a larger effect size in RT analysis, $d_z = .55$, than in the analysis of error rates, $d_z = .30$. Also, the advantage of ascending numerical order revealed for both performance outcomes. While it had a medium effect size in the analysis of error rates, $\eta_p^2 = .12$, it had a large effect size in the analysis of RT, $\eta_p^2 = .16$. The STSE in RT was especially pronounced and highly significant in trials of ascending numerical order, $r = .82$.

2.10 Experiment 10: Volume Comparisons

Experiment 10 tested the STSE for another domain of magnitude comparisons. Stimuli in this experiment were quadrats of different size. As in the previously presented experiments, participants performed a speeded 2AFC task identical to the task employed in Experiment 6. H_6 was formulated on the basis of H_5 that had been affirmed by the evidence of Experiments 6 to 9 (see section 4).

2.10.1 Open Science Statement and Power Analysis

The preregistration of Experiment 10 as well as materials and data are available online in the Open Science Framework at https://osf.io/5kb6z. Power and data analyses equaled Experiments 6, 8 and 9. The data analysis for Experiment 10 (see section 2.10.3) deviates from the 2 by 2 by 2 ANOVA, that was preregistered parallel to the analyses of Experiments 6, 8 and 9. The a priori power analysis for Experiment 10 was the same as for Experiment 6 to 8. For a preset power 80% and an estimated medium to large effect size of f = .40 on the basis of the d_z = .90 for the effect on error rates and d_z = .42 for RT found in Experiment 1, resulted in a required sample size of N = 10. The study was arbitrarily overpowered with a preregistered N = 100, a smaller sample size than preregistered for Experiments 8 and 9 as the magnitude assessment and comparison task of Experiment 10 was expected to be easier than for the complex stimuli of Experiments 8 and 9.

2.10.2 Method

Subjects. 125 students participated in a 30 minutes test battery in laboratories at the campuses of the universities of Cologne and Würzburg. Demographic data for 8 cases was missing, 86 indicated to identify as a woman, 21 as men and 10 as diverse gender. Their mean age was 23 years. They received 4 Euro or a course credit for compensation.

Materials. In this experiment square frames of different sizes were used as stimuli. Stimuli could have one of nine different sizes. Target and standard differed between participant in the thickness of the outlines of the square frames. To one half of the participants the targets were introduced as the quadrats framed by a bold line, while standard quadrats were framed by a thin line, to the other half of the participants the targets were introduced as the quadrats framed by a thin line and standards as framed by a thick line. In this experiment, to avoid ties, the pairing of quadrates was controlled, so that every possible pair of target and standard was presented once in the ‘standard preceding target’ (ST) within participant condition and once in the ‘target preceding standard’ (TS) condition. Programming resulted in 9 times 8 trials per condition of the Sequence factor. Every participant had to perform a total amount of 144 trials, 72 ST trials of ascending volume order, 72 ST trials of descending volume order, 72 TS trials of ascending volume order and 72 TS trials of descending volume order. Trials of all within participant conditions were presented intermixed in randomized order. Figure 38 displays an example for a trial in Experiment 10.

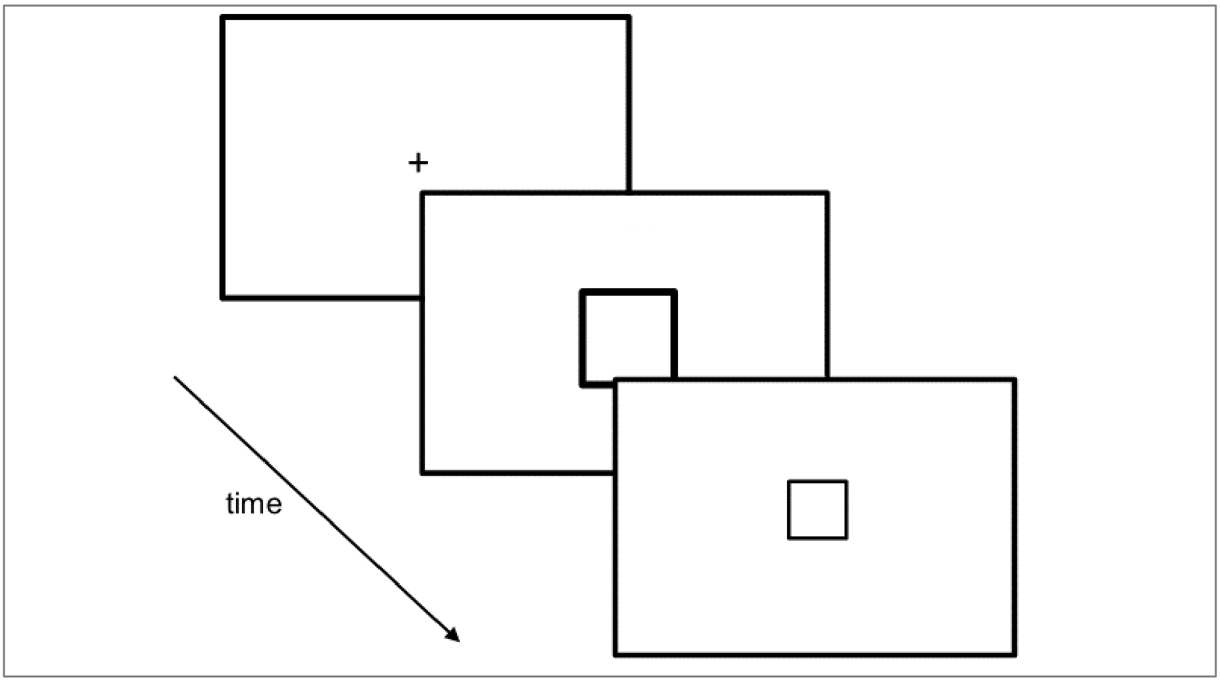

Figure 38. Example for a trial of Experiment 10. Thickness of quadrats’ frames marked standard and target. For description of the temporal progression of a trial see section 4.6.2.

Procedure. Procedure equaled Experiment 6, except for a mandatory break after 72 trials (compare Experiments 7 and 9).

2.10.3 Results

Data of 17 cases was excluded form data analyses because they performed worse than the minimal criterion of 70% correct. RT and error rates were statistically analyzed with a remaining sample of 108 cases.

Trials with a reaction time that deviated more than 2.5 times the standard deviation from the mean were discarded from further analyses. This concerned 1.9 % of all trials over all 108 cases. The mean response time of the trials that were included in the data analysis was M = 745.51 ms (SD = 349.28 ms).

Error rates. 6.2% of all trials were erroneous. To check for the STSE on error rates, means of ST trials and TS trials were compared in a paired t-test. The STSE was significant with medium effect size , $t(107) = 5.43$; $p < .001$, $d_z = .52$.

To control for the effect volume order (ascending vs. descending), a 2 by 2 rmANOVA (Sequence x Order) was run. Sequence, as in the t-test, showed a significant effect of a small size, $F(1,107) = 29.22$, $p < .001$, $\eta_p^2 = .21$. Order also had a significant effect on error rates, $F(1,107) = 8.11$, $p = .005$, $\eta_p^2 = .07$. The interaction of both factors included in the analysis was not significant, $F(1,107) = 1.71$, $p = .190$. The effects of Sequence and Order were in the expected directions, trials that presented standards before targets were performed more often correct than trials in which targets preceded standards. Trials with an ascending order of volume (smaller quadrat followed by larger quadrat) were performed more often correctly than trials in which larger quadrats preceded smaller quadrats. The trials that matched the STSE and the ascending order had the lowest mean error rate, M = .24 (SD = .05).

RT. For the analysis of RT the erroneous trials were discarded and the natural logarithmized RT scores were used. The paired t-test comparing the means of the ST trials and of the TS trials revealed a highly significant STSE with a large effect size (Cohen, 1988), $t(107) = 12.10$; $p < .001$, $d_z = 1.16$.

A 2 by 2 rmANOVA (Sequence x Order) resulted in a highly significant STSE with a large effect size, $F(1,107) = 145.32$, $p < .001$, $\eta_p^2 = .58$, order affected RT highly significant as well, $F(1,107) = 25.88$, $p < .001$, $\eta_p^2 = .20$. The interaction of both factors was significant, $F(1,107) = 8.41$, $p = .005$, $\eta_p^2 = .07$. Simple effects analyses revealed that order had no significant effect within the TS trials, $F(1,107) = 1.31$, $p = .290$, but within the ST trials, $F(1,107) = 30.97$, $p < .001$, $r = .47$. The STSE was highly significant within both factor levels of Order [descending: $F(1,107) = 39.86$, $p < .001$, $r = .52$; ascending: $F(1,107) = 105.18$, $p < .001$, $r = .70$]. The interaction plot displayed in Figure 39 reveals that both effects unfold in the as predicted directions over all conditions. The trials that matched the STSE and had an ascending order of volume on average showed the fastest performance.

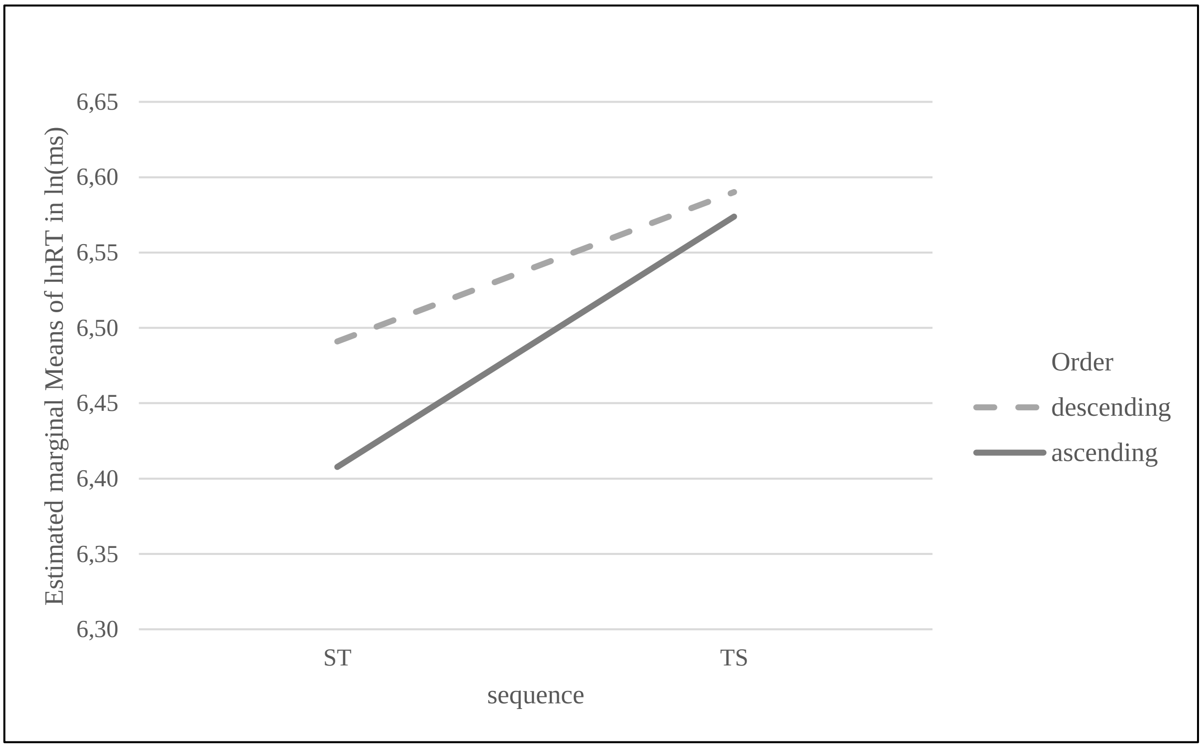

Figure 39. Interaction plot of a 2x2 rmANOVA of RT in Experiment 10.

2.10.4 Discussion

The data analyses of Experiment 10 revealed the same effects as Experiments 8 and 9. The performance advantage for standard-target-sequences compared to target-standard-sequences could be found for error rates and RT. The STSE had a larger effect size in RT analysis, $d_z = 1.16$, than in the analysis of error rates, $d_z = .52$. Furthermore, the effect of order of volume (ascending vs. descending) was observed in both performance outcomes. While it had a medium effect size in the analysis of error rates, $\eta_p^2 = .07$, it had a large effect size in the analysis of RT, $\eta_p^2 = .58$. Again, the STSE in RT was slightly more pronounced in trials of ascending numerical order, $r = .70$, compared to trials of descending order, $r = .52$. Divergent from the results of Experiments 8 and 9, the STSE did not moderate the order effect in this Experiment. Both effects, STSE and order, revealed the largest effect sizes in the series of experiments presented in this work. This could be explained by the directly processable magnitude information in contrast to the (contextualized) numerical magnitudes that had to be compared in the other experiments presented in this work. In Experiment 10, there was no interference with number processing or a applied context the responses had to be transferred into (compare Banks & Flora, 1977).

2.11 Experiment 11: Investigating the Underlying Mechanism

The purpose of Experiment 11 was to explore the underlying mechanism of the STSE by varying the position of the task’s instruction (compare Banks & Flora, 1977; Holyoak, 1978; Agostinelli et al., 1986). To find out at which level of information processing – stimuli encoding, mental arithmetic, decision making, response selection or response execution – the sequence of standard and target provides an processing advantage, the first step was to locate the effect on the encoding level or the response level. For this purpose, the experimental setting was modified in the tradition of expectancy theory (Banks & Flora, 1977), reference point account (Holyoak, 1978) and the feature matching model (Agostinelli et al., 1986), to name only a few examples: The comparative instructions were randomly altered from trial to trial, placing them at the end of each trial. This approach follows the assumption, that the participants could not engage in instruction specific encoding. If the STSE did not occur under such conditions, this would be interpreted as evidence for the previously reported effect (see Experiment 1, 2, 3, 5, 6, 8, 9 and 10) to occur on the encoding level. If the STSE still occurred under such conditions, this would be interpreted as evidence for the effect originating on the response level. If the effect of the sequence of standard and target reversed in such an experimental set up, in terms of a performance advantage for 'target preceding standard' trials, this would indicate a hybrid mechanism of encoding and working memory processes similar to the feature matching model (Tversky & Gati, 1978; Agostinelli et al., 1986), providing insight into the mental maintenance of the order of the to be compared stimuli. The inversion of the STSE would confirm Agostinelli et al.'s (1986) finding: when comparative instructions were placed at the end of stimulus encoding, the stimulus presented last became the standard of the comparison.

2.11.1 Method

There was no preregistration for the exploratory study realized in Experiment 11. The sample size was set according to the previous power analyses and an estimated high drop off rate of cases due to high error rates in the task employed in Experiment 11. The task, in which two stimuli had to be remembered (see *Procedure*), was considered to be more difficult than the task in that only one stimulus per trial had to be remembered (see Experiments 6 to 10).

Subjects. 137 students, 106 women, 28 men and 3 of diverse gender identification, participated in a 15 minutes test battery at the campus of the University of Cologne. Their mean age was 24.

Materials. Materials of Experiment 11 equaled Experiment 6, except that target and standard did not differ in their physical appearance. They were marked by the instruction comparable to the procedure in Tversky & Gati (1978), that varied from trial to trial indicating which of the two sequentially presented stimuli was the target and which the standard. A comparative construction for a trial could by for example, “Is the first stimulus bigger than the second?”, indicating that the first stimulus was the target and the second was the standard. Stimuli and comparative instructions were presented in the same font in the middle of the screen as in Experiments 2 to 9. Figure 40 illustrates an example for a trial of Experiment 11. Like in Experiment 6, stimuli were one digit numbers ranging from 1 to 9 and systematically paired avoiding ties. There were 72 possible pairings and 144 trials in total because every possible pairing was presented once in the ‘standard preceding target’ (ST) within participant condition and once in the ‘target preceding standard’ (TS) condition (compare 4.6.2). Trials of both within participant conditions were presented randomized order.

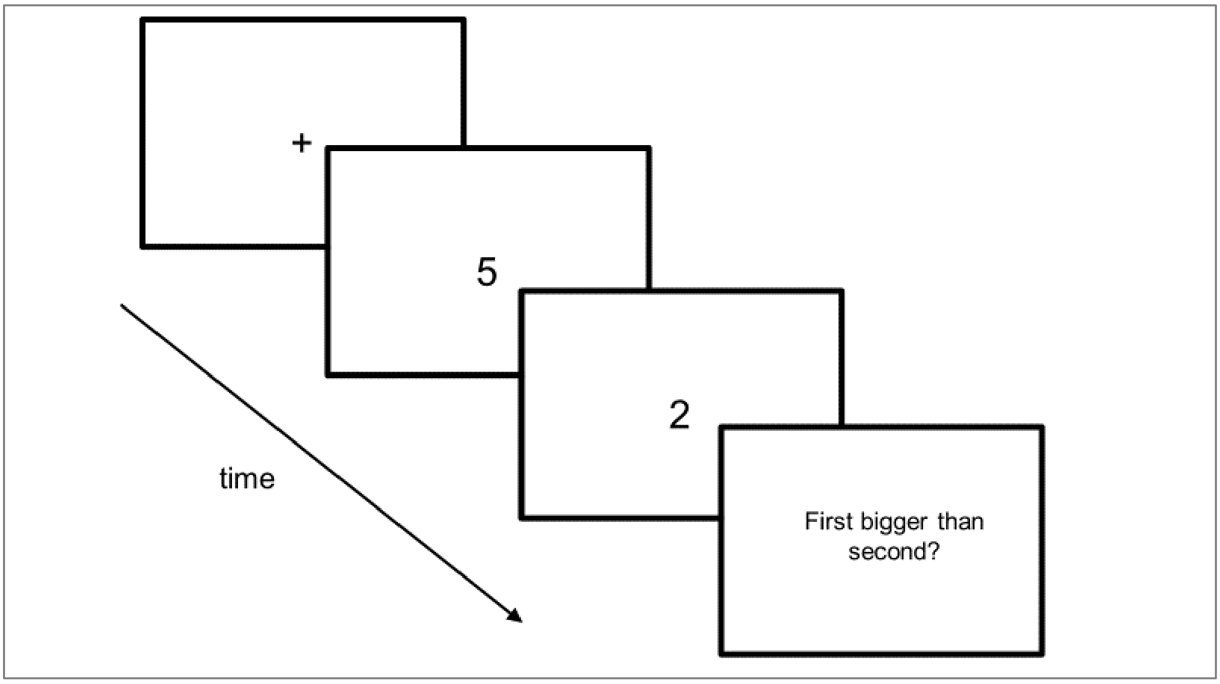

Figure 40. Example for a trial of Experiment 11. The size ratio of stimulus and frame is altered here compared to the laboratory setting in favor of improved readability.

Procedure. Before the task started, the instructions were displayed. Participants were told that in the following experiment they would need to repetitively engage in comparisons of two sequentially presented numbers and that they would have to judge the size of one number relative to the other as fast and as correct as possible. Which one of the two numbers had to be judged would always be displayed right after the two to be compared numbers had been displayed. A trial started with a fixation cross for 500 ms, followed by the first stimulus for 1000 ms and then the second stimulus for 1000 ms. The following comparative question for each trial appeared immediately after the second stimulus had disappeared and stayed on the screen until the participant had indicated their response. RT was measured form the onset of the comparative construction of each trial. The allocation of the participants into the four between participant groups, that resulted from the permutation of the two counterbalancing factors Key Assignment (right CTRL key indicating ‘smaller’ vs. right CRTL-key indicating ‘larger’) and the magnitude marker in the instruction (larger vs. smaller), was randomized. Like in Experiments 6 to 10 there was a test phase consisting of 5 trials with feedback on erroneous responses.

2.11.2 Results

Seven cases did not meet the inclusion criterion of 70% correct and therefore were excluded from further analyses, RT and error rates were statistically analyzed with a remaining sample of 130 cases.

Trials with an RT that deviated more than 2.5 times the standard deviation from the mean were discarded from further analyses. This concerned 1.4 % of all trials. The mean RT of the trials, that were included in the data analyses, was *M* = 1,839.76 ms (*SD* = 1,524.71 ms).

A next step of data cleaning was the deletion of trials with response primes. Response primes were defined as trials that started with a 1 or a 9.

Error rates. 9.2% of all trials were erroneous. To check for the STSE on error rates measured within participants, ST trials and TS trials were compared in a paired t-test. The STSE was significant, $t(129) = 2.38$; $p = .019$, $d_z = .21$. To control for the effect of numerical order, a 2 by 2 rmANOVA (Sequence x Order) was run. Sequence, as in the t-test, showed a significant effect, $F(1,129) = 5.64$, $p = .019$, $\eta_p^2 = .04$, while order had a highly significant effect, $F(1,129) = 33.79$, $p < .001$, $\eta_p^2 = .21$, the interaction of both factors was highly significant as well, $F(1,129) = 15.47$, $p < .001$, $\eta_p^2 = .11$, and had the largest of all three effect sizes. The interaction plot, displayed in Figure 41, and the simple effects analyses revealed that there was an overall performance advantage for descending numerical order, however, it was only significant within the TS trials, $F(1,129) = 57.31$, $p < .001$, $r = .55$. The STSE was only significant in the ascending order trials, $F(1,129) = 15.88$, $p < .001$, $r = .33$.,

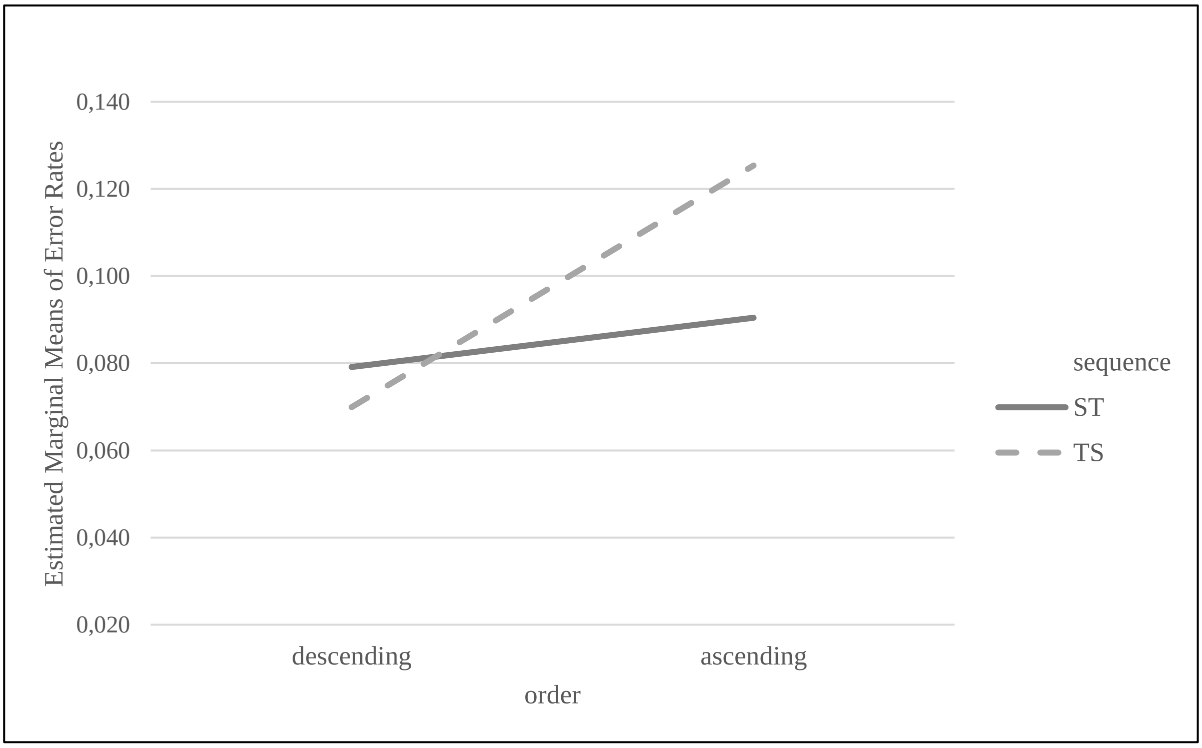

Figure 41. Interaction plot of a 2x2 rmANOVA of error rates in Experiment 11.

and even turned, but not significant, in the descending order trials. So, order modulated the sequence effect.

The data analyses for the STSE and the order effect revealed data patterns that indicated an effect of the targets' relative magnitude on the error rates: larger targets than standards were descriptively responded to more often correctly. This effect was tested controlling for its interaction with the magnitude marker (between participants) given in the instructions. This was tested in a 2 by 2 mixed rmANOVA with the within participant factor Relative Target Magnitude (RTM, 'larger than standard' vs. 'smaller than standard') and the between participant factor Magnitude Marker (MM, larger vs. smaller). RTM, had a significant effect with a small effect size, $F(1,128) = 6.02$, $p = .016$, $\eta^2 = .06$, MM had a highly significant effect with a small effect size, $F(1,128) = 6.56$, $p < .001$, $\eta^2 = .04$, the interaction of both factors was not significant, $F(1,128) = 2.97$, $p = .09$. Condition means are displayed in Table 3.

Table 3

Condition Means of Error Rates (Exp. 11)

	MM smaller			MM larger		
RTM	*n*	*M* (*SD*)	95% CI	*n*	*M* (*SD*)	95% CI
smaller	64	.10 (.007)	[.08, .11]	64	.11 (.007)	[.09, .12]
larger	64	.08 (.008)	[.07, .10]	64	.08 (.01)	[.06, .09]

Note. CI = confidence interval.

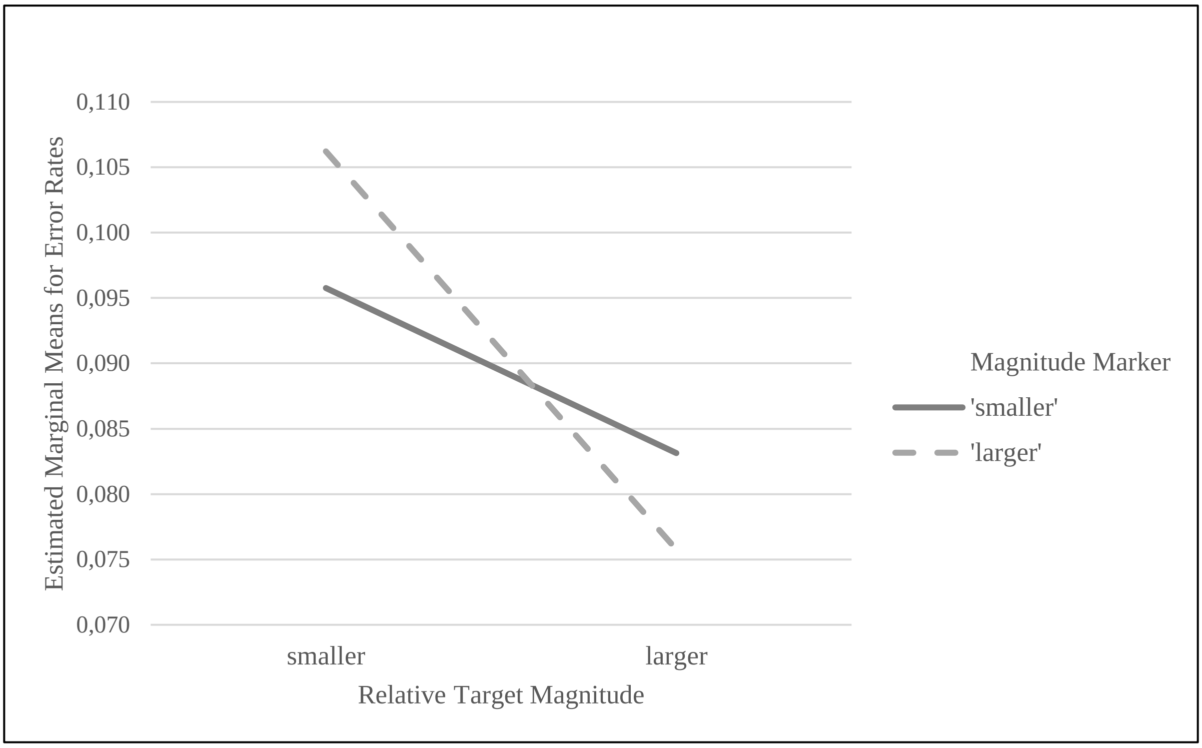

Figure 42. Interaction plot of a 2x2 mixed rmANOVA of error rates in Experiment 11.

The interaction plot in Figure 42 reveals a not significant but descriptively present congruency effect of larger targets being detected more often correctly in ‘larger’ instructions.

RT. For analysis of RT the erroneous trials were discarded, and the natural logarithmized RT scores were used for the parametric tests. The paired t-test for the STSE revealed no difference of mean RT between ST and TS trials, *t*(129) = 0.21, *p* = .835. The 2 by 2 rmANOVA (Sequence x Order) resulted in no STSE, *F*(1,129) = 0.02, *p* = .879, a just not significant effect of numerical order, *F*(1,129) = 3.52, *p* = .063, and a highly significant interaction, *F*(1,129) = 8.78 , *p* < 0.01, η_p^2 = .06. Simple effect analyses revealed that the effect of sequence was significant in both factor levels of Order, but in opposite directions (see interaction plot in Figure 42). Within the numerically descending trials, there was a TS advantage, *F*(1,129) = 5.55, *p* = .020, *r* = .20, while in the numerically ascending trials, there was a ST advantage, *F*(1,129) = 5.12, *p* = .025, *r* = .20. The effect of order was only

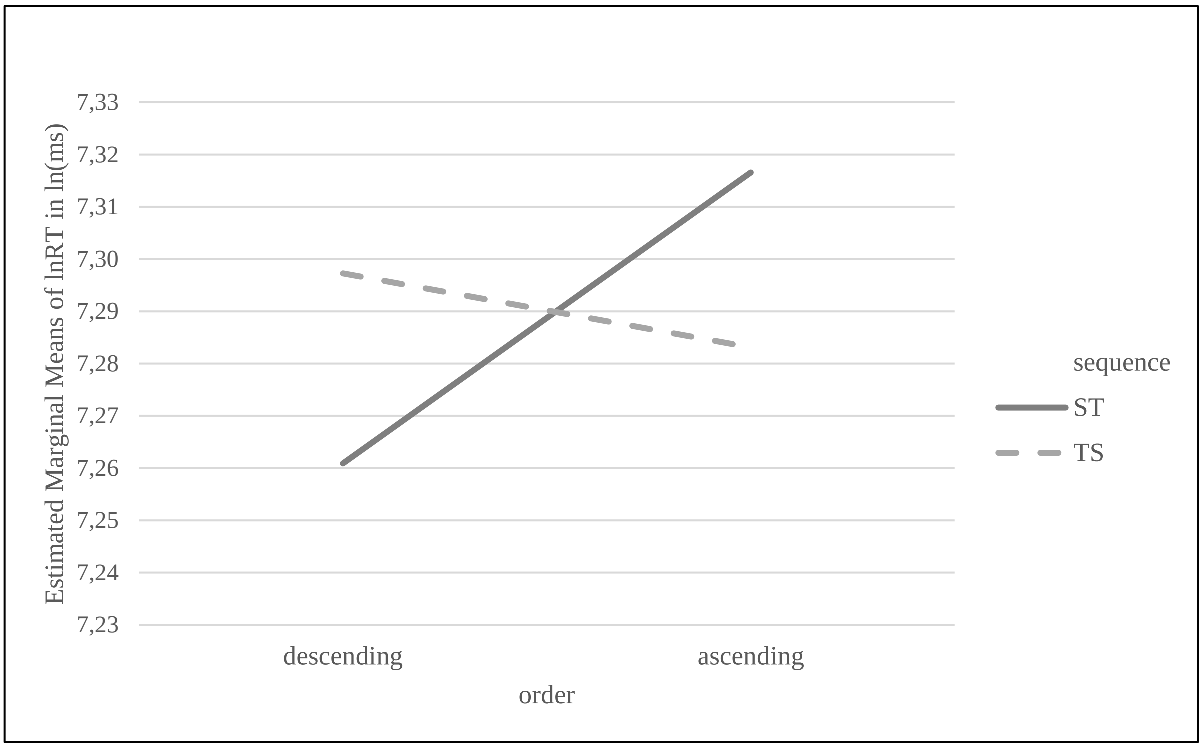

Figure 43. Interaction plot of a 2x2 rmANOVA of RT in Experiment 11.

significant in the ST trials, $F(1,129) = 16.10$, $p < .001$, $r = .33$, in terms of a descending order advantage, but not within the TS trials, $F(1,129) = 0.61$, $p = .437$.

Like in the analyses for error rates the data pattern of RT suggested a main effect of the RTM, here in the opposite direction: smaller targets than standards being responded to faster. This effect was tested controlling for its interaction with the magnitude marker given in the instructions. This was tested in a 2 by 2 mixed rmANOVA with the within participant factor Relative Target Magnitude (RTM, ‘larger than standard’ vs. ‘smaller than standard’) and the between participant factor Magnitude Marker (MM, larger vs. smaller). RTM, had a significant effect with a small effect size, $F(1,128) = 16.01$, $p < .001$, $\eta_p^2 = .11$, MM had a highly significant effect with a small effect size, $F(1,128) = 16.01$, $p < .001$, $\eta_p^2 = .11$, the interaction of both factors was highly significant as well, $F(1,128) = 135.92$, $p < .001$, $\eta_p^2 = .52$. Condition means are displayed in Table 4.

Table 4

Condition Means of Reaction Times (Exp. 11)

	MM smaller			MM larger		
RTM	*n*	*M* (*SD*)	95% CI	*n*	*M* (*SD*)	95% CI
smaller	64	2,022 (765)	[523; 3521]	64	1,443 (505)	[453; 2433]
larger	64	1,892 (684)	[551; 3232]	64	1,626 (511)	[624, .2628]

Note. Reaction time in ms. CI = confidence interval.

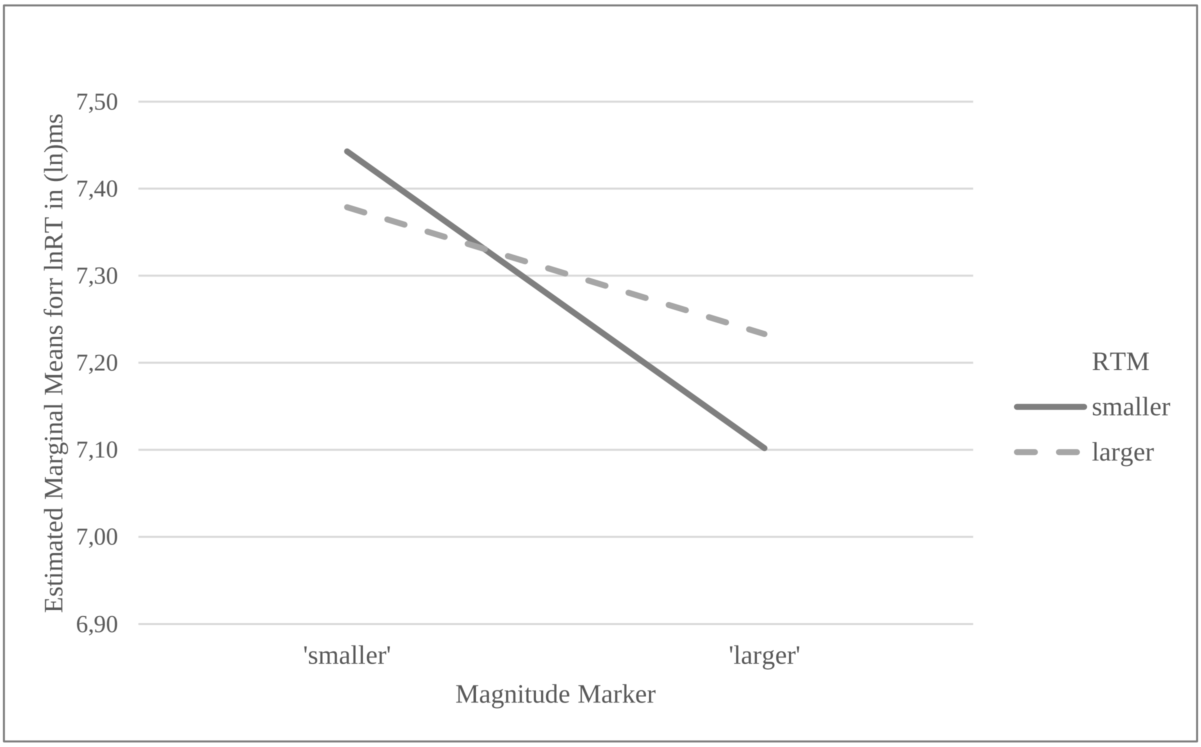

Figure 44. Interaction plot of a 2x2 mixed rmANOVA of RT in Experiment 11.

The interaction plot in Figure 44 reveals an inverted congruency effect of larger targets being detected more often correctly in ‘smaller’ instructions.

*2.11.*3 *Discussion*

In Experiment 11, the comparative instructions were given after each trial to avoid the encoding of the to be compared numbers as standard and target. For RT, the interaction of order and sequence reached significance, a local STSE was found for numerically descending sequences, while neither the main effect of sequence nor the main effect of order was significant. The descriptively global STSE in error rates was moderated by order but reached significance only in the numerical ascending pairs. Descriptively a global performance advantage of numerically descending pairs was reported, only significant in error rates. While the effects of order and the interaction of sequence and order had the same direction in RT and error rates, a speed-accuracy-trade-off is likely to have happened within the target preceding standard trials. These ambiguous data patterns of the analyses of sequence and order effects pointed at a global speed-accuracy-trade-off with regard to an effect of the RTM (relative target magnitude) and a congruency of the MM (magnitude marker in the instructions) and the RTM (relative target magnitude). This could be supported statistically. For RT an inverted congruency effect was found, $\eta_p^2 = .52$, with a speed advantage for smaller targets was found, $\eta_p^2 = .11$, while for error rates a descriptively present classical congruency effect was found and a significant accuracy advantage for larger targets, $\eta_p^2 = .11$. A potential confound of the experimental design (see section 2.11.2) was that the magnitude marker in the instructions (`Was the first [second] number LARGER [SMALLER]?*')* was held constant between participants (detailed description of the experimental design in section 2.11.2). The stable MM between participants in Experiment 11 were utilized to ensure effects comparable to the within participants stable key assignments of Experiments 6 to 10 (left [right] key indicating smaller [larger] targets). This could have provoked systematic congruency coding, in terms of matching MM and RTM. In contrast to the other experiments presented in the work at hand, the RT in Experiment 11 correlated with the trial number, r $_{RT*Trial}$ = -.30, *p* = .002, indicating that the RT became lower as the task preceded (for correlation analyses for RT for selected experiments see Appendix B3). According to Relative Judgment Theory (Link & Heath, 1975), this supports the notion of a systematic response behavior that is improved by training.

The reversed congruency effect in RT indicated that participants reacted faster on targets that did not meet the magnitude marker in the post-encoding instructions. Therefore "no" answers were given faster: participants were faster in detecting numerical sequences that did not fit the comparative sentence of the comparative instructions. The general advantage for error detection in 2AFC task could potentially be explained by the error detection system involved reinforcement learning, recently discussed in behavioral neuroscience discussed (e.g. Hoffmann & Beste, 2015). The error detection advantage could possibly indicate a response strategy, especially primed by the training trials in advance of the experimental task with error feedback. As RT were analyzed only for the correct trials, the speed advantage for incongruent MM and RTM is only evident within the correctly performed trials. In error rates a performance advantage for larger targets than standards was reported. The RTM effect and the effect for order showed equal effect sizes (η_p^2 = .11) in separate analyses. An even larger main effect was observed for order in terms of an overall descending order advantage in the error rates (η_p^2 = .22). To draw conclusions on the data pattern of error rates – whether it indicates a speed-accuracy-trade-off or an (compared to the previous gathered data) inverted interaction of sequence and order – further studies are needed.

Appendix B4 provides an overview of the variance analyses of Experiment 11. Besides the effects mentioned above, the analyses revealed that Sequence interacted significantly with Response Side (a factor that had been counterbalanced between participants) – the only effect that was consistent between RT and error rates. Participants reacted significantly faster with their right hand on standard-preceding-targets trials, whilst on target-preceding-standard trials they reacted faster with their left hand. This effect had a medium to large size in error rates, η_p^2 = .14, and a large effect size in RT, η_p^2 = .21. A SNARC-like congruency effect for the interaction of MM and Response Side could not be found.

3. Summary of Results

The results presented in Chapter 2 provide empirical evidence for a performance advantage for the sequence of standard preceding target in speeded 2AFC magnitude comparisons of prices, mere numbers, multi-digit scores on diverse dimensions and of the volume of geometrical figures. Performance was measured in RT and error rates. The STSE was stronger in RT, ranging from d_z = .38 in Experiment 6 (numerical comparisons) to d_z = 1.16 in Experiment 10 (comparisons of volume), than in error rates, ranging from d_z = .22 in Experiment 6 to d_z = .90 in Experiment 1 (vertical bargains with intense marking of standards). Besides the STSE the numerical order of the sequential stimuli affected participants` performance and modulated the STSE.

3.1 Tests of Hypotheses

$H_{1},$ assuming an enhanced bargain detection for paired prices fitting a SNARC compatible presentation format, had to be rejected. The hypothesized effect, derived from the findings of vertical SNARC effects in magnitude comparisons (Itto & Hatta, 2004; Ben Nathan et al., 2009; Petrusic et al., 2011) indicating a spatial numerical association of larger numbers presented above smaller numbers, could not be found. Instead, the results of Experiment 1 revealed patterns in both performance outcomes that pointed at a processing advantage for sequences of standard (previous price) standing above target (actual price) independent from the numerical relations of the two stimuli (RT: d_z = .42; error rates: d_z = .90).

Adapted to the results of Experiment 1, H_2 assumed a processing advantage for the sequence of standard preceding target in an experiment where previous and actual sales prices were presented sequentially instead of simultaneously. Differing from Experiment 1, the stimuli in Experiment 2 were one-digit prices, noise-reduced to their basic magnitude information. In Experiment 1, standards had been intensely marked by red crosses and the prices had been displayed underneath picture of a product, with two decimal places each. *H_2* was supported by the results of Experiment 2, with a smaller effect size in the error rates analysis, η_p^2 = .42, than for RT, η_p^2 = .82. The results were defined as an STSE (standard-target-sequence effect).

H_3 , assuming the STSE was evident for spatial arrangements of bargains, was partly supported. Experiment 3 investigated a horizontal arrangement of standard prices on left (right) positions and targets in right (left) positions. Experiment 4 investigated a vertical arrangement of standard prices von top (bottom) position and targets in bottom (top) positions. The noise-reduced stimuli from Experiment 2 were employed, but presented simultaneously instead of sequentially. Experiment 3 revealed a smaller STSE in RT than in Experiment 2, η_p^2 = .25, and no STSE in error rates. In Experiment 4, no STSE, neither in RT, nor in the error rates, was found; only the Bargain factor revealed significant effects. Hence, *H_3* was supported for horizontal price arrangement and was rejected for vertical price arrangement. Besides that, the results of Experiments 3 and 4 supported the rejection of *H_1* – in spatial arrangement of prices no SNARC-like effect was found. The divergent results of Experiments 1, 3 and 4 should be discussed in terms of their diagnostic value for the underlying mechanisms (see sections 4.1.1 and 4.1.4).

In order to test the STSE against a performance advantage for the familiar temporal sequence of ‘yesterday followed by today’, the next hypothesis was formulated. *H_4* assumed a performance advantage for sequences of a future standard followed by an actual target compared to sequences of a future target followed by an actual standard. The results of Experiment 5 partly supported *H_4*: the STSE for prospective bargains could be found for RT, η_p^2 = .25, but not for error rates.

Regarding the results of Experiments 1 to 5, the bargain-judgement-paradigm appeared to be insufficient to disentangle STSE-modulating effects of numerical order and the interaction of bargain and sequence. From Experiment 6 on, paired magnitude comparisons were tested for the STSE beyond the context of price comparisons. H_5 assumed the STSE in numerical comparisons in general. It was tested in Experiment 6 and 7 for contextless numbers (one-digit to four-digit numbers) and in Experiments 8 and 9 for numbers with various contextual frames. H_5 was supported for RT, with a medium effect size ranging from d_z = .38 to .50 and a small to medium effect size for error rates ranging from d_z = .22 to .30. There was only one exception: In Experiment 7, where stimuli were completely randomized four-digit numbers, the STSE could not be found in error rates. The exceptions of the STSE should be discussed with regard of the limits of the STSE and its underlying mechanism (see section 4.1).

To generalize the findings of the STSE in numerical comparisons to magnitude comparisons in general, H_6 assumed an STSE in the comparison of volumes of quadrates. Experiment 10 was designed to test H_6, keeping the design of the previous studies and employing squared frames of various volumes as stimuli. The results supported H_6 and revealed the highest effect size of the STSE throughout the series of experiments of the work at hand for RT, d_z = 1.16, and a large effect for error rates, d_z = .52.

3.2 The Interaction of the STSE with other Effects

3.2.1 Order Effects

An ascending order advantage could be reported in both performance parameters in the experiments that investigated contextualized number comparisons and the volume of geometrical figures (Experiment 8: η_p^2 = .06 in error rates, η_p^2 = .10 in RT; Experiment 9: η_p^2 = .12 for error rates, η_p^2 = .16 for RT; Experiment 10: η_p^2 = .07 for error rates, η_p^2 = .20

for RT). These order effects in error rates were moderated by the sequence of target and standard. Within TS-trials a descending order advantage was observed.

In Experiment 6, the ascending order advantage revealed only descriptively in the RT and a local descending order advantage in the target-preceding-standard-trials was observed in the error rates, whereas in the standard-preceding-target trials order had no effect. A speed-accuracy-trade-off for order could be observed, while participants responded faster on ascending orders, they made more mistakes in these conditions compared to pairs of numbers of descending order.

In Experiment 7, the order effect only reached significance in error rates, $\eta_p^2 = .06$. The ascending order advantage suppressed the STSE effect, whereas a local STSE was reported in the in the descending numerical pairs.

In the experiments that employed the bargain-judgement-paradigm (Experiment 1 to 5) the ascending order advantage could not be found. In Experiments 2 and 3 a local descending order advantage was reported in ST-trials. An overall main effect for descending numerical order was reported in Experiment 5 for error rates. The interaction of sequence and order in Experiments 1 to 5 could not be disentangled from the factor Bargain, that was the explicit decision criterium for correct responses.

In Experiment 11 an interaction of order and sequence was observed indicating an opposite hierarchy of the effects compared to the interaction effects reported in Experiments 6 to 10 in error rates. The inversion of the previous reported interaction of order effect and sequence effect has to be discussed (see section 4.2). The results of Experiment 11, with regard to the interaction of order and sequence, may be of limited diagnostic value due to a conceptual confound that enabled the participants to employ a responding strategy according to error detection (see section 2.11.3).

3.2.2 Effects of Response Side, Congruency and Distance

In the first ten experiments presented in the work at hand response side effects and SCE had been prevented with the help of between participant counterbalancing factors. RT and error rates included in the data analyses had been aggregated over all participants to avoid a suppression and/or modulation of the effects of numerical order and sequence by the classical discontinuities in symbolic magnitude comparisons (see section 1.2). Examples for the instruction texts can be found in Appendix A. An exemplary analysis controlling for the response side effect (as defined by Dehaene et al.,1990) can be found in Appendix B4.

In Experiment 11, where Magnitude Marker had not been counterbalanced between participants and a congruency effect in the error rates was observed. It was modulated by the effects of sequence and order (for a discussion see section 2.11.3), while order had the largest effect on response patterns (indicating a descending order advantage)

Exemplary correlational analyses for the distance effect are provided in Appendix B3.2. For all experiments presented in Chapter 2 no distance effects could be found.

4. General Discussion

The empirical part of the work at hand reports a series of studies that combined aspects of different lines of research that were presented in Chapter 1 – an effect known form DL measurements in psychophysics and similarity judgements of complex stimuli (Tversky & Gati, 1978; Agostinelli et al. 1986, see section 2.4) was reported for numerical and symbolic magnitudes in speeded pairwise comparisons. The STSE, in psychophysics referred to as the negative Type B effect, was defined as an increased discrimination performance for standards preceding targets, was reported in speeded 2AFC tasks with sequentially presented one digit numbers, three-digit numbers, range dependent four-digit numbers, prices and geometrical figures. In most of the experiments the effect was more pronounced in RT than in error rates. This suggests it might be an effect of intuitive decision making rather than elaborate processing. In tasks that could be performed intuitively without engaging in higher level comparative processing (comparisons of volume of geometric figures) and had an applied contextual frame (judgement of bargains fitting realistic marketplace adverts, comparisons of persons IQ scores or monthly income) a robust STSE in error rates and RT was found (compare Houston et al., 1989). Several findings supported the notion that the STSE originates at the encoding stage of the comparison process. First, the more direct the stimuli could be encoded as standard and target, the more pronounced the STSE was. Second, the effect size of the STSE revealed to be modulated by the relative saliency of standard and target during task performance (Experiments 1 to 10). Third, in Experiment 7, where encoding exceeded the mental rehearsal span of working memory, and in Experiments 11, where position-based magnitude encoding was not possible, the STSE was marginalized.

Additionally, next to the introduction of the STSE, interesting new perspectives of the processing of order information were pointed out. First, the ascending order advantage in pairwise magnitude judgment tasks was modulated by the sequence of standard and target when rather elaborate processing was required (Experiments 5, 6 and 7). Second, a performance advantage for descending numerical order was found in a task where comparative instructions were given after encoding of stimuli (Experiment 11).

4.1 The Mechanism Underlying the STSE

To conclude on the origin of the STSE, the results summarized in Chapter 3 are interpreted with regard to the experimental variations that influenced the STSE. The STSE-enhancing or suppressing contextual factors can be identified by comparing the effect sizes and the changes of the effect hierarchy between the different experimental set ups.

4.1.1 Intuitive Decision Making and Relative Saliency

The STSE in the empirical work of this dissertation was especially pronounced in experiments that required less elaborate processing or mental arithmetic and rather encouraged intuitive decision making (Experiments 8 to 10). Among these, the STSE unfolded the largest effect size in the RT (d_z = 1.16) and a large effect size in error rates (d_z = .52) of Experiment 10, where direct magnitude information in terms of volume of quadrats had to be compared. In tasks that required rather elaborate processing (like bargain judgment tasks, comparisons of more digit numbers or instructions presented after stimuli) other cues – not the sequence of target and standard – or higher-level decision criteria were likely to be used by the participants. More difficult tasks produced smaller STSE effects and partly ambiguous data patterns of RT and error rates. For example, in Experiment 5, influent temporal relations of standard and actual prices (future standard, present target) might have required more elaborate processing weakening the STSE in the RT and revealing no STSE in the error rates. In contrast, the most naturalistic frame of bargains, used in Experiment 1, produced the largest effect sizes of the STSE for the bargain-judgement-paradigm (RT: d_z = .42; error rates: d_z = .90), albeit smaller (in RT) than in the pairwise number comparisons

(Experiments 6 to 9). This supports the assumption that the STSE is especially relevant in pairwise comparisons where a lesser degree of mental arithmetic, mental rearrangement and mental assorting of standard and target had to be done.

Experiments 1 to 5, in general, required a more elaborate processing in terms of an additional reframing of the encoded magnitude- and position information, because the correctness criterium was the detection of bargains. Therefore, the error rates and RT of the experiments employing the bargain-judgment-paradigm should not be put in direct reference to the performance of the tasks to judge the targets magnitude.

The assumption that the STSE is especially relevant for tasks of intuitive decision making was supported by the observation that the effect was especially pronounced in and more frequently reported for RT, although the participants had been instructed to respond as fast and as accurate as possible. This indicated an increased ease of processing of standard-target sequences. This needs further investigation, e.g. studying the processing fluency (Topolinski & Strack, 2009) of standard-preceding-target sequences (compare Tversky & Gati, 1978).

Exceptions from the more pronounced STSE in RT compared to error rates were observed in Experiments 1 and 11. The large effect sizes reported in Experiment 1 are especially interesting because, in a very similar setting, in Experiment 4 (vertical stimuli arrangement, both stimuli present at the point of responding) no STSE was found. The crucial difference of Experiment 1 and 4 was the intense, differentiating marking of standard and target in Experiment 1. In Experiment 4, stimuli were visually reduced to their essential magnitude information. The possible explanation, that the different saliency of standard and target in Experiment 1 could have promoted the STSE, corresponds to the focusing hypothesis causing an asymmetry of difference- and similarity judgements proposed by Tversky (1977;

Tversky & Gati, 1978). This also stresses the assumption that the STSE, reported in this work, originated at the encoding level of the comparisons. Furthermore, in Experiment 11, where the stimuli could not be encoded as target and standard, there was no clear evidence for the STSE. Instead, participants seemed to have employed another (implicit) heuristic to respond as fast and as accurate as possible (see section 2.11.3).

4.1.2 Direct Processing

The STSE was enhanced when standard and target were visually marked during encoding (Experiments 1, 6 to 10), without having to conclude on semantic meanings of markers (e.g. experiments employing the bargain-judgement paradigm). The shortest RT and lowest error rate of all experiments were found for the comparisons of volume of quadrates, target and standard being marked by a bold (thin) surrounding line (Experiment 10). This confirmed the assumption of other researchers and theorists, that stimuli, which can be processed directly, have speed and accuracy advantages compared to symbolic stimuli (e.g. Dehaene, 1989). Banks & Flora (1977) reported shorter RT and higher accuracy for pictures, compared to words, in pairwise magnitude comparison task. They concluded that the more direct the magnitude information could be extracted from the stimuli, the easier the comparison could be performed. The finding that the STSE was even more pronounced directly perceivable magnitudes stressed the assumption of the effect happening at the encoding stage of the comparison process.

4.1.3 Working Memory Load

Some of the presented findings argue for an impairment of the STSE under increased working memory load. For instance, divergent findings regarding the effect of sequence on speed and accuracy of four-digit number comparisons were reported. In Experiment 7, where the STSE was absent in error rates, a complete variation of the to be compared numerical stimuli was employed, while in Experiment 9, with a limited range of four-digit stimuli, a strong STSE was reported. It can be concluded that in Experiment 9 participants' response behavior could be automatized to a higher degree (compare section 4.1.1), because numbers only varied in the second digit while the other digits were held constant. On the opposite, comparisons of four-digit numbers with completely randomized digits required more elaborate processing and working memory capacity.

In Experiment 11, where both stimuli of a pair had to be remembered to mentally arrange them in the direction of post-encoding instructions, the STSE was not significant, at least suppressed by other effects. The performance advantage (lower error rate) for larger targets could have potentially resulted from an interaction of order and sequence, or could indicate a size effect – previously reported only for RT (Henik & Tzelgov, 1982; Krajcsi et. al., 2018). Therefore, based on the results of Experiment 11, one cannot discard a potential origin of the STSE in post-encoding working memory operations. However, because of the conceptual confound of Experiment 11, the diagnostic value of the absence of the STSE in paired magnitude comparison task with higher working memory load is limited.

4.1.4 Grounded Cognition Approach

There were contradicting findings for the influence of simultaneous (not sequential) stimulus presentation on the STSE. The STSE did not show up in Experiment 4 with vertical simultaneous stimulus presentation and in Experiment 3 with horizontal stimulus presentation. only in the RT, modulated by order. The setting of Experiment 3 fitted the reading direction of western languages, which could explain why the sequence of standard presented on the left side and target on the right side was facilitated comparable to the sequential presentation of prices as in Experiment 2. Taking the perspective of the grounded cognition approach on this finding, the sequence of standard preceding target could be associated with the spatial dimension from left to right just like time and magnitude (Walsh, 2003; Müller & Schwarz, 2008; Huber, Klein, Moeller & Willmes, 2016). This approach would be an alternative explanation or extension of the ease of processing-hypothesis for the performance advantage for standard-preceding-target sequence (see section 4.1.1). In Experiment 11, an interaction of sequence and response side was found for both performance outcomes (see Appendix B4). The reaction in standard-preceding-target trials was faster and more often correct with the right response side (right hand responses), while it was faster and most accurately with the left side in the target-standard-sequence trials. This supports a grounded cognition approach, in addition to the ease of processing hypothesis, in terms of an association of standard-preceding-target sequence with proceeding from left to right. Referring to the results reported in Experiment 4, simultaneous vertical stimulus arrangement appears to be a boundary condition of the association between standard-target-sequence, magnitude and space. The observation that the effect of ascending order and the STSE had the largest effect sizes in settings that provoked both effects, might indicate shared underlying mechanisms (discussed in section 4.2) and stresses the assumption of the sequence of standard and target being an implicit sequence of progress into time and space (just like ascending numerical orders).

4.1.5 Relative Encoding and Counterfactual Operation

Summarizing the interpretation gathered from the boundary conditions and modulating contextual effects mentioned above, the STSE is likely to occur at the encoding stage of the comparison processes and to depend on working memory capacity. The flipside of the effect is the impaired comparative processing of and responding to target-standard sequences, possibly resulting from a counterfactual operation. In all experiments presented in the work at hand, participants were aware during encoding that they would have to engage in a comparative judgement from trial to trial and that they would have to assess the stimuli in relation to each other. In experiments where the STSE was observed, targets and standards were differently marked, identifiable directly during encoding. To perform the task correctly, the absolute magnitude of the standard had to be assessed before the magnitude of the target in relation to the standard could be judged. In a trial, the stimulus encountered first could not be assessed relatively because no previous information about the size of the other stimulus was given. Hence, the first stimulus could only be encoded absolutely; while the second stimulus of a pair was encoded relatively. While in standard-preceding-target trials the answer could be given according to this modus of encoding, in target-preceding-standard trials the extra operation in working memory of mentally going back to the target, to perform the relative magnitude judgement in the right direction, costed speed and accuracy.

In Experiment 11 stimulus encoding could not depend on the target-standard-assignment, because the direction of comparison was only determined after stimulus encoding. As here, the STSE could not be found, participants seemed to engage in a different processing mode. The congruency effect, reported in the error rates of Experiment 11, suggests that participants first engaged in undirected relative magnitude encoding of both stimuli and then, according to the instructions, mentally arranged the stimuli due to the direction of comparison. No sequence of standard and target had a processing advantage. Corresponding to the reference point theory of Holyoak (1978; Chen et al., 2014), accounting for the SCE in undirected pairwise comparisons, the reference point set by the constant magnitude criterium in instructions was the only available anchor during stimulus encoding that could help to perform the task as fast and as accurate as possible. In contrast to that, in Experiments 1 to 10, except for Experiment 4, position based relative magnitude coding seems to have caused the ST-advantage in RT and error rates. This hypothesis, in turn to the reference point theory, corresponds to the SPBD account in serial decision tasks (Jou et al., 2020) and order decision tasks (Turconi et al., 2006). In the tasks employed in the empirical part of this work, this process does not lead to higher accuracy under increased working memory load (like in Experiment 7), because the capacity needed for magnitude encoding and information maintenance was exceeded.

The STSE was observed in tie trials as well, with lower RT and higher error rates over all conditions, this supports the assumptions that speed and accuracy costs in target-preceding-standard trials were based on a counterfactual operation and the assumed process hierarchy of magnitude coding depending on position coding.

4.2 Competing Effects of Order and Congruency

In Experiment 6 to 10 an ascending order advantage was observed, which had been reported before for diverse experimental settings of pairwise magnitude comparisons (see section 1.3.3). Former studies reported an ascending order advantage for a (memorized) numerical selection task with horizontal stimulus arrangement (Turconi et al., 2006), as well as for a magnitude selection task with sequentially stimuli presentation and the recent stimuli being present during responding of an ascending order advantage (Müller & Schwarz, 2008). And in directed pairwise magnitude comparisons with vertical stimulus presentation and standards in top positions (Ben Nathan et al., 2009). The results presented in this work do not only contribute to a further generalization of the ascending order advantage, but also to the definition of boundary conditions.

The ascending order advantage revealed its largest effect sizes in experiments in that the STSE was especially pronounced (see section 3.2.1) – namely conditions with contextual frames that stressed the magnitude comparison and the comparison’s direction during encoding. This suggests that both effects share underlying mechanisms or at least require shared capacities. A boundary condition of this assumption appeared to be working memory load, as in the error rates of Experiment 7 the ascending order advantage was observed but no significant effect of sequence. And the opposite data pattern was observed in the error rates of

Experiment 6, where the lowest working memory capacity was engaged in task performance. The impaired task performance under increased working memory load could be caused by an impaired magnitude coding, which is most relevant for the order effect. Taking this interpretation valid, would support the notion for the STSE being an effect on the encoding stage, because working memory influences on task performance in the empirical part of this work were caused by impaired magnitude coding.

In Experiments 1 to 10 the STSE moderated the order effect, stressing the position-based magnitude coding introduced in 4.1.5 and the process hierarchy of magnitude coding depending on position coding. In Experiment 11, where position-based magnitude coding was not possible, this moderation effect turned: order moderated sequence. for RT, in descending trials the STSE showed up, whereas in ascending pairs the reverse sequence effect was found. The diagnostic value of these results with regard to the modulating effects of numerical order might be limited as the data pattern in RT could be explained best by an advantage for error detection (see section 2.11.3). For error rates a general descending order advantage was found, the difference between standard-preceding-target and target-preceding-standard trials was only significant in numerically ascending pairs. In this experiment, participants most likely engaged in magnitude instead of position-based magnitude coding, increasing the influence of numerical magnitude information during encoding.

Turconi et al. (2006) and Müller and Schwarz (2008) concluded on different processes of order judgements and numerical magnitude comparisons. While the assessment of numerical order required serial search within the paired stimuli of a trial, magnitude comparisons required magnitude coding. In both studies, the authors focused on the interaction of order with other effects of pairwise magnitude comparisons. Turconi et al. (2006) found that the distance effect revealed only in numerically descending pairs, whereas comparisons numerically ascending pairs were performed best when they consisted of consecutive numbers. Turconi et al. concluded on a detection of consecutive ascending pairs (due to the familiar numerical order of two consecutive numbers) that had a performance advantage over magnitude processing. In Experiment 11 of the work at hand an ascending order advantage for consecutive pairs could be found as well whereas for non-consecutive pairs a performance advantage for descending number could be found. Taken together with the findings of Turconi et al. this supports the assumption that familiar orders are processed prioritized or are automatized to a higher degree than magnitude processing.

The comparative instruction given after encoding (Experiment 11) had not been realized in a study investigating effects of magnitude order so far. Under this condition the frequently reported order effect was inverted into a descending order advantage. In the experiments where the instructions had been given before encoding (6 to 10), there was a moderation effect of sequence on order. Müller & Schwarz (2008) reported that the effect of temporal numerical order was modified by the "specific task and response requirements" (Müller & Schwarz, 2008, p. 147) and that it moderated the SNARC effect and the SCE. Comparing the results of Experiment 10 to the experiments that investigated the effects of order and sequence with numerical stimuli, indicates that position-based magnitude encoding and order based magnitude coding especially compete for attentional and cognitive resources when numbers were used as stimuli. The main effects for order and sequence were the largest in Experiment 10 while the interaction of both factors did not modulate the predicted effects

In experiments employing the bargain-judgement-paradigm, especially in prospective bargains (Experiment 5), the order effect was reversed as well. In the error rates of Experiments 3 and 5 there was a highly significant order effect while Sequence, Bargain and their interaction revealed no significant effects. Nevertheless, in these experiments the order effect should not be interpreted without considering the task's instruction to detect bargains. According to the usual real marketplace definition of a bargain – a product that has a lower price today than the higher price of yesterday, a bargain was defined in the experiments as a descending numerical order from standard to target. The numerically descending temporal development of prices, that makes a bargain and which participants should look for to perform the task correctly, could explain the performance advantage of descending orders in the experiments employing the bargain-judgement-paradigm.

4.3 Limitations and Further Research

The promoting and inhibiting factors of the STSE that have been discussed above, need further empirical exploration to conclude on their diagnostic value.

The STSE was rather weak when elaborate processing was needed: in Experiment 7, where the absence of an STSE in error rates is hypothesized to be caused by a working memory load; in bargain judgement experiments where the STSE was weakened (compared to Experiment 6 to 10); in Experiment 5, where prospective bargains had to be judged, or in Experiment 11, where the comparison had to be performed on the basis of the memory traces of target and standard. An experiment that employs a classical dual task paradigm (compare van Dijck & Fias, 2011) should be conducted to test for the influence of working memory load on the STSE.

The role of the saliency of the standard for the STSE in pairwise number comparisons should be tested by systematically varying the saliency of the standard and the target using intense visual markers, comparable to the stimuli employed in Experiment 1, where only the standards had been marked intensely.

The degree of arbitrariness, that already had been theorized as a boundary condition for the SNARC (e.g. Ben Nathan et al. 2009; van Dijk & Fias, 2011) and for the bowed serial position effect (Shoben et al., 1989a) should further be explored for the STSE, as in Experiment 7, where a larger range of numerical stimuli was employed than in the other experiments, the effect could not be reported. For example, in an experiment that randomizes all numbers between 1 and 9999, controlling for the distance effect.

Agostinelli et al. in 1986, found a turn of the sequence of target and standard when the instructions were given after the encoding of the comparative stimuli. In their set up, they found that the second stimulus became the standard of the comparison. Besides the confound in Experiment 11, mentioned above (see section 2.11.3), a crucial difference between the repeated measurement design of Experiment 11 and Agostinelli and colleagues' study is that in Agostinelli et al.'s study, participants were completely naïve during encoding. In Experiment 11, participants knew that they would have to engage in pairwise comparison during encoding and according to this, engaged in relative encoding of the stimuli. The comparison of study designs of Experiment 11 and the experiments of Agostinelli et al. discloses a conceptual boundary of the repeated measurement design with magnitude comparisons being performed one after another: the naivety of participants cannot be maintained throughout the experiment. Therefore, in the tradition of Agostinelli et al. (1986) standard position effects in magnitude comparisons should be tested with larger samples and only one point of measurement. This would serve a deconfoundation of the test of the role of the position of comparative instructions as in Experiment 11. At least, a spin-off of Experiment 11 should be conducted, that randomizes within (instead of between) participants the relative magnitude maker (larger vs. smaller) in the comparative instructions provided after stimulus encoding. Another option to test for the STSE under the condition of naïve encoding, would be a selection task instead of a directed comparison task.

Learning effects, systematic response behavior and adaption levels during repeated task performance could be additionally reduced by filler trials unassociated with the relevant task, for example, judging the color of numbers.

As in psychophysics the opposite of the negative Type B effect, the positive Type B effect, namely the performance advantage for target-preceding-standard sequences, could be found for brief standard presentation (50 ms) and short ISI (300 ms; e.g. Hellström, 2020), this should be tested for the STSE as well.

A serious limitation of magnitude stimuli studying the STSE is the effect of numerical order that partly corrupts the effect of sequence in pairwise number comparisons. In general, the conditions of the hierarchy and mutual moderation of these two effects need to be further investigated. As Müller and Schwarz (2008) already stated for their experiment on the temporal number line, the diverse discontinuities of pairwise comparisons are difficult to study in isolation and therefore, should not be interpreted in isolation. An elaborate exploration of the moderation effects that depend on task peculiarities is needed. To disentangle the STSE from order effects (and position encoding processes from magnitude coding processes), studies with various stimulus material and tasks beyond judgements of magnitude, difference and similarity (e.g. a parity judgement task, compare Dehaene et al. 1990) should be realized.

To conclude on the STSE being an implicit sequence of relative judgements, due to familiarity, the processing fluency of standard-preceding-target comparisons could be investigated by asking participants which of the comparative sequences they preferred (for the measurement of processing fluency see Topolinski & Strack, 2009). Tversky and Gati (1978) had already realized such a study for their linguistic asymmetry of similarity and difference judgements and found preferences for standard-preceding- target sentences (see section 1.4.2).

In Experiments 1 and 11 of this work, a faster reaction on non-compatible trials (trials in which participants correctly responded with “no”) could be reported. This general detection advantage should be explored further with regard to an error detection system (e.g. Hoffmann

& Beste, 2015) that means a higher sensibility for incompatible trials to conclude on a systematic response strategy or rule it out respectively to further interpret the interaction of sequence and order effects in these experiments. The faster error detection could also be sample dependent. All the data for this work was gathered with academic samples in a university like test setting, that could have primed error detection.

In general, other samples from diverse populations should be tested to generalize the STSE effect or define it as a sample specific effect.

To further conclude on implicit or explicit response heuristics in future studies participants could be asked to report their response strategy after having performed a repeated directed 2AFC-task.

Taking the perspective of the grounded cognition approach on the side finding of standard-target-sequences being associated with the right response side (see Appendix B4), the sequence of standard preceding target could be associated with the spatial dimension from left to right just like proceeding time (Sekuler, Tyann & Levinson, 1973), ascending magnitude ; (Walsh, 2003; Winter et al., 2015) and ascending order (Müller & Schwarz, 2008; Huber et al., 2016). This could be investigated further in a study with crossed hands (compare Dehaene et al., 1993) and stimuli beyond magnitudes to disentangle effects of spatial and temporal magnitude coding from the STSE.

Another interesting aspect to explore would be the culture dependency of the STSE. Since the STSE was present in the horizontal stimuli presentation (Experiment 3) but not in the vertical stimulus presentation (Experiment 4), one could suggest that the western culture reading direction from left to right accords to the implicit sequence of standard first, respectively at the left side position, and target second, respectively at the right-side position. For this purpose, Experiments 3 and 4 could be replicated a Japanese sample, for example, to test the STSE effect in a culture of a reading direction from top to bottom (compare Ito & Hatta, 2004). Additionally, such a study would help to rule out the linguistic fluency of the syntax of subject preceding object in western languages, which Tversky suggested as the underlying mechanism for the asymmetry of similarity judgements.

With regard to applied research questions, it would be interesting to measure the influence of the sequence of standard and target on evaluations that base on paired comparisons, comparable to Houston et al. (1989, see section 1.4.4). Houston et al. (1989) failed to find an asymmetry of target-standard-sequences in preference judgements, most likely because they let participants perform preference judgments directly after comparative stimulus encoding without instructing the participants to perform the comparison as correctly as possible in the first place. A study with a two-step comparison and rating design would clarify if the sequence of standard and target affects evaluations of targets and standards.

4.4 Scientific Relevance of the STSE

The roots of pairwise comparison, that are the red line throughout the theoretical part and the empirical part of this work, lie in the measurement of discrimination sensitivity of physical stimuli. The investigation of the origins of the discontinuities of the performance of pairwise comparisons with numerical stimuli resulted in the conclusion that the discrimination of expressions on magnitude continua follow the Weber fraction – the logarithmic compression of the experience of magnitude. As the discrimination of physical and symbolic magnitudes resulted in concordant data patterns, experimental psychologists, like Stanislas Dehaene (1992), first concluded on a sense of numbers and later on a general sense of magnitude, also comprising numerosity (Leibovich et al., 2017; Krajsci et al., 2018). According to the theories of Banks (1977), Dehaene (1989, 1990, 1993), Holyoak (1978) it shares information processing characteristics with the five sensual modalities evident for humans. According to this perspective, humans not only use a general dimension of magnitude to verbally express their experiences of intensities of physical percepts, of intensities of experienced emotions, of their assessment of heights, volume, duration, social and physical distance – but magnitude appears to be a general principle of information acquisition and organization. Recently, other researchers stressed the role of working memory functions for the pairwise comparison of magnitudes and stated that task relevant orders (or stimulus ranges), implicitly or explicitly learned (e.g. Jou et al., 2020; Schroeder et al., 2017) or built up during task performance (e.g. Colling et al., 2020; compare SW model, Hellström, 1979) are the sense-general dimension used to acquire, organize and compare magnitude information. The relation of the sense of magnitude and the *sense of order*, and whether they are two manifestations of the same underlying dimension has to be investigated further. The STSE indicates an implicit order that is even more relevant for intuitive magnitude comparisons than the previously studied sense of magnitude.

The grounded cognition approach connects the senses of magnitude and/or order with bodily experiences of time and space, and claims information on these four dimensions to be sensationally close. The cognitive CMT and the neuropsychological ATOM argue for one system for all four dimensions (magnitude, order, space and time) rooted in the most important information for humans, the distance of something from one's person and body and its magnitude in relation to one's body. The SNARC effect (Dehaene et al., 1990; Dehaene at al., 1993), that was assumed to be a prominent example for the grounded cognition approach over three decades, recently, was interpreted as a finding resulting from the general principle of magnitude and order encoding associated with the temporal proceeding form left to right in western cultures (Walsh, 2003; Winter et al., 2015).

The performance advantage for pairwise magnitude comparisons for the sequence of standard preceding target, introduced in this dissertation, could be interpreted as an innate (or early acquired with the comprehension of time) heuristic of temporal occurrence – the standard always comes first, the target is always the unknown, new item, occurring after the standard. This had already been a conclusion of logical reasoning stated by the Relative Judgement Theory but so far had not been empirically explored for symbolic magnitudes.

While psychophysicists studied the relevance of the sequence of standard and target (Type B effect), researchers of symbolic magnitude and numerical comparisons had not yet questioned the influence of the temporal sequence of standard and target on discrimination sensitivity and comparison performance. In most studies on numerical comparisons, just like in psychophysics until 2009 (Ulrich & Vorberg, 2009), the standard was defined as the stimulus that was constant in terms of expression and position. The randomization of the standards' expression and position, comparable to the experimental design of the studies of this work, had not been reported so far, neither in psychophysical studies nor in studies of symbolic magnitude comparisons. According to the empirical work of this dissertation, this experimental setup revealed to be promising for the investigation of the influence of the sequence of standard and target and the underlying mechanism of the Type B effect. The constraint of psychophysical research, that a total randomization of standard and target is not possible, stresses the relevance of the investigation of the Type B effect with the help of research on the newly defined STSE in numerical pairwise magnitude comparisons. The recently established scientific view of magnitude and number processing in terms of a *sense of magnitude* supports the idea to transfer the findings of the study of the STSE in magnitude comparisons to psychophysical research on discontinuities of the assessment of physical magnitudes.

Furthermore, the few studies on effects of the relative position of standards and targets on outcomes of similarity judgements (Tversky & Gati, 1978) and detection of change

(Agostinelli et al, 1986) stress the broader relevance of the STSE and its interpretation to be a domain general phenomenon.

As of today, the ideas on the origin of the STSE introduced in section 4.1 are not based on sufficient empirical evidence. The interpretation of the results of this dissertation to locate the STSE at the encoding stage of comparative processing can be supported by previous findings: on the one hand, the Type B effect so far had only been reported for physical stimuli that cannot be *recalled* in an absolute sense; on the other hand, the reversion of sequence of target and standard depending on the position of instructions, as interpreted by Agostinelli et al. (1986). Due to an influence of the effect’s size in error rates and its suppression in RT in tasks of increased working memory load, working memory capacity is hypothesized to be involved in the relative encoding of sequence of standard and target, at least when it comes to magnitude comparisons.

Several findings have been reported, that argue for a contribution of working memory capacities on the performance advantage of standards preceding targets: debating on the origin of the Type B effect, the exact magnitude of physical percepts cannot be derived and encoded from the stimulus itself (Thurstone, 1927; Michels & Helson, 1954) to perform an accurate comparison between two percepts. The distinct timing hypothesis (Rammsayer, 2008; Troche & Rammsayer, 2009) explained the absence of a Type B effect in DL measurements for acoustic stimulus duration for very short tones with the missing requirement of the cognitive mechanism that underlay the Type B effect, only observed for rather long duration of tones. Also, the feature matching of multi-feature objects, as suggested by Tversky (supported by Tversky & Gati, 1978; Agostinelli et al., 1986) requires working memory capacity.

A counterfactual operation in working memory was suggested to be responsible for the impaired performance of target-standard sequences (see section 4.1.5). This account is similar to what had been stated by the feature matching model of Tversky (1977; Tversky & Gati, 1978). According to Tversky's model, the stimulus encountered first is processed feature-based in the first place, then matching its features with the features of the second stimulus. With the help of the research presented in this work, this model could be reformulated and expanded to unidimensional stimuli than can be processed absolutely, like numbers.

The reference point models introduced by researchers of psychophysics and symbolic magnitude comparisons (Michels & Helson, 1954; Michels & Helson, 1957; Holyoak, 1978; Chen et al., 2014) stated that humans cannot perform absolute judgements, or at least for domains in that absolute judgements are possible (numerical comparisons), absolute encodings are of no relevance for comparative judgements. These models combine mechanisms of stimulus encoding and working memory operations to explain standard position effects as well as congruency effects, the distance effect and end effects of magnitude comparisons. A reference point model could account for congruency effect found for the error rates in Experiment 11 of this work, where the direction of the comparison was unknown at the point of encoding (see section 4.2). The SW model (Hellström, 1979, 1985, 2000, 2003), explaining discontinuities in DL measurements (pairwise comparisons) of physical intensities due to the relative position of standard and target, stated a weighting of both stimuli and their relative encoding due to a reference point or subjective criteria. Also, the recursive random walk models (compare Buckley & Gillman, 1974; Leth-Steensen & Marley, 2000; Page et al., 2004) are maximally flexible to account for effects that occur when magnitude coding is prioritized by the task but is insufficient to explain the STSE.

4.5 Conclusion

Human attention can only focus on one stimulus at a time. The inevitable consequence of this attention focus is sequential processing. The nature of the human mind is to work in an energy saving and complexity reducing manner, it builds upon comparisons. In line with these two basic assumptions about human cognition, and as their consequence, the research presented in this dissertation could identify a strong performance advantage for standards preceding targets in pairwise comparisons of magnitudes. The STSE could serve as a general mechanism of several unresolved questions of comparative processing, as discussed above. It combines and extends the approach of a general cognitive system of magnitude and order by the temporal aspect of the sequence of standards and targets and stresses and expands the CMT approach for magnitude, time and space processing by the (possibly phylogenetically, linguistically or semantically) more accessible order of standard preceding target. The actual approach in cognitive research to stress the causal role of order in comparative thinking (explaining the SNARC effect, e.g. Schröder & Nuerk, 2017; Krajsci et al., 2018; the serial position effect, the distance effect and the SCE, Jou et al., 2018; 2020) is supported by the finding of the STSE. It could be interpreted as an important order, just like the mental number line, the alphabet and the month of the year. As the Type B effect, known in psychophysics, is evident as the performance advantage for standards preceding targets in pairwise presented stimuli that are processed directly without involving higher level cognition, the sequence of standard and target could even be more deeply rooted than the mentioned learned orders and based on an innate positional coding.

References

Agostinelli, G., Sherman, S. J., Fazio, R. H., & Hearst, E. S. (1986). Detecting and identifying change: Additions versus deletions. *Journal of Experimental Psychology: Human Perception and Performance, 12*(4), 445–454. doi: 10.1037/0096-1523.12.4.445

Aspinwall, L. G., & Taylor, S. E. (1993). Effects of social comparison direction, threat, and self-esteem on affect, self-evaluation, and expected success. *Journal of Personality and Social Psychology, 64,* 708-722.

Audley, R. J., & Wallis, C. P. (1964). Response instructions and the speed of relative judgments: I. Some experiments on brightness discrimination. *British Journal of Psychology, 55,* 59-73.

Banks, W. P. (1977). Encoding and processing of symbolic information in comparative judgments. In G. H. Bower (Ed.), *The psychology of learning and motivation, 11*, 101–159.

Banks, W. P., & Flora, J. (1977). Semantic and perceptual processes in symbolic comparisons. *Journal of Experimental Psychology:Human Perception and Performance, 3,* 278-290.

Banks, W. P., & Flora, J. (1977). Semantic and perceptual processes in symbolic comparisons. *Journal of Experimental Psychology: Human Perception and Performance, 3,* 278-290.

Banks, W. P., Fujii, M., & Kayra-Stuart, F. (1976). Semantic congruity effects in comparative judgments of magnitudes of digits. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 435-447.

Banks, W.P. & Root, M. (1979). Semantic congruity effects in judgments of loudness. *Perception & Psychophysics*, *26*, 133–142. doi: 10.3758/BF03208307

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Science*, *22*, 577–660.

Bausenhart, K. M., Dyjas, O., & Ulrich, R. (2015). Effects of stimulus order on discrimination sensitivity for short and long durations. *Attention, Perception, & Psychophysics, 77,* 1033–1043 doi: 10.3758/s13414-015-0875-8

Ben Nathan, M., Shaki, S., Salti, M., & Algom, D. (2009). Numbers and space: Associations and dissociations. *Psychonomic Bulletin & Review, 16*(3), 578-582. doi: 10.3758/PBR.16.3.578

Berkovits, I., Hancock, G. R., & Nevitt, J. (2000). Bootstrap Resampling Approaches for Repeated Measure Designs: Relative Robustness to Sphericity and Normality Violations. *Educational and Psychological Measurement*, *60*(6), 877–892. doi: 10.1177/00131640021970961

Biernat, M, Manis, M., & Nelson, T. E. (1991). Stereotypes and standards of judgment. *Journal of Personality and Social Psychology, 60,* 485-499.

Biernat, M. (2005). *Standards and expectancies: Contrast and assimilation in judgments of self and others.* New York: Psychology Press.

Biernat, M., & Manis, M. (1994). Shifting standards and stereotype-based judgment. *Journal of Personality and Social Psychology, 66,* 5–20.

Buckley, P. B., & Gillman, C. B. (1974). Comparisons of digits and dot patterns. *Journal of Experimental Psychology*, *103*(6), 1131–1136. doi: 10.1037/h0037361

Cattell, J.M. (1902). The time of perception as a measure of differences in intensity. *Festsch., 19,* 63–68.

Chen, D., Lu, H. & Holyoak, K. J. (2014). The discovery and comparison of symbolic magnitudes. *Cognitive Psychology*, *71*, 27-54, doi: 10.1016/j.cogpsych.2014.01.002

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Colling, L. J., Szűcs, D., De Marco, D., Cipora, K., Ulrich, R., Nuerk, H.-C., Soltanlou, M., Bryce, D., Chen, S.-C., Schroeder, P. A., Henare, D. T., Chrystall, C. K., Corballis, P. M., Ansari, D., Goffin, C., Sokolowski, H. M., Hancock, P. J. B., Millen, A. E., Langton, S. R. H., McShane, B. B. (2020). Registered Replication Report on Fischer, Castel, Dodd, and Pratt (2003). *Advances in Methods and Practices in Psychological Science,* 143–162. doi: 10.1177/2515245920903079

Collins, R. L. (1996). For better or worse: the impact of upward social comparison on self-evaluations. *Psychol Bull., 119,* 51–69. doi: 10.1037/0033-2909.119.1.51

de Hevia, M. D., & Spelke, E. S. (2010). Number-Space Mapping in Human Infants. *Psychological Science*, *21*(5), 653-660. doi: 10.1177/0956797610366091

Deese, J., & Kaufman, R. A. (1957). Serial effects in recall of unorganized and sequentially organized verbal material. *Journal of Experimental Psychology, 54*(3), 180–187. doi: 10.1037/h0040536

Dehaene S., Dupoux E., & Mehler J. (1990). Is numerical comparison digital? Analogical and symbolic effects in two-digit number comparison. *J Exp Psychol Hum Percept Perform*, *16*(3), 626-41. doi: 10.1037//0096-1523.16.3.626

Dehaene, S. (1989). The psychophysics of numerical comparison: A reexamination of apparently incompatible data. *Perception & Psychophysics, 45*, 557–566. doi: 10.3758/BF03208063

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition*, 4. doi: 10.1016/0010-0277(92)90049-N.

Dehaene, S. (1997). *The number sense: How the mind creates mathematics*. Oxford University Press.

Dehaene, S. (2003). The neural basis of the Weber-Fechner law: a logarithmic mental number line. *TRENDS in Cognitive Science, 7*(4), 145-147.

Dehaene, S., Bossini, S, & Giraux, P. (1993). The Mental Representation of Parity and Number Magnitude. *Journal of Experimental Psychology: General, 122*(3), 371-396.

Deng, Z., Chen, Y., Zhu, X. et al. The effect of working memory load on the SNARC effect: Maybe tasks have a word to say. *Memory & Cognition, 45*, 428–441 (2017). doi: 10.3758/s13421-016-0676-x

Dyjas, O., & Ulrich, R. (2014). Effects of stimulus order on discrimination processes in comparative and equality judgements: Data and models. The Quarterly Journal of Experimental Psychology, 67(6), 1121–1150. doi: 10.1080/17470218.2013.847968

Dyjas, O., Bausenhart, K. M., & Ulrich, R. (2012). Trial-by-trial updating of an internal reference in discrimination tasks: Evidence from effects of stimulus order and trial sequence. *Attention, Perception, & Psychophysics, 74*, 1819–1841.

Ebbinghaus, H. (1913). *On memory: A contribution to experimental psychology*. New York: Teachers College.

Ellinghaus, R., Ulrich, R., & Bausenhart, K. M. (2018). Effects of stimulus order on comparative judgments across stimulus attributes and sensory modalities. Journal of Experimental Psychology: Human Perception and Performance, 44(1), 7–12. doi: 10.1037/xhp0000495

Faul, F., Erdfelder, E., Lang, A.G. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191. doi: 10.3758/BF03193146

Fechner, G. T. (1860). *Elemente der Psychophysik* [*Elements of psychophysics*]. Leipzig, Germany: Breitkopf & Härtel.

Festinger, L. (1954) A theory of social comparison processes. *Human Relations, 7*, 117-140.

Fischer M. H., & Shaki S. (2014). Spatial associations in numerical cognition: from single digits to arithmetic. *Quarterly Journal of Experimental Psychology, 67*, 1461–1483. doi: 10.1080/17470218.2014.927515

Fischer, M., Castel, A., Dodd, M., & Pratt, A. (2003). Perceiving numbers causes spatial shifts of attention**.** *Nature Neuroscience*, *6*(6), 555-556. doi: 10.1038/nn1066

Gallistel, C. R., & Gelman,R. (1992). Preverbal and verbal counting and computation. *Cognition*, *44*, 43-74.

Gevers, W., & Lammertyn, J. (2005). The hunt for the SNARC, *Psychology Science, 47*, 10-21.

Gevers, W., Reynvoet, B., & Fias, W. (2004). The mental representation of ordinal sequences is spatially organized: Evidence from the days of the week. *Cortex: A Journal Devoted to the Study of the Nervous System and Behavior, 40,* 171–172. doi: 10.1016/S0010-9452(08)70938-9

Guida, A., Megreya, A. M., Lavielle-Guida, M., Noël, Y., Mathy, F., van Dijck, J.P., & Abrahamse, E. (2018). Spatialization in working memory is related to literacy and reading direction: Culture "literarily" directs our thoughts. *Cognition*, *2*(13), 96-100. doi: 10.1016/j.cognition.2018.02.013.

Hellström, Å. (1979). Time errors and differential sensation weighting. Journal of Experimental Psychology: Human Perception and Performance, *5,* 460–477. doi: 10.1037/0096-1523.5.3.460

Hellström, Å. (1985). The time-order error and its relatives: Mirrors of cognitive processes in comparing. *Psychological Bulletin, 97*, 35–61. doi: 10.1037/0033-2909.97.1.35

Hellström, Å. (2000). Sensation weighting in comparison and discrimination of heaviness. *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 6–17. doi: 10.1037/0096-1523.26.1.6

Hellström, Å. (2003). Comparison is not just subtraction: Effects of timeand space-order on subjective stimulus difference. *Perception & Psychophysics*, *65*, 1161–1177. doi: 10.3758/BF03194842

Hellström, Å., Patching, G. R., & Rammsayer , T. H. (2020). Sensation weighting in duration discrimination: A univariate, multivariate and varied-desig study of presentation-order effects. *Attention, Perception, & Psychophysics, 82*, 3196–3220. doi: /10.3758/s13414-020-01999-z

Hellström, Å., & Rammsayer, T. H. (2004). Effects of time-order, interstimulus interval, and feedback in duration discrimination of noise bursts in the 50- and 1000-ms ranges. *Acta Psychologica, 116*, 1–20. doi: 10.1016/j.actpsy.2003.11.003

Hellström, Å., & Rammsayer, T. H. (2015). Time-order errors and standard-position effects in duration discrimination: An experimental study and an analysis by the sensation-weighting model. *Attention, Perception, & Psychophysics, 77*, 2409–2423. doi: 10.3758/s13414-015-0946-x

Henik, A., & Tzelgov, J. (1982). Is three greater than five: the relation between physical and semantic size in comparison tasks. *Memory & Cognition*, *10,* 389–395.

Higgins, E. T. (1996). *Knowledge activation: Accessibility, applicability, and salience.* In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (p. 133–168). The Guilford Press.

Hoffmann, S., & Beste, C. (2015). A perspective on neural and cognitive mechanisms of error commission. *Frontiers in behavioral neuroscience, 9*, 50. doi: 10.3389/fnbeh.2015.00050e

Holyoak, K. (1978). Comparative Judgements with numericlal reference points. *Cognitive Psychology, 10*, 203-243.

Houston, D. A., Sherman, S.J., & Baker, S.A. (1989). The Influence of Unique Features and Direction of Comparison on Preferences. *Journal of Experimental Social Psychology*, *25*, 121-141.

Huber, S., Klein, E., Moeller, K., & Willmes, K. (2016). Spatial-Numerical and Ordinal Positional Associations Coexist in Parallel. *Frontiers in psychology, 7*(438). doi: 10.3389/fpsyg.2016.00438

Ishihara, M., Keller, P. E., Rossetti, Y., & Prinz, W. (2008). Horizontal spatial representations of time: evidence for the STEARC effect. *Cortex*, *44*, 454–461.

Ito, Y., & Hatta, T. (2004). Spatial structure of quantitative representation of numbers: Evidence from the SNARC effect. *Memory & Cognition, 32,* 662–673. doi: 10.3758/BF03195857

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General, 110*(3), 306–340. doi:10.1037/0096-3445.110.3.306.

Jamieson, D. G., & Petrusic, W. M. (1975). Presentation order effects in duration discrimination. *Perception & Psychophysics, 17*(2), 197–202.
doi: 10.3758/BF03203886

Jou, J. (2010). The Serial Position, Distance, and Congruity Effects of Reference Point Setting in Comparative Judgments. *The American Journal of Psychology*, *123*(2), 127-136. doi: 10.5406/amerjpsyc.123.2.0127

Jou, J., Escamilla, E. E., Torres, A. U., Ortiz, A., & Salasar, P. (2018). Where does the congruity effect come from in memorial comparative judgments: A serial-position-based distinctiveness account. *Journal of memory and Language, 103*. doi: 10.1016/j.jml.2018.08.003

Jou, J., Matos, M., Martinez, M., Sierra, F., Guzman, C., & Hut, A. (2020). Redefining the Congruity Effect in Comparative Judgments: A Review of the Theories and a Further Test. *The American Journal of Psychology, 133*(2), 221-239. doi: 10.5406/amerjpsyc.133.2.0221

Kaan, E. (2005). Direction effects in number word comparison: an event-related potential study. *NeuroReport, 16*(16), 1853-1856. doi: 10.1097/01.wnr.0000185016.21692.50

Keus, I.M., Schwarz, W. (2005). Searching for the functional locus of the SNARC effect: Evidence for a response-related origin. *Memory & Cognition,* 33**,** 681–695. doi: 10.3758/BF03195335

Krajcsi, A., Lengyel G., & Kojouharova, P. (2018). Symbolic Number Comparison Is Not Processed by the Analog Number System: Different Symbolic and Non-symbolic Numerical Distance and Size Effects. *Frontiers in Psychology, 9*(124). doi: 10.3389/fpsyg.2018.00124

Krajcsi, A., & Kojouharova, P. (2017). Symbolic Numerical Distance Effect Does Not Reflect the Difference between Numbers. *Frontiers in Psychology, 8*. doi: 10.3389/fpsyg.2017.02013

Krueger, L.E. (1989). Reconciling Fechner and Stevens: Toward a unified psychophysical law. *Behavioral and Brain Science, 12*, 251–267.

Lakoff, G. (1993). The contemporary theory of metaphor. In Ortony (Ed.), *Metaphor and thought* (pp. 203-251). New York, NY: Cambridge University Press.

Lakoff, G. (2008). The neural theory of metaphor. In R. W. Gibbs (Ed.), *Cambridge handbook of metaphor and thought* (pp. 17-38). Cambridge, MA: Cambridge University Press.

Lakoff, G. (2012). Explaining embodied cognition results. *Topics in Cognitive Science, 4*, 773-785.

Leek, M.R. (2001). Adaptive procedures in psychophysical research. *Perception & Psychophysics, 63*, 1279–1292 doi: 10.3758/BF03194543

Leibovich, T., Katzin, N., Harel, M., & Henik, A. (2017). From "sense of number" to "sense of magnitude": The role of continuous magnitudes in numerical cognition. *Behavioral and Brain Sciences, 40*. doi: 10.1017/S0140525X16000960

Leth-Steensen, C., & Marley, A. A. J. (2000). A model of response time effects in symbolic comparison. *Psychological Review, 107*(1), 62–100. doi: 10.1037/0033-295X.107.1.162

Liesefeld, H.R., & Janczyk, M. (2019). Combining speed and accuracy to control for speed-accuracy trade-offs(?). *Behav Res 51*, 40–60. doi: 10.3758/s13428-018-1076-x

Link, S. W., & Heath, R. A. (1975). Sequential Theory Of Psychological Discrimination. *Psychometrika, 40*(1).

Mantel, S. P., & Kardes, R., R. (1999). The Role of Direction of comparison, Attribute based processing and Attitude based processing in consumer preference. *Journal of Consumer Research, 25*(4), 335-352.

Marks, D. F. (1972). Relative judgment: A phenomenon and a theory. *Perception and Psychophysics, 11,* 156-160.

Marschark, M., & Paivio, A. (1979).Semantic congruity and lexical marking in symbolic comparisons: An expectancy hypothesis. *Memory* &: *Cognition,* 7, 175-184.

Marsh, H. W.(1986). Verbal and math self-concepts: An internal/external frame of reference model. *American Educational Research Journal, 23,* 129– 149.

Martin, L. J., & Müller, G. E. (1890). *Zur Analyse der Unterschedsempfindlichkeit* [On the analysis of discrimination sensitivity]. Tübingen, Germany: Laupp.

Michels,W. C., & Helson, H. (1954). A quantitative theory of time-order errors. *American Journal of Psychology, 67,* 327–334. doi: 10.2307/1418635

Moyer, R. S. (1973). Comparing objects in memory: Evidence suggesting an internal psychophysics. *Perception & Psychophysics, 13*(2), 180–184. doi: 10.3758/BF03214124

Moyer, R. S. (1973). Comparing objects in memory: Evidence suggesting an internal psychophysics. *Perception* & *Psychophysics, 13,* 180-184.

Moyer, R. S., & Bayer, R. H. (1976). Mental comparison and the symbolic distance effect. *Cognitive Psychology, 8,* 228-246.

Moyer, R. S., & Dumais, S. T. (1978). Mental Comparison. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation*. New York: Academic Press.

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgments of numerical inequality. *Nature, 215,* 1519–1520.

Mullen, B. & Hu, L. (1989). Perceptions of Ingroup and Outgroup Variability: A Meta-Analytic Integration. *Basic and Applied Social Psychology*, *10*(3), 233-252. doi: 10.1207/s15324834basp1003_3

Müller, D., & Schwarz, W. (2008). “1–2–3”: Is There a Temporal Number Line? Evidence from a Serial Comparsion Task. *Experimental Psychology*, *55*(3), 143-150.

Mussweiler, T. (2003). Comparison processes in social judgment: Mechanisms and consequences. *Psychological Review*, *110*(6), 472–489. doi:10.1037/0033-295X.110.3.472.

Mussweiler, T., (2009). Comparison, in Fristz Strack & Jens Förster (Eds.), *Social Cognition – the Basis of Human Interaction.* New York: Psychology Press.

Newell, B. R., & Shanks, D. R. (2007). Recognising what you like: Examining the relation between the mere-exposure effect and recognition. *European Journal of Cognitive Psychology*, *19*(1), 103–118. doi*:*10.1080/09541440500487454. ISSN 0954-1446

Pagano, R. R. (2010). *Understanding statistics in the behavioral sciences* (9th ed.). Australia, Belmont, CA: Thomson Wadsworth.

Page, R., Izquierdo, E. Saal, A., Codnia, J., & El Hasi, C. (2004). A response time model for judging order relationship between two symbolic stimuli. *Perception & Psychophysics*, *66*(2), 196-207.

Petrusic, W. M., Lucas, J. A., & Leth-Steensen, C. (2011). Vertical SNARC with positive and negative numbers. *Proceedings of Fechner Day*, *27*(1), 131–136.

Petzschner, F. H., Glasauer, S., & Stephan, K. E. (2015). A Bayesian Perspective On Magnitude Estimation. *Trends In Cognitive Science, 19*(5), 285-293.

Proctor, R. W., & Cho, Y. S. (2006). Polarity correspondence: A general principle for performance of speeded binary classification tasks. *Psychological Bulletin*, *132*(3), 416–442.

Proctor, R.W., & Cho, Y.S. (2003). Effects of response eccentricity and relative position on orthogonal stimulus-response compatibility with joystick and keypress responses. *The Quarterly Journal of Experimental Psychology, 56*(2), 309–327.

Rammsayer, T. H. (2008). Neuropharmacological approaches to human timing. In S. Grondin (Ed.), *Psychology of time* (pp. 295–320). Bingley, England: Emerald.

Reber, R., Winkielman, P., & Schwarz, N. (1998). Effects of Perceptual Fluency on Affective Judgments. *Psychological Science,* 9, 45-48. doi: 10.1.1.232.8868. doi*:*10.1111/1467-9280.00008

Salkind, N. J. (2010). *Encyclopedia of Research Design* (Vol. 2). Los Angeles: Sage.

Sanbonmatsu, D. M., Kardes, F. R., & Gibson, B. D. (1991). The role of attribute knowledge and overall evaluations in comparative judgment. *Organizational Behavior and Human Decision Processes, 48*(1), 131–146. doi: 10.1016/0749-5978(91)90009-I

Santiago, J., & Lakens, D. (2014). Can conceptual congruency effects between number, time, and space be accounted for by polarity correspondence?. *Acta Psychologica*, *156*, 179-191.

Schroeder, P. A., Nuerk, H.-C., & Plewnia, C. (2017). Space in Numerical and Ordinal Information: A Common Construct? *Journal of Numerical Cognition, 3*(2), 164-181.

Schwarz, W. & Stein, F. (1998). On the Temporal Dynamics of Digit Comparison Processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(5) , 1275-1293.

Sekuler, R., Tynan, P., & Levinson, E. (1973, April 13). Visual temporal order: A new illusion. *Science, 180*, 210–212.

Shaki, S. & Petrusic, W.M. (2005). On the mental representation of negative numbers. Context dependent SNARC effects with comparative judgments. *Psychonomic Bulletin and Review*, *12*, 931-937.

Shaki, S. A., Petrusic W. M., & Leth-Steensen (2012). SNARC Effects With Numerical and Non-Numerical Symbolic Comparative Judgments: Instructional And Cultural Dependencies. *Journal of Experimental Psychology: Human Perception and Performance, 38*(2), 515–530.

Shaki, S., Petrusic, W. M. & Leth-Steensen, C. (2012). SNARC Effects With Numerical and Non-Numerical Symbolic Comparative Judgments: Instructional And Cultural Dependencies. *Journal of Experimental Psychology: Human Perception and Performance, 38*(2), 515–530.

Shoben, E. J., Čech, C. G., Schwanenflugel, P. J., & Sailor, K. M. (1989a). Serial position effects in comparative judgments. *Journal of Experimental Psychology: Human Perception and Performance, 15*(2), 273–286. doi: 10.1037/0096-1523.15.2.273

Shoben, E. J., Sailor, K. M., & Wang, M.-Y. (1989b). The role of expectancy in comparative judgments. *Memory & Cognition, 17*(1), 18–26. doi: 10.3758/BF03199553

Simon, J. R. & Rudell, A. P. (1967): Auditory S-R compatibility: The effect of an irrelevant cue on information processing. *Journal of Applied Psychology*, 51, 300–304.

Simon, J. R. (1969). Reactions toward the source of stimulation. *Journal of Experimental Psychology*, *81*(1), 174–176. doi: 10.1037/h0027448

Stevens, S. (1957). On the psychophysical law. *Psychological Revue*, *64*(3), 153-181. doi:10.1037/h0046162.

Sudevan, P., & Taylor, D. A. (1987). The cuing and priming of cognitive operations. *Journal of Experimental Psychology: Human Perception and Performance, 13*(1), 89–103. doi: 10.1037/0096-1523.13.1.89

Suls, J., & Wheeler, L. (2012*)*. Social comparison theory. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (pp. 460–482). Sage Publications Ltd. doi: 10.4135/9781446249215.n23

Topolinski, S., & Strack, F. (2009). The architecture of intuition: Fluency and affect determine intuitive judgments of semantic and visual coherence, and of grammaticality in artificial grammar learning. Journal of Experimental Psychology: General, 138, 39-63.

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

Thurstone, L.L. (1929). The Measurement of Psychological Value. In T.V. Smith and W.K. Wright (Eds.), *Essays in Philosophy by Seventeen Doctors of Philosophy of the University of Chicago*. Chicago: Open Court.

Thurstone, L.L. (1959). *The Measurement of Values*. Chicago: The University of Chicago Press.

Troche, S. J., & Rammsayer, T.H. (2009a). The influence of temporal resolution power and working memory capacity on psychometric intelligence. *Intelligence, 37*, 479–486.

Turconi, E., Campbell, J. I. D., & Seron, X. (2006). Numerical order and quantity processing in number comparison. *Cognition, 98*(3), 273-285. doi: 10.1016/j.cognition.2004.12.002

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*(4), 327–352. doi: 10.1037/0033-295X.84.4.327

Tversky, A. & Gati, I. (1978). Studies of similarity. In Eleanor Rosch & Barbara Lloyd (eds.), *Cognition and Categorization*. Lawrence Elbaum Associates.

Tversky, A. & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*,*185*(27), 1124–1131. *doi:10.1126/science.185.4157.1124*

Ulrich, R., & Vorberg, D. (2009). Estimating the difference limen in 2AFC tasks: Pitfalls and improved estimators. *Attention, Perception, & Psychophysics, 71*, 1219–1227.

van Dijck, J. P., & Fias, W. (2011). A working memory account for spatial-numerical associations. *Cognition*, *119*(1), 114–119. doi: 10.1016/j.cognition.2010.12.013

Van Opstal, F., Gevers, W., De Moor, W., & Verguts, T. (2008). Dissecting the symbolic distance effect: Comparison and priming effects in numerical and nonnumerical orders. *Psychonomic Bulletin & Review, 15*(2), 419–425. doi: 10.3758/PBR.15.2.419

Vasey, M. W., & Thayer, J. F. (1987). The Continuing Problem of False Positives in Repeated Measures ANOVA in Psychophysiology: A Multivariate Solution. *Psychophysiology, 24*(4), 479–486. doi: 10.1111/j.1469-8986.1987.tb00324.x

Walsh, V. (2003). A theory of magnitude: common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences, 7,* 483-488.

Weber, E. H. (1851). Index programmatum, librorum et commentationum ab editorum. In: *Annotationes anatomicae et physiologicae,* 3. Heft (p. 117 f.). Leipzig.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing* (*3rd ed.). Statistical modeling and decision science.* Amsterdam, Boston: Academic Press.

Winter, B. Marghetis, T., & Matlock, T. (2015). Of magnitudes and metaphors: Explaining cognitive interactions between space, time, and number. *Cortex, 64,* 209 – 224.

Woodworth, R. S., & Schlosberg, H. (1954). *Experimental psychology*. New York: Academic Press.

Appendix A: Examples for Task Instructions

Between participant counterbalancing variations are inserted in brackets.

A1.1 Instructions of Experiment 6 in German (original):

Willkommen zu diesem Experiment!

In dieser Aufgabe untersuchen wir grundlegende Zahlenvergleiche.

Sie sollen im Folgenden immer jeweils zwei Zahlen miteinander vergleichen, die nacheinander gezeigt werden.

Dabei gibt es immer Ihre Zahl und eine Vergleichszahl.

Beurteilen Sie immer so schnell wie möglich, ob die Ihre Zahl GRÖßER [KLEINER] ist als die Vergleichszahl.

Die Vergleichszahl ist immer fett [dünn] gedruckt.

IHRE ZAHL ist immer dünn [fett] gedruckt.

Bitte reagieren Sie so schnell wie möglich, indem Sie eine der beiden STRG-Tasten drücken.

Wenn IHRE Zahl GRÖßER [KLEINER] ist als die andere Zahl, drücken Sie bitte die LINKE [RECHTE] STRG-Taste.

Wenn IHRE Zahl KLEINER [GRÖßER] ODER GLEICH der anderen Zahl ist, drücken Sie bitte die RECHTE [LINKE] STRG-Taste.

Legen Sie Ihre beiden Zeigefinger bitte auf die beiden STRG-Tasten der Tastatur vor Ihnen.

Drücken Sie nun bitte eine beliebige STRG-Taste, um zur nächsten Seite zu gelangen.

Vielen Dank!

A1.2 Instructions of Experiment 6 in English (translation):

Welcome to this experiment!

In this task we will investigate basic number comparisons.

In the following, you are to always compare two numbers with each other, which are shown one after the other.

There is always your number and a comparison number.

Always judge as quickly as possible whether your number is LARGER [SMALLER] than the comparison number.

The comparison number is always in bold [thin] type.

YOUR NUMBER is always printed in thin [bold] type.

Please react as soon as possible by pressing one of the two CTRL keys.

If YOUR NUMBER is LARGER [SMALLER] than the other number, please press the LEFT [RIGHT] CTRL key.

If YOUR number is SMALLER [LARGER] OR EQUAL to the other number, please press the RIGHT [LEFT] CTRL key.

Please place your two index fingers on the two CTRL keys on the keyboard in front of you.

Now please press any CTRL key to move to the next page.

Thank you very much!

A2.1 Instructions of Experiment 8 in German (original):

Willkommen zu diesem Experiment!

In dieser Aufgabe sollen Sie die IQ Werte zweier Personen miteinander vergleichen.

Dabei gibt es immer einen IQ Wert, den Sie beurteilen sollen

und einen IQ Wert mit dem Sie diesen vergleichen sollen.

Beurteilen Sie immer SO SCHNELL WIE MÖGLICH, ob der IQ Wert von Person A [Person B] größer oder kleiner oder gleich ist als der von Person B [Person A].

Bitte reagieren Sie so schnell wie möglich, indem Sie eine der beiden STRG-Tasten drücken.

Wenn Person As Wert GRÖßER [KLEINER] ist als Person Bs [Person As], drücken Sie bitte die LINKE [RECHTE] STRG-Taste.

Wenn Person As]Person Bs] Wert KLEINER [GRÖßER] ODER GLEICH Person Bs [Person As] Wert ist, drücken Sie bitte die RECHTE [LINKE] STRG-Taste.

Legen Sie nun Ihre beiden Zeigefinger bitte auf die beiden STRG-Tasten der Tastatur vor Ihnen.

Drücken Sie jetzt bitte eine beliebige STRG-Taste, um zur nächsten Seite zu gelangen.

Vielen Dank!

A2.2 Instructions of Experiment 8 in English (translation):

Welcome to this experiment!

In this task you have to compare the IQ values of two persons.

There is always one IQ value, which you have to evaluate

and an IQ value with which you have to compare it.

Always judge AS FAST AS POSSIBLE whether the IQ value of Person A [Person B] is

is greater or less than or equal to that of Person B [Person A].

Please respond as quickly as possible by pressing one of the two CTRL keys.

If Person A's [Person B’s] value is GREATER [LESS] than Person B's [Person A’s], please press the LEFT [RIGHT].

If Person A's [Person B’s] value is LESS [GREATER] OR EQUAL to Person B's [Person A’s] value, please press the RIGHT [LEFT] CTRL key.

Now please place your two index fingers on the two CTRL keys on the keyboard in front of you.

Now please press any CTRL key to move to the next page.

Thank you very much!

A3.1 Instructions of Experiment 11 in German (original):

Willkommen zu diesem Experiment!

In dieser Aufgabe untersuchen wir grundlegende Zahlenvergleiche.

Sie sollen im Folgenden immer jeweils zwei Zahlen miteinander vergleichen, die nacheinander gezeigt werden.

Auf welche Weise Sie die beiden Zahlen miteinander vergleichen sollen,

erfahren Sie nachdem Sie die beiden Zahlen gesehen haben.

Sie sollen entweder die erste Zahl oder die zweite Zahl beurteilen.

Nachdem zwei Zahlen gezeigt wurden, erscheint eine Frage auf dem Bildschirm,

die Sie entweder BEJAEN oder VERNEINEN.

Bitte drücken Sie die LINKE [RECHTE] STRG-TASTE, wenn Sie die Frage BEJAEN wollen und

die RECHTE [LINKE] STRG-TASTE, wenn Sie die Frage VERNEINEN wollen.

Es gibt immer eine eindeutig richtige Antwort.

Bitte lesen Sie die Frage in jedem Durchgang aufmerksam und reagieren Sie erst,

wenn Sie diese gelesen haben und die richtige Antwort wissen.

Antworten Sie, indem Sie eine der beiden STRG-Tasten drücken.

Bitte reagieren Sie so schnell wie möglich, sobald Sie die richtige Antwort kennen.

Legen Sie nun bitte Ihre beiden Zeigefinger auf die beiden STRG-Tasten der Tastatur vor Ihnen.

Drücken Sie nun bitte eine beliebige STRG-Taste, um zur nächsten Seite zu gelangen.

Vielen Dank!

A3.2 Instructions of Experiment 11 in English (translation):

Welcome to this experiment!

In this task we will investigate basic number comparisons.

In the following, you are to always compare two numbers with each other, which are shown one after the other.

In which way you should compare the two numbers with each other,

you will find out after you have seen the two numbers.

You are to judge either the first number or the second number.

After two numbers are shown, a question appears on the screen,

which you have to answer either in the affirmative or in the negative.

Please press the LEFT[RIGHT] CTRL-BUTTON if you want to AFFIRM the question and

the RIGHT [LEFT] CTRL key if you want to DENY the question.

There is always one unambiguously correct answer.

Please read the question carefully in each pass and do not respond until,

when you have read it and know the correct answer.

Answer by pressing one of the two CTRL keys.

Please respond as soon as you know the correct answer.

Now please place your two index fingers on the two CTRL keys on the keyboard in front of you.

Now please press any CTRL key to move to the next page.

Thank you very much!

Appendix B: Additional Statistical Analyses

B1 Preregistered Mixed Model Experiment 6

Table 5

2x2x2 mixed rmANOVA (RTM x Sequence x Order) for RT (Exp. 6)

Predictor	df_{Num}	df_{Den}	SS_{Num}	SS_{Den}	*F*	*p*	η_p^2
(Intercept)	1	88			70910.40	<.001	.99
Sequence	1	88	.253	1.439	15.117	<.001	.15
Order	1	88	.012	.558	1.327	.172	.02
Sequence x Order	1	88	.013	1.027	1.069	.284	.012
RTM	1	88	.005	20.012	.022	.882	.99
Sequence x RTM	1	88	.006		.749	.389	.004
Order x RTM	1	88	.007		.015	.996	<.001
Sequence x Order x RTM	1	88	.012		8.48	.357	.10

Note. MM = Magnitude Marker. df_{Nom} indicates degrees of freedom numerator. df_{Den} indicates degrees of freedom denominator. SS_{Num} indicates sum of squares numerator. SS_{Den} indicates sum of squares denominator.

Table 6

2x2x2 mixed rmANOVA (RTM x Sequence x Order) for error rates (Exp. 6)

Predictor	df_{Num}	df_{Den}	SS_{Num}	SS_{Den}	F	p	η_p^2
(Intercept)	1	88			200.71	<.001	.67
Sequence	1	88	0.022	0.418	0.15	.903	<.01
Order	1	88	0.002	0.227	0.001	.974	<.01
Sequence x Order	1	88	0.022	0.248	7.766	.007	.08
RTM	1	88	0.772	19.490	1.29	.256	.01
Sequence x RTM	1	88	0.001	0.418	0.015	.903	<.01
Order x RTM	1	88	0.001	0.227	0.251	.618	<.01
Sequence x Order x RTM	1	88	0.003	0.248	0.003	1.17	.01

Note. MM = Magnitude Marker. df_{Nom} indicates degrees of freedom numerator. df_{Den} indicates degrees of freedom denominator. SS_{Num} indicates sum of squares numerator. SS_{Den} indicates sum of squares denominator.

B2 Data Analyses without ties for Experiment 8

Table 7

2x2 rmANOVA (Sequence x Order) for RT (Exp. 8)

Predictor	df_{Num}	df_{Den}	SS_{Num}	SS_{Den}	F	p	η_p^2
(Intercept)	1	129	23,579.62	33.69	90,29	<.001	.99
Sequence	1	129	0.503	1.714	37.87	<.001	.23
Order	1	129	0.130	1.245	13.47	<.001	.10
Sequence x Order	1	129	0.295	1.665	22.83	<.001	.15

Note. MM = Magnitude Marker. df_{Nom} indicates degrees of freedom numerator. df_{Den} indicates degrees of freedom denominator. SS_{Num} indicates sum of squares numerator. SS_{Den} indicates sum of squares denominator.

Table 8

2x2 rmANOVA (Sequence x Order) for error rates (Exp. 8)

Predictor	df_{Num}	df_{Den}	SS_{Num}	SS_{Den}	F	p	η_p^2
(Intercept)	1	129	3.817	3.364	146.36	<.001	.532
Sequence	1	129	0.042	0.852	6.330	.013	.05
Order	1	129	0.032	0.515	8.068	.005	.06
Sequence x Order	1	129	0.004	0.660	0.716	.399	.01

Note. MM = Magnitude Marker. df_{Nom} indicates degrees of freedom numerator. df_{Den} indicates degrees of freedom denominator. SS_{Num} indicates sum of squares numerator. SS_{Den} indicates sum of squares denominator

B3.1 Performance Improvement with Task Proceeding

Table 9

Exemplary Correlations between Performance Outcomes and Trial Order (Exp. 1, 2, 6, 10 and 11)

	RT			error rates		
Exp.	$r_{\textit{trial order, ln(RT)}}$	*p*	95% CI	$r_{\textit{trial order, error}}$	*p*	95% CI
1	-.29**	<.001	[-.32; -.28]	-.06**	<.001	[-.08; -.04]
2	-.13**	<.001	[-.15; -.10]	-.06**	<.001	[-.09; -.03]
6	-.07**	<.001	[-.09; -.06]	<.01	.72	[-.01; .02]
10	-.132	<.001	[-.14; -.12]	.02	.14	[-.03; -.01]
11	-.30**	<.001	[-.33; -.29]	-.04	<.001	[-.06; -.02]

Note. CI = confidence interval. ** = highly significant

B3.2 Distance Effect

Table 10

Exemplary Correlations between RT and Numerical Distance (Exp. 3, 8 and 11)

Exp.	$r_{\textit{distance, ln(RT)}}$	*p*	95% CI
3	-.02*	.039	[-.04; -.00]
8	-.10**	<.001	[-.13; -,08]
11	-.02*	.05	[-.04; -.02]

Note. CI = confidence interval. * = significant; ** = highly significant

B4 Extra Analyses Experiment 11

Table 11

2x2x2x2 mixed rmANOVA (Sequence x Order x Responseside x MM) for RT (Exp. 11)

Predictor	df_{Num}	df_{Den}	SS_{Num}	SS_{Den}	F	p	η_p^2
(Intercept)	1	123	26,930.55	56.014	59,136.22	<.001	.99
Sequence	1	123	0.02	1.130	<0.01	.96	.15
Order	1	123	.053	1.333	4.85	.029	.04
MM	1	123	6.690	56.014	14.69	.284	.01
Responseside	1	123	.132	56.014	0.291	.591	<.01
Sequence x Order	1	123	.006	1.333	17.35	.539	<.01
Sequence x MM	1	123	.033	1.130	3.62	.06	.03
Sequence x Responseside	1	123	.306	1.130	33.30	<.001	.21
Order x MM	1	123	.255	1.333	23.53	<.001	.16
Order x Responseside	1	123	.066	1.333	6.06	.015	.05
Sequence x Order x MM	1	123	1.135	1.083	128.997	<.001	.51
Sequence x Order x Responseside	1	123	1.135	1.083	0.30	.59	<.01
MM x Responseside	1	123	0.001	56.014	<.01	.970	<.01
Sequence x Order x MM x Responseside	1	123	0.004	1.083	0.44	.51	<.01

Note. MM = Magnitude Marker. df_{Nom} indicates degrees of freedom numerator. df_{Den} indicates degrees of freedom denominator. SS_{Num} indicates sum of squares numerator. SS_{Den} indicates sum of squares denominator.