On having a 200% chance: Additivity neglect in probability estimates

Anine Riege

Department of Psychology, University of Oslo

Series of dissertations submitted to the
Faculty of Social Sciences, University of Oslo
No. 576

ISSN 1564-3991

Cover: Hanne Baadsgaard Utigard.
Print production: Reprosentralen, University of Oslo.

Acknowledgements

There is a common proverb stating that "it takes a village to raise a child", I would like to venture here that the same can be said of completing a PhD. Firstly, I want to thank the village chiefs, aka my supervisors: Karl Halvor (Teigen), when gathering all my courage (as a 2nd year BA-student) to ask for your advice regarding a project, I never imagined that we would be able to fill so many years with exciting, interesting and fun research projects. You are the best! (At what, you ask? Everything!). And Geir (Kirkebøen), there would be no PhD without you. Thank you both, so much.

The rest of the village goes as follows: Thank you, Unni (Sulutvedt) for teaching me how to use the eye-tracker, extract the data, interpret the data, and for co-authoring the paper (in addition to being a great friend). And many thanks to Bruno Laeng and Siri Leknes for giving me helpful advice during the eye-tracking experiment, I really appreciated your input.

During my PhD I was also given the opportunity to work with a great colleague and friend, Jon Anders Lone. I had so much fun on our project, and thank you for all your support, read-throughs, listening to my presentations (you might have, in effect, taken all my classes), and our many, many discussions.

I would also like to thank my great colleagues Erik Løhre, Mo Mowinkel (both read drafts of this dissertation), Marie Juanchich (great project), Gro Hege Haraldsen Nordbye (another fun project), Kjersti Walle (you truly are one of my favorite people), Petra Filkukova (for our many talks over dinner), and the rest of our great group members (Torleif, Alf Børre, Sigrid, Miro and Magne). A special thank you goes to Karen Hauge, I don't know what I would have done without our weekly writing-meetings. Thomas Schubert, you really have helped me with so many things, always offering good advice, and I am very grateful. Also, Beate Seibt, you are the best "partner in crime" a girl can have (no actual crimes were committed, of course). Many thanks to all the great people on "the third floor" (Irmelin, especially – I was so lucky to have the same start date as you) for all our lunches, discussions, and good times. I also want to thank all the participants for their time and effort – there would truly be no PhD without you!

Of course my family has been (forced to) live in the PhD-village, and I really want to thank my parents (they both proof-read this dissertation), Liv and Kjell, for all their support and love (always). I also want to thank my family-in-law, especially Gunvor, Halvor, Kristel and Anne (yes, there are many more), for all your support and cheering. Lastly, and most importantly, I want to thank my lovely husband Bjørn Tore: thank you for supporting me, believing in me, and for learning to cook during my PhD. For that, and for all that you are – I love you.

And there you have it – extensive proof that it takes a village to complete a PhD.

Table of Contents

Summary

Several studies have shown that when people are asked to estimate the probabilities for a set of exclusive and exhaustive events they often produce probabilities that add up to more than 1 or 100% (Robinson & Hastie, 1985; Teigen, 1974b, 1983), thus violating the additivity principle stated by formal probability theory. The present dissertation aims to further investigate the determining factors of additivity neglect and the underlying cognitive processes.

Paper I - The aim of Paper I was to investigate several determinants of additivity neglect. We wanted to investigate the notion that this bias is related to people's lack of mathematical skills, by giving participants a numeracy test. We were also interested in examining how people understand and predict disjunctive sets of outcomes, where the total ought to sum to more than 100%. Lastly, several pilot studies had suggested that answering format affected people's additivity, where a written format (writing estimates in an empty slot next to each outcome in the set) seemed to prompt more additive responses than estimates given on a scale format (circling numbers on 0-100% horizontal rating scales for each outcome in the set).

The overall results showed that numeracy (Experiments 1 and 3) was positively related to additive responses. In addition, varying the presentation of the Numeracy scale (before vs. after the estimation tasks), revealed that answering the Numeracy test prior to the probability tasks "primed" participants (mainly those with high numeracy) to answer according to mathematical principles. It is thus not sufficient to have high numeracy; one must also be reminded that mathematical rules apply. We also found a clear tendency for estimates given in the written format to be lower than estimates in scale format (Experiments 2, 3 and 4). The written format in the single outcome conditions also yielded more additive responses. Lastly, participants estimates of disjunctive probabilities were, in both formats, quite close to the estimates of single outcomes (Experiment 1 and 2), and the probability of winning was judged to be higher than the probability of becoming second best (Experiment 2 and 3). The probability estimates for the disjunctive outcome tasks were unrelated to participants' numeracy and the mathematical mindset priming.

Paper II – The aim of Paper II was to further investigate the difference in answering format found in Paper I, by using a process measure in the form of eye-tracking. Monitoring eye movements provides information about the decision process by registering what participants are looking at (fixations), repeated inspections of the same material (revisits),

and cognitive load (fixation durations). We predicted that fixations, fixation durations, and revisits would differ based on the two answering formats, due to participants assigning probabilities one by one, in a case-based manner, in the Scale format, or in a compensatory manner, in the Self-generated format.

The results showed that participants in the Self-generated condition had more fixations, and on average almost twice as many revisits between the alternatives compared to the Scale condition, indicating more comparisons between the alternatives in the set in the Self-generated condition. We also expected to find that the participants in the Self-generated condition had longer fixation durations than participants in the Scale condition, however, this was not the case. Overall, the results from Paper II indicate that the Scale format might prompt a selective evaluation of the alternatives, and thereby discourage comparisons between alternatives. The Self-generated estimates facilitate a class-based approach and make people engage in more comprehensive comparisons.

Paper III – The aim of Paper III was to compare additivity neglect to another type of bias that commonly occurs in referent-dependent judgments, namely the nonselective superiority bias. The nonselective superiority bias (NSSB) is the consistent evaluation of individual members of a positive set of items (e.g., five good movies) as superior to most other members in the set (Giladi & Klar, 2002). Both biases violate basic formal constraints, as a set of attractive candidates cannot all be rated as better than the group mean (NSSB), and are thus “unbalanced”; while the probabilities of a set of exhaustive events cannot add up to more than 100% (additivity neglect), and are thus non-additive.

Participants in three experiments were asked to give both probability estimates and comparative judgments in separate tasks. The results from all experiments indicated several similarities between the nonselective superiority bias and additivity neglect. Both biases seem to be about equally widespread, as a majority of participants´ probability sums far exceeded 100%, and mean ratings were significantly greater than zero (normative mean for items in a set), even when presented with the full set of alternatives. The two biases were also related, as participants who gave additive probability estimates gave more balanced distributions of ratings to the NSSB tasks. However, NSSB seems to be more robust, in the sense that the degree of bias could not be reduced by changing the answering format to a Self-generated format.

List of Papers

I. Riege, A. H., & Teigen, K. H. (2013). Additivity neglect in probability estimates: Effects of numeracy and response format. *Organizational Behavior and Human Decision Processes, 121*(1), 41-52. doi: 10.1016/j.obhdp.2012.11.004[1]

II. Riege, A. H., Sulutvedt, U., & Teigen, K.H, (2014). Format dependent probabilities: An eye-tracking analysis of additivity neglect. *Polish Psychological Bulletin, 45*(1), 12-20. doi: 10.2478/ppb-2014-0003

III. Riege, A. H., & Teigen, K. H. (2015, manuscript submitted for publication). Everybody will win, and all must be hired: Comparing additivity neglect with the nonselective superiority bias.

[1] Please note that a mistake in Paper I (Table 3) is corrected in a corrigendum placed immediately after Paper I in the present dissertation.

Introduction

Our lives are fraught with uncertainty: what will happen to the key policy rate? Which political party will win the next election? Which football team will win the next world cup? Which summer month will have the highest number of sunny days? Why isn’t my car starting? Moreover, all of these uncertain events have more than one possible outcome, but the different outcomes are often not equally likely. For example: at the present time, the Norwegian Bank’s key policy rate is set at 1.25%. In the future it can decrease (<1.25%), stay the same (1.25%), increase by a little (1.30% - 1.60%), or increase with a lot (>1.60%). When trying to decide your price range for purchasing a new apartment, these possible outcomes will affect your mortgage differently. We thus need to consider the chances of these future outcomes before we make our decision. However, when people are asked to estimate the probabilities for a set of mutually exclusive and exhaustive outcomes, they often produce probabilities that add up to more than 100% (or $p > 1$) (Robinson & Hastie, 1985; Teigen, 1974b, 1983), thus violating the additivity principle stated by formal probability theory. We call judgments involving the relative standing of several candidate events *referent-dependent judgments*, and the particular error in people’s judgment *additivity neglect*.

The present work investigates why people fail to adhere to the principle of additivity in three papers. In Paper I we examined the role of numeracy in judgments of both single outcome events and disjunctive outcome events, and the effect of reminding participants that mathematical rules apply to the probability tasks. We found evidence of numeracy being related to additive responses in the single outcome events, but not in the disjunctive outcome events. Further, giving participants the numeracy test prior to the probability tasks led to more additive responses, particularly for the more numerate participants. Paper I also revealed that different answering formats affect additivity neglect, where writing the probabilities in empty slots next to the alternatives (Self-generated format) lead to more additive responses than circling numbers on a scale (Scale format). In Paper II we further investigated the difference in answering formats and the claim of case-based processing by monitoring participants eye-movements whilst making their judgments. The results revealed that the Self-generated format encourages a more comparative processing thus leading to more additive responses. As referent-dependent judgments need not only be judgments of probabilities, Paper III compared additivity neglect to another bias that can arise in referent-dependent judgments, namely the nonselective superiority bias (NSSB). This bias is the systematic evaluation of individual members of a positive set of items (e.g., five desirable vacation destinations) as superior to most other members in the set (Giladi & Klar, 2002), thus yielding an “unbalanced” set of judgments, as a set of attractive candidates cannot all be rated as better than the group mean. The results show that the same participants often display both biases. In addition, verbal reports demonstrated that people fail to consider normative constraints due to considering alternatives one-by-one (case-based), or because the constraints would lead to judgments not representing their true beliefs.

In the following, I will first briefly introduce referent-dependent judgments. As additivity neglect can be characterized as a type of subadditive judgment, I will explain these two related concepts, the differences in how they are investigated, before I review specific findings related to additivity neglect. Next, some relevant theoretical perspectives will be introduced. Firstly I will discuss support theory and its shortcomings in explaining additivity neglect. Subsequently, I will introduce important terms such as numeracy, case-based processing, and the nonselective superiority bias. The next part will summarize the results from the three papers included in the dissertation, followed by a discussion of methodological considerations. Lastly, the General Discussion will discuss possible explanations, theoretical implications and future directions of the present work.

Referent-dependent judgments

A judgment is an assessment or evaluation of something or someone, and can be rankings, estimates, predictions, ratings, categorizations, etc. (e.g., Hardman, 2009). Normatively speaking, one could argue that all judgments are referent-dependent (Sanbonmatsu, Posavac, Kardes, & Mantel, 1998), in the sense that they presuppose comparisons with a standard, a default value, or with non-focal alternative objects or outcomes. This is also the case with probability judgments. Firstly, the probability of an event occurring has at least one counterpart, which is the event not occurring. Thus, if we are considering the chance of the key policy rate increasing next year, we should also consider the chances of the key policy rate *not* increasing. Secondly, there will often be more than one counterpart, as demonstrated above. In general, any type of judgment that requires dividing a fixed pie into separate pieces can be considered a referent-dependent judgment (Windschitl, Conybeare, & Krizan, 2008; Windschitl, Rose, Stalkfleet, & Smith, 2008). In other words, such judgments not only require people to consider the *target* (the outcome in focus), but also the *referents* (the other possible outcomes in the set).

A frequent error in referent-dependent judgments is inaccurate judgments of numerical probabilities. There are two ways in which numerical probability judgments can be wrong: they can be inaccurate according to a *correspondence criterion*, or according to a *coherence criterion* (Carlson & Yates, 1989; Winman, Juslin, Lindskog, Nilsson, & Kerimi, 2014). The correspondence criterion is satisfied when a person's judgments attain observed accuracy, like for example the relative frequency of an event observed in the real world; the coherence criterion is achieved when a person's judgments are consistent compared to logical, mathematical, or statistical rules, such as the principle of additivity.

Numerous studies have shown that people's judgments can be normatively incorrect, and often systematically deviate from what is predicted by probability and utility theory, such as conjunction and disjunction fallacies (Carlson & Yates, 1989; Tversky & Kahneman, 1983) and base-rate neglect (Bar-Hillel, 1980; Barbey & Sloman, 2007; Kahneman & Tversky, 1972). The systematic deviation from the norm is important; if the errors in judgments were random they would cancel each other out. However, as judgments are systematically biased towards either overestimations or underestimations, most judges will make the same mistake. Needless to say, this can be problematic.

Subadditivity and Additivity neglect

Additivity neglect is a type of subadditive judgment, and subadditivity is often defined as "the probability of the whole is judged to be less than that of the sum of its parts" (Hastie & Dawes, 2010, pp. 174). Although both the name and the definition of subadditivity seems to imply that "the whole" is underestimated it would be more correct to say that the individual parts are overestimated (Koehler, 2000). Subadditivity is a robust finding that has been demonstrated in many studies, using both students and experts as participants, such as experienced physicians (Redelmeier, Koehler, Liberman, & Tversky, 1995), lawyers (Fox & Birke, 2002), chess players (Nordbye & Teigen, 2014), basketball fans (Fox, 1999), rescue workers (Hill, 2012), and options traders (Fox, Rogers, & Tversky, 1996). Subadditivity research has been conducted using several different designs, focusing on violations of mainly two different mathematical norms, namely non-extensionality and non-additivity. It is beyond the present dissertation to make claims to whether or not these types of subadditivity are variants of the same phenomenon or if they are separate biases with different underlying cognitive processes. Nonetheless, one way of organizing the literature is according to which normative rules are being violated and the differences in experimental setup.

The first way of investigating subadditive judgments is as systematic overestimations of all (or most) specified targets, and is sometimes referred to as *generic subadditivity* (e.g., Fox & Birke, 2002). Such studies often use a between-subject design, where participants receive one alternative belonging to an exhaustive set of outcomes, but are not asked to give estimations for the whole set (Fox & Birke, 2002; Redelmeier et al., 1995). The estimates given across the conditions are then added up, and if these estimates exceed 100%, participant's judgments are seen as subadditive. An example of this paradigm is found in Redelmeier et al. (1995). In this study four groups of physicians were asked to estimate one of four prognoses (dying during admission; surviving the admission, but dying within one year; living for one year, but less than ten years; and surviving for more than ten years) for a 67 years old patient with acute myocardial infarction. When the physicians' estimates in each condition were added up, the probabilities for all the patient's prognoses reached 164% rather than 100%, as required by the additivity principle. However, each participant can be said to only being faulted with an overestimation of his or hers specific target, akin to a focusing illusion (Kahneman, Krueger, Schkade, Schwarz, & Stone, 2006; Schkade & Kahneman, 1998; Wilson, Wheatley, Meyers, Gilbert, & Axsom, 2000). Only when adding participants' estimates does the violation of additivity become transparent.

The second way of studying subadditive judgments is as violations of the mathematical principle of extensionality, which states that events with the same extension must be assigned the same probability. This type of subadditivity is sometimes called *implicit subadditivity* (e.g., Fox & Birke, 2002; Rottenstreich & Tversky, 1997). Research investigating violations of extensionality often compare judgments of *packed* versus *unpacked* hypotheses. A packed hypothesis will contain a nested set of alternatives and ask for a judgment of for example, "dying of unnatural causes". The unpacked hypothesis will ask for individual judgments of all (or some) alternatives belonging to the packed event, like for example, being murdered, dying in a car accident, drowning, etc. An unpacked hypothesis thus contains an explicit subset of some, or all, alternatives belonging to the uncertain event, whereas the packed hypothesis will contain a nested set of unspecified alternatives. In order to not violate extensionality, the probability of a packed hypothesis should be equal to or greater than the sum of an unpacked one. However, many studies have demonstrated that people frequently overestimate the probability of unpacked sets, rendering probability estimates far exceeding that of the packed set (Fox & Birke, 2002; Fox & Clemen, 2005; Rottenstreich & Tversky, 1997; Tversky & Koehler, 1994). An example of an early such study is the fault tree experiment by Fischhoff, Slovic, and Lichtenstein (1978). A fault tree is a visual organization of possible sources of a problem, in this case, reasons why a car would not start. Participants were not given the same fault trees, with the expectation that participants would detect the missing alternatives on their trees. However, the results showed that participants were rather insensitive to the alternatives left out of their fault tree, as their judged probabilities for the outcomes that were packed (i.e., probable alternatives "missing" from the fault tree) were small. Two important notes should be made about this type of studies: Firstly, they mostly use a between-subjects design, where participants either judge the packed set of outcomes or the unpacked set of outcomes (for an exception see Hill, 2012). Secondly, participants may be reminded to make sure their estimates added to 100%, as adherence to the principle of additivity for an exhaustive set is not necessarily the focus of such studies.

This leads to the third way of studying study subadditive judgments, namely as violations of the principle of additivity, using a within-subjects design and giving participants the full set of outcomes. However, despite such studies using a within-subjects design, presentation of the outcomes is often either sequential or includes distractor items, making it difficult for participants to parse out which outcomes belong together in a set (e.g., Dougherty & Hunter, 2003; Fox & Tversky, 1998; Rottenstreich & Tversky, 1997). For example, a study asking participants to judge the winning probabilities of eight teams in the NBA basketball quarterfinals, presented the outcomes in lists containing both single and disjunctive outcomes, 14 outcomes in total (Fox & Tversky, 1998). Such designs often originate from the assumption that presenting all alternatives together will make the additivity requirement transparent and thus eliminate the bias (Koehler, Brenner, & Tversky, 1997; Macchi, Osherson, & Krantz, 1999). However, some early studies by Teigen (1974a, 1974b, 1983, 1988) and Robinson/Van Wallendael and Hastie (1985; 1990) demonstrated that people often produce probability sums exceeding 100%, even in a within-subjects design. Given that participants have problems remaining within the 100% limit when the outcomes are presented together as an exhaustive set, it is perhaps not surprising that participants fail to do so in studies using a more complex layout of the alternatives.

In the present studies all the alternatives are presented on the same page (except Experiment 4, Paper I), asking participants to give estimates for each outcome. We have called this type of global subadditivity *additivity neglect*, as the term more aptly describes the bias (Riege & Teigen, 2013). For example, even though only one team can win the

World Cup, people's judgments of the individual team's chances often sum up to more than 100%, thus neglecting the principle of additivity.

Additivity neglect

Within psychology, non-additive probability judgments was originally documented by Teigen (1974a, 1974b), and earlier studies of subjective probability estimates assumed that people would adhere to the rule, instructing people to make sure their estimates added up to 1 or 100% (Kahneman & Tversky, 1972; Peterson, Ducharme, & Edwards, 1968) . However, Teigen (1974b) showed that when allowing participants to give "unrestricted" estimates they frequently violate the principle of additivity. People can violate the additivity axiom in two ways, they can overestimate the total probabilities, or they can underestimate the total probabilities which is often referred to as *superadditivity* (e.g., Sloman, Rottenstreich, Wisniewski, Hadjichristidis, & Fox, 2004). Superadditivity are probabilities that sum to less than 100%. Such systematic underestimations are rare, but can occur when events are atypical (Sloman et al., 2004), unpacked in great detail (Redden & Frederick, 2011), or outside participants field of knowledge (Macchi et al., 1999). The present work concerns judgments that add to more than 100%.

Overall, studies investigating additivity neglect have shown that the sum of participants' probability estimates increase with the number of alternative outcomes even when each set is exhaustive (Robinson & Hastie, 1985; Teigen, 1983). When participants are given two outcomes, most estimates will be additive (Robinson & Hastie, 1985; Teigen, 1983; Tversky & Koehler, 1994; Van Wallendael, 1989; Van Wallendael & Hastie, 1990, for an exception see McKenzie, 1999). Increasing the number of alternatives leads to an increase in probabilities. For example, Teigen (1983, Exp.II) showed that by increasing the number of outcomes to four; seven; and ten alternatives, across three tasks, increased participants means from 130.1% for four alternatives, 190.9% for seven alternatives, to 227.4% for ten alternatives. Across the three tasks, only 8.8% of the participants consistently gave estimates adding to 100%. Moreover, in Experiment IV, the experimenter added two new alternatives to the set of outcomes during the data collection; yet only 16% reduced their prior estimates, even when the experimenter explicitly allowed them to change their original estimates. Most participants thought the two new outcomes had roughly the same chances as the first four. This demonstrates another important finding, specifically, that participants seem to lack an understanding of the complementarity of the outcomes. Adding two new, and apparently plausible, alternatives to the set should have resulted in a downward adjustment of the judged chances of first four alternatives. However, very few changed their original estimates. In a similar study, participants playing an experimental version of a murder mystery game, were asked to revise the estimates of all the suspects for each new clue they were given. Participants rarely, or insufficiently, revised their estimates for the referent suspects when changing the estimate of a focal suspect in light of new evidence (Robinson & Hastie, 1985). This meant that evidence implying a suspect's guilt was treated as if it had no bearing on the guilt of the other suspects. Thus, when one suspect's probability increased, the others did not decrease. Further, when one suspect was eliminated by a strong "innocent" clue, the other suspects' probabilities did not increase. One explanation for additivity neglect has therefore been that people either fail to use their mathematical skills when making probabilistic judgments of uncertain events, or that they altogether lack the mathematical skills necessary to give normative responses (Robinson & Hastie, 1985; Teigen, 1983). Another explanation is that the bias occurs due to a case-based approach to the probability judgments, where people consider one alternative at the time, thus failing to take the complementarity of the alternatives into account (Sanbonmatsu, Posavac, & Stasney, 1997; Teigen & Brun, 2011).

Theoretical background

Support theory

The broadest theoretical account of subadditivity, and thus additivity neglect, is support theory (Tversky & Koehler, 1994). Support theory is a descriptive account of subjective probabilities, developed to account for biases in numeric probability judgments, and one of the foundations is subadditive judgments. The theory is mainly concerned with explaining why people violate the extensionality principle (implicit subadditivity), and suggests that the failure to adhere to the axiom of extensionality has two main sources. The first is limitations of memory. People cannot be expected to recall or list all possible subsets of a category/event, although they will be able to recognize them when explicitly reminded. An explicit description might therefore remind people of relevant cases, despite not being able to list them. The second reason for non-extensionality is that different descriptions of the same event might direct people's attention to different aspects of the outcome and thus affect the saliency of the descriptions. Support theory states that probability estimates are based on descriptions of events, which within the support theory framework is called *hypotheses*, as opposed to the actual events (Fox & Birke, 2002; Tversky & Koehler, 1994).

Support theory further suggests that a probability estimate will depend on the perceived evidence, or *support*, for the focal hypothesis, relative to the support for the referent hypotheses (Idson, Krantz, Osherson, & Bonini, 2001; Rottenstreich & Tversky, 1997; Tversky & Koehler, 1994). Support theory thus suggests that participants' estimates are formed based on the perceived balance between supporting vs. non-supporting evidence. This might be done using actual statistical information or by mental shortcuts such as the availability or representativeness heuristic (Tversky & Kahneman, 1974). It is possible that participants use heuristics such as representativeness or availability when making their estimations, as both heuristics are said to substitute a difficult question ("what are the chances of the next Nobel Peace Prize laureate being Asian?") with an easier question ("how many potential future Asian Nobel Peace Prize laureates can I think of?") (Kahneman, 2011; Kahneman & Frederick, 2002, 2005). Subadditivity thus arises due to people expressing their beliefs, which can be based on both correct and incorrect evidence, and their probability judgments reflect the perceived weight of the evidence for each outcome. Support theory claims that people's beliefs need not adhere to mathematical constraints, and thus allows probability judgments to be both non-extensional and non-additive.

However, the heuristics and bias tradition has been criticized for being "one word explanations" that do not specify the cognitive process underlying, for example, the representativeness and availability heuristics (Gigerenzer, 1998). In addition, it is argued that heuristics lack predictive validity, as it is difficult to predict in advance which heuristic participants will make use of (Anderson, 1991; Kahneman & Frederick, 2002), though some attempts have been made (Braga, Ferreira, & Sherman, 2015). Thus, support theory does not fully account for the cognitive mechanisms involved in additivity neglect. Further, as support theory sidesteps participants lack of additivity by "strength of beliefs" not needing to comply with classical probability theory, it does not fully account for why people fail to adhere to the principle of additivity. Accordingly, support theory fails to explain that even when nothing is "hidden" from participants, they grossly overestimate the likelihood of each individual alternative. The questions of why people give belief based estimates disregarding the laws of probability and which cognitive mechanisms involved in such judgments thus still remains.

Numeracy and understanding probability theory

In order to investigate if people's failure to use formal mathematical theory is rooted in a lack of mathematical knowledge, we decided to investigate the relationship between additivity neglect and numeracy in Paper I. Numeracy is the ability to understand, use and reason with numbers (Peters, 2012; Peters et al., 2006). Moreover, numeracy as a construct does not merely reflect pure mathematical skills, but refers to mathematical or quantitative literacy (Ghazal, Cokely, & Garcia-Retamero, 2014; Nelson, Reyna, Fagerlin, Lipkus, & Peters, 2008; Reyna, Nelson, Han, & Dieckmann, 2009). In fact, the term numeracy was coined in 1959 by Geoffrey Crowther of the U.K. Committee on Education as a word for "numeral literacy" (Reyna et al., 2009). The most basic levels of numeracy involve understanding "the real number line, time, measurement and estimation" and "the ability to perform simple arithmetic operations and compare numerical magnitudes". Higher levels of numeracy concerns "basic logic and quantitative reasoning skills, knowing when and how to perform multistep operations, and an understanding of ratio concepts like fractions, proportions, percentages and probabilities" (Reyna et al., 2009, p. 945).

Given that several heuristics and bias tasks can be solved mathematically, it is reasonable to assume that individual differences in numeracy might be associated with correct responses to such tasks. Even though there are, to our knowledge, no other studies investigating the relationship between numeracy and additivity neglect (or subadditivity), numeracy has been found to predict susceptibility to several other biases (Liberali, Reyna, Furlan, Stein, & Pardo, 2012; Peters, 2012; Peters & Levin, 2008; Winman et al., 2014). For example, Peters et al. (2006, Experiment I) found that participants with a higher numeracy score were less affected by attribute framing effects than those with a lower numeracy score. However, they found weak/no effects of numeracy on effects of risky choice framing (Peters & Levin, 2008). Several studies have also found that the more numerate participants perform better in Bayesian reasoning tasks (Chapman & Liu, 2009; Galesic, Gigerenzer, & Straubinger, 2009; Sirota & Juanchich, 2011), and make fewer conjunction and disjunction errors (Liberali et al., 2012; Winman et al., 2014). However, other studies have failed to find an effect of numeracy on conjunction errors (Wedell, 2011). The relationship between high numeracy and normative responses to classical heuristics and bias tasks is thus not clear cut.

Over the years, several tests designed to measure numeracy have been developed. One of the earliest tests included only three items and was developed in order to assess individual differences in ability to use numerical information about the benefit of mammography screening (Schwartz, Woloshin, Black, & Welch, 1997). The Lipkus numeracy test (Lipkus, Samsa, & Rimer, 2001) was based on Schwartz et al.'s (1997) test with eight tasks added for a finer grained test. This test has been extensively used in educated samples, and we thus chose the Lipkus test for Experiment 1 (Paper I). However, in our student sample the Lipkus test yielded highly skewed scores, where almost 2/3 of our participants attained a score of 10 or 11 out of 11 possible, thus rendering a negatively skewed distribution of answers. During the data collection for Paper I we became aware of a newly developed numeracy test, namely the Berlin Numeracy Test (BNT) allowing for a better discernment between highly educated samples (Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012). Experiment III thus used the BNT in order to further replicate the findings from Experiment I. The results from both experiments showed that numeracy was related to probability sums in the probability tasks.

Some participants may not have the necessary mathematical skills in order to give additive responses. However, another explanation could be that they simply do not think about mathematical rules applying to the probability tasks, particularly for tasks where the probabilities cannot be calculated, but have to be given as subjective estimates. As providing additive responses is dependent on realizing that a mathematical rule is applicable to the situation, we were also interested in whether the numeracy test itself could function as "calculation priming". Previous findings show that people's mindset can affect their judgments (Bless, Betsch, & Franzen, 1998; Hsee & Rottenstreich, 2004). For example, Hsee and Rottenstreich (2004) found that priming people with either *feeling* or *calculation* affected their evaluation of second-hand box sets containing 5 or 10 Madonna CDs. Participants primed with calculation gave prices where the quantities of the items were taken into account, whereas those primed with feeling disregarded the set size, giving prices reflecting how much they liked the items. We thus decided to vary the order of the numeracy test, by either presenting it before or after the probability tasks.

Differences in answering format

Differences in answering format emerged as an unexpected factor in explaining variations in additivity neglect (Riege, 2011; Paper I). However, it is well known in the literature that different contexts can affect people's judgments (Bless et al., 1998; Brun & Teigen, 1988; Gigerenzer & Hoffrage, 1995; Hsee & Rottenstreich, 2004; Reeves & Lockhart, 1993). For example, providing participants with information about protagonists in a case-specific manner leads to more conjunction and disjunction errors compared to frequentistic information (Reeves & Lockhart, 1993); and expressing probabilities as frequencies can reduce participants base-rate neglect (Gigerenzer & Hoffrage, 1995).

Other studies have also shown that judgments can be influenced by answering formats. For example, overconfidence in probabilistic judgments can be dependent on response formats (Juslin, Wennerholm, & Olsson, 1999; Klayman, Soll, Gonzalez-Vallejo, & Barlas, 1999; Teigen & Jorgensen, 2005), such as half-range formats leading to under-confidence, full-range format leading to more calibrated responses, and an interval format leading to overconfidence (Juslin et al., 1999). Within other domains there are also several studies showing that the way rating scales are constructed can influence participants answers (e.g., Schwarz, 1999). For example, when asking participants to report their success in life, varying the numbers of the 11-point scales ranging either from -5 to 5, or from 0 to 10, produced considerably different answers, where almost three times as many participants used the left side of the scale in the first version compared to the second version (Schwarz, Knäuper, Hippler, Noelle-Neumann, & Clark, 1991). The authors argued that including the negative numbers made participants think not only of their success, but also of their failures. Thus, seemingly small differences in the task context can sometimes affect the way people think about the task at hand, and the answers they give.

Case-based versus class-based approach

Several studies have suggested that people can adopt two different approaches to probability judgments, namely a *case-based* or *class-based* approach (Fox & Rottenstreich, 2003; Fox & Ülkümen, 2011; Kahneman & Tversky, 1982; Reeves & Lockhart, 1993; Teigen & Brun, 2011). Within psychology, the distinction originates from Kahneman and Tversky (1982), who distinguish between singular versus distributional approaches (Kahneman & Tversky, 1982; Klar, Medding, & Sarel, 1996; Reeves & Lockhart, 1993). It is beyond the present thesis to parse out the similarities and differences between these terms. The main issue is that these are two different ways of viewing external uncertainty (uncertainty related to events people cannot control). According to *class-based* (distributional) approach, the event in question is seen as one of several events or outcomes, and the relative frequencies of the event is known or can be estimated. Imagine a job hiring scenario with four applicants, and your job is to assess Bob's chances of being hired. Without knowing anything about the applicants, allotting Bob a 25% chance would be a good guess. However, Bob has an excellent grade point average and 10 years of experience from a similar job. What is Bob's chance of being hired? Given this case-specific information it might be tempting to estimate Bob's chances to be higher than 25%. This would be compatible with a class-based approach as long as the other applicants were adjusted downwards in a complementary fashion. A purely case-based (singular) approach would be to assess Bob's probabilities of being hired based on his assets alone, disregarding the number and skills of his competitors. *Case-based* assessments are typically based on the perceived propensities of the particular event or case. It has been argued that people use the two approaches in different situations, favoring the class-based approach only when no individuating or case-specific information is available. For example, when judging vulnerability to negative life events, participants mostly considered statistical information, such as base-rates, when evaluating a generalized target (e.g., the average peer), while making use of case-specific information when considering a familiar target (e.g., self or close other) (Klar et al., 1996). The case-based approach is often favored by people in real life as case-specific information is often readily accessible (Kahneman & Tversky, 1982).

Using a case-based approach also implies considering each alternative individually without much comparison between them, and as a consequence, allotting probabilities unaware of, or disregarding, the 100% "rule". Further, understanding the complementary of the set of outcomes is a fundamental prerequisite for an additive response, making comparisons between the alternatives necessary. A class-based approach would therefore involve comparing all, or at least some of the alternatives, whilst keeping a mental tab of the 100% "rule". We argue in Paper I that the Self-generated format facilitated a class-based approach to the probability tasks, and we wanted to further investigate this claim in Paper II using eye-tracking methodology allowing us to monitor participant's eye-movements during the judgment tasks. Process tracing methods offer a means to gain insight to the decision process itself, instead of inferring the process based on the end result (the judgments). It has thus become increasingly common to use process measures in judgment and decision making research, as a means to further understand the underlying cognitive processes (Glaholt & Reingold, 2011; Schulte-Mecklenbeck, Kuehberger, & Ranyard, 2010; Schulte-Mecklenbeck, Kuhberger, & Ranyard, 2011).

We also hypothesize that a class-based approach will be more cognitively demanding than the case-based approach, because one has to combine the evaluation of case-specific information with a strict numerical rule. Interestingly, Robinson and Hastie (1985, Experiment 1) taught one group of participants probability theory prior to the (aforementioned) murder mystery tasks. Aside from these participants giving additive responses, verbal reports revealed that staying within the normative limit was a difficult task. Participants also seemed to narrow the number of suspects early, probably as a way of reducing the number of alternatives they had to consider. This indicates that giving additive responses with multiple comparisons between the alternatives involves a high cognitive load. This hypothesis could be investigated further by means of an eye-tracker, as the equipment also monitors fixation durations which are often used as proxies for cognitive load (Findlay & Kapoula, 1992; Horstmann, Ahlgrimm, & Gloeckner, 2009; Velichkovskiĭ, 1999).

As Paper II indicated that the two answering formats led to (somewhat) different cognitive processes, where the Self-generated format prompted more comparisons, we decided to include a different process measure in Paper III, namely a retrospective verbal report. In a previous study, investigating unrealistic optimism in comparative risk judgments, participants were asked to provide written accounts of their thought processes (Klar et al., 1996, Experiment 3). The results indicated that participants often used case-specific information in their decisions, indicating a case-based approach to the probability judgments. As we were interested in the violation of the mathematical rules, we asked participants why they had, or in most cases, why they had not, adhered to the normative rules. Our aim was to investigate if some participants would describe a case-based assessment of the alternatives, by for example answering that they considered the alternatives one-by-one.

The literature cited so far belongs at large to a cognitive approach to subjective probability judgments. Interestingly, within the social psychology literature one can find a distinction similar to case-based vs. class-based approaches to referent-dependent judgments, labeled selective vs. comparative processing (Kardes, 2013; Sanbonmatsu et al., 1998; Sanbonmatsu et al., 1997; Sanbonmatsu, Vanous, Hook, Posavac, & Kardes, 2011). Selective processing resembles the case-based approach in that people have a tendency to search for evidence related to the target or focal event, rather than evidence related to the referents (Sanbonmatsu et al., 1997). A large body of research has indicated that several biases or judgmental errors can result from selective processing (Kardes, 2013; Sanbonmatsu et al., 1998; Sanbonmatsu et al., 2011).

The nonselective superiority bias

As mentioned above, a large body of research has indicated that several biases or judgmental errors can result from selective processing (or a case-based approach), amongst them both the nonselective superiority bias and subadditivity (Kardes, 2013; Posavac, Brakus, Cronley, & Jain, 2009; Posavac, Brakus, Jain, & Cronley, 2006; Sanbonmatsu, Posavac, Kardes, & Mantel, 1998; Sanbonmatsu, Vanous, Hook, Posavac, & Kardes, 2011). However, these biases are rarely studied together, despite an assumption of their shared foundation. In Paper III we wanted to compare additivity neglect to the nonselective superiority bias in a parallel design, and involving the same participants.

The nonselective superiority bias (NSSB) is the phenomenon where participants consistently judge individual members of a positive set of items (e.g., five attractive vacation spots, five pleasant smelling soaps) as superior to most other members in the set (Bruchmann et al., 2013; Giladi & Klar, 2002; Klar, 2002; Krizan & Suls, 2008; Suls et al., 2010; Windschitl, Conybeare, et al., 2008). This systematic judgment of all items in a set as superior to each other violates elementary logic, as some members cannot be better unless others are worse.

The similarity between additivity neglect and the nonselective superiority bias can be observed in several aspects: firstly, both biases have at large been studied in between-group designs, presumably under the assumption that using a within-group design would reduce the bias. However, as with additivity neglect, a few studies have indicated that merely asking participants to judge the full set of outcomes does not alleviate the bias (Chambers, 2010, Experiment 2; Klar, 2002, Experiment 4). Secondly, both types of referent-dependent judgments seem to be made in a non-complementary fashion, that is, participants appear to view selected items as unrelated to the other items in the set. Lastly, some of the mechanisms posited to explain NSSB are similar to those suggested for additivity neglect. Despite these similarities, few attempts have been made to study these two biases together (for a recent integrated treatment, see Smith, Windschitl, & Rose, 2015).

Methodological considerations

Participants

The participants in the three papers were either students recruited from various universities, or participants recruited on Amazon Mechanical Turk (MTurk). All participants in Papers I and II were students, mainly from the University of Oslo, but also from the Universities of Bergen and Tromsø, and from BI Norwegian Business School. Except for the students in Experiment 3 in Paper I, the students did not receive any compensation for their participation. Participants in Paper III were all recruited from MTurk (www.MTurk.com), which is an online marketplace offering people small jobs they can complete in their own homes. The site thus contains the main elements required to conduct research: a large participant pool with an integrated system for compensation and a streamlined process of participant recruitment and data collection. In general, MTurk workers are diverse, but several studies have indicated that they, as is the case with student populations (Peterson, 2001; Sears, 1986), differ from the general population in several characteristics. For example, MTurk workers tend to have a lower income, and are younger, less religious, and more liberal than the general population (Berinsky, Huber, & Lenz, 2012; Paolacci, Chandler, & Ipeirotis, 2010). Studies have also indicated that the MTurk samples differ in some personality traits and are less extraverted than students and the general public (Goodman, Cryder, & Cheema, 2013). Investigations of cognitive abilities have found no difference between MTurkers' and students' numeracy scores (Paolacci et al., 2010), and a recent study indicates that MTurkers are more attentive to instructions than students and thus show larger effects in response to subtle manipulations (Hauser & Schwarz, 2015a). However, MTurk samples are more diverse than student samples (Buhrmester, Kwang, & Gosling, 2011), particularly in regards to demographic variables. Regardless of the relative merits of these two populations, neither can be considered as representative of the general population. Thus, the current findings may not be valid in all samples or in all cultures. Nonetheless, some of the effects reported in the present work have been found in student samples all across Norway, and in MTurk samples located in the US, yielding some generalizability.

External validity of Questionnaire-based experiments

All studies have been questionnaire-based experiments, conducted either in a classroom setting (Experiments 1, 2, and 3, in Paper I), as online experiments (Experiment 4 in Paper I, Experiments 1, 2, and 3, in Paper III), or in a laboratory setting (Paper II). There are several unnatural aspects of such experimental settings, which might cause people to consider the tasks differently than they would in a more natural environment.

The physical setting is perhaps most artificial in the classroom and laboratory settings, although this might not be so unnatural for students who frequent such places often. Regardless, in lecture halls there are many students as data collection is often done during a break, and this might reduce concentration. In the laboratory participants were in a small room together with the experimenter, giving estimations whilst trying to sit still and not to blink too much, which is an unnatural situation. The MTurk workers can have been anywhere, which could have affected their responses. However, all MTurk studies included an attention check (Oppenheimer, Meyvis, & Davidenko, 2009), presented at the end of the questionnaire (Hauser & Schwarz, 2015b) , asking participants a multiple choice question where the correct answer was provided in the instructions. Most importantly, despite participants' location or experimental setting, the same pattern of judgments was found in classrooms, the lab, and in online experiments. In addition, subadditivity has been replicated in numerous samples including various expert populations with a high external validity (e.g., search and rescue workers using maps in Hill, 2012).

The tasks and the topics of the tasks include vignettes of both real and hypothetical situations, and topics. Not everyone has knowledge about the selected topics, nor an interest in politics, football, hiring situations, possible reasons for car trouble, etc. However, the numerous tasks given to participants cover a wide array of topics, all showing a similar pattern of results.

Eye-tracking

Though process-tracing methods such as eye-tracking have become more common in psychological research, the interpretations of the results are still contingent on auxiliary hypotheses about what the specific eye-tracking measures mean. The most fundamental problem with eye-tracking is the assumption that the viewer's attention is fixed at the gaze point, commonly referred to as the eye-mind hypothesis (Just & Carpenter, 1980). We assume that the recorded fixations give information about what participants are looking at, though people are perfectly capable of directing their attention to their peripheral visual field, without moving their gaze (Posner, Snyder, & Davidson, 1980). However, gaze and attention are usually closely linked, with attention being directed at a target shortly before the gaze (Hoffman & Subramaniam, 1995). In addition, detaching attention away from the gaze requires effort, thus making it less likely that participants would make such an effort in our experimental setting. However, if the visual stimuli is “cluttered” making the Areas of Interest (AOI) small and closely spaced together, it is possible for participants to see several AOI’s in their (semi)peripheral vision, making it difficult to know what they have really been attending to (Orquin, Ashby, & Clarke, 2015). A related problem is that even though the eye-tracker is calibrated with great care at the beginning of each data collection, participants might move their head, blink too frequently, etc., thus reduce the accuracy of the measurement. Again, if the AOI’s are small or closely spaced together, even small deviations in the calibration can affect the results (Orquin et al., 2015). However, recent advances in eye-tracking technology has increased the precision in terms of collecting both temporal and spatial information (Glaholt & Reingold, 2011), and we used a 9-point calibration to make sure our measurements were as accurate as possible.

Another auxiliary assumption concerns fixation durations, which are seen as an indirect measure of cognitive load, where longer fixations reflect more cognitive effort than short fixations (Findlay & Kapoula, 1992; Horstmann et al., 2009). However, some studies have shown that long fixation durations do not always indicate high cognitive effort. For example, people driving in monotone and boring landscapes have very long fixation-durations (Holmqvist, 2011). In such cases it is easy to understand that the fixation durations are not connected with high cognitive load. However, other situations may be less transparent thus leading to incorrect interpretations. In Paper II we found some evidence that longer fixation durations were associated with additive responses. As additive responses are contingent on mental calculations, it is reasonable to assume that the longer fixation durations indicate a higher cognitive load. We also asked our participants informally, both after completing the eye-tracking experiment and other experiments, how they experienced our tasks. Several participants have mentioned that they found it taxing to adhere to the 100% - rule.

Lastly, it is important to consider whether introducing a process method such as eye-tracking can alter the decision making process. Several comparative studies have shown that some process methods influence the decision process by for example adding extra cognitive demand (Glaholt & Reingold, 2011; Lohse & Johnson, 1996). For example, a comparison between eye-tracking and Mouse Lab showed that participants’ needed more time to gather information and that information acquisition behavior tended to be more systematic in Mouse Lab compared to eye-tracking (Lohse & Johnson, 1996), thus indicating that eye-tracking as a method affected participants behavior less than Mouse Lab. Further, as eye movements require little deliberate effort, it is unlikely that the process tracing actually alters the decision making process (Glaholt & Reingold, 2011). More importantly, participants in Paper II demonstrated the same additivity neglect and answering format differences as participants in Papers I and III. These measures were independent of the eye-tracker indicating that the eye-tracker did not affect the outcomes of participants' judgments.

Summary of papers

Paper I - Additivity neglect in probability estimates: Effects of numeracy and response format

Authors: Anine H. Riege and Karl Halvor Teigen

Several studies have shown that people who are asked to estimate the probabilities for an exhaustive set of more than two events often produce probabilities that add up to more than 1 or 100% (Robinson & Hastie, 1985; Teigen, 1974b, 1983). We draw a distinction between studies of *local* additivity where participants are given a disjunctive subset of outcomes, and studies of *global* additivity which entail a total exhaustive set of outcomes. Both biases are often referred to as subadditive judgments, but to distinguish between the two, we call the latter additivity neglect. The aim of Paper I was to investigate several determinants for such additivity neglect.

A series of pilot experiments, using both real life and hypothetical vignettes, were conducted in order to investigate possible determinants of additivity neglect. Two of the pilot experiments, Pilot A and B, had three interesting findings that warranted further investigation (for a full description, see Riege, 2011). Firstly, Pilot A and B used different answering formats: whereas Pilot A asked participants to write down their own probabilities, Pilot B used scales ranging from 0% to 100%, with increments of 10. Participants in the latter gave relatively fewer additive responses compared to the former, indicating that the answering format affected participants' additivity neglect. Secondly, both Pilot A and Pilot B used numeracy tests and probability sums were in both experiments negatively correlated with numeracy scores. These correlations were, however, only significant for the vignettes in Pilot A. In Pilot A the numeracy test was introduced on the last page. Inspection of the questionnaires revealed that several participants had changed their responses. If these changes had occurred after seeing the numeracy test, it was possible that the numeracy test itself could serve as a reminder of a mathematical rule being relevant when answering the probability tasks. Lastly, both pilots had given participants disjunctive probability tasks, asking participants to estimate the chances for five individual teams to be No.1 or No. 2 in a group tournament. Disjunctive probabilities should, on average, be twice the probabilities of becoming No. 1, but were judged to be only 10–15% higher. Performance on the disjunctive tasks appeared to be unrelated to numeracy.

Experiment 1 gave participants five probability tasks pertaining to current events such as a forthcoming political election and the qualification playoffs of Group H in UEFA

EURO 2012. Each task had 4 or 5 outcomes, and participants were asked to judge the probabilities of all the alternatives. The aim of Experiment 1 was to investigate the effects of priming participants with a “calculation” mindset by giving one group the Lipkus (2001) numeracy test before performing the estimation tasks, while another group performed the estimation task first. The estimation tasks and the response format (rating scales) were identical in both groups. In addition, we wanted to further investigate the underestimation of disjunctive probabilities observed in the pilot studies. In the pilot experiments these estimates were made by participants in different conditions. As a means to enhance the difference between single and disjunctive estimates, Experiment 1 gave the same participants three single outcome tasks and two disjunctive outcome tasks.

The results showed that additive responses were twice as common in the Numeracy first as in the Numeracy last condition, indicating that the Numeracy scale “primed” participants to answer according to mathematical principles. However, the priming appeared to be most effective for participants with a high numeracy score, indicating that it is not sufficient to have high numeracy; one must also be reminded that mathematical rules might apply. Most participants gave higher probability sums for the disjunctive tasks, but this increase was not sufficient according to normative expectations. The probability estimates for the disjunctive outcome tasks were unrelated to participants’ numeracy and the mathematical mindset priming manipulation.

Experiment 2 was designed to investigate the format difference observed in the pilot experiments. In addition, there were three outcome conditions that asked participants to assess the probabilities for each of five teams becoming No. 1, No. 2, or among the two best (No. 1 or No. 2). Within each outcome condition, probability estimates were given either by selecting a number on separate scales (scale format), or by filling in appropriate percentages in empty slots where the teams were listed (written format). Differing results in the pilot experiments suggested that the second arrangement may facilitate more additive responses. The present experiment was set up to test this conjecture. The results showed a clear tendency for estimates in the written format to be lower than estimates in the scale format. There was also a tendency for participants to judge the probability of winning to be higher than probability of becoming second best, and the estimates of disjunctive probabilities were, in both formats, quite close to the estimates of single outcomes. Thus participants in this experiment perceive the probabilities of being among the two best and the probabilities of becoming the winner as very similar, and far from being the sum of becoming No. 1 and No. 2. Lower probabilities in the written format were associated with a higher frequency of additive responses, and participants in the single outcome conditions using the written format produced far more additive sums than participants in the scale conditions.

One of the limitations with the Lipkus scale is that it produces a negatively skewed distribution of answers with a large proportion of high numeracy scores (in Experiment 1 almost 2/3 of student participants attained a score of 10 or 11 out of 11 possible). **Experiment 3** was designed to further investigate the role of numeracy and to replicate the two main findings in the previous experiments, using a more appropriate measure of numeracy. We used a recently developed and more advanced numeracy test, namely the Berlin Numeracy Test (Cokely, Galesic, Schulz, Garcia-Retamero, & Ghazal, 2012). Participants were given three probability tasks. Two tasks were about the four remaining teams (at the time) in the Norwegian football championship where one task asked participants to predict the chances for each team becoming No.1 and the second task the chances of becoming No.2. The last task was about an upcoming political election. The experiment utilized two manipulations; the presentation order of the numeracy test (numeracy scale first vs. last) and response format (scale vs. written). As in Experiment 2, the written format led to a much higher number of additive responses than the scale format, for all vignettes. The Berlin Numeracy Test led to a positively skewed distribution of scores, and the participants probability estimates were inversely related to numeracy scores on the Berlin test. The correlations between numeracy scores and mean probability estimates only reached significance for the participants who received the numeracy test first, again indicating that the effects of numeracy are amplified when mathematical skills are activated.

Experiment 4 was designed to disentangle one of the differences between the answering formats. The written format lead to twice as many additive responses as the scale format, but this difference could either be due to visual display (the alternatives are spaced further apart in the scale format because they are separated by the scales themselves) or to response factors (writing in the probabilities vs. circling numbers on a scale). Participants were given three probability tasks with four or five different outcomes each. Experiment 4 contained two conditions where all alternatives were displayed on one screen, with responses given either in the written format or in scale format. In the two other conditions, the alternatives were presented sequentially on separate screens, requiring either self-generated responses (written format) or ratings on 11 point scales (scale format). The results showed that the advantage of a written format persisted, albeit slightly reduced, even when the estimates were given on separate screens. This suggests that self-generated estimates require participants to think more carefully about the task, perhaps encouraging an analytical approach. In contrast, a row of numbers displayed along a rating scale might appear to license the choice of any number that happens to match one's gut feelings. Despite the apparent "ease" of rating scale responses, few were additive, indicating that additivity is not simply obtained by alleviating the cognitive load.

Overall, the results of Paper I show that response format plays a decisive role in additivity neglect. Numerical estimates written by participants themselves appear to facilitate additive responding, whereas estimates made by circling numbers on separate scales for each alternative seem to promote unconstrained case-based responses. Also, the effect of numeracy on additive probability estimates is greatest for highly numerate participants when primed to think in a mathematical manner. Very few participants adjusted their estimates sufficiently in the disjunctive outcome tasks regardless of their answers in the single outcome tasks, the answering format or numeracy. The kind of reasoning that facilitates additive responding in the single outcome condition does therefore not necessarily imply a deeper understanding of the principles of probability calculation.

Paper II - Format dependent probabilities: An eye-tracking analysis of additivity neglect

Authors: Anine H. Riege, Unni Sulutvedt, and Karl Halvor Teigen

In this paper we wanted to further explore the difference in answering format found in Paper I, where writing estimates in an empty slot next to each specified outcome in the set generated more additive estimates than giving estimates by circling numbers on 0-100% horizontal rating scales. It is important to investigate this finding further as such answering formats are often used interchangeably in probability estimation tasks. In Paper I we suggested that the difference in answering format could elicit different cognitive processes; Paper II thus used an eye-tracking measure to attempt to investigate the underlying judgment process.

Eye movements are naturally occurring behaviors and a generally accepted measure of attention, information acquisition, and as a means to infer cognitive processes (Glaholt & Reingold, 2011; Russo, 2011; Schulte-Mecklenbeck et al., 2011). Monitoring eye movements provides information about the decision process by registering what participants are looking at (fixations) and their search strategies by recording the direction of these fixations. Moreover, information about cognitive load can be inferred by looking at individual fixations' fixation durations within and between tasks (Findlay & Kapoula, 1992; Horstmann, Ahlgrimm, & Glöckner, 2009). It is also possible to monitor participants repeated inspections of the same material, by counting the number of revisits between (pre)defined *Areas of Interest* (AOI) within each task. We predicted that fixations, fixation durations, and repeated inspections of outcomes (revisits) will differ based on whether participants assign probabilities by judging the alternatives one by one, in a case-based manner, or additively, as members of a set, which requires a distributive approach where alternatives are compared to each other.

Participants were given ten probability tasks pertaining to both real-life and hypothetical uncertain events, and two control tasks describing chance events where probabilities could be obtained by calculation. Participants were randomly assigned to either the Self-generated answering format or the Scale format. The results showed that participants in the Self-generated condition had significantly more additive responses than participants in the Scale condition, replicating the results from Paper I. However, this finding did not extend to the control tasks where almost all participants gave probability estimates within the 100% limit. Also, participants spent more time completing the tasks in the Self-generated condition, indicating that they might need more time to revise their responses making sure they adhered to the 100%-rule. Despite the visual display in the Scale condition offering participants more to look at, participants in the Self-generated conditions had more fixations. However, as participants in the Self-generated condition spent more time answering the tasks, it is not unreasonable that they also had more fixations; the longer you look at something, the more fixations you will invariably have. Interestingly, the opposite pattern was observed in the control tasks, where participants in the Scale condition had more fixations than those in the Self-generated condition, although there was no difference in the time spent on the control tasks between the two conditions.

As predicted, participants in the Self-generated condition had on average almost twice as many revisits between the alternatives compared to the Scale condition. This indicates that participants in the Self-generated condition were conducting more comparisons between the alternatives, rather than evaluating each alternative individually. It is easy to imagine that comparisons are important for accomplishing an additive response pattern, as it requires seeing the alternatives as a complete set. We also expected to find that the participants in the Self-generated condition had longer fixation durations than participants in the Scale condition. Although the percentages of “long” fixations (≥ 500 ms) were considerably higher than reported elsewhere in eye-tracking literature on decision making, there was no difference between the conditions. This indicates that our probability estimation tasks, regardless of answering format, require considerable cognitive effort. A closer inspection of the standard deviations did, however, show that the fixation durations were more variable in the Self-generated condition. A multiple regression analysis showed that fixation durations were positively related to the number of additive outcomes. However, these results do not allow any strong inferences to be drawn about the causal relationship between search pattern and additivity. Overall, the results from Paper II indicate that the Scale format might prompt a selective evaluation of the alternatives, and thereby discourage comparisons between alternatives. The Self-generated estimates facilitate a class-based approach and make people engage in a more comprehensive search pattern.

Paper III – Everybody will win, and all must be hired: Comparing additivity neglect with the nonselective superiority bias.

Authors: Anine H. Riege and Karl Halvor Teigen

There are several types of biases that can occur in referent-dependent judgments, additivity neglect being but one. However, they are rarely studied together or compared. As an attempt to compare two such biases, Paper III looked at both the nonselective superiority bias and additivity neglect, as both biases violate basic formal constraints. The nonselective superiority bias (NSSB) is the consistent evaluations of individual members of a positive set of items (e.g., five attractive vacation spots) as superior to most other members in the set (Giladi & Klar, 2002); additivity neglect is the overestimations of the probabilities for most alternatives in a set of mutually exclusive and exhaustive outcomes (Riege & Teigen, 2013). The errors in judgments lies in that a set of attractive candidates cannot all be rated as better than the group mean, and the probabilities of a set of exhaustive events cannot add up to more than 100%. We distinguish between the two judgments by calling the former a “comparative judgment” and normative answers “balanced”. The latter is referred to as probability judgments and normative answers are additive.

It is interesting to compare these two biases as they both seem to involve a consistent overestimation of people, items or outcomes in a positively valued, exhaustive set.

Also, participants' judgments seem to be made in a non-complementary fashion as if one alternative's chances or superiority does not affect the relative standing of the others in the group. This implies that both biases might depend on selective processing (Kardes, 2013). Further, other mechanisms proposed in order to explain the biases are also similar. Finally, these biases share a similar methodological research design, by both being mostly studied in between-group designs. The aim of Paper III was to empirically compare additivity neglect and the nonselective superiority bias in a within-participant design. Our hypothesis was that people who exhibit these biases fail to realize the normative requirement, which leads to a failure to control the sums of their probability estimates or the means of their comparative ratings.

Participants in **Experiment 1** were asked to give both probability estimates and comparative judgments, two tasks for each type, for sets of five Oscar nominees in various categories. In all tasks, the five nominated actors/actresses in the category were listed on the same page. We also attempted to induce comparative processing by asking participants in one condition to rank-order the alternatives before answering the probability tasks. The results showed that the two biases were related, and that participants who gave additive responses also gave more balanced ratings. There was, however, no effect of the rank-ordering, rendering further evidence to the difficulty of inducing comparative processing.

In **Experiment 2** participants gave comparative judgments and probability estimates to sets of four shortlisted job applicants. The comparative judgments were ratings regarding the applicants' qualifications, and the probability judgments were each applicant's chances of being hired. In order to allow a direct comparison between the two types of judgments, participants were given a prepared, "balanced" set of unbiased qualification ratings, and asked to "translate" these ratings into probabilities. They were also given a prepared, additive set of probability estimates and asked to translate these into relative qualification ratings. The results again showed that additive respondents in general gave more balanced ratings, regardless of condition. However, there was a negative correlation between the average ratings given to the two types of tasks. In the presented tables, probabilities had to be low (around 25%) to satisfy the additivity requirement. Many participants felt that this indicated rather poor (below average) qualifications. In other words, despite an ignorance prior of 25% for 4 job applicants, a 30% chance was translated into an applicant with less than average qualifications. At the same time, these unbiased ratings also gave participants a hint of the normative requirements of such judgments. The order of the two types of tasks were counterbalanced, and the results showed that participants who were presented with additive probability sets first (to be "translated" into comparative judgments), produced more additive probability sets when translating qualification ratings in the second part of the experiment, compared to participants who received the tasks in reverse order. This might indicate that some participants were "reminded" of the normative rule when exposed to additive sets of probability.

Previous studies of additivity neglect have shown a difference in additivity based on the answering format, where participants are more additive when given a Self-generated format. In Experiments 1 and 2, all probability tasks used the Self-generated format, while the qualification ratings used a scale format. In **Experiment 3** participants were again given both types of tasks, but with a between-group manipulation of the answering format where half of the participants were given the Scale format and the other half was given the Self-generated format. In addition, participants were asked if they had thought about the normative constraints, and to give a reason why they had / had not decided to answer in accordance with these requirements. The results again showed that participants who produced additive probability estimates gave more balanced NSSB ratings, than those who gave non-additive estimates. Despite the answering format not affecting participants' NSSB, the correlation between participants mean ratings and mean probabilities in the Scale format reached a correlation of $r(84) = .60$, $p < .001$, compared to more moderate correlations in the Self-generated condition and in Experiment 1. Further, the results revealed that overall only 31.5% of the participants had thought about the 100% rule or about balancing their ratings for at least some of the tasks. The verbal data indicated that beside reasons such as "it wasn't in the instructions", or "I didn't think about it", there were two main reasons for participants' overestimations: 1) they reported that they had considered the alternatives individually; 2) they thought it was unfair or incorrect, to give good job applicants a low chance or a low rating.

The three experiments reported in Paper III indicate several commonalities between additivity neglect and the nonselective superiority bias. Both biases seem to be about equally widespread, occurring in a majority of participants even when presented with the full set of alternatives. The two biases were also related, as participants who gave additive probability estimates gave more balanced distributions of ratings to the NSSB task. However, NSSB seems to be more robust in the sense that the degree of bias could not be reduced by changing the answering format to a Self-generated format. Previous studies of both biases have focused on the differences between selective and comparative processing and some studies suggest that inducing comparative processing by asking participants to evaluate all the alternatives may alleviate bias in referent-dependent judgments. This was not the case in the present studies. Explanations of biases in referent-dependent judgments must give an account of two prominent and separable findings: (1) why the formal rules are neglected, and (2) why the distributions of answers are positively skewed.

General Discussion

The present dissertation investigated subadditive judgments that violate the principle of additivity stating that the probability of a mutually exclusive and exhaustive set of outcomes cannot exceed 1 or 100%. In Paper I we found evidence of two relevant factors that affected participants' additivity neglect. Firstly, we found that participants with high numeracy produced more additive responses in the single outcome events, indicating that mathematical knowledge can reduce participants' bias. Also, "priming" participants by giving them the numeracy test before the probability tasks led to more additive responses, indicating that it might not be apparent to all participants that mathematical norms apply to the situation. Secondly, Paper I revealed that participants additivity neglect is affected by the answering formats they are given: writing the probabilities in empty slots next to the alternatives (Self-generated format) lead to more additive responses than circling numbers on a scale (Scale format). Paper II found that the Self-generated format seems to encourage a more class-based, or comparative, processing with more comparisons between the alternatives in the set. Paper III compared additivity neglect to another bias that can arise in referent-dependent judgments, namely the nonselective superiority bias (NSSB). The results demonstrated that the same participants often display both biases. In addition, several people reported their failure to consider normative constraints as due to considering alternatives independently, or because the constraints would lead to "unfair" or incorrect judgments. In the General discussion we will take a closer look at the mechanisms that may be responsible for these effects, and report results from three follow-up experiments and a reanalysis of the data from Paper II, that were designed to examine the source of format effects.

Numeracy and cognitive ability

As participants' answers to our probabilistic tasks are evaluated based on whether they comply with the mathematical norm of additivity, we wanted to investigate if correct answers to these tasks is related to numeracy in Paper I. We predicted that participants with a high numeracy score would give lower probability estimates and be less prone to additivity neglect compared to participants with a low numeracy score. We also wanted to investigate the boundaries of peoples' understanding of probabilities by including tasks with disjunctive outcome events.

The results from Paper I demonstrated that when the numeracy test is introduced prior to the probability tasks, participants with a high numeracy particularly benefitted from the "reminder" and gave more additive responses. Performance on the disjunctive tasks, on the other hand, appeared to be unrelated to numeracy. Even though most participants gave higher probability sums for the disjunctive tasks, the increase was not sufficient according to normative expectations. However, the relationship between numeracy and the estimates for the disjunctive outcome events was only investigated using the Lipkus (2001) test. As already noted, the Lipkus test is subject to a ceiling effect in educated samples. As we did not investigate the disjunctive outcome events with the Berlin Numeracy Test (BNT), which is better at discerning the numeracy level in student populations, we cannot exclude the possibility of numeracy being a factor in understanding disjunctive outcome events. It is also likely that understanding the set requirements for disjunctive outcome events require higher level of numeracy than that measured by the Lipkus test. Future studies could investigate the role of numeracy in disjunctive outcome events using the BNT.

Given that both "priming" participants with a numeracy test and the Self-generated answering format leads to more additive responses, there are clearly certain contexts that aids normative responses. We also have some evidence that these contexts particularly aid those with higher numeracy. There has been much debate about the individual differences involved in predicting normative responses to heuristics and biases tasks. Investigations have included general ability or intelligence, thinking dispositions (e.g., measures such as the cognitive reflection test) and numeracy (Klaczynski, 2014; Sinayev & Peters, 2015; Stanovich & West, 2008; Toplak, West, & Stanovich, 2013). Some show that normative responses are relatively unrelated to cognitive abilities (Stanovich & West, 2008), others claim that a high general ability is needed, but not sufficient without a favorable thinking disposition and high numeracy (Klaczynski, 2014). Some studies claim that thinking dispositions have a high predictive validity (Frederick, 2005; Toplak et al., 2013), whereas other studies indicate that such measures explain less variance than "pure" numeric ability (Sinayev & Peters, 2015).

However, it has recently been suggested that the predictive power of numeracy does not necessarily come from individual differences in mathematical aptitude, but individual differences in metacognition (Ghazal et al., 2014). Metacognition is often defined as “thinking about thinking”, and includes both the knowledge and beliefs about cognition, and the regulation and control of cognition (Brown, 1987; Garofalo & Lester, 1985). Metacognitive strategies have been described as sequential processes used to control cognitive activities, as a means to ensure reaching a cognitive goal. Because normative responses to referent-dependent judgments require complementarity and staying within limits, adhering to these constraints requires either continuous updates and adjustments, or going over one’s responses after all judgments have been performed. Regardless, such normative constraints require participants to be willing to take a second look and revise their estimates if needed, which can be seen as a metacognitive strategy. Some studies have also suggested that metacognitive experiences of difficulty or disfluency can prompt participants to take a more analytic approach to the task at hand, and at times even correct the more intuitive responses (Alter, Oppenheimer, Epley, & Eyre, 2007). Further support for the need to take a second look is illustrated in Paper II where participants in the Self-generated condition spent more time answering the additivity tasks compared to participants in the Scale condition. It is difficult to determine whether it is participants’ numerical skills or their metacognitive abilities that underlie our findings in Paper I. However, different superior cognitive skills are often connected as individuals with a high numeracy often also have superior metacognitive abilities, thus making it difficult to parse out the effect of each skill. Future studies of additivity neglect, and other biases in referent-dependent judgments, should investigate the role of metacognition as a determinant for normative responses.

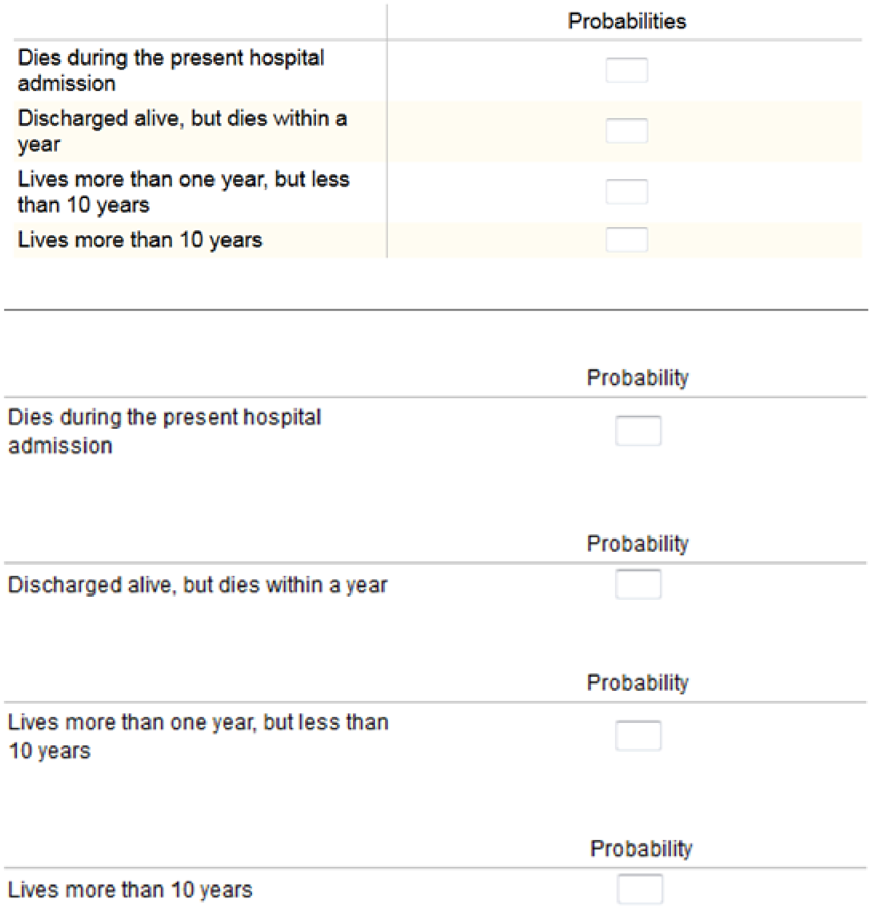
Effects of answering format

Several studies have shown that contexts, such as the way questions are posed, or the way rating scales are constructed, can strongly influence the answers obtained (e.g., Schwarz, 1999). All three papers report that participants display less additivity neglect when they are given the Self-generated answering format. Paper II also shows that this difference is reflected in different visual patterns for participants in the two formats, where the Self-generated format has almost twice the number of revisits between the alternatives compared to the Scale format. However, this does not explain why the difference occurs, and there are several competing explanations. The alternative explanations are explored in three new experiments from a manuscript in preparation, and a reanalysis of the data from Paper II.

Lastly, I also report a follow-up study on answering format and the nonselective superiority bias.

Differences in cognitive processing. One explanation is that the process of generating the probabilities without the "aid" of the scale leads to a deeper processing (Paper I), that makes participants realize that the alternatives are part of a "fixed pie" and that you can't have your pie and eat it too. We also found that when the numeracy test was given prior to the probability tasks it acted as a reminder to the more numerate participants that a mathematical rule applied to the situation. Others have found that contexts facilitating normative responses given in a within-Ss design produces a carryover effect to tasks without the facilitating aid (e.g., Reeves & Lockhart, 1993). Taken together, one would expect that participants given the Self-generated format first will become more additive even if the answering format changes. In Experiment 1 in an unpublished manuscript (Riege, Sulutvedt, Bjørgfinsdottir, & Miljeteig, 2015), we gave 76 MTurk workers six probability tasks, similar to those reported in the present dissertation. Participants in Condition 1 were presented with the Self-generated format for the three first tasks, followed by three tasks using the Scale format. Participants in Condition 2 were given the same tasks in opposite order, i.e., they received the Scale format first, followed by the Self-generated format. The results showed that in the group that was given the Self-generated format first (Condition 1), 34% of the participants gave at least one additive response to the Scale format tasks, versus 19% in the group that got the Scale tasks first. Out of these, seven gave additive responses to all three Scale tasks in Condition 1 (Self-generated first), versus only one in Condition 2 (Scale first). These numbers suggest a small effect on additivity neglect by being presented with the Self-generated format first. However, a 2 x 2 mixed ANOVA using number of additive responses for each format (Self-generated vs. Scale) as the within-subjects factor and condition as a between-Ss factor (Self-generated first vs. Scale first), revealed no effect of condition ($F(1,74) = 1.94$, $p = .168$, $\eta^2 = .026$). When the same ANOVA was run using participants mean probability sums for each format (Self-generated vs. Scale) as the within-subjects factor there was a marginally significant main effect of condition ($F(1,74) = 3.48$, $p = .06$, $\eta^2 = .045$). Thus, there might be a slight benefit to receiving the Self-generated format prior to the Scale format for some participants, as it leads to slightly lower probability sums, but being presented with the Self-generated format first does not change the number of additive responses in the subsequent Scale format. We cannot completly rule out that the Self-generated format serves as a reminder, though it seems unlikely. One possibility is that as the additivity principle is not explicitly stated, participants could decide not to further uphold the 100% - rule when switched to the Scale condition, due to a lack of motivation or exhaustion. In Paper II we found that participants in both conditions had a significant proportion of long fixation durations ($M_{(Self\text{-}generated)}$ = 11% and $M_{(Scale)}$ = 13% of fixations that were longer than 500 ms), indicating that our judgment tasks are taxing. Alternatively, there could be something about the Scale format that makes it even more difficult to adhere to the 100% - rule.

Differences in visual display. In Paper II, we suggested that the differences between the two answering formats might arise from the differences in the visual display. For example, there is a difference in how far the alternatives are spaced from each other. An early eye-tracking study showed that participants often chose to compare pairs of alternatives based on spatial proximity (Russo & Rosen, 1975), suggesting that participants might strive to minimize attention costs when comparing the alternatives (Orquin & Mueller Loose, 2013). Paper II indicated that participants in the Self-generated condition had significantly more revisits between the alternatives. As mental addition requires some effort it is possible that participants use eye-movements as a means to ease the working memory load. This might be more difficult when the alternatives are spaced further apart. Additional evidence comes from results showing that the de-biasing effect of the Self-generated format is strongly diminished when participants give their responses to each alternative on separate pages (Paper I, Experiment 4). We thus decided to investigate the differences in spacing by manipulating the space between the alternatives in the Self-generated format. In Experiment 2 (Riege et al., 2015), we allocated 61 Amazon MTurk workers to one of two conditions. Condition 1 and 2 both used the Self-generated answering format, but differed in the space between alternatives (see Figure 1), where the alternatives were spaced further apart in Condition 2 than in Condition 1. The results however, revealed no differences (t(59) = -0.56, p = .58) between Condition 1 (M = 126.57) and Condition 2 (M = 132.40), indicating that this difference in visual display might not lie at the root of the difference (either).



	Probabilities
Dies during the present hospital admission	□
Discharged alive, but dies within a year	□
Lives more than one year, but less than 10 years	□
Lives more than 10 years	□

	Probability
Dies during the present hospital admission	□

	Probability
Discharged alive, but dies within a year	□

	Probability
Lives more than one year, but less than 10 years	□

	Probability
Lives more than 10 years	□

Figure 1. The two versions of the Self-generated format: the traditional version is displayed in the top panel, and the one with the alternatives spaced further apart is in the bottom panel.

Another visual difference between the two answering formats is that the Scale format could draw the attention towards the middle of the scale, possibly anchoring estimates on 50% as the middle, and conceivably “neutral” value, thus prompting participants to adjust their estimates up or down from 50%. On the other hand, participants in the Self-generated would “start” their estimates from zero, relying on the mental representation of numbers from low to high (Paper II). We decided to explore this idea in Experiment 3 (Riege et al., 2015) by giving 176 MTurk workers five probability tasks using either the traditional Scale format (Condition 1), a slider format starting at zero (Condition 2), or a slider starting at 50 (Condition 3).

Table 1

Probability means, mean number of additive responses (out of five), and standard deviations for the three conditions with different scale formats.

	Scale	Slider zero	Slider 50
Mean probabilities	175.10	169.79	196.01
(SD)	(41.45)	(48.19)	(47.01)
Number of additive responses	0.79	0.74	0.39
(SD)	(1.29)	(1.2)	(0.95)

As seen in Table 1, the results showed no difference between the participants' probability sums in Condition 1 and 2, but participants in Condition 3 performed significantly worse than participants in Conditions 1 and 2, $F(2,175) = 5.27$, $p = .006$, $\eta^2 = .057$ (Bonferroni corrected). A similar pattern is observed for the number of additive responses (estimates between 90 – 110% were deemed additive), however, there are no significant differences between the conditions. It is difficult to establish where participants in the Scale format "start out", but when we make them start out at 50%, as an anchor, the probability estimates become even higher.

Lastly, we have speculated whether the scales themselves are contributing to the observed effect. In order to investigate this we re-analyzed the eye-tracking data from Paper II using more fine-grained Areas of Interest (AOI). The differences between the AOI's can be observed in Figure 2, which displays the original AOI's in the top panel, and the new AOI's in the bottom panel. The results are shown in Table 2.

Table 2

Mean number of fixations and revisits for the alternatives in two conditions for the three sets of AOI's (SD in parentheses), based on data from Riege, Sulutvedt & Teigen (2014)

		Self-generated	Scale	*t*(25)	*p**	Cohen's *d*
Number of fixations						
Original AOI's		128.6 (46.3)	100.1 (33.4)	1.8	.081	0.73
Fine-grained AOI's	Text	50.8 (26.3)	25.0 (11.8)	3.2	.003	1.28
	Scale	9.9 (8.2)	58.0 (21.2)	-7.9	.000	-3.16
Revisits to alternatives						
Original AOI's		43.6 (20.9)	22.0 (12.0)	3.3	.003	1.32
Fine-grained AOI's	Text	28.3 (15.2)	11.6 (6.3)	3.7	.001	1.48
	Scale	5.8 (6.0)	21.6 (11.3)	-4.6	.000	-1.84

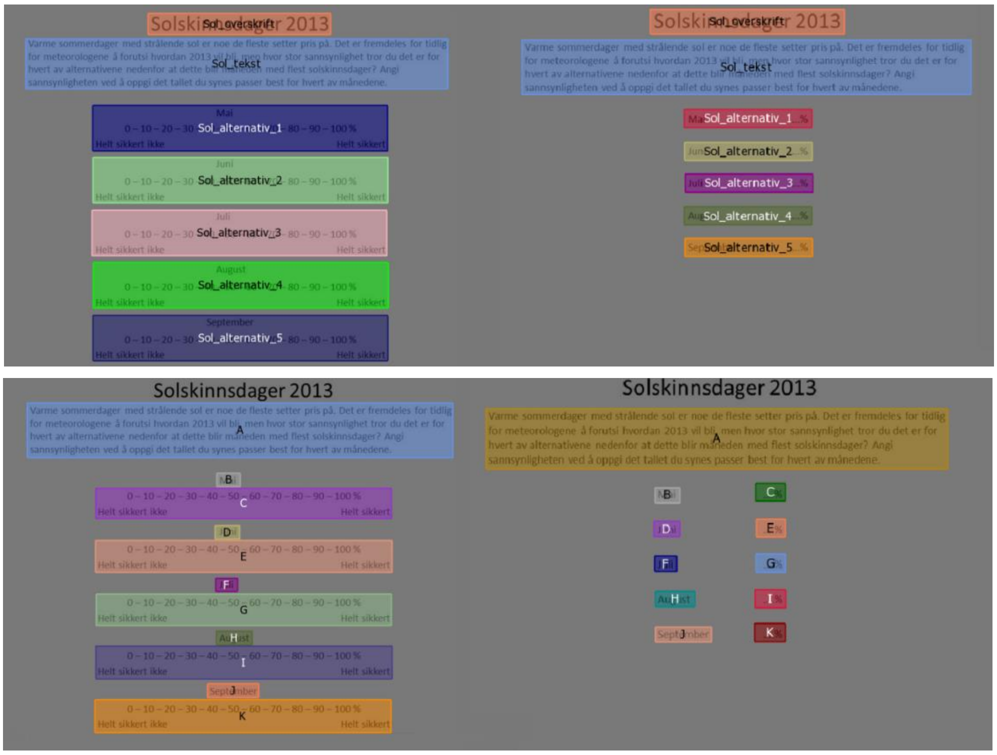

Solskinnsdager 2013
Sol_overskrift
Sol_tekst
Varme sommerdager med strålende sol er noe de fleste setter pris på. Det er fremdeles for tidlig for meteorologene å forutsi hvordan 2013 vil bli, men hvor stor sannsynlighet tror du det er for hvert av alternativene nedenfor at dette blir måneden med flest solskinnsdager? Angi sannsynligheten ved å oppgi det tallet du synes passer best for hvert av månedene.
Mai
Juni
Juli
August
September
0 – 10 – 20 – 30 – 40 – 50 – 60 – 70 – 80 – 90 – 100 %
Helt sikkert ikke
Helt sikkert
Sol_alternativ_1
Sol_alternativ_2
Sol_alternativ_3
Sol_alternativ_4
Sol_alternativ_5
A
B
C
D
E
F
G
H
I
J
K

Figure 2. The original AOI's are displayed in the top panel. The new AOI's are displayed in the bottom panel. The Scale formats are on the left side, and Self-generated format on the right.

The results show that participants spend a significant amount of their fixations on the scales themselves. Although the overall patterns of results are similar to the original results, the fine-grained AOI's reveal that participants in the Scale condition have more fixations and repeated inspections of the scales (the numbers) compared to the alternatives (text). In comparison, participants in the self-generated format are only able to focus on the alternatives, as they have no scales to look at. This is perhaps not so surprising, however, it indicates that scales in the Scale-format might become a "choice within a choice" (Orquin & Mueller Loose, 2013), where the participant focuses on the choice options *within* the scale (the numbers), rather than the alternatives (the text) themselves. This may in turn lead to less focus on the alternatives in the whole set; encouraging a case-based approach and thus as a result inflated probability estimates.

Answering format and the nonselective superiority bias. In Paper III we investigated whether the "debiasing" effect of the Self-generated answering format could reduce participants' nonselective superiority bias. However, this did not affect NSSB-ratings. Still, the verbal data collected in Paper III suggested that some participants found it difficult or unfair to give good job applicants negative ratings. Studies of the nonselective superiority bias often use scales from -4 to 4, where -4 is labelled "*Much worse than the others*"; 0 is labelled "*As good as the others*"; and 4 is labelled: "*Much better than the others*". As such, these labels clearly communicate what the respective values mean. Even so, the items in the local groups are all superior items compared to a general group, perhaps making it difficult for participants to choose negative values for an item. For example, it might be difficult to rate a popular vacation destination such as Las Vegas in a negative manner, regardless of a judge finding Las Vegas less appealing than Hawaii, rendering the illogical evaluation of both vacation destinations as being better than the other. In an unpublished follow-up study, we therefore gave 44 participants in the department's research pool two tasks regarding the 2015 Oscar nominations. Half the participants were given the traditional NSSB scale ranging from -4 (much less chance than the others), to 4 (much higher chance than the others), whereas the other half were given scales ranging from 1 to 9 (same anchors). The results showed that participants who were given the latter gave on average much more balanced ratings (*M* = 5.35, corresponding to *M* = 0.35 on the original scale) than those given the original scales (*M* = 1.11) (*t*(42) = 2.56, *p* = .014), indicating that participants might find it easier to place some of the alternatives on the left side (below the midpoint) of the scale when the numbers are positive.

Communicating beliefs

Support theory states that the strength of people's perceived evidence and the weighting of this evidence mediate the description of the event and the judgment. In other words, support theory claims that the judged probabilities are an expression of people's beliefs which are not subject to normative constraints (Tversky & Koehler, 1994). Despite the problems with, for example, the non-independence of judgments of support and judgments of probabilities (e.g., Pleskac, 2012), there are several other theories suggesting that peoples' probability judgments are expressions of beliefs (Robinson & Hastie, 1985). Interestingly, some of our participants expressed that they thought adhering to the normative constraints would be "unfair" or "wrong" in the verbal data (Paper III, Exp. 3). As mentioned in Paper I, some studies have shown that additivity is also consistently violated when participants are allowed to express their judgments in words rather than numbers (Teigen, 1988; Windschitl & Wells, 1996). Several studies of verbal probabilities indicate that translations of positive verbal probability estimates such as "good chance", "likely", or "probable", are often translated into numbers above .6 or 60% (Brun & Teigen, 1988; Reagan, Mosteller, & Youtz, 1989; Theil, 2002). In addition, using such positively valenced words is a way of communicating participants' beliefs of the event happening (Teigen & Brun, 1999). If we consider the case of Bob, who is one of four applicants applying for a job, telling him that he has a "good chance" of getting the job will be seen as indicative of the speaker being optimistic about Bob's chances. However, telling Bob that he has a 40% chance of getting the job might not be interpreted as an optimistic or positive remark, despite a 40% chance rendering, on average, only a 20% chance for the remaining 3 applicants. Numerical probabilities thus lack the pragmatic information inherent in most expressions of verbal probabilities, thereby making them more ambiguous and difficult to interpret (Teigen & Brun, 1999, 2000). Also, many participants might not find it natural to express themselves using numerical probabilities (Windschitl, 2002). Fuzzy-trace theory claims that people often simplify by exchanging exact (verbatim) numbers for simpler gist-based distinctions, such as "some" or "many" (Reyna & Brainerd, 1991). Taken together, it is possible to imagine that the numerical estimates found in subadditivity literature, at least for some participants, are "mental translations" of their verbal representations of their beliefs, rendering high numerical estimates as a means to communicate judges' beliefs about the event in question. As such, situations enforcing additivity may not get a true picture of people's perceived chances of various events.

Cognitive mechanisms in probability judgments

The present work distinguishes between a case-based approach and a class-based approach towards referent-dependent judgments, where the former is seen as a main contributor to biased responses. Participants' verbal explanations of violating additivity showed that considering the alternatives independently was indeed common, despite being presented with a full set of outcomes and instructed to make comparative judgments.

The present work clearly indicates that asking participants to judge the full set of outcomes does not alleviate additivity neglect or the nonselective superiority bias. However, there are some studies showing that biases arising from selective processing will be reduced when comparative processing is encouraged (e.g., Posavac, Brakus, Cronley, & Jain, 2009). For example, asking participants to allocate 100 dollars between all four national parks specified in the experiment, resulted in less "overspending" compared to participants who were asked to allocate money to one randomly drawn focal park (Posavac et al., 2009). This indicates that being reminded of the whole set can reduce inflated donations. However, all studies in the present dissertation asked participants to make judgments for a full set of outcomes, as have others, without resulting in additive responses. In addition, attempts to induce comparative processing by making participant's rank-order the alternatives from most favorable to least favorable did not affect additivity neglect nor NSSB (Bruchmann et al., 2013; Paper III, Experiment 1), adding further evidence to the robustness of both these biases.

It is possible that allocating money is different from allocating probabilities or evaluations of superiority, as it might be easier to understand that money is a zero-sum entity and thus cannot be spent twice. However, another distinction between our work and that of Sanbonmatsu, Posavac and colleagues is the possibility for external information search. Although we provide participants with some information in our tasks, there is little tangible information or evidence to direct participants in their judgments. Studies on selective processing on the other hand often provide some information about the targets. For example, Sanbonmatsu et al. (1997, Experiment 2) using similar tasks as those in Paper III (Experiments 2 and 3), gave participants eight statements for each of the four job applicants, and asked to estimate one randomly chosen job applicant chances of being hired. After the judgment was made the information about the job applicants was removed. Participants were then given five minutes to recall as many statements as possible for all the job applicants, as a measure of the information gathered and used to form the judgment. The results showed that participants remembered significantly more statements regarding their target applicant than statements about the referent applicants. This indicates that participants were focusing more on their target candidate than the referent candidates. This might also to an extent explain the findings of Posavac et al. (2009), as participants asked to allocate money to only one park may have neglected to read much information about the referent parks, thus overestimating the deservingness of the target park.

Nonetheless, an important question, which is beyond the scope of this dissertation, is whether there is a qualitative difference between estimating one alternative versus the full set. Different cognitive mechanisms might be involved when asked to estimate one alternative compared to estimating the full set. On the one hand one can argue that in everyday life, people do not evaluate full sets of exhaustive and mutually exclusive outcomes or items. More often the judgments encountered in life are related to whether a particular outcome will or will not happen, not which outcome will happen (Houghton & Kardes, 1998; Sanbonmatsu et al., 1998). In other words, many judgments are considered as yes/no questions, instead of a set of possible outcomes, thus subject to case-based reasoning based on case-specific information. Studying how people make such judgments is therefore important. Studies investigating participants' hypothesis generation generally indicate that participants stop their generation of hypotheses prematurely (e.g., Gettys & Fisher, 1979), and are overconfident in the hypotheses they do generate (e.g., Mehle, Gettys, Manning, Baca, & Fisher, 1981). One can thus argue that there is an advantage to giving participants the full set of outcomes, as this procedure eliminates any possible confusion regarding the partitioning of the probabilities or items included in the judgment (e.g., Gigerenzer, Hertwig, van den Broek, Fasolo, & Katsikopoulos, 2005). Also, asking participants to judge the full set of outcomes or items is a more stringent test of disregard for erroneous logic or mathematics, as opposed to asking participants to judge one or a subset of alternatives. Displaying the different outcomes in such a manner, gives participants a fair chance of making additive estimations. This allows for "full disclosure", thus avoiding the "hidden trap" of unspecified outcomes, strengthening the robustness of the bias.

Implications and potential future directions of the present work

Assessing the probabilities of future events is both an important and common task. Every time we decide to add to or spend our savings, play the lottery, cast our vote at an election, or choose which restaurant to eat, we are making a forecast about how the future will proceed, often as a means to affect our odds of a favorable outcome. Accordingly, considerable research efforts have gone into understanding how people make such judgments and how these judgments can be improved. The present research has several implications for future research in this field.

Firstly, the present work demonstrates that contrary to implicit and explicit conventions within the field, people still give non-additive responses in situations where the complete set of outcomes is presented on the same page. Thus, the assumption that people will correct themselves in contexts containing such "extensional cues" is not necessarily the case. Studies investigating subjective probability judgments or asking people for predictions of uncertain events should thus take care when deciding how to ask their questions. People will have trouble staying within limits and depending on the information the researcher wants, one can either explicitly remind participants that they have to stay within limits (e.g., De Bruin, Manski, Topa, & van der Klaauw, 2011), or accept estimates that add to more than 100%. The downside of the former is that when asking participants to stay within limits, their answers might not reflect their true beliefs. Further, some answers may simply adhere to the requirement of coherence by simply giving ignorance prior responses (e.g., 25% - 25% - 25% - 25%), thus failing to distinguish between the alternative outcomes. This can be circumvented using the latter strategy, and perhaps scaling the judgments down for analysis if coherence is an important concern. However, one cannot be sure that people's estimates are relative to all the outcomes in the set, but rather a set of absolute judgments for some participants, and for others a set of comparative judgments.

The effect of answering format also has implications for how researchers design their studies. Rating scales and self-generated studies are often used interchangeably, but might affect people's answers, particularly in the case of probability judgments. The finding that Self-generated answers led to lower probability estimates and more coherent responses indicates that this format might be preferable to rating scales when coherence is important. The present work cannot determine if Self-generated estimates are consistently more coherent or consistently lower than responses on a rating scale in other contexts, but adds to the growing literature of how contexts affect people's probability judgments and responses to questionnaires in general (Schwarz, 1999). Further, the finding that the self-generated format facilitates more comparative judgments (Paper II) shows that in situations where comparisons between alternatives are important, rating scales might interrupt comparisons, thus resulting in absolute or independent judgments for each alternative, despite explicit instructions to make comparative judgments (Paper III).

The finding of the role of numeracy in Paper I have implications for research on numeracy. The role of numeracy in heuristics and biases tasks has been mixed. The present results indicate that numeracy is indeed related to additivity neglect, but perhaps more importantly, reminding participants that mathematical rules might apply led to more coherent responses, particularly for those with higher numeracy. Perhaps some of the confusing findings in the numeracy literature are based on participants not realizing that they could make use of their math skills. Consequently, using measures to remind people might make the role of numeracy clearer. On the other hand, some people have a difficult relationship with math, and may not only receive no benefit from the reminder, but possibly perform worse as a consequence.

An interesting implication of numeracy research is the possibility of serving as an argument for all teenagers who feel that learning math is irrelevant, often arguing that "they will never need this (math) in real life". The present research shows that having the ability to reason with numbers and mathematical concepts, and the ability to realize when these skills can be used, seems to aid coherence in probabilistic predictions of future events.

Lastly, the findings in Paper III have several implications for future research on referent-dependent judgments. Paper III indicated that there might be similar underlying mechanisms for errors in referent-dependent judgments. Future research should expand on this by for example including self-other judgments (better than average effect), preferably with the aim of developing an overarching theoretical framework for errors in referent-dependent judgment. Further, the failure to understand both "rules" and complementarity of referent-dependent judgments should be investigated. For example, several studies have found subadditivity in judgments of frequency (Bearden & Wallsten, 2004; Mulford & Dawes, 1999; Tversky & Koehler, 1994), using a between- subjects design. Similarly, a study investigating allocations of market shares, also found that participants' estimates of percentage market shares for a set of exhaustive companies added to, on average, 192% (Houghton & Kardes, 1998). Both frequencies and market shares seem intuitively easier to understand than probabilities, though people seem to struggle with staying within the limits in such cases as well. Investigating the boundary conditions of people's understanding of complementarity should be investigated using within-subjects judgments, asking people to allocate various entities, such as money, expected frequency of occurrence for various events, as well as probability judgments and superiority ratings. As people make such judgments on a daily basis it is essential to collate the information already available, as well as establishing an overall framework and boundary conditions for referent-dependent judgments.

Conclusions

In three papers we have explored determinants and cognitive underpinnings of additivity neglect. We found that individual differences in numeracy predict additive responses, where numeracy is inversely related to additivity neglect in single outcome events. In addition, inducing a mathematical mindset, reminding participants that mathematical rules may apply, leads to more additive responses. However, the mathematical mindset does not aid responses to disjunctive outcome events, nor does numeracy predict normative responses to such tasks. We also found that different answering formats affect additivity neglect, where writing the probabilities in empty slots next to the alternatives (Self-generated format) lead to more additive responses than circling numbers on a scale (Scale format). The Self-generated format seems to encourage a more comparative (or class-based) processing thus leading to more additive responses, where the

Scale format hiders normative responses by making comparisons between the alternatives harder to perform, as indicated by participants eye-movements whilst making their judgments. The present work also compared additivity neglect to another bias in referent-dependent judgments, namely the nonselective superiority bias (NSSB). The results show that the same participants often display both biases. In addition, verbal reports demonstrated that people fail to consider normative constraints due to considering alternatives one-by-one (case-based), or because the constraints would lead to judgments not representing their true beliefs.

References

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General, 136*(4), 569-576. doi: 10.1037/0096-3445.136.4.569

Anderson, N. H. (1991). *Contributions to information integration theory* (Vol. I). Hillsdale, N.J.: Erlbaum.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44*(3), 211-233. doi: 10.1016/0001-6918(80)90046-3

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences, 30*(3), 241-+. doi: 10.1017/s0140525x07001653

Bearden, J. N., & Wallsten, T. S. (2004). MINERVA-DM and subadditive frequency judgments. *Journal of Behavioral Decision Making, 17*(5), 349-363. doi: 10.1002/bdm.477

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis, 20*(3), 351-368. doi: 10.1093/pan/mpr057

Bless, H., Betsch, T., & Franzen, A. (1998). Framing the framing effect: The impact of context cues on solutions to the 'Asian disease' problem. *European Journal of Social Psychology, 28*(2), 287-291.

Braga, J. N., Ferreira, M. B., & Sherman, S. J. (2015). The effects of construal level on heuristic reasoning: The case of representativeness and availability. *Decision, 2*(3), 216-227. doi: 10.1037/dec0000021

Brown, A. (1987). Metacognition, executive control, self-regulation, and other more mysterious mechanisms. In F. E. Weinert & R. H. Kluwe (Eds.), *Metacognition, motivation and understanding*. Hillsdale, N.J: Lawrence Erlbaum.

Bruchmann, K., Suls, J., Lee, S., Rose, J. P., Krizan, Z., & Windschitl, P. D. (2013). Searching for the limits and explanations of the nonselective superiority bias. *Social Psychological and Personality Science, 4*(1), 124-130. doi: 10.1177/1948550612443387

Brun, W., & Teigen, K. H. (1988). Verbal probabilities - Ambiguous, context-dependent, or both. *Organizational Behavior and Human Decision Processes, 41*(3), 390-404. doi: 10.1016/0749-5978(88)90036-2

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*(1), 3-5. doi: 10.1177/1745691610393980

Carlson, B. W., & Yates, J. F. (1989). Disjunction errors in qualitative likelihood judgment *Organizational Behavior and Human Decision Processes, 44*(3), 368-379. doi: 10.1016/0749-5978(89)90014-9

Chambers, J. R. (2010). Why the parts are better (or worse) than the whole: The unique-attributes hypothesis. *Psychological Science, 21*(2), 268-275. doi: 10.1177/0956797609359509

Chapman, G. B., & Liu, J. J. (2009). Numeracy, frequency, and Bayesian reasoning. *Judgment and Decision Making, 4*(1), 34-40.

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin numeracy test. *Judgment and Decision Making, 7*(1), 25-47.

De Bruin, W. B., Manski, C. F., Topa, G., & van der Klaauw, W. (2011). Measuring consumer uncertainty about future inflation. *Journal of Applied Econometrics, 26*(3), 454-478.

Dougherty, M. R. P., & Hunter, J. (2003). Probability judgment and subadditivity: The role of working memory capacity and constraining retrieval. *Memory & Cognition, 31*(6), 968-982. doi: 10.3758/bf03196449

Findlay, J. M., & Kapoula, Z. (1992). Scrutinization, spatial attention, and the spatial programming of saccadic eye movements. *The Quarterly Journal of Experimental Psychology Section A, 45*(4), 633-647. doi: 10.1080/14640749208401336

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1978). Fault trees - Sensitivity of estimated failure probabilities to problem representation. *Journal of Experimental Psychology-Human Perception and Performance, 4*(2), 330-344.

Fox, C. R. (1999). Strength of evidence, judged probability, and choice under uncertainty. *Cognitive Psychology, 38,* 167–189.

Fox, C. R., & Birke, R. (2002). Forecasting trial outcomes: Lawyers assign higher probability to possibilities that are described in greater detail. *Law and Human Behavior, 26*(2), 159-173. doi: 10.1023/a:1014687809032

Fox, C. R., & Clemen, R. T. (2005). Subjective probability assessment in decision analysis: Partition dependence and bias toward the ignorance prior. *Management Science, 51*(9), 1417-1432. doi: 10.1287/mnsc.1050.0409

Fox, C. R., Rogers, B. A., & Tversky, A. (1996). Options traders exhibit subadditive decision weights. *Journal of Risk and Uncertainty, 13*(1), 5-17.

Fox, C. R., & Rottenstreich, Y. (2003). Partition priming in judgment under uncertainty. *Psychological Science, 14*(3), 195-200. doi: 10.1111/1467-9280.02431

Fox, C. R., & Tversky, A. (1998). A belief-based account of decision under uncertainty. *Management Science, 44*(7), 879-895.

Fox, C. R., & Ülkümen, G. (2011). Distinguishing two dimensions of uncertainty. In W. Brun, G. Keren, G. Kirkebøen & H. Montgomery (Eds.), *Perspectives on thinking, judging, and decision making: a tribute to Karl Halvor Teigen* (pp. 21-35). Oslo: Universitetsforl.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25-42. doi: 10.1257/089533005775196732

Galesic, M., Gigerenzer, G., & Straubinger, N. (2009). Natural frequencies help older adults and people with low numeracy to evaluate medical screening tests. *Medical Decision Making, 29*(3), 368-371. doi: 10.1177/0272989x08329463

Garofalo, J., & Lester, F. K., Jr. (1985). Metacognition, cognitive monitoring, and mathematical performance. *Journal for Research in Mathematics Education, 16*(3), 163-176. doi: 10.2307/748391

Gettys, C. F., & Fisher, S. D. (1979). Hypothesis plausibility and hypothesis generation. *Organizational Behavior and Human Performance, 24*(1), 93-110. doi: http://dx.doi.org/10.1016/0030-5073(79)90018-7

Ghazal, S., Cokely, E. T., & Garcia-Retamero, R. (2014). Predicting biases in very highly educated samples: Numeracy and metacognition. *Judgment and Decision Making, 9*(1), 15-34.

Gigerenzer, G. (1998). Surrogates for theories. *Theory & Psychology, 8*(2), 195-204. doi: 10.1177/0959354398082006

Gigerenzer, G., Hertwig, R., van den Broek, E., Fasolo, B., & Katsikopoulos, K. V. (2005). "A 30% chance of rain tomorrow": How does the public understand probabilistic weather forecasts? *Risk Analysis, 25*(3), 623-629. doi: 10.1111/j.1539-6924.2005.00608.x

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction - Frequency formats. *Psychological Review, 102*(4), 684-704.

Giladi, E. E., & Klar, Y. (2002). When standards are wide of the mark: Nonselective superiority and inferiority biases in comparative judgments of objects and concepts. *Journal of Experimental Psychology-General, 131*(4), 538-551. doi: 10.1037//0096-3445.131.4.538

Glaholt, M. G., & Reingold, E. M. (2011). Eye movement monitoring as a process tracing methodology in decision making research. *Journal of Neuroscience, Psychology, and Economics, 4*(2), 125-146. doi: 10.1037/a0020692

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making, 26*(3), 213-224. doi: 10.1002/bdm.1753

Hardman, D. (2009). *Judgement and decision making: Psychological perspectives*. Chichester: BPS Blackwell.

Hastie, R., & Dawes, R. M. (2010). *Rational choice in an uncertain world: The psychology of judgement and decision making*. Los Angeles: Sage.

Hauser, D. J., & Schwarz, N. (2015a). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*. doi: 10.3758/s13428-015-0578-z

Hauser, D. J., & Schwarz, N. (2015b). It's a Trap! Instructional manipulation checks prompt systematic thinking on "tricky" tasks. *SAGE Open, 5*(2). doi: 10.1177/2158244015584617

Hill, K. A. (2012). Cognition in the woods: Biases in probability judgments by search and rescue planners. *Judgment and Decision Making, 7*(4), 488-498.

Hoffman, J. E., & Subramaniam, B. (1995). The role of visual-attention in saccadic eye-movements. *Perception & Psychophysics, 57*(6), 787-795. doi: 10.3758/bf03206794

Holmqvist, K. (2011). *Eye tracking: a comprehensive guide to methods and measures*. Oxford: Oxford University Press.

Horstmann, N., Ahlgrimm, A., & Gloeckner, A. (2009). How distinct are intuition and deliberation? An eye-tracking analysis of instruction-induced decision modes. *Judgment and Decision Making, 4*(5), 335-354.

Houghton, D., & Kardes, F. (1998). Market share overestimation and the noncomplementarity effect. *Marketing Letters, 9*(3), 313-320. doi: 10.1023/A:1008028407624

Hsee, C. K., & Rottenstreich, Y. (2004). Music, pandas, and muggers: On the affective psychology of value. *Journal of Experimental Psychology: General, 133*(1), 23-30. doi: 10.1037/0096-3445.133.1.23

Idson, L. C., Krantz, D. H., Osherson, D., & Bonini, N. (2001). The relation between probability and evidence judgment: An extension of support theory. *The Journal of Risk and Uncertainty, 22*, 227-249.

Juslin, P., Wennerholm, P., & Olsson, H. (1999). Format dependence in subjective probability calibration. *Journal of Experimental Psychology-Learning Memory and Cognition, 25*(4), 1038-1052. doi: 10.1037/0278-7393.25.4.1038

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*(4), 329-354. doi: 10.1037/0033-295X.87.4.329

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, G. Dale & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49-81). Cambridge: Cambridge University Press.

Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 267-294). Cambridge: Cambridge University Press.

Kahneman, D., Krueger, A. B., Schkade, D., Schwarz, N., & Stone, A. A. (2006). Would you be happier if you were richer? A focusing illusion. *Science, 312*(5782), 1908-1910. doi: 10.1126/science.1129688

Kahneman, D., & Tversky, A. (1972). Subjective probability - Judgment of representativeness. *Cognitive Psychology, 3*(3), 430-454. doi: 10.1016/0010-0285(72)90016-3

Kahneman, D., & Tversky, A. (1982). Variants of uncertainty. *Cognition, 11*(2), 143-157.

Kardes, F. R. (2013). Selective versus comparative processing. *Journal of Consumer Psychology, 23*(1), 150-153. doi: 10.1016/j.jcps.2012.10.003

Klaczynski, P. A. (2014). Heuristics and biases: interactions among numeracy, ability, and reflectiveness predict normative responding. *Frontiers in Psychology, 5*. doi: 10.3389/fpsyg.2014.00665

Klar, Y. (2002). Way beyond compare: Nonselective superiority and inferiority biases in judging randomly assigned group members relative to their peers. *Journal of Experimental Social Psychology, 38*(4), 331-351. doi: 10.1016/s0022-1031(02)00003-3

Klar, Y., Medding, A., & Sarel, D. (1996). Nonunique invulnerability: Singular versus distributional probabilities and unrealistic optimism in comparative risk judgments. *Organizational Behavior and Human Decision Processes, 67*(2), 229-245.

Klayman, J., Soll, J. B., Gonzalez-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes, 79*(3), 216-247. doi: 10.1006/obhd.1999.2847

Koehler, D. J. (2000). Probability judgment in three-category classification learning. *Journal of Experimental Psychology-Learning Memory and Cognition, 26*(1), 28-52.

Koehler, D. J., Brenner, L. A., & Tversky, A. (1997). The enhancement effect in probability judgment. *Journal of Behavioral Decision Making, 10*(4), 293-313.

Krizan, Z., & Suls, J. (2008). Losing sight of oneself in the above-average effect: When egocentrism, focalism, and group diffuseness collide. *Journal of Experimental Social Psychology, 44*(4), 929-942. doi: 10.1016/j.jesp.2008.01.006

Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M., & Pardo, S. T. (2012). Individual Differences in Numeracy and Cognitive Reflection, with Implications for Biases and Fallacies in Probability Judgment. *Journal of Behavioral Decision Making, 25*(4), 361-381. doi: 10.1002/bdm.752

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making, 21*(1), 37-44.

Lohse, G. L., & Johnson, E. J. (1996). *A comparison of two process tracing methods for choice tasks.* Paper presented at the System Sciences, 1996., Proceedings of the Twenty-Ninth Hawaii International Conference on.

Macchi, L., Osherson, D., & Krantz, D. H. (1999). A note on superadditive probability judgment. *Psychological Review, 106*(1), 210-214. doi: 10.1037//0033-295x.106.1.210

McKenzie, C. R. M. (1999). (Non) complementary updating of belief in two hypotheses. *Memory & Cognition, 27*(1), 152-165. doi: 10.3758/bf03201221

Mehle, T., Gettys, C. F., Manning, C., Baca, S., & Fisher, S. (1981). The availability explanation of excessive plausibility assessments. *Acta Psychologica, 49*(2), 127-140. doi: http://dx.doi.org/10.1016/0001-6918(81)90024-X

Mulford, M., & Dawes, R. M. (1999). Subadditivity in memory for personal events. *Psychological Science, 10*(1), 47-51. doi: 10.1111/1467-9280.00105

Nelson, W., Reyna, V. F., Fagerlin, A., Lipkus, I., & Peters, E. (2008). Clinical implications of numeracy: Theory and practice. *Annals of Behavioral Medicine, 35*(3), 261-274. doi: 10.1007/s12160-008-9037-8

Nordbye, G. H. H., & Teigen, K. H. (2014). Responsibility judgments of wins and losses in the 2013 Chess Championship. *Judgment and Decision Making, 9*(4), 335- 348.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45*(4), 867-872. doi: http://dx.doi.org/10.1016/j.jesp.2009.03.009

Orquin, J. L., Ashby, N. J. S., & Clarke, A. D. F. (2015). Areas of interest as a signal detection problem in behavioral eye-tracking research. *Journal of Behavioral Decision Making*. doi: 10.1002/bdm.1867

Orquin, J. L., & Mueller Loose, S. (2013). Attention and choice: A review on eye movements in decision making. *Acta Psychologica, 144*(1), 190-206. doi: http://dx.doi.org/10.1016/j.actpsy.2013.06.003

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5*, 411-419.

Peters, E. (2012). Beyond Comprehension: The Role of Numeracy in Judgments and Decisions. *Current Directions in Psychological Science, 21*(1), 31-35. doi: 10.1177/0963721411429960

Peters, E., & Levin, I. P. (2008). Dissecting the risky-choice framing effect: Numeracy as an individual-difference factor in weighting risky and riskless options. *Judgment and Decision Making Journal, 3*(6), 435-448.

Peters, E., Vastfjall, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological Science, 17*(5), 407-413.

Peterson, C. R., Ducharme, W. M., & Edwards, W. (1968). Sampling distributions and probability revisions. *Journal of Experimental Psychology, 76*(2P1), 236-&. doi: 10.1037/h0025427

Peterson, R. A. (2001). On the use of college students in social science research: Insights from a second-order meta-analysis. *Journal of Consumer Research, 28*(3), 450-461. doi: 10.1086/323732

Pleskac, T. J. (2012). Comparability effects in probability judgments. *Psychological Science, 23*(8), 848-854. doi: 10.1177/0956797612439423

Posavac, S. S., Brakus, J. J., Cronley, M. L., & Jain, S. P. (2009). On assuaging positive bias in environmental value elicitation. *Journal of Economic Psychology, 30*(3), 482-489. doi: 10.1016/j.joep.2008.07.007

Posner, M. I., Snyder, C. R. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology-General, 109*(2), 160-174. doi: 10.1037//0096-3445.109.2.160

Reagan, R. T., Mosteller, F., & Youtz, C. (1989). Quantitative meanings of verbal probability-expressions. *Journal of Applied Psychology, 74*(3), 433-442. doi: 10.1037//0021-9010.74.3.433

Redden, J. P., & Frederick, S. (2011). Unpacking unpacking: Greater detail can reduce perceived likelihood. *Journal of Experimental Psychology: General, 140*(2), 159-167. doi: 10.1037/a0021491

Redelmeier, D. A., Koehler, D. J., Liberman, V., & Tversky, A. (1995). Probability judgment in medicine - Discounting unspecified possibilities. *Medical Decision Making, 15*(3), 227-230.

Reeves, T., & Lockhart, R. S. (1993). Distributional versus singular approaches to probability and errors in probabilistic reasoning. *Journal of Experimental Psychology: General, 122*(2), 207-226. doi: 10.1037/0096-3445.122.2.207

Reyna, V. F., & Brainerd, C. J. (1991). Fuzzy-trace theory and framing effects in choice: Gist extraction, truncation, and conversion. *Journal of Behavioral Decision Making, 4*(4), 249-262. doi: 10.1002/bdm.3960040403

Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin, 135*(6), 943-973. doi: 10.1037/a0017327

Riege, A. H. (2011). *Additivity neglect in probability judgments and the role of numeracy*. (Master of Philosophy in Psychology), University of Oslo, DUO. Retrieved from http://urn.nb.no/URN:NBN:no-28969

Riege, A. H., Sulutvedt, U., Bjørgfinsdottir, R., & Miljeteig, K. (2015). *An eye for probabilities: The effect of answering format on Additivity neglect*. Manuscript in preparation.

Riege, A. H., & Teigen, K. H. (2013). Additivity neglect in probability estimates: Effects of numeracy and response format. *Organizational Behavior and Human Decision Processes, 121*(1), 41-52. doi: 10.1016/j.obhdp.2012.11.004

Robinson, L. B., & Hastie, R. (1985). Revision of beliefs when a hypothesis is eliminated from consideration. *Journal of Experimental Psychology-Human Perception and Performance, 11*(4), 443-456.

Rottenstreich, Y., & Tversky, A. (1997). Unpacking, repacking, and anchoring: Advances in support theory. *Psychological Review, 104*(2), 406-415. doi: 10.1037//0033-295x.104.2.406

Russo, J. E. (2011). Eye Fixations as a Process Trace. In M. Schulte-Mecklenbeck, A. Kuehberger & R. Ranyard (Eds.), *A Handbook of Process Tracing Methods for Decision Research*. New York: Psychologt Press.

Russo, J. E., & Rosen, L. D. (1975). Eye fixation analysis of multialternative choice. *Memory & Cognition, 3*(3), 267-276. doi: 10.3758/bf03212910

Sanbonmatsu, D. M., Posavac, S. S., Kardes, F. R., & Mantel, S. P. (1998). Selective hypothesis testing. *Psychonomic Bulletin & Review, 5*(2), 197-220. doi: 10.3758/bf03212944

Sanbonmatsu, D. M., Posavac, S. S., & Stasney, R. (1997). The subjective beliefs underlying probability overestimation. *Journal of Experimental Social Psychology, 33*(3), 276-295. doi: http://dx.doi.org/10.1006/jesp.1996.1321

Sanbonmatsu, D. M., Vanous, S., Hook, C., Posavac, S. S., & Kardes, F. R. (2011). Whither the alternatives: Determinants and consequences of selective versus comparative judgemental processing. *Thinking & Reasoning, 17*(4), 367-386. doi: 10.1080/13546783.2011.625659

Schkade, D. A., & Kahneman, D. (1998). Does living in California make people happy? A focusing illusion in judgments of life satisfaction. *Psychological Science, 9*(5), 340-346. doi: 10.1111/1467-9280.00066

Schulte-Mecklenbeck, M., Kuehberger, A., & Ranyard, R. (2010). *A handbook of process tracing methods for decision research*. Hoboken: Taylor & Francis.

Schulte-Mecklenbeck, M., Kuhberger, A., & Ranyard, R. (2011). The role of process data in the development and testing of process models of judgment and decision making. *Judgment and Decision Making, 6*(8), 733-739.

Schwartz, L. M., Woloshin, S., Black, W. C., & Welch, H. G. (1997). The role of numeracy in understanding the benefit of screening mammography. *Annals of Internal Medicine, 127*(11), 966-972.

Schwarz, N. (1999). Self-reports - How the questions shape the answers. *American Psychologist, 54*(2), 93-105. doi: 10.1037/0003-066x.54.2.93

Schwarz, N., Knäuper, B., Hippler, H.-J., Noelle-Neumann, E., & Clark, L. (1991). Rating scales numeric values may change the meaning of scale labels. *Public Opinion Quarterly, 55*(4), 570-582. doi: 10.1086/269282

Sears, D. O. (1986). College sophomores in the laboratory - Influences of a narrow database on social-psychology view of human-nature. *Journal of Personality and Social Psychology, 51*(3), 515-530. doi: 10.1037/0022-3514.51.3.515

Sinayev, A., & Peters, E. (2015). Cognitive reflection versus calculation in decision making. *Frontiers in Psychology, 6*. doi: 10.3389/fpsyg.2015.00532

Sirota, M., & Juanchich, M. (2011). Role of numeracy and cognitive reflection in Bayesian reasoning with natural frequencies *Studia Psychologica, 53*(2), 151-161.

Sloman, S., Rottenstreich, Y., Wisniewski, E., Hadjichristidis, C., & Fox, C. R. (2004). Typical versus atypical unpacking and superadditive probability judgment. *Journal of Experimental Psychology-Learning Memory and Cognition, 30*(3), 573-582. doi: 10.1037/0278-7393.30.3.573

Smith, A. R., Windschitl, P. D., & Rose, J. P. (2015). *Biases in probability and comparative judgments: Under one theoretical framework*. Manuscript submittet for publication.

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology, 94*(4), 672-695. doi: 10.1037/0022-3514.94.4.672

Suls, J., Chambers, J. R., Krizan, Z., Mortensen, C. R., Koestner, B., & Bruchmann, K. (2010). Testing four explanations for the better/worse-than-average effect: Single- and multi-item entities as comparison targets and referents. *Organizational Behavior and Human Decision Processes, 113*(1), 62-72. doi: 10.1016/j.obhdp.2010.03.003

Teigen, K. H. (1974a). Overestimation of subjective probabilities. *Scandinavian Journal of Psychology, 15*(1), 56-62.

Teigen, K. H. (1974b). Subjective sampling distributions and additivity of estimates. *Scandinavian Journal of Psychology, 15*(1), 50-55.

Teigen, K. H. (1983). Studies in subjective-probability III - The unimportance of alternatives. *Scandinavian Journal of Psychology, 24*(2), 97-105. doi: 10.1111/j.1467-9450.1983.tb00481.x

Teigen, K. H. (1988). When are low-probability events judged to be 'probable'? Effects of outcome-set characteristics on verbal probability estimates. *Acta Psychologica, 67*(2), 157-174. doi: http://dx.doi.org/10.1016/0001-6918(88)90011-X

Teigen, K. H., & Brun, W. (1999). The directionality of verbal probability expressions: Effects on decisions, predictions, and probabilistic reasoning. *Organizational Behavior and Human Decision Processes, 80*(2), 155-190. doi: 10.1006/obhd.1999.2857

Teigen, K. H., & Brun, W. (2000). Ambiguous probabilities: When does p=0.3 reflect a possibility, and when does it express a doubt? *Journal of Behavioral Decision Making, 13*(3), 345-362. doi: 10.1002/1099-0771(200007/09)13:3<345::aid-bdm358>3.0.co;2-u

Teigen, K. H., & Brun, W. (2011). Responsibility is divisible by two, but not by three or four: Judgments of responsibility in dyads and groups. *Social Cognition, 29*(1), 15-42.

Teigen, K. H., & Jorgensen, M. (2005). When 90% confidence intervals are 50% certain: On the credibility of credible intervals. *Applied Cognitive Psychology, 19*(4), 455-475. doi: 10.1002/acp.1085

Theil, M. (2002). The role of translations of verbal into numerical probability expressions in risk management: a meta-analysis. *Journal of Risk Research, 5*(2), 177-186. doi: 10.1080/13669870110038179

Toplak, M. E., West, R. F., & Stanovich, K. E. (2013). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning, 20*(2), 147-168. doi: 10.1080/13546783.2013.844729

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty - Heuristics and biases. *Science, 185*(4157), 1124-1131.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning - The conjunction fallacy in probability judgment. *Psychological Review, 90*(4), 293-315. doi: 10.1037/0033-295x.90.4.293

Tversky, A., & Koehler, D. J. (1994). Support Theory - A nonextensional representation of subjective-probability. *Psychological Review, 101*(4), 547-567. doi: 10.1037//0033-295x.101.4.547

Van Wallendael, L. R. (1989). The quest for limits on noncomplementarity in opinion revision. *Organizational Behavior and Human Decision Processes, 43*(3), 385-405. doi: http://dx.doi.org/10.1016/0749-5978(89)90044-7

Van Wallendael, L. R., & Hastie, R. (1990). Tracing the footsteps of Sherlock-Holmes - Cognitive representations of hypothesis-testing. *Memory & Cognition, 18*(3), 240-250.

Velichkovskiĭ, B. M. (1999). From levels of processing to stratification of cognition: Converging evidence from three domains of research. In B. H. Challis & B. M. Velichkovskiĭ (Eds.), *Stratification in cognition and consciousness* (pp. 203-226). Amsterdam: J. Benjamins.

Wedell, D. H. (2011). Probabilistic reasoning in prediction and diagnosis: Effects of problem type, response mode, and individual differences. *Journal of Behavioral Decision Making, 24*(2), 157-179. doi: 10.1002/bdm.686

Wilson, T. D., Wheatley, T., Meyers, J. M., Gilbert, D. T., & Axsom, D. (2000). Focalism: A source of durability bias in affective forecasting. *Journal of Personality and Social Psychology, 78*(5), 821-836. doi: 10.1037/0022-3514.78.5.821

Windschitl, P. D. (2002). Judging the accuracy of a likelihood judgment: The case of smoking risk. *Journal of Behavioral Decision Making, 15*(1), 19-35. doi: 10.1002/bdm.401

Windschitl, P. D., Conybeare, D., & Krizan, Z. (2008). Direct-comparison judgments: When and why above- and below-average effects reverse. *Journal of Experimental Psychology: General, 137*(1), 182-200. doi: 10.1037/0096-3445.137.1.182

Windschitl, P. D., Rose, J. P., Stalkfleet, M. T., & Smith, A. R. (2008). Are people excessive or judicious in their egocentrism? A modeling approach to understanding bias and accuracy in people's optimism. *Journal of Personality and Social Psychology, 95*(2), 253-273. doi: 10.1037/0022-3514.95.2.253

Windschitl, P. D., & Wells, G. L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied, 2*(4), 343-364. doi: 10.1037/1076-898x.2.4.343

Winman, A., Juslin, P., Lindskog, M., Nilsson, H., & Kerimi, N. (2014). The role of ANS acuity and numeracy for the calibration and the coherence of subjective probability judgments. *Frontiers in Psychology, 5,* 15. doi: 10.3389/fpsyg.2014.00851