To appear in *Journal of Behavioral Decision Making*

**Everybody will win, and all must be hired:
Comparing additivity neglect with the nonselective superiority bias**

Anine H. Riege¹ and Karl Halvor Teigen^{1,2}

¹Department of Psychology, University of Oslo

²Simula Research Laboratory, Oslo

Total word count: 10486

Corresponding author:

Anine H. Riege

University of Oslo

P.b. 1094 Blindern, NO-0317 Oslo, Norway

Telephone: +47 22 84 50 56

a.c.riege@psykologi.uio.no

Abstract

Two streams of research looking at referent-dependent judgments from slightly different angles are subadditivity research and research on the nonselective superiority bias. Both biases violate basic formal constraints: the probabilities of a set of exclusive events cannot add up to more than 100%, and a set of attractive candidates cannot all be rated as superior to the group mean. We examine in three experiments how these two biases are related, by asking the same participants to perform both kinds of tasks on the same material. Both biases appear to be widespread, even for sets where all alternatives are presented together, but they differ in the way they are affected by response format and experimental setup. Thus, presenting participants with an unbiased set of ratings will reduce, but not normalize their probability estimates of the same alternatives; while presenting them with an unbiased (additive) set of probabilities will make most alternatives appear *inferior* to the group mean, inverting the superiority bias. Self-reports reveal that additivity neglect and the nonselective superiority bias can be based on two main response-strategies: (1) considering each alternative independently, or (2) comparing alternatives, while neglecting their complementarity. In both cases, assessments will be the outcome of a compromise between the perceived “absolute” merits of each alternative, its standing relative to referents, and properties of the response scale.

Key words: Additivity neglect, nonselective superiority bias, comparative judgments, selective processing, subadditivity, referent-dependent judgments.

Everybody will win, and all must be hired:

Comparing additivity neglect with the nonselective superiority bias

The Dodo-bird verdict from Alice in Wonderland (“Everybody has won, and all must have prizes”) has been adopted by psychologists who claim that all psychotherapies, regardless of their specific components, produce equivalent outcomes (Rosenzweig, 1936). However, it also summarizes, perhaps even more aptly, the essence of two widespread judgmental biases: additivity neglect and the nonselective superiority bias.

People perform quality assessments and probability judgments, at least implicitly, several times every day. Such judgments are often comparative, or referent-dependent, as they involve the relative standings of several candidate objects or events. For example, when trying to decide which computer to buy, or when evaluating one’s chances of winning a friendly game of golf, the judgments should take the relative merits of several computers or fellow golfers into account. In other words, such judgments do not only require people to consider the *target* (the item or outcome in focus), but also the *referents* (the other items or outcomes in the set).

However, such judgments are often inconsistent and appear to violate fundamental principles of logic and mathematics. For instance, when people are asked to estimate the winning chances of each member in an exhaustive set of candidates, they often break the 100% limit. In other words, they disregard basic requirements of *additivity*. Likewise, when they are asked to assess the relative merits of a strong set of job applicants, people tend to consider most of them above the group average, despite the fact that the task calls for a *balanced* distribution of judgments around the group mean. Thus, both biases reveal a tendency to weigh the features of the target (the focal computer or your own golf skills) more heavily than those of the referents (the relative merits of other computers, or your friends’ golf skills).

Evidence for these judgmental biases comes from two streams of research, namely from studies of non-complementary probability estimates, chiefly subadditivity, on one hand, and from social psychological research on the so-called nonselective superiority/inferiority bias (NSSB/NSIB), including the Better Than Average-effect, on the other. Research on “nonunique vulnerability” has included probability (risk) judgments of single individuals vs. the average group member (Klar, Medding, & Sarel, 1996), leading to research on “direct” vs. “indirect” comparisons of risk (Price, Pentecost, & Voth, 2002; Rose, 2010). However, this research did not explicitly consider the additivity issue.

The similarity between additivity neglect and imbalanced judgments can be observed in several aspects: firstly, both biases seem to involve a consistent overestimation of items or alternatives in a positively valued, exhaustive set. Secondly, judgments appear to be made in a non-complementary fashion, that is, participants seem to think that selected items can be preferred without dispreferring other items accordingly. Thirdly, they have mostly been studied in between-group designs, as there seems to be tacitly supposed that giving participants the full set would alert them to the underlying logic, thereby making them less biased. Lastly, some of the mechanisms posited to explain NSSB are similar to those suggested for additivity neglect. Both biases appear to depend on selective processing (Kardes, 2013), where target outcomes are assessed one-by-one, without proper consideration of the alternatives. A large body of research has indicated that several biases or judgmental errors can result from selective processing, amongst them both the nonselective superiority bias and subadditivity (or “unpacking effects”) (Kardes, 2013; Posavac, Brakus, Cronley, & Jain, 2009; Posavac, Brakus, Jain, & Cronley, 2006; Sanbonmatsu, Posavac, Kardes, & Mantel, 1998; Sanbonmatsu, Vanous, Hook, Posavac, & Kardes, 2011). Despite these similarities, few attempts have been made to study these two biases together. In a recent integrated treatment, Smith, Windschitl, and Rose (2015) suggest that both

types of judgments are dependent upon an initial impression of the alternatives (as strong or weak), as well as a more controlled aggregation process (checking the judgments against task constraints). The authors predict that both biases will be affected by impression strength and time pressure, but that only additivity is affected by set size (as this bias is revealed when estimates are added together, whereas NSSB estimates are averaged) .

A closer analysis of the two biases suggests that they are similar in the sense that both involve neglect of elementary mathematical operations, namely averaging (in the case of NSSB) and adding (in additivity neglect), either because people don't think of them or dismiss them as irrelevant in situations where individuating (case) information is available. They differ, however, in the way the rating scales provide valid reference points for the individual judgment. In the case of NSSB tasks the default estimate (the ignorance prior) is explicitly provided as the midpoint of the rating scale (the zero point on a bipolar scale). Yet, individual ratings tend to cluster above, rather than around this reference value, presumably because they are also (implicitly) compared with a more general population standard to which they are all superior. Thus the bias presupposes, as the name suggests, a set of *strong* alternatives.

In contrast, the scale for probability estimates in additivity tasks does not indicate a valid default value, unless calculated by the participants themselves as $1/n$, where n = number of alternatives. Thus with four applicants for a single job opening, their chances should cluster around 25% as the ignorance prior. This value is, however, rarely made salient on the response scales. Responses are typically given on a probability scale centered on 50% as the most conspicuous reference value, but this value works as a valid default only in the binary case. Sums of probability estimates have been found to exceed the 100% limit even for sets without an outstanding ($> 50\%$) candidate. Thus additivity neglect is not confined to sets of strong alternatives.

Thus a comparison between the two biases suggests that they may be committed by the same individuals, yet affected (increased or decreased) by different considerations, which do not make them completely interchangeable. The strategy followed in the present studies is to use the same material and answering format in both tasks, to ensure comparability. We also add to previous research by giving both tasks, including complete sets of alternatives, to the same participants, to maximize task transparency and allowing us to compare the two biases directly.

The nonselective superiority bias

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) 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, & Krizan, 2008). This indiscriminate superiority violates elementary logic, as some members cannot be better unless others are worse. Several theories have been suggested to account for such findings. *Focalism* accounts suggest that more attention is allotted to the target than to the referents, enhancing its attractiveness (Kahneman, Krueger, Schkade, Schwarz, & Stone, 2006; Schkade & Kahneman, 1998; Wilson, Wheatley, Meyers, Gilbert, & Axson, 2000; Windschitl, Kruger, & Simms, 2003). According to the *unique attributes* account (Chambers, 2010), people focus excessively on features that positively distinguish the target from the referents in the local group (i.e., Las Vegas has considerably more entertainment options than most other destinations, whereas Hawaii is an ideal place for watersports), which might lead to an over-evaluation of the target. In contrast, the *LOGE* model (Giladi & Klar, 2002; Klar, 2002) posits that participants solve such tasks by making a “bifocal” comparative judgment, partly comparing the target to the referents, while simultaneously making a logically irrelevant comparison between the target and a general standard.

Subadditivity

Non-additive probability estimates have been frequently observed within both economics and psychology (e.g., Gilboa & Schmeidler, 1994). Subadditivity is the tendency to judge the probability of the whole set of outcomes to be less than the total probabilities of its parts (Tversky & Koehler, 1994). This means that individual outcomes in an exhaustive set will add up to more than 100%. Subadditivity has been demonstrated in many studies, using both students and experts as participants, such as experienced physicians (Redelmeier, Koehler, Liberman, & Tversky, 1995), chess players (Nordbye & Teigen, 2014), rescue workers (Hill, 2012), and options traders (Fox, Rogers, & Tversky, 1996). Subadditivity has been studied by comparing packed vs. unpacked lists of outcomes (Tversky & Koehler, 1994), by adding up between-subjects' probability estimates for individual outcomes in an exhaustive set (Redelmeier et al., 1995), and by adding up within-subjects' probability estimates for a complete list of outcomes (Dougherty & Hunter, 2003a; Dougherty & Hunter, 2003b; Fox, 1999; Posavac, Kardes, & Brakus, 2010; Riege, Sulutvedt, & Teigen, 2014; Riege & Teigen, 2013; Robinson & Hastie, 1985; Teigen, 1974, 1983). For example, participants asked to estimate the chances of four job applicants competing for one opening, will often produce probabilities that taken together greatly exceed 100%, even when all applicants are listed on the same page. This phenomenon might be called “additivity neglect” (Riege & Teigen, 2013), as many people seem to disregard the additivity requirement of an exhaustive set of independent outcomes.

Subadditive probability judgments have been given a variety of explanations. Support theory (Tversky & Koehler, 1994) suggests that the probability assigned to an event reflects evidence favouring one alternative relative to evidence favouring other alternatives, which are often more poorly specified and less available. An extension of support theory also includes comparison between the target and a default (or general) standard (Idson, Krantz, Osherson, &

Bonini, 2001). Focalism accounts view the overestimations as a result of people focusing on the relevant information concerning the target outcome, giving it excessive impact on the judgment, while disregarding information relevant to the referents (Fox et al., 1996; McKenzie, 1998, 1999; Windschitl et al., 2003). However, such mechanisms may not be sufficient to explain the persistence of additivity neglect in situations where all alternatives are presented together and thus made equally available. Additivity neglect could also arise from a case-based approach, where each alternative is evaluated in isolation, according to its inherent strengths, irrespective of the number and strength of other alternatives. Some studies have accordingly suggested that biases that arise from selective processing will be reduced when comparative processing is encouraged (Kardes, 2013; Posavac et al., 2009; Posavac et al., 2010; Sanbonmatsu et al., 1998). For example, asking participants to allocate a proportion of a 100 dollars to a list of four national parks resulted in less “overspending” compared to participants who were asked to allocate money to one randomly drawn focal park (Posavac et al., 2009). However, several studies have shown that comparisons among alternatives do not by themselves lead to additive judgments. Judgments must also be seen as complementary, that is, that a high probability of a favoured alternative requires the probabilities of less favoured alternatives to be correspondingly small. Additive judgments presuppose a class-based approach, where the total probability (100%) is distributed amongst the individual outcomes, adjusting their probabilities upwards or downwards according to available evidence, while keeping the estimates within the limit. The need to keep within limits might be clearer when allocating dollars than probabilities.

The present research

The main goal of the present research is to compare additivity neglect and NSSB empirically in a parallel design, where the same participants receive both kinds of tasks. We also increase the transparency of the tasks by asking participants to make assessments of all items in

the set. This allows us to test the robustness of the effects, their relative strength, and to examine whether the same individuals exhibit both biases to the same degree.

Normative referent-dependent judgments should adhere to the complementarity principle: If the probability/superiority of one alternative increases, the mean probabilities/superiority of the other alternatives must decrease accordingly (and vice versa). We therefore posit that people who exhibit additivity neglect and the nonselective superiority bias somehow fail to realize the full implications of this normative requirement, and do not think it imperative to check the sums of probability estimates or the means of the comparative ratings. The observed tendency to overestimate such “unconstrained” judgments could be a result of considering each alternative independently (Riege & Teigen, 2013; Sanbonmatsu et al., 1998; Sanbonmatsu et al., 2011), comparing them to general standards in their respective domains rather than to the other alternatives (Giladi & Klar, 2002; Idson et al., 2001). Alternatively, comparisons could be made pairwise between individual alternatives in the set rather than with the group as a whole (Windschitl & Wells, 1998).

Another source of bias could be related to the response format. The rating scales used in studies of comparative judgments are typically symmetrical around a zero midpoint, dividing the responses into a positive and a negative region. Probabilities are, in contrast, typically rated on a visual scale ranging from 0 – 100% with increments of ten, where only the extremes are clearly defined. However, even on such scales people will often define the 50% midpoint as a kind of neutral or default value, sometimes simply used to indicate “I don’t know” (Fischhoff & Bruine de Bruin, 1998), with higher probabilities perceived as good chances and lower probabilities perceived as small chances. In line with this interpretation, Riege and Teigen (2013) found more additivity neglect for probabilities estimated on a 0-100% rating scale with 50% as the midpoint, than when participants were asked to generate written estimates without this scale.

In Experiment 1, the same participants performed both comparative judgments and probability estimates on sets of five nominees to win the Oscar for best actor or best actress rewards. This design allowed comparison of both biases, in a full set of outcomes, for equivalent tasks. In addition, Experiment 1 attempted to induce comparative processing by asking participants in one condition to rank-order the alternatives before answering the probability tasks.

In Experiment 2 and 3 participants gave comparative judgments and probability estimates to sets of four shortlisted job applicants. Again, participants were asked to evaluate all candidates, one by one, either according to their qualifications for the job, relative to the others in the selected group, or according to their chances of being hired. Experiment 2 allowed direct comparisons of both kinds of ratings by presenting participants with a prepared set of qualification ratings, which they were asked to “translate” into probabilities, and vice versa: by giving them a prepared set of probabilities and ask for corresponding qualification ratings. By presenting participants with unbiased (balanced, additive) sets we gave them an additional hint of the normative requirements of such judgments. This procedure also provides some insight into scale usage and reference points, by indicating whether qualification judgments below or above the midpoint – zero – on the comparative scale will correspond to probability judgments below or above the midpoint – 50% – of the probability scale.

Experiment 3 further explored the issue of different response formats, by having participants in separate conditions submit their responses either as tick marks on a visual scale or by typing appropriate numbers. Finally, participants were asked why they had, or, in most cases, why they had not, balanced their ratings or made their estimates add to 100%. In a previous study of unrealistic optimism in comparative risk judgments, participants were asked to provide written accounts of their thought processes (Klar et al., 1996), and similarly, Posavac et al. (2010) asked

participants *how* they came to their judgments. However, we have found no previous study where the *why* question has been asked.

Experiment 1

Experiment 1 was designed with three objectives: To examine (1) to which extent people exhibit additivity neglect and NSSB when they are asked to evaluate multiple targets presented side by side in the same session; (2) whether instructions requiring participants to compare the alternatives with each other will reduce these two biases; and (3) the relationship between these two biases; specifically, whether participants who exhibit additivity neglect also tend to exhibit NSSB.

Most studies of subadditivity have used a between-Ss design. There seems to be an implicit assumption that judgments of all items together will make the task transparent and suggest a balanced set of judgments, but, as mentioned in the introduction, additivity neglect has also been demonstrated in studies where the same judge is asked to estimate the whole set in the same session (Fox & Tversky, 1998; Riege et al., 2014; Riege & Teigen, 2013; Robinson & Hastie, 1985; Teigen, 1983, 1988). With some exceptions (e.g., Chambers, 2010, Experiment 2; Klar, 2002, Experiment 4), all NSSB studies to date have presented different items to different people.

The experiment included a manipulation where half the participants rank-ordered the alternatives prior to the additivity task. This procedure would counteract a tendency to view each alternative in isolation, and thus might facilitate additive responses. The role of comparisons is supported by an eye-tracking study of additivity neglect that indicated that participants indeed moved their gaze more frequently between alternatives when answering in an additive manner (Riege et al., 2014). Rank orderings have, to our knowledge, never been studied with additivity tasks. However, in a previous NSSB study Bruchmann et al. (2013) made participants' rank-order

the alternatives from most favorable to least favorable, predicting less biased responses. Contrary to expectation, these investigators found no effect of rank-ordering on NSSB.

Method

Participants. Participants were 107 Amazon MTurk workers located in the US (of originally 131 submitted questionnaires, 24 were discarded because they were incomplete, contained ratings instead of probability estimates, or wrong answers to the attention control questions). There were 40.2% women, median age 29.5 years. Approximately 45% of the participants had four years of College, and only 15% had no College education. They were randomly assigned to one of two conditions.

Design and questionnaires. The data collection was collected less than two weeks before the 86th Academy Awards (also known as the Oscars), which was scheduled on March 2nd, 2014, at the Dolby Theatre in Los Angeles. The questionnaire briefly described five actors and five actresses nominated for the best actor / actress in a leading role, and the five actors / actresses nominated for the best actor / actress in a supporting role.

All participants received two probability estimation tasks followed by two NSSB tasks. In the two probability estimation tasks they were given one list of five actors, and another of five actresses, and asked to estimate the chances of each nominee winning the Oscar for best actor [actress] in a leading role. Participants were asked to give their answers as numbers between 0-100%.

For the two NSSB tasks participants were given two other lists, one of five actresses, and one of five actors, and asked to estimate how much each nominee deserved winning the Oscar for a supporting role, using a scale from -4 to 4, where -4 was labelled: “Much less than the others”; 0 was labelled: “As much as the others”; and 4 was labelled: “Much more than the others”. Thus the probability estimation tasks asked about chances and the NSSB tasks asked about

deservingness, both using formats representative of typical studies within these separate research traditions.

Participants in Condition 2 were in addition asked to rank-order the nominated actors/actresses prior to their probability estimations, forcing them to make comparisons between alternatives, whereas participants in Condition 1 were not given any prior task.

All participants were finally asked which of the nominated movies they had seen and how much they liked/disliked the movies, how interested they were in movies and the Oscars, followed by an attention control question (Goodman, Cryder, & Cheema, 2013; Oppenheimer, Meyvis, & Davidenko, 2009).

Results

Probability estimates for the sets of actors/actresses were added to detect non-additive responses. For the NSSB tasks, ratings were averaged to reveal unbalanced sets (means above 0). We defined in the current studies response sums between 90% and 110% as additive, and sets with means between -0.1 and + 0.1 as balanced, to include answers with minor inaccuracies. Despite this lenient criterion, a large proportion of participants exhibited additivity neglect: 59 % in best actor and 41 % in best actress vignette; and an even greater number produced unbalanced sets: 86 % gave unbalanced ratings for best actor and 79 % for best actress. Table 1 shows that on average, probability sums for five candidates were between 150 and 200%. The distributions of NSSB ratings were positively skewed with mean scores higher than 0.5.

<Insert Table 1 about here>

The rank-ordering manipulation introduced in Condition 2 was not successful in facilitating additive responses. On the contrary, the sums of the probability estimates were higher in this condition than in Condition 1 (the difference is only significant for the set of actors, with $t(89.34) = -2.09, p = .04$), whereas the NSSB tasks showed a slight, but non-significant reduction

of mean ratings in Condition 2, in the predicted direction. Table 1 further shows that male actors were estimated to have better chances than actresses, in all conditions.

To examine the relationship between additivity neglect and NSSB, we compared mean ratings of additive responders (participants with probability sums in the 90-110% range) and non-additive responders (participants with probability sums outside of – mostly above – this range). Non-additive responders gave generally higher NSSB ratings ($M = 0.84$) than the additive responders ($M = 0.39$); $t(105) = -3.365$, $p = .001$, Cohen's $d = 0.67$, rendering evidence of a connection between these two tasks. Probability sums and mean ratings were positively correlated, $r(105) = .375$, $p < .001$, indicating that subadditive probability estimates and positively skewed ratings may share some common mechanisms.

Discussion

Experiment 1 demonstrated that the participants' nonselective superiority bias prevailed, even when given all the alternatives in the set, confirming the results from Chambers' (2010) Experiment 2. Participants in Condition 2, who were asked to rank order the alternatives, did no better on the additivity task than the others. Thus the biases demonstrated in the present study are not simply caused by a failure to consider the alternatives, mirroring Bruchmann et al.'s (2013) findings with additivity neglect. The experiment showed a positive correlation between the two biases, indicating underlying communalities. However, all participants received the tasks in the same order (probability judgments before deservedness ratings), which might have influenced the rating tasks. Also participants' own preferences for movies and actors might have made them less attentive to the formal requirements of the tasks. In the following experiments the order of the tasks is counterbalanced, and the judgments made in the two tasks are of a more "objective" nature, concerning the chances and qualifications of job applicants according to descriptions given in the text.

Experiment 2

Experiment 1 showed that participants who are asked to assign probabilities to a set of selected candidates (Oscar nominees) will violate the 100% limit. They will also rate most of them as having better chances than the mean of the set, even when all candidates are presented together. Experiment 2 was designed to investigate the relationship between these two types of judgments more directly by asking participants to suggest qualification ratings based on probabilities of success, and vice versa: to estimate probabilities of success based on qualification ratings. Material for this study was sets of four fictitious job seekers applying for the same job openings. They should, on average, have a 25% chance each of getting the job, and have qualifications that would place them, on average, around the zero point of a -4 to +4 scale. In line with these requirements, we constructed sets of probabilities that did not violate additivity norms, by giving, for instance, one applicant a 40% chance, another 30%, and the last two a 15% chance each. How would this additive set be “translated” into qualification ratings? We could envisage three possibilities:

- (1a) Participants will continue to produce positively skewed ratings, as they did in Experiment 1, suggesting that most candidates in a highly qualified group will be judged as better than the average group member (with a relative standing reflecting their chances of being hired).
- (1b) Presented with an additive set of probabilities, participants may realize that also their qualification ratings should conform to normative constraints, and produce more balanced sets of ratings than they would have done otherwise.

(1c) Since “unconstrained” participants often produce subadditive (exaggerated) probability estimates, they might find the provided additive probability estimates to be rather low, and hence suggest rather poor qualifications. This might lead to negatively skewed distributions of ratings, where most candidates are perceived as *worse* than the average group member.

Similarly, we constructed sets of balanced qualification ratings (e.g., +3, +1, -2, and -2 for four candidates, respectively). How would ratings from such “balanced” sets be translated into probabilities for becoming hired? Again, three hypotheses could be posed:

(2a) Participants might continue to make subadditive probability judgments, with sums of four candidates exceeding 100%, as in Experiment 1.

(2b) The balanced sets of ratings might serve as a “hint”, suggesting complementary probability estimates as well, leading to more additive estimates.

(2c) Based on the fact that people usually give more positive than negative ratings, they may find too many negative ratings in our “balanced” distributions, which in their minds might be associated with low rather than high probabilities. This could lead to a superadditive set of probability estimates (i.e., estimates that add up to *less* than 100%).

All these hypotheses are, in our opinion, to some extent plausible, and may be valid for some rather than all participants. If this is the case, we should on average expect reduced probability estimates (fewer participants with additivity neglect), and rating sets that are less positively skewed than in Experiment 1. Whether the effects sketched in hypotheses (b) or (c) above are strong enough to counteract, or reverse the standard superiority/subadditivity biases, remain to be seen.

Method

Participants. Participants were 77 Mechanical Turk workers (excluding one who had misunderstood the task and was omitted from the analysis), 45.5% women, median age 33 years. All participants had finished high-school, and 55% had at least four years of college. They were randomly allocated to two conditions, which differed according to the order of tasks.

Materials and design. Participants received six written announcements of job vacancies (as senior attorney, regional sales manager, university lecturer, accountant, real estate manager, and mechanical engineer) based on real job announcements and checked by an experienced recruiter. They were also presented with fictitious assessments from a recruitment company of the four best applicants to each job. These assessments were said to be based on the applicants' GPA, previous job experience, personality, and impression from interviews, and were presented in one of two formats: either as probabilities of being hired, or as job qualification ratings on -4 to +4 rating scales.

Participants in Condition 1 were first given three vignettes where the individual applicants' chances of getting the jobs were listed as probabilities, forming an additive set. For instance, the probabilities of Steven, Carol, Joseph, and Barbara of getting the job as a university lecturer were set to 5%, 30%, 25%, and 40%, respectively. Participants were on this basis asked to assess the candidates' qualification scores as ratings from -4 to 4; thus, they were required to "translate" numerical probabilities into qualification ratings.

The next three vignettes followed an inverse procedure. Each of these three jobs were accompanied by four nine-point rating scales with candidates placed according to their relative qualifications, from -4 (much less qualified than the others) to +4 (much more qualified than the others). These provided scores were constructed so as to form balanced distributions of ratings, for instance Kenneth, Margaret, Michael, and Susan were given relative qualification scores of 1,

4, -3, and -2 for the job as a mechanical engineer. Participants were then asked to suggest for each candidate his or her probability of getting hired, as a number between 0 and 100%; they were in other words asked to “translate” a given set of qualification scores into numerical probabilities.

Participants in Condition 2 received these two tasks in reverse order. They started with three vignettes where they translated qualification scores into numerical probabilities, before being given the NSSB task where they were asked to translate probabilities into qualification scores.

Results

When asked to translate a balanced set of qualification ratings into probabilities, non-additive estimates were rather frequent, with sums exceeding 100% in 31 of 77 cases (overall $M = 132.3\%$). As seen in Table 2, the two condition differed, as participants in Condition 1 gave significant lower probability estimates, $t(60) = -3.37, p = .001$, and produced a greater number of additive sets. Observe that these participants had started with the NSSB task, where they had been shown three sets of additive probabilities. This might have alerted several participants to the 100% rule, which they continued to follow in the second part of the experiment.

Additive probability estimates were, in turn, translated into qualification ratings that yielded negatively skewed distributions, with means *lower* than the scale midpoint of 0 (Table 2, left panel; overall $M = -0.55$). Only 12 out of 77 participants produced balanced sets (defined as mean ratings between -0.1 and +0.1). Again, participants in Condition1 produced mean NSSB ratings closer to 0 than participants in Condition 2, but this difference did not reach significance; $t(52) = 1.72, p = .09$.

<Insert Table 2 about here>

To examine the relationship between additivity neglect and NSSB, mean qualification ratings of additive responders (participants with probability sums in the 90-110% range) were compared to ratings produced by non-additive responders, as in Experiment 1. Non-additive probability responders gave generally more imbalanced sets of qualification ratings ($M = -1.27$) than the additive responders ($M = -0.10$); $t(49.5) = -6.544, p < .001$, Cohen's $d = 1.37$. For participants in Condition 1 the correlation between qualification ratings and mean probability estimates was $r(39) = -.31, p < .05$, while in Condition 2 it was even stronger, $r(34) = -.84, p < .001$, rendering further evidence of the relationship between the two types of biases.

Discussion

The results support the prediction that participants presented with a balanced set of qualification ratings will reduce their probability estimates compared to the exaggerated estimates they usually produce when such ratings are not provided (as in Experiment 1). However, they were still more subadditive than superadditive, giving more support to hypotheses 2a and 2b than to 2c.

Qualification ratings were even more strongly affected. After being presented with an additive set of probabilities, participants judged many job applicants' qualifications to be below rather than above the mean, producing distributions that were negatively skewed. Thus these results were more in line with hypotheses 1b and 1c than 1a.

These results indicate that the effect of qualification ratings upon probability estimates and the effects of probability estimates upon qualification ratings are not symmetrical. People oppose the idea that a candidate can have a "low" probability ($p < 50\%$) of being hired and still have better than average qualifications. They find it more conceivable that a candidate with low or moderate qualifications (compared to others in the group) can have a reasonably high probability of getting hired. These asymmetries reveal that participants who produced additive

probability responses did not, as a rule, produce balanced qualification distributions, but those few who gave balanced ratings tended to give additive probabilities as well.

Experiment 3

This experiment was designed with three objectives in view.

(1) First, as a control for findings in Experiment 2, where we claimed that the “translation task” had reduced the additivity neglect and inverted the NSSB bias. This called for a control study using the same material without asking for translations.

(2) Previous studies of additivity neglect have indicated that probability estimates can be strongly influenced by answering format (Riege & Teigen, 2013; Riege et al., 2014). Specifically, participants who are asked to come up with their own written or spoken estimates (“self-generated format”), appear to produce more additive responses than those using the “scale format”, who submit their estimates by simply choosing a number from 0 to 100 along a horizontal scale. In the two previous experiments, probability estimates were typed in as self-generated numbers, whereas ratings were submitted as tick marks on a rating scale. Participants in the present experiment answered both tasks either in the self-generated or in the scale format. This would make their answers more strictly comparable, and form a better basis for assessing the correlation between the two tasks. Further it might reveal whether additivity and NSSB tasks are influenced by response format to the same extent.

(3) Studies of additivity and NSSB have rarely asked people to explain their own biased responses, despite some studies asking participants to provide written or oral accounts of their thought processes whilst performing their judgments (e.g., Klar et al., 1996). In Experiment 3 we asked participants to offer reasons for giving or not giving additive and balanced distributions of responses, probing their understanding of the normative principles involved. Do they violate these principles inadvertently, or intentionally, as the result of a deliberate choice?

Method

Participants. Altogether 200 Mechanical Turk workers were recruited to fill in an online (Qualtrics) questionnaire. Of these, 162 participants finished the questionnaire and answered the attention control question correctly (56.2% women, median age 33 years). More than half the participants had at least four years of college.

Materials and design. Participants received six vignettes regarding job hiring scenarios, identical to those used in Experiment 2. However, this time they were not asked to “translate” ratings into probabilities, or vice versa, but to assess chances and qualifications from short thumbnail descriptions of each candidate, in the form of GPA scores, personality, and work experience. Three vignettes asked for qualification ratings of each of four shortlisted candidates (NSSB tasks), and three other vignettes asked for the candidates’ probabilities of being hired (additivity tasks). Half of the participants received the NSSB tasks first followed by the additivity tasks, and vice versa for the other half. No order effects were found, so in the subsequent analyses, data from both orders are pooled.

Participants were randomly assigned to two format conditions. In Condition 1, the Self-generated format condition, participants gave their answers by typing appropriate numbers in designated slots on the screen. For the NSSB tasks they were asked to type numbers between -4 and 4, where -4 meant: “Much less qualified than the other applicants”; 0 meant: “As qualified as the other applicants”; and +4 meant: “Much more qualified than the other applicants”. Answers to the additivity tasks were given by typing probability estimates between 0 and 100.

In Condition 2, the Scale condition, participants gave their estimates to both kinds of tasks as tick marks on horizontally arranged scales. The NSSB tasks were accompanied by 9-point scales ranging from -4 to +4, labelled in the same way as in Condition 1. The probability

estimates in the additivity tasks were given on 11-point scales from 0% to 100%, with increments of 10%.

Self-reports. After completing all judgment tasks, participants were asked to look back on their probability estimates, and asked if they had considered adding these estimates to 100%. They were offered four response alternatives: *Yes, for all the tasks*; *Yes, for some of the tasks*; *No, I did not*; or *Don't remember*. Subsequently they were asked to give a reason for why they had or had not considered this. Those who responded “Yes” were further asked to state how difficult it had been to make the probability estimates sum to 100%, and all were asked how difficult it had been to come up with probability estimates. Similar questions were asked for the NSSB-ratings.

Results

The participants' probability estimates were summed for each additivity task, and their mean qualification ratings were calculated for each NSSB task. As seen in Table 3, participants' mean estimates were in both cases highly biased. Overall, only 30 out of 162 participants (18.5%) gave consistently additive responses (sums in the 90-110% range), the rest (except one) produced probabilities adding up to more than 110%. In the NSSB task, only a single participant produced three balanced sets of ratings, using the same criterion as in Experiment 1, whereas a majority of 142 (87.7%) produced distributions that were on average positively skewed.

Format differences were demonstrated for probability estimates, replicating earlier findings. In the scale condition (Condition 2), only 10.5% of the participants gave one or more additive responses, whereas in Condition 1, where participants typed in their answers, as many as 35.5% gave one or more additive responses. In contrast, NSSB ratings appeared not to be affected by response format.

<Insert Table 3 about here>

Participants who produced additive probability estimates gave more balanced NSSB ratings ($M_{\text{NSSB}} = 0.65$) than those who showed additivity neglect ($M_{\text{NSSB}} = 1.16$), $t(160) = -2.949$, $p = .004$, Cohen's $d = 0.61$, indicating a relationship between the two tasks. Overall, participants' mean probability estimates and mean NSSB ratings were positively correlated, $r(160) = .36$, $p < .001$, replicating the findings from Experiment 1. The correlation was especially strong in the Scale format condition, with $r(84) = .60$, $p < .001$, in Condition 2, against $r(74) = .23$, $p < .05$, for Condition 1.

Previous studies have shown that the Self-generated answering format makes participants more additive in their probability estimates (Riege & Teigen, 2013; Riege et al., 2014). Barring this “de-biasing” format the correlation between these two biases seems to increase, indicating that participants who overestimate all candidates' chances are often those who consider all candidates to be better than the group average regarding qualifications.

Self-reports. Most participants (68.5%) claimed they had not considered the additivity or balancing principles at all. Those who did, felt that adhering to these rules were easier with self-generated estimate than in the scale condition. Participants produced a variety of written explanations for (not) balancing or (not) adhering to the additivity principle. For the present purpose, they were coded in four distinguishable categories. (1) Rule-based assessments; mostly from unbiased participants. (2) Case-based assessments, from participants who explained that they had considered each applicant separately. (3) Assessments from participants who felt it would have been “unfair” or an incorrect assessment of the candidate's qualifications or chances had they balanced their ratings or adhered to the 100% rule. (4) Various other reasons, many of these from participants who explained “how” they had made their assessment rather than “why” they had followed or not followed the normative rules. Descriptions and examples of reasons within each category are provided in Table 4.

The written explanations were coded by two independent raters. After one round of practice, a randomly selected subset of statements (22%) were coded by both raters, and the inter-rater agreement between them had a Cohen's kappa = .89.

As seen in Table 4, Case-based explanations were twice as frequent for probability estimates as for qualification ratings ($p = .003$ by a binomial test). After all, the instructions explicitly required the qualification ratings to be comparative (much less / much more qualified than the other applicants). In contrast, fairness or “correctness” were more often given as reasons for imbalanced qualification ratings than for subadditive probabilities ($p = .008$ by a binomial test). These different explanations point to subtle differences between the tasks. Ratings below the midpoint on a -4 to +4 scale clearly indicate a negative assessment, which people think should be avoided in a group of “superior” job applicants.

<Insert Table 4 about here>

General Discussion

The three experiments reported in the present paper indicate several commonalities between additivity neglect and the nonselective superiority bias. Both biases appeared to be about equally widespread, and occurred in a majority of participants even when presented with the full set of alternatives. They were also related, as participants who gave additive probability estimates gave more balanced distributions of ratings to the NSSB task. Experiment 1 and 3 revealed positive correlations between the two biases, but in Experiment 2 the correlations were negative, as the ratings in this study were negatively biased rather than positively biased.

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

(Kardes, 2013; Koehler, Brenner, & Tversky, 1997; Posavac et al., 2009; Posavac et al., 2010).

However, this was not sufficient in the present experiments. Experiment 1 attempted to further induce comparative processing by asking participants to rank-order the alternatives; however, this did not lead to less additivity neglect, but rather the opposite in one condition.

Experiment 2 indicated that probability estimates up to 30-35% were typically perceived as low, warranting negative ratings, despite the fact that with a set of four applicants, 25% indicates equal chances. Applicants with an assigned probability higher than 25% should accordingly be rated as better, not worse than average. Experiment 3 corroborated these findings indicating that many thought that adhering to the 100% rule or balancing the ratings would be an “unfair” or incorrect assessment of the candidates.

Together, these findings demonstrate that these two biases are not simply a result of a minority producing erroneous responses in the same direction, as suggested by Krueger and Funder (2004), but are routinely performed by a majority so as to constitute the rule and not the exception (Klar & Levi, 2004). It has indeed been difficult to find limits to the NSSB (Bruchmann et al., 2013; Klar & Ayal, 2004), thus the findings in Experiment 2 that “low” but appropriate probabilities can reverse this effect are remarkable.

NSSB and additivity neglect tasks are structurally similar, in that they both ask people to perform judgments of an exhaustive set of alternatives. They also share an implicit normative requirement, namely that judgments should be complementary and stay within bounds. However, the constraints are of a different nature. In additivity tasks, judgments are constrained by a fixed sum (100%), whereas in comparative judgment tasks, ratings are constrained by a fixed mean (0). Adherence to or violations of these constraints must either be continuously updated and adjusted, or be checked retrospectively, after all judgments have been performed, so they can be aggregated (in the first case) or averaged (in the second). Either way, they can only be respected

by participants who are willing to take a second look and adjust their individual first estimates accordingly. It may not come as a surprise that most participants failed to take this second look to fulfil a requirement that was never explicitly pointed out to them. Some participants in Experiment 3 explained that they had failed to produce 100% sums of probabilities because “it was not mentioned in the instructions”.

Explanations of additivity neglect and NSSB must give an account of two prominent and distinguishable findings: (1) why the formal rules are neglected, and (2) why the distributions of answers are positively skewed.

(1) People who fail to understand basic mathematical concepts and principles are often described as low in numeracy, defined as the ability to reason with numbers, percentages and similar mathematical concepts (Peters et al., 2006). The present experiments did not include a numeracy measure. However, previous studies have shown that more numerate participants are less susceptible to additivity neglect, particularly when reminded of thinking in a mathematical manner, for instance by being presented with a numeracy test before rather than after the additivity task (Riege & Teigen, 2013). Thus, it may not be enough to have acquired the appropriate mathematical skills; one also has to realize that they are relevant and applicable to the task in question. Results from Experiment 2 showed that participants in Condition 1, who were initially presented with additive probability sets (to be “translated” into comparative judgments), produced more additive probability sets in the second part of the experiment, compared to participants who received the tasks in reverse order. Thus some Condition 1 participants (perhaps the more numerate ones) profited from being “primed” with additive sets of probability.

(2) However, rule neglect, whatever its origins, is not sufficient to explain why judgments of both kinds were strongly biased towards the positive end of the scale. Probability estimates were unrealistically high, and qualification judgments were mostly above average. Such biases

can arise from three sources, pertaining to biased perceptions (a) of the targets, (b) of the tasks, and (c) of the response scales.

(a) When judgments are taken at face value, it appears that participants tend to overestimate, or exaggerate the merits of almost all alternatives indiscriminately. This hypothesis, which can find some support in studies of focalism and related accounts, like the unique alternatives hypothesis (Chambers, 2010), may account for a part of the bias, but is hardly enough to explain its persistence and magnitude when all targets are presented together and evaluated one after the other, as in the present studies.

(b) A “positivity” bias could also arise if participants did not comply with the task, but judged each target from its absolute qualities rather than relative to the others. This may seem to be an unlikely strategy, especially for the NSSB task, which explicitly asked for comparative judgments. But also judgments of probabilities are commonly seen as dependent on comparisons rather than being fixed properties of a target outcome. For instance, most people presumably know that one’s winning chances are reduced when a stronger competitor enters the scene, and that one’s chances are improved when tasks are easy and competitors weak. Moreover, a failure to understand the comparative nature of the task would be alleviated by asking people to rank order the alternatives, as in Experiment 1. However, this did not lead to lower probabilities in the probability estimation task.

A weaker version of this hypothesis might still be tenable. It could be the case that even people who actually compare alternatives place more weight on the strengths and shortcomings of the target, than on those of its referents, even when all outcomes are given target status one by one. Such a strategy could be chosen for several reasons. In everyday life, one usually knows more about the qualities of the target outcome than those of an undefined set of “other” candidates. Even in a clearly defined set, as in the present studies, the concept of “others” is more

fuzzy and requires more effort to be summarized (averaged or aggregated) than the target. To ease cognitive processing demands, probability estimates are often made as comparisons between individual targets rather than with the group as a whole (Windschitl & Wells, 1998), or by bringing in more general referents in addition to the local comparison set, as claimed by the LOGE account (Giladi & Klar, 2002) and extended support theory (Idson et al., 2001). Including a general standard as a second reference-class (e.g., average job applicants) would give most members in a superior set, when judged sequentially, a better score and a better chance in comparison to the more obscure and ambiguous category of “others”.

(c) Response scales are not neutral carriers of information, but provide responders with reference values over and above those supplied by the referents in the comparison set. Ratings scales with a zero point in the middle, as those used in many NSSB tasks, might lead to higher ratings than scales without this midpoint, for the simple reason that participants tend to think of scores below zero as “bad” in a more absolute sense (Schwarz, Knäuper, Hippler, Noelle-Neumann, & Clark, 1991). Probability scales usually come without a predefined neutral reference point, but people who are used to think about probabilities in a binary setting often regard the 50/50 value as the “default”, or their ignorance prior (Fox & Rottenstreich, 2003). They will accordingly give preferred candidates probabilities that are too high, neglecting the fact that the neutral reference value for four candidates is 25% rather than 50%. Results from Experiment 2 indicated that probabilities around 25% were indeed judged to reflect poor (negatively valued) qualifications, and several participants in Experiment 3 argued that additive probability estimates would paint an “unfair” picture of the individual candidates. Experiment 3 also replicated previous studies showing that ready-made 0-100 probability scales (with 50% as the scale midpoint) lead more often to inflated probability estimates, than self-generated estimates made without the “aid” of such scales.

In the introduction we suggested that rating scales with an explicit reference point (0) may result in different judgments from probability scales, where the default depends on inferences made by the responder. If, for simplicity's sake, we distinguish between class-based (normative) and case-based responders, both will use the 0-point of the rating scale as their reference value, but the meaning will differ. Class-based responders will define it as the average of the selected group, whereas case-based responders will think of this score as describing an individual who are neither well nor badly qualified. As a result, case-based responders will give most shortlisted job seekers positive ratings, just because they are good. On the 0-100 probability scale, their default values might differ. Class-based responders will place the reference point at $1/n$ (25% in the job example) giving a job-seeker the same chance as the others in a set of four. Case-based responders may set their reference point at 50% indicating that this job-seeker has the same chance of getting as not getting the job.

When these two types of responders are provided with a (balanced) set of qualification ratings clustered around 0, as in our Experiment 2, class-based responders will translate them into probabilities around 25%, whereas case-based responders might place them, on average, in the 25%-50% interval (perhaps below 50%, because a score of 0 is not that good, but above 25%, because such a value is for them clearly substandard). Such probabilities will add up to sums in the 100-200% range, that is, they may still produce a considerable number of inflated probability estimates, as indeed we found in Experiment 2 (Table 2). In contrast, case-based participants who are given an (additive) set of probability ratings clustered around 25% will think that most of them are poorly qualified and thus produce qualification ratings below 0. Thus the case-based approach will have different implications for qualification ratings and probability estimates because of scale asymmetries.

In their recent presentation of a common framework for comparative and probability judgments, RDJF (Referent-Dependent Judgment Framework), Smith et al. (2015) distinguish between a more immediate impression phase and a more deliberate and controlled comparison and aggregation phase. Target stimuli are in the impression phase judged as strong or weak based on their individual merits and on spontaneous comparisons with expectations or general standards within the domain. In the more controlled comparison/aggregation stage these judgments are adjusted to satisfy implicit and explicit task requirements (like comparing targets to the group mean, or distributing probabilities between the individual members of the set). Biases can arise in the impression stage according to whether targets are perceived to be generally strong or weak. They can also arise from insufficient adjustments in the controlled comparison/aggregation stage, for instance due to cognitive load or insufficient processing time.

The research discussed in present paper fits well into the RDJF account by confirming the primacy of the impression stage both for comparison judgments and probability estimates. In contrast with the experiments by Smith et al. (2015), we did not manipulate target strength, but made it clear that the selected candidates, by being Oscar nominees or shortlisted job applicants, belonged to superior groups of targets. Nor did our experiments manipulate set size, which in Smith et al.'s studies was shown to affect probability judgments, but not comparative judgments. On the other hand the present studies go beyond those reported by Smith et al. by asking for judgments of the complete set of target items, by using a within-Ss design, and also by asking participants to perform “translations” of comparison judgments into probabilities, and vice versa, and finally by asking them to explain their choices in the aggregation/comparison stage. These results provide further insights into the nature of “comparison and aggregation” biases by highlighting the role played by response scales, as outlined above. It appears that response scales, even those that are explicitly designed as relative, tend to be treated as absolute measures of the

underlying traits. Thus the “neutral” midpoint of a comparative scale is perceived as a partition between strong and weak candidates; similarly, the 50/50 value on a probability scale is taken as a dividing line between “likely” and “unlikely” winners, in an absolute rather than a relative sense.

References

- 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
- 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
- Dougherty, M. R. P., & Hunter, J. (2003a). Probability judgment and subadditivity: The role of working memory capacity and constraining retrieval. *Memory & Cognition*, 31(6), 968-982.
- Dougherty, M. R. P., & Hunter, J. E. (2003b). Hypothesis generation, probability judgment, and individual differences in working memory capacity. *Acta Psychologica*, 113(3), 263-282. doi: 10.1016/s0001-6918(03)00033-7
- Fox, C. R. (1999). Strength of evidence, judged probability, and choice under uncertainty. *Cognitive Psychology*, 38, 167-189.
- 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.
- 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
- Gilboa, I., & Schmeidler, D. (1994). Additive Representations of Non-Additive Measures and the Choquet Integral. *Annals of Operations Research*, 51, 43-65.
- 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
- 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.

- 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.
- 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
- Kardes, F. R. (2013). Selective versus comparative processing. *Journal of Consumer Psychology*, 23(1), 150-153. doi: 10.1016/j.jcps.2012.10.003
- 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., & Ayal, S. (2004). Event frequency and comparative optimism: Another look at the indirect elicitation method of self-others risks. *Journal of Experimental Social Psychology*, 40(6), 805-814. doi: 10.1016/j.jesp.2004.04.006
- Klar, Y., & Levi, U. (2004). Not just a passion for negativity. *Behavioral and Brain Sciences*, 27(3), 349.
- 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.
- 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
- Krueger, J. I., & Funder, D. C. (2004). Towards a balanced social psychology: Causes, consequences, and cures for the problem-seeking approach to social behavior and cognition. *Behavioral and Brain Sciences*, 27(3), 313-+. doi: 10.1017/s0140525x04000081
- McKenzie, C. R. M. (1998). Taking into account the strength of an alternative hypothesis. *Journal of Experimental Psychology-Learning Memory and Cognition*, 24(3), 771-792. doi: 10.1037/0278-7393.24.3.771

- McKenzie, C. R. M. (1999). (Non) complementary updating of belief in two hypotheses. *Memory & Cognition*, 27(1), 152-165. doi: 10.3758/bf03201221
- 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: 10.1016/j.jesp.2009.03.009
- Peters, E., Vastfjall, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological Science*, 17(5), 407-413.
- 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
- Posavac, S. S., Brakus, J. J., Jain, S. P., & Cronley, M. L. (2006). Selective assessment and positivity bias in environmental valuation. *Journal of Experimental Psychology-Applied*, 12(1), 43-49. doi: 10.1037/1076-898x.12.1.43
- Posavac, S. S., Kardes, F. R., & Brakus. (2010). Focus induced tunnel vision in managerial judgment and decision making: The peril and the antidote. *Organizational Behavior and Human Decision Processes*, 113, 102-111.
- Posavac, S. S., Kardes, F. R., & Brakus, J. J. (2010). Focus induced tunnel vision in managerial judgment and decision making: The peril and the antidote. *Organizational Behavior and Human Decision Processes*, 113(2), 102-111. doi: 10.1016/j.obhdp.2010.07.002
- Price, P. C., Pentecost, H. C., & Voth, R. D. (2002). Perceived Event Frequency and the Optimistic Bias: Evidence for a Two-Process Model of Personal Risk Judgments. *Journal of Experimental Social Psychology*, 38(3), 242-252. doi: 10.1006/jesp.2001.1509
- Redelmeier, D., Koehler, D. J., Liberman, V., & Tversky, A. (1995). Probability judgment in medicine: Discounting unspecified alternatives. *Medical Decision Making*, 15, 227-230.
- Riege, A., Sulutvedt, U., & Teigen, K. (2014). Format dependent probabilities: An eye-tracking analysis of additivity neglect. *Polish Psychological Bulletin*, 45(1), 12-20. doi: 10.2478/ppb-2014-0003

- Riege, A. H., Sulutvedt, U., & Teigen, K. (2014). Format dependent probabilities: An eye-tracking analysis of additivity neglect. *Polish Psychological Bulletin*, 45(1), 12-20. doi: 10.2478/ppb-2014-0003
- 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.
- Rose, J. P. (2010). Are direct or indirect measures of comparative risk better predictors of concern and behavioural intentions? *Psychology & Health*, 25(2), 149-165. doi: 10.1080/08870440802340164
- Rosenzweig, S. (1936). Some Implicit Common Factors in Diverse Methods of Psychotherapy. *American Journal of Orthopsychiatry*, 6, 412-415. doi: 10.1111/j.1939-0025.1936.tb05248.x
- 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., 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
- 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
- Smith, A. R., Windschitl, P. D., & Rose, J. P. (2015). *Biases in probability and comparative judgments: Under one theoretical framework*. Manuscript submitted for publication.
- 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. (1974). 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: 10.1016/0001-6918(88)90011-x
- 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
- Wilson, T. D., Wheatley, T., Meyers, J. M., Gilbert, D. T., & Axson, 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., 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., Kruger, J., & Simms, E. N. (2003). The influence of egocentrism and focalism on people's optimism in competitions: When what affects us equally affects me more. *Journal of Personality and Social Psychology*, 85(3), 389-408. doi: 10.1037/0022-3514.85.3.389
- Windschitl, P. D., & Wells, G. L. (1998). The alternative-outcomes effect. *Journal of Personality and Social Psychology*, 75(6), 1411-1423.

Table 1

Means and standard deviations for probability sums (additivity tasks) and mean ratings (NSSB tasks), Experiment 1.

Set	<u>Probability sums, additivity tasks</u>		<u>Mean ratings, NSSB tasks</u>	
	Condition 1	Condition 2	Condition 1	Condition 2
Five actors	156.3 (88.8)	198.4 (115.2)	0.74 (0.93)	0.64 (0.71)
Five actresses	154.4 (83.5)	166.7 (95.9)	0.56 (0.85)	0.50 (0.56)

Table 2

Means and standard deviations for probability sums (additivity tasks) and mean ratings (NSSB tasks) of job applicants, Experiment 2.

Set	<u>Probability sums, additivity tasks</u>		<u>Mean ratings, NSSB tasks</u>	
	Condition 1	Condition 2	Condition 1	Condition 2
Mean	115.60	151.21	-0.38	-0.75
(SD)	(36.16)	(53.69)	(0.61)	(1.15)

Table 3

Mean sums of probability estimates, mean NSSB ratings, and mean number of additive responses (SD) in two conditions, Experiment 3.

	Condition 1:	Condition 2:	<i>t(df)</i>	<i>p</i>
	Self-generated	Scale format		
Number of additive responses (of 3)	1.01 (1.40)	0.24 (0.78)	4.24 (114)	.000
Sums of probability estimates	204.4 (88.0)	261.5 (66.3)	-4.61 (138)	.000
Mean NSSB ratings	1.08 (0.90)	1.05 (0.84)	0.16 (160)	ns

Table 4

Explanations of responses with number of participants in each coding category by task

	Code	Description	Example	Qualification ratings	Probability estimates
1.	Rule-based assessment	Mathematically or logical rule-based decisions.	<p>“because that’s how percentages work”</p> <p>“Due to only one position being available between the candidates.”</p> <p>“Because 100% represents the entire potential for hire and cannot be exceeded. You can’t be 110% chance of a hire.”</p> <p>“Because they were being compared to one another. If all were stellar than they would have all been marked 0 because they were equal to one another”.</p> <p>“Because if they weren’t balanced it would be hard to describe the points of comparison and explain why candidates received the ratings they did.”</p>	19	29
2.	Case-based assessment	People describing that they considered each applicant separately	<p>“I graded each person separately, not part of a sum total.”</p> <p>“I judged them individually against the ideal in the job description.”</p> <p>“Because their chance isn’t related to how much other’s chances were.”</p> <p>“I rated them separately.”</p> <p>“I looked at the ratings one by one and marked them what I felt accordingly.”</p>	24	51
3.	“Wrong” or “incorrect” assessment	People describing that they thought balancing or adhering to 100% rule would have been unfair or an incorrect representation of the candidates’ actual chances or qualifications.	<p>“I felt like it would have been unfair to some of the applicants to rate them so low. It would have made them seem less adequate than they were.”</p> <p>“It was too much mental effort and would have forced me to make decisions that I didn’t necessarily believe were true.”</p> <p>“I tried to rate on a positive scale. I didn’t feel any of them deserved negative ratings”</p> <p>“I looked at them as compared to each other, but also as compared to an ideal.”</p>	42	20
4.	Other	All other reasons	<p>“because it is irrelevant to the qualification evaluation”</p> <p>“I judge people based on job not numbers”</p> <p>“I am lousy at stats. I am no good at all in it.”</p> <p>“I would look at the GPA and the experience and try to give them the benefit of the doubt considering they would make better candidates.”</p> <p>“It was not mentioned in the instructions.”</p> <p>“It just never occurred to me to do that”</p> <p>“I tried to provide an accurate rating”</p> <p>“Sometimes I thought that people with more of a personality would get the job, regardless of their experience, because they had a better interaction with the employers.”</p>	77	62