Original Papers

Polish Psychological Bulletin
2014, vol 45(1), 12-20
DOI - 10.2478/ppb-2014-0003

*Anine H. Riege *,****
*Unni Sulutvedt **
*Karl Halvor Teigen *,***

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

Abstract: *When people are asked to estimate the probabilities of uncertain events, they often neglect the additivity principle, which requires that the probabilities assigned to an exhaustive set of outcomes should add up to 100%. Previous studies indicate that additivity neglect is dependent on response format, self-generated probability estimates being more coherent than estimates on rating scales. The present study made use of eye-tracking methodology, recording the movement, frequency and duration of fixations during the solution of ten additivity problems and two control tasks. Participants produced more non-additive estimates in the Scale format than in the Self-generated format. Self-generated estimates also led to longer decision time and a higher number of repeated inspections, suggesting a deliberate comparison process. In contrast, the Scale format seemed to encourage a case-based approach where each outcome is evaluated in isolation.*

Key words: *Probability judgments, Additivity neglect, Response format, Subadditivity, Eye-tracking.*

When people are asked to estimate the probabilities of uncertain events, they often violate basic assumptions of probability theory. Thus they often fail to respect the additivity principle, which requires that the probabilities assigned to an exhaustive set of mutually exclusive outcomes should add up to 1, or 100%; no more, no less. For instance, if you think that your favorite singer has a 70% chance of winning the European Song Contest, the other participants cannot have more than a 30% chance to share between them, regardless of how many they are and the quality of their performances. Similarly, a player's probability of winning raffle should correspond to the percentage of tickets he or she owns, making up a total of 100% for all players. However, whereas people (students) are relatively good at calculating the probabilities of random events, like the winning probabilities in a raffle, they often neglect the 100% rule in contexts where chances are unevenly distributed and non-random factors play a part. Students who were asked to estimate the winning chances for 20 countries participating in the European Song Contest gave them a mean probability of 27.7%, requiring five and a half rather than just one winner. Only three students out of 31 produced estimates adding up to 100% (Teigen, 1988).
[1] - see page 13

Previous studies of additivity violations have concluded that they tend to be subadditive, i.e., the probability of a complete set of events is often smaller than the probabilities assigned to the individual events in the set. For an exhaustive set, where the total probability is normatively equal to 100%, this total is typically smaller than the sum of its constituent parts, or, put differently, there is a tendency to overestimate the probabilities of individual outcomes. It has been suggested that this occurs because it is easier to recruit supporting evidence for alternatives that are specified compared to unspecified alternatives (Tversky & Koehler, 1994), but subadditive probabilities also occur in situations where all outcomes in the set are estimated in the same session and by the same judge (Fox & Tversky, 1998; Riege & Teigen, 2013; Robinson & Hastie, 1985; Teigen, 1983, 1988). Additivity neglect is most frequent for sets of

* **Department of Psychology, University of Oslo**
** **Simula Research Laboratory, Oslo**
*** **University of Oslo, P.b. 1094 Blindern, NO-0317 Oslo, Norway, tel: +47 22 84 50 56, a.c.riege@psykologi.uio.no**

alternatives greater than two; with only two outcomes, most people will realize that the alternatives are complementary. A "probable" winner cannot at the same time be regarded as a "probable" loser. With more alternatives, the situation grows more complicated, and it becomes less obvious how each alternative should be compared to the rest of the set. One may accordingly have several favourites in a contest like ESC who can all be judged as "probable" winners. This requires that the individual candidates are not primarily viewed as members of a class of outcomes, with an average ignorance prior of p = 1/n, but instead as separate cases, each to be judged on its own. Individual probabilities might be estimated from each event's perceived match with characteristics of the parent population, as suggested by the representativeness heuristic (Kahneman & Tversky, 1972; Teigen, 2004), or from the balance between supporting and non-supporting evidence, as suggested by support theory (Tversky & Koehler, 1994).

To achieve a coherent set of estimates, participants have to adapt their case-based judgments to fit within the 100% frame dictated by probability theory. Such adjustments are sometimes enforced by explicit instructions (Haran, Moore, & Morewedge, 2010), and sometimes encouraged by introducing more subtle "additivity prompts" (Koehler, Brenner, & Tversky, 1997) or "extensional cues" (Kahneman & Tversky, 1996), like presenting the complete set of outcomes jointly to the same participants. However, it can be shown that a joint presentation format reduces, but does not eliminate, additivity neglect (Riege & Teigen, 2013). Moreover, Riege and Teigen found that additivity neglect is format dependent: participants who produced their estimates by writing numbers in an empty slot next to each alternative generated more additive estimates than participants who made their estimates by circling numbers on 0-100% horizontal rating scales. In their final experiment, participants who generated and wrote their own estimates achieved 57% more additive responses compared to those who used rating scales. Written estimates and estimates obtained on rating scales have previously been used interchangeably in probability estimation tasks (often with no information given about how estimates have been obtained). It is accordingly important to establish format dependence as a replicable finding and to investigate more closely the processes responsible for this finding. The present research was conducted to compare probabilities obtained by both methods for a greater set of tasks than those used previously, and under stricter laboratory control. To uncover aspects of the deliberation process during probability estimation we made in the present study use of eye-tracking methodology.

Eye movements during problem solving offer insights into natural shifts in attention and information acquisitions through fixations and saccades (Schulte-Mecklenbeck, Kühberger, & Ranyard, 2011; 1999). Eye movements are naturally occurring behaviour and a generally valid measure of attention, information acquisition and a means to infer cognitive processes (Glaholt, 2011; Russo, 2011; Schulte-Mecklenbeck et al., 2011). Recorded fixations give information about what participants are looking at, and the direction of fixations can indicate their search strategies within a display of information. In addition, fixation durations (in milliseconds) have been used as an indirect measure of cognitive effort, with longer fixations being less common, and reflecting a heavier cognitive load than short fixations (Findlay & Kapoula, 1992; Horstmann, Ahlgrimm, & Glöckner, 2009). Eye-tracking methodology provides a means of investigating repeated inspections of the same material, by counting the number of revisits between (pre)defined Areas of Interest (AOI) on a screen, which in the present study comprises the problem text and the set of individual alternative outcomes. Fixations, inspection time, repeated inspections of outcomes (revisits), and fixation durations are predicted to differ according to 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.

Specifically, the following predictions were made:

1. Participants who are asked to generate their own probability estimates without the aid of rating scales (henceforth: the Self-generated condition) will produce more additive responses than participants who are asked to pick probabilities displayed on horizontal scales (henceforth: the Scale condition), replicating the finding of Riege and Teigen (2013).
2. Participants in the Self-generated condition will use more time per task than participants in the Scale condition, because they have a more complex task to do: Not only do they evaluate the alternatives; they also have to consider the 100% rule by mentally keeping tabs on the sum of their predictions.
3. Participants in the Self-generated condition will have more fixations, partly due to longer deliberation time. This prediction is less obvious than those above, as

[1] A chance to replicate the experiment with experts arose more recently when a Norwegian TV production team decided to create a popular science program on mathematics in everyday life, and as part of this, invited four pop musicians and disk jockeys to listen to 20 songs submitted to the 2009 ESC finals, with the purpose of estimating the winning chances for each song. Their probability assessments were written on a blackboard, where the host of the show, a mathematician, proceeded to add up the numbers in front of the puzzled participants. The sums turned out to be 790%, 560%, 295%, and 975%, for the four individual experts, respectively. The show was repeated before the 2010 ESC finals in Norway, with four new experts (music professionals), who evaluated the chances of 25 participating countries, ending up with sums of 1301%, 186%, 977%, and 676%. (There were, in fact, more than 25 countries participating, so this was not a completely exhaustive set.)

Despite large individual variability, these results indicate a robust additivity neglect in all samples. A class-based approach would have yielded probabilities for individual songs around 5% (in 1986 and 2009) or 4% (in 2010), provided no winning chances for countries outside the list. Instead they obtained average winning probabilities of 27.7% (1986), 32.7% (2009), and 31.3% (2010). The experts were no better than the students, believing that they participated in an experiment about their "scent for music" (rather than their scent for probabilities). Evidently, the winning potential of each song was judged on its own merits, according to a case-based approach, with no consideration given to the number of competitors. (Thanks to producer Petter Nome in Teddy TV and television host Jo Røislien for inviting one of the present authors to the show and sharing the results with us.)

participants in the Scale condition have more to look at, namely the scales, which may by itself increase the number of fixations in this condition.

4. Participants in the Self-generated condition will be more engaged in comparing alternatives, and accordingly have more repeated information inspections (revisits) than participants in the Scale condition.
5. Participants in the Self-generated condition will have longer fixations, reflecting a higher cognitive load.

Method

Participants and design. Thirty students recruited from the Department of Psychology at the University of Oslo participated in the study (23 women and 7 men; median age 22.5 years), without receiving any compensation for their participation. They were randomly assigned to one of two format conditions displayed in Figure 1. Participants in the Scale format condition were presented with a list of probabilities from 0 to 100% lined up to form a horizontal scale, whereas those in the Self-generated probabilities condition had to produce their own estimates, rather than picking the most appropriate number from a list.

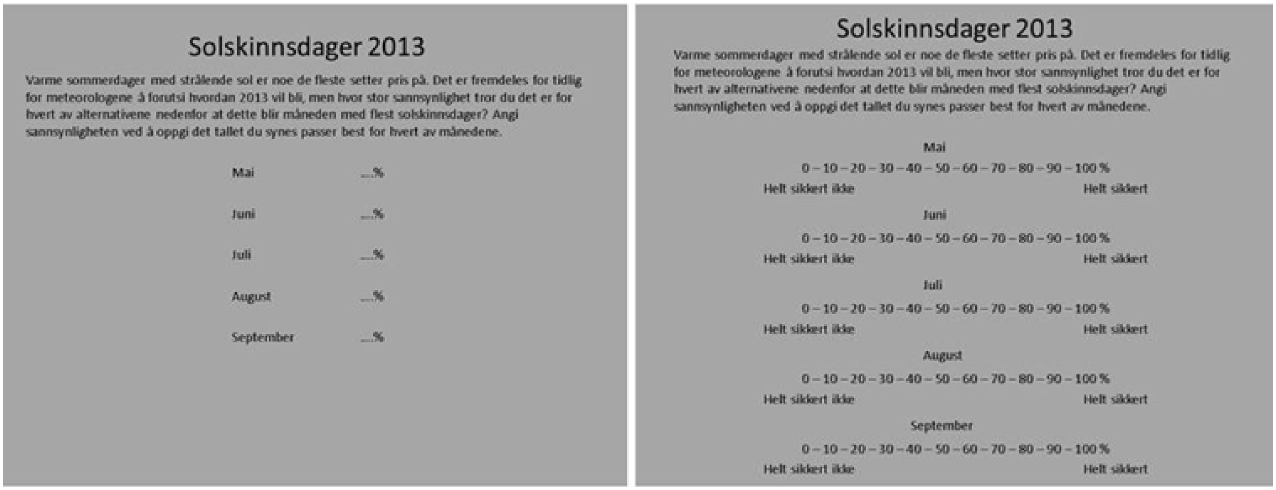
Material. The participants were given 10 experimental and 2 control tasks, presented in random order (the same for all participants). The experimental tasks were probability estimation problems, eight of which were adapted from previous additivity studies. Half of the tasks had four potential outcomes and the other half had five potential outcomes. For example, one vignette asked participants to state their probability estimates of the outcomes of the upcoming (2013) general election in Norway, with four alternatives listed, namely: A social-democratic majority or minority cabinet and a non-socialist majority or minority cabinet (Riege & Teigen, 2013). Other vignettes asked them to predict the probability of five different exam grades for a hypothetical student (Teigen, 1983), four different reasons for a car that won't start (Fischhoff, Slovic, & Lichtenstein, 1978), four different consequences for a patient suffering a heart attack (Redelmeier, Koehler, Lieberman, & Tversky, 1995), and so on. Answers to the experimental tasks had to be based on personal judgment, whereas the two control tasks, which were placed at the end of the session, described random events (the outcomes of dice throws and lottery draws) that could be answered by simple mental calculation. The latter tasks were included to control for potential format differences affecting all responses and not just additivity tasks.

Apparatus and procedure. Each problem was presented on a LCD monitor, with a resolution of 1680 x 1050 pixels (due to technical errors two participants had a resolution of 1440 x 900, but inspection of visual stimuli and data revealed no differences). The stimuli was created using PowerPoint® software, showing text and outcomes on the same screen, including the 0-100% scales in the Scale condition, and blanks in the Self-generated response condition (see Figure 1). Each trial was preceded by a blank screen (1 s), followed by the judgment task. To make responses in the two conditions comparable, participants gave all answers orally, to be written down by the experimenter. However, participants were told that they could ask to have their responses read back as many times as they wanted within each task, but only few participants took advantage of this opportunity. This was done to ease their memory load and to make the tasks as similar as possible to the paper-and-pencil versions studied previously (Riege & Teigen, 2013).

Eye-tracking methodology. Binocular eye-movements were recorded thorugh I-View Software© using the Remote Eye Tracking Device (RED) from SensoMotoric Instruments®, Teltow, Germany. The RED System recorded the eye tracking data at a rate of 60 Hz from a distance of 0.5-1.0 m, with a resolution better than 0.1 degree. Prior to starting the experiment a 9-point calibration procedure was performed.

The number of fixations, revisits (repeated inspections of the same information), and fixation durations were extracted for each participant and for each task, using

Figure 1. The visual display of the two formats, as presented to participants. The question (in Norwegian) in this task is to estimate the probabilities for each of five months to have the highest number of "sunny days", using a Self-generated format (left panel) or a Scale format (right panel).



Solskinnsdager 2013

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 ...%

Solskinnsdager 2013

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
0 – 10 – 20 – 30 – 40 – 50 – 60 – 70 – 80 – 90 – 100 %
Helt sikkert ikke Helt sikkert

Juni
0 – 10 – 20 – 30 – 40 – 50 – 60 – 70 – 80 – 90 – 100 %
Helt sikkert ikke Helt sikkert

Juli
0 – 10 – 20 – 30 – 40 – 50 – 60 – 70 – 80 – 90 – 100 %
Helt sikkert ikke Helt sikkert

August
0 – 10 – 20 – 30 – 40 – 50 – 60 – 70 – 80 – 90 – 100 %
Helt sikkert ikke Helt sikkert

September
0 – 10 – 20 – 30 – 40 – 50 – 60 – 70 – 80 – 90 – 100 %
Helt sikkert ikke Helt sikkert

two predefined sets of non-overlapping areas of interest (AOIs). The first set divided the screen in two major AIOs, one around the text, and another containing all the alternatives. The second set included one AOI for each alternative. We thus had four AOIs for tasks containing four alternatives and five AOIs for tasks with five alternatives. This was done in order to get the number of fixations and revisits for each alternative. For both sets the AOIs containing alternatives were larger in the Scale condition, due to the area requirements of the scales themselves, as shown in Figure 1. If anything, this might allow for more fixations in the Scale conditions (the more there is to look at, the more one looks), working against our prediction of a higher number of fixations in the Self-generated condition.

Results

Probability estimates. Probability sums were calculated for each task by adding the probability estimates for all four or five alternatives. Cronbach's $\alpha = .78$ for tasks with four alternatives, and Cronbach's $\alpha = .92$ for tasks with five alternatives. Sums of 100% were defined as additive. Such responses occurred, as predicted, most frequently in the Self-generated condition. Participants in this condition gave on average 53% additive estimates, against 15% in the Scale condition. All tasks yielded approximately the same number of additive responses. Most non-additive estimates added up to much more than 100%, i.e., they were subadditive, as seen in Table 1. These results replicate previous findings both by showing that the sums of probability estimates increase with number of alternatives, and more additivity neglect in the Scale condition. A 2 x 2 mixed ANOVA of probability sums with number of alternatives (4 vs. 5) as a within-Ss factor and condition as a between-Ss factor revealed significant main effects both of numbers of alternatives, $F(1,28) = 10.86$, $p = .003$, $\eta^2 = .28$, and of condition, $F(1,28) = 16.80$, $p< .001$, $\eta^2 = .375$. There was also a significant interaction effect, $F(1,28) = 8.80$, $p = .006$, $\eta^2 = .239$, indicating that the effect of number of alternatives mainly occurred within the Scale condition. The two control tasks were answered in an additive fashion by most participants in both conditions (90% and 77% additive answers in the lottery and dice vignette, respectively).

Response time. Time per task for participants in the two conditions is presented in the two bottom rows of Table 1. Participants that were asked to produce their own estimates spent, on the average, 36.5% more time on the experimental problems than those who simply picked their numbers from the scales. On the control problems, no such difference was found. This is in line with our hypothesis that self-generated probability estimates are perceived as more demanding, and require more deliberation than estimates performed as ratings on a probability scale.

Fixations. The mean numbers of fixations on the alternatives for both types of tasks are presented in the first two rows in Table 2. From the sheer extension of the spatial layout one would expect a greater number of fixations on the alternatives in the Scale condition than in the Self-generated condition where no scales were displayed. This was indeed the case for the two control problems, which led to nearly twice as many fixations in the Scale condition than in the Self-generated condition. However, for the experimental problems a different pattern emerged, with a higher number of fixations in the Self-generated conditions. A mixed analysis of variance with condition as a between-subjects factor and type of problem (experimental vs. control) as a within-subjects factors confirms a highly significant interaction between these two factors, $F(1, 28)= 13.05$, $p = .001$, $\eta^2 = .32$.

More detailed information can be obtained by the Search Index (SI) originally developed by Payne (1976) for information search in a vertical/horizontal matrix. A search index based on eye-movements is calculated by subtracting vertical from horizontal transitions, divided by the total number of fixations. If participants mainly move their gaze across (within) each alternative, their search index will have a value from 0 to +1. If they mainly move their gaze between different alternatives the search index will take on values between –1 and 0. More extreme values indicate a more dominating search strategy (Franco-Watkins & Johnson, 2011). Search index means were $M_{SI} = .34$ ($SD_{SI} = .13$) in the Self-generated condition against $M_{SI} = .57$ ($SD_{SI}= .09$) in the Scale condition. These means are significantly different from each other, with $t(28) = 5.25$, $p< .001$, Cohen's $d = 2.07$, indicating that participants move their gaze differently in the two conditions, with horizontal movements being especially predominant in the Scale condition.

Table 1. Mean number of additive responses, mean sums of probability estimates, and mean time per task in two conditions (SD in parentheses).

	Self-generated	Scale	*t*(28)	*p*	Cohen's d
Total additive responses (of 10)	5.33 (3.56)	1.47 (2.75)	3.33	<.005	1.22
Probability sums					
Tasks with four outcomes	120.8% (22.6)	155.1% (33.8)	-3.26	<.005	-1.19
Tasks with five outcomes	122.8% (36.0)	192.9% (54.9)	-4.14	<.001	-1.51
Control tasks	98.7% (3.4)	108.3% (19.4)	-1.89	ns	- 0.69
Time per task (sec)					
Additivity problems	75.9s (21.8)	55.6s (16.9)	2.84	<.005	1.04
Control problems	54.8s (13.2)	50.9s (21.4)	0.46	ns	0.22

Table 2. Mean number of fixations and revisits for additivity and control tasks in two conditions (SD in parentheses)

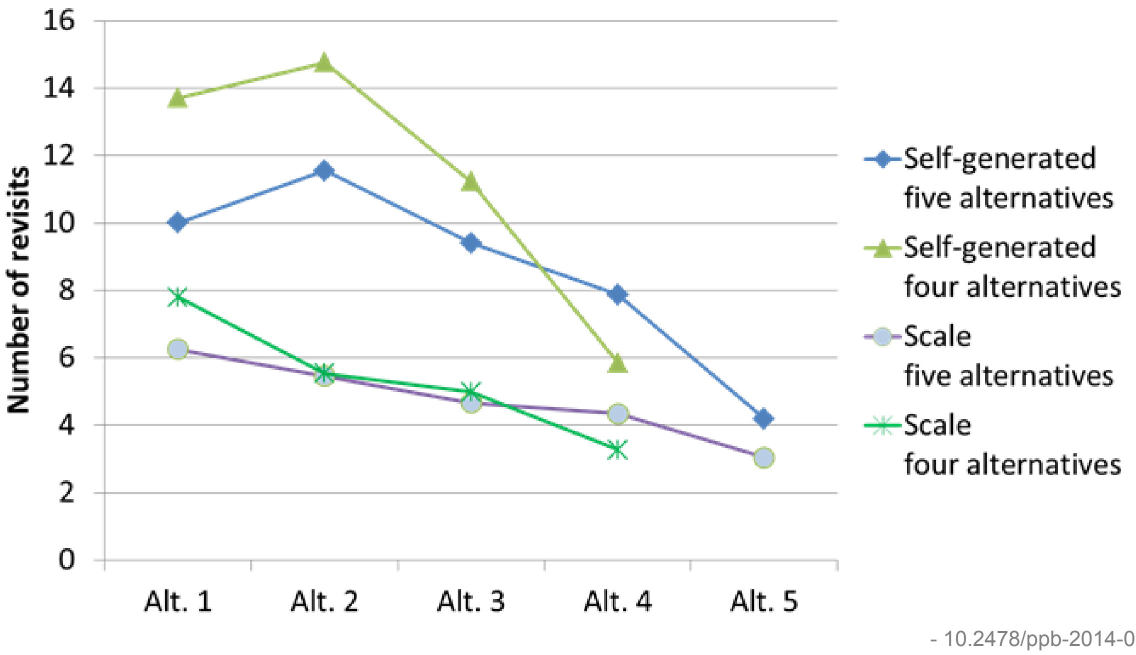
	Self-generated	Scale	*t*(28)	*p*	Cohen's d
Number of fixations					
additivity tasks	130.2 (45.0)	101.9 (31.5)	1.99	.056	0.73
control tasks	40.8 (18.5)	70.7 (53.6)	-2.04	.051	-0.75
Revisits to text					
additivity tasks	8.6 (4.4)	5.6 (1.4)	2.48	.019	0.92
control tasks	9.3 (4.9)	12.9 (5.1)	-1.99	.056	-0.72
Revisits to alternatives					
additivity tasks	44.3 (20.3)	22.7 (11.2)	3.62	.001	1.30
control tasks	15.7 (9.1)	20.6 (18.9)	-0.90	.376	-0.33

The search index is based on relative rather than absolute frequencies, which makes it difficult to tell whether the higher SI values in the Scale condition is a function of fewer comparisons between alternatives, or simply due to a greater number of fixations along the horizontally arranged display of numbers accompanying each alternative. To capture a central aspect of the deliberation process, we counted the number of revisits between different areas of interest. Revisits are repeated inspections of an AOI that do not follow each other in time. Revisits between alternatives and text may indicate a need to check one's understanding of the problem while estimating the probabilities involved. Such revisits were more common in the Self-generated condition than in the Scale condition, but only for the experimental tasks, as shown in the middle two rows of Table 2. Revisits within the set of alternatives are perhaps even more informative, by indicating a comparison of alternatives. As seen in the last two rows of Table 2, the two conditions differed in the number of revisits between alternatives in the experimental, but not in the control tasks. For the additivity tasks, the participants in the Self-generated condition had twice as many revisits than the participants in the Scale condition. In other words, participants were in this condition looking back and forth between the alternatives, indicating a comparison process between alternatives, whereas participants in the Scale condition tended to focus more exclusively on one alternative at the time.

A more detailed picture of the pattern of revisits is given in Figure 2, showing that the number of revisits was consistently higher for all alternatives in the Self-generated condition. The figure also shows that participants' revisits, in both conditions, declined systematically from the alternatives listed first to the alternatives listed last. This is a natural consequence of people's preference to read from top to bottom (Orquin & Mueller-Loose, 2013), and in line with other studies showing that items at the top of a list are given more attention than items at the bottom (Shi, Wedel, & Pieters, 2013; Sütterlin, Brunner, & Opwis, 2008). The line graph shows, in addition, that the tendency of revisiting the first two alternatives is especially strong in the Self-generated condition with four alternatives, which incidentally is the condition with the largest number of additive responses.

Fixation durations have by some investigators been related to cognitive load (Findlay & Kapoula, 1992; Horstmann et al., 2009; Velichkovsky, 1999). Fixation

Figure 2. Number of revisits per alternative in both conditions, for additivity tasks with four and five alternatives.



- 10.2478/ppb-2014-0003

durations were longer for fixations on alternatives than for fixations within the text area. In the Self-generated condition the mean durations were 283.9 ms for the alternatives and 222.8 ms for the text, and in the Scale condition mean fixation durations were 300.1 ms (alternatives) and 220.0 ms (text). A 2 x 2 mixed ANOVA, with alternatives and text as the within-Ss factor and condition as the between-Ss factor, revealed no significant main effect of condition, $F(1,28) = .13$, $p = .73$, $\eta^2 = .005$, nor a significant interaction effect of fixation duration and condition, $F(1,28) =.56$, $p = .46$, $\eta^2 = .019$. However, the difference between viewing the alternatives and reading the text was significant, $F(1,28) = 30.82$, $p< .001$, $\eta^2 = .524$, suggesting a higher cognitive effort while viewing the alternatives than during reading.

Discussion

The results of this study provide strong confirmations of four of the five predictions listed in the introduction.

1. First, we replicated the finding by Riege and Teigen (2013) concerning the effects of response format on additivity neglect. Participants who are asked to assess the probabilities of an exhaustive set of outcomes fail to obtain a coherent set of estimates (i.e., sums of 100%) when using rating scales, but succeed more often to provide additive values when generating their estimates without this "aid". We used in this experiment a larger set of tasks than in previous studies, each with four or five mutually exclusive outcomes. It turned out that participants could not simply be divided dichotomously into a group of "additive" and another group of "non-additive" responders, as only two participants in the Self-generated condition gave additive responses to all ten tasks (and only one gave none). In the Scale condition, in contrast, most responders (11 out of 15) gave non-additive responses to all ten tasks.

We also demonstrated that additivity neglect did not extend to tasks involving chance events (the control tasks), where correct answers could be calculated according to elementary probability rules as the ratio between "favorable" and total number of cases. In the previous study by Riege and Teigen (2013), the two conditions also differed in the way answers were submitted: whereas participants in the Scale condition responded by circling an appropriate number on the scale, they were in the Self-generated condition asked to write or type their estimates on an assigned empty slot, perhaps imparting to them a stronger feeling of *accountability* for their responses, which has in other contexts been shown to attenuate judgmental biases (for an overview, see Lerner & Tetlock, 1999). In the present experiment, answers were conveyed to the experimenter in the same way in both conditions, namely by speaking them out aloud, perhaps strengthening the perceived accountability for participants in both conditions. Other than that, their spontaneous comments were not recorded, but the debriefing session revealed that several participants had been in doubt whether the 100% rule applied or not (even among those in the Scale condition who decided to neglect it).

2. Secondly, we found, as predicted, that responses to the additivity tasks took longer time in the Self-generated condition, despite the fact that the Scale condition provided a richer visual material to watch. This is compatible with Horstmann et al.'s (2009) *integrated process assumption*, which is a dual-process theory postulating common automatic processes that underlie all types of decision making, but that these automatic processes are supplemented with additional deliberate processing steps when needed. Such supplementary adjustments, which here imply a revision of intuitive probabilities to satisfy additivity requirements, will by necessity slow down the time needed for emitting a response.

3. Despite the visual display, which offered participants in the Scale conditions more to look at (cf. Figure 1), participants in the Self-generated conditions had more fixations, especially when fixations to the experimental and control tasks are being compared: Participants in the Self-generated condition had 3.2 times more fixations on the additivity tasks than on the control tasks, whereas participants in the Scale condition had only 1.4 times more fixations on additivity tasks than on the controls.

4. The number of revisits was, as predicted, far greater in the Self-generated than in the Scale condition, indicating that participants in the former condition were more strongly engaged in comparing the alternatives, rather than evaluating each alternative in isolation. The visual layout of the screen could have contributed to this effect by facilitating vertical eye movements in the condition without scales, or discouraging such movements in the condition where alternatives are separated by horizontal rating scales, increasing the distance between the alternatives (as seen in Figure 1). Participants who strive to minimize attention costs (Orquin & Muller-Loose, 2013) will choose to compare pairs of alternatives based on spatial proximity (Russo & Rosen, 1975). It has also been suggested that eye movements can replace working memory as a storage and retrieval system due to the easy way the eyes gather information (Droll & Hayhoe, 2007), perhaps making such comparisons easier in the Self-generated condition.

Comparisons appear to be crucial for achieving an additive response pattern. An analysis across items (the ten additivity tasks) showed that mean number of revisits was consistently larger for participants who produced additive responses than for those that did not. Participants in the Self-generated condition who produced additive sums had on the average 49.7 revisits per task, against 38.3 revisits for those who made non-additive estimates, $t(9) = 3.15$, $p = .012$, Cohen's $d = 2.06$. A parallel difference could be observed in the Scale condition, with a mean of 33.1 revisits per task for those (very few) who produced additive sums, against 20.9 for those who did not, $t(9) = 4.01$, $p = .003$, Cohen's $d = 2.82$.

5. The hypothesis of longer fixation durations in the Self-generated condition was the only of our predictions that was not confirmed. It appears that the percentages of "long" fixations ($\geq$ 500 ms) were in both conditions considerably higher than those reported by Horstmann et al. (2009), with $M_{(Self\text{-}generated)} = 11\%$ and $M_{(Scale)} = 13\%$, against approximately

2% in Horstmann et al.'s city size task. This indicates that the probability estimation tasks in both conditions required considerable cognitive effort. However, fixation durations were more variable in the Self-generated condition, with *SD* = 103.2 vs. *SD* = 41.9 in the Scale condition ($F = 13.33$, $p = .001$), possibly reflecting a relationship between fixation duration and additive response.

To test the potential effects of fixation durations, a multiple regression analysis was conducted with condition (dummy coded: 1 = Self-generated, 2 = Scale), mean revisits, and fixation durations as independent predictor variables and the number of additive responses as the dependent variable. This analysis revealed significant simple effects of condition (Beta = −.368, $t = -2.25$, $p = .033$), mean revisits (Beta = .373, $t = 2.26$, $p = .033$), and mean fixation durations (Beta = .434, $t = 3.16$, $p = .004$). These results indicate that all three predictor variables are related to the number of additive responses. However, these results do not allow any strong inferences to be drawn about the causal relationship between search pattern and additivity. On one hand, it is reasonable to assume that an additive approach requires more mental effort, leading to longer fixation durations. It will also take more time and require repeated inspections of the alternatives, causing many revisits to occur. On the other hand, a search pattern with multiple revisits and comparisons of alternatives will increase the chances of obtaining an additive set of responses. Some studies suggest that that fixation processes could have a causal effect on both the comparison process and decision process (Armel, Beaumel, & Rangel, 2008; Krajbich, Armel, & Rangel, 2010; Shimojo, Simion, Shimojo, & Scheier, 2003). Most likely, the questionnaire format lies at the root of both effects: the Scale format suggests that alternatives might be evaluated individually, according to a case-based approach, and discourages comparisons between alternatives, whereas Self-generated estimates facilitate a class-based approach and make people engage in a more comprehensive search pattern.

The present research was not intended to uncover all factors contributing to additivity neglect. From prior research we know that number of alternatives is inversely related to additivity; the higher the number of alternatives, the smaller the number of participants who manage to distribute probabilities between alternatives in an additive fashion (Teigen, 1983). Numeracy, defined as a person's ability or skill to reason with numbers and mathematical concepts, can also play a role. High numeracy scores have been found to be related to additive responses, but mostly when the numeracy test was given prior to the probability tasks (Riege & Teigen, 2013), prompting participants to apply their mathematical skills to the probability judgments. It has further been suggested that numeracy is related to working memory capacity (Cokely & Kelly, 2009), which has also been claimed to be essential for additive responding (Dougherty & Hunter, 2003). Measures of numeracy and working memory capacity were not included in the present experiment, as our focus was on the effects of presentation mode and response format rather than individual differences.

These results have implications for our understanding of how people make probability estimates. On one hand one could argue that people in real life rarely encounter situations where they are given the full set of outcomes and that this might make tasks such as these slightly unrealistic. At the same time, the simultaneous display of all outcomes should facilitate comparison between alternatives, and might be expected to encourage additive responding. This is probably the first study of additivity where participants are given as many as ten problems in a row, all with a similar structure, which could give them an opportunity to develop a common strategy based on distributional thinking in the course of the experiment. Yet, even participants who occasionally produced additive estimates did not repeat this response pattern consistently on consecutive problems. Estimates of singular events (which are more common outside the laboratory) would probably be even less restricted by distributional considerations, and as a consequence, severely overestimated.

The present research also shows that the way we ask participants to respond can make their estimates more or less consistent with mathematical norms. Asking participants to generate their own probabilities seems to make them think a little harder, as indicated by the eye tracking measures. Format differences might accordingly activate different cognitive processes. The present study could thereby have implications for studies of other probabilistic biases, as for instance overconfidence, base rate neglect, and the conjunction fallacy, where responses are sometimes given as free numerical estimates and sometimes as ratings along a numerical scale. If self-generated numbers require more deliberate thought, the use of this response format might affect or attenuate some of these biases as well.

The present findings may also have relevance for rating scales applied to other domains. Several studies have shown that the way questions are posed, and the way rating scales are constructed, can strongly influence the answers obtained (e.g., Schwartz, 1999). For the past decade it has been increasingly common within this field to use eye-tracking as a means to investigate how people respond to surveys (Galesic, Tourangeau, Couper, & Conrad, 2008; Graesser, Cai, Louwerse, & Daniel, 2006; Redline & Lankford, 2001). The present study adds to this literature by showing that even numeric responses on a 0-100 probability dimension will be answered differently depending upon whether the numbers are picked from a list arranged on a horizontal scale, or produced by the participants themselves. The study also shows that this difference is mirrored in the pattern of information search suggested by these two questionnaire formats.

The extensive use of rating scales in psychology is probably due to a need of standardizing people's judgments of quantities; investigators may also share the implicit belief that predefined scales somehow make the task "easier" for participants. While it may be difficult to come up with numerical estimates reflecting, for instance, a persons' degree of responsibility for a decision or the percentage of one's time spent on work or leisure activities, it might appear easier to emit one's answers by simply circling numbers on an appropriate scale. The present data suggests that this "ease" can have an objective counterpart in shorter response time, fewer fixations, and a smaller number of repeated inspections between alternatives. This ease might come with a cost: participants who spend fewer resources on deliberation and comparison processes, might turn out to be more vulnerable to contextual influences (for instance from superficial features of the response scale). They may also yield less reliable results, and fail to reflect the participant's "true" or carefully considered opinion. An interesting suggestion that might be considered in further studies is whether self-generated estimates can be considered as more cautious and perhaps less extreme than estimates given on rating scales where all values are made equally accessible. Alternatively, responses on a rating scale may be anchored in the middle, apparently "neutral" value, whereas numerical estimates produced without this aid may be more dependent upon one's mental representation of numbers (e.g., from low to high), and of probabilities (e.g., likely versus unlikely).

References

Armel, K. C., Beaumel, A., & Rangel, A. (2008). Biasing simple choices by manipulating relative visual attention. *Judgment and Decision Making, 3(5)*, 396-403.

Cokely, E. T., & Kelley, C. M. (2009). Cognitive abilities and superior decision making under risk: A protocol analysis and process model evaluation. *Judgment and Decision Making, 4(1)*, 20-33.

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

Droll, J. A., & Hayhoe, M. M. (2007). Trade-offs between gaze and working memory use. *Journal of Experimental Psychology-Human Perception and Performance, 33(6)*, 1352-1365. doi: 10.1037/0096-1523.33.6.1352

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,* 330-344.

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

Franco-Watkins, A. M. & Johnson, J. (2011). Applying the decision moving window to risky choice: Comparison of eye-tracking and mouse-tracing methods. *Judgment and Decision Making, 6*, 740-741.

Galesic, M., Tourangeau, R., Couper, M. P., & Conrad, F. G. (2008). Eye-tracking data: New insights on response order effects and other cognitive shortcuts in survey responding. *Public Opinion Quarterly, 72(5),* 892-913. doi: 10.1093/poq/nfn059

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

Graesser, A. C., Cai, Z. Q., Louwerse, M. M., & Daniel, F. (2006). Question Understanding Aid (QUAID): A Web facility that tests question comprehensibility. *Public Opinion Quarterly, 70(1)*, 3-22. doi: 10.1093/poq/nfj012

Haran, U., Moore, D. A., & Morewedge, C. K. (2010). A simple remedy for overprecision in judgment. *Judgment and Decision Making, 5(7),* 467-476.

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

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430–454.

Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review, 103*, 582-591.

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

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience, 13(10)*, 1292–1298.

Lerner, J.S., & Tetlock, P.E. (1999). Accounting for the effects of accountability. *Psychological Bulletin, 125*, 255-275.

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

Payne, J. W. (1976). Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational Behavior and Human Performance, 22*, 17-44.

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

Redline, C. D., & Lankford, C. P. (2001). Eye-movement analysis: A new tool for evaluating the design of visually administered instruments (paper and web). *Proceedings of the Survey Research Methods Section of the American Statistical Association*. Available at www.amstat.org/Sections/Srms/Proceedings/y2001/Proceed/00248.pdf.

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*, 41-52. doi:10.1016/j.obhdp.2012.11.004. [Corrigendum, OBHDP, 121, 278]

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*, 443-456.

Russo, J. E. (2011). Eye fixations as a process trace. In M. Schulte-Mecklenbeck, A. Kühberger & R. Ranyard (Eds.), *A Handbook of Process Tracing Methods for Decision Research* (pp. 43-64). New York: Psychology Press.

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

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

Schulte-Mecklenbeck, M., Kühberger, A., & Ranyard, R. (2011b). *A handbook of process tracing methods for decision making research: A critical review and user's guide.* New York: Psychology Press.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*, 93–105.

Shi, S. W., Wedel, M., & Pieters, F. G. M. (2013). Information acquisition during online decision making: A model-based exploration using eye-tracking data. *Management Science, 59(5)*, 1009-1026. doi:10.1287/mnsc.1120.1625

Shimojo, S., Simion, C., Shimojo, E., & Scheier, C. (2003). Gaze bias both reflects and influences preference. *Nature Neuroscience, 6(12)*, 1317-1322. doi: 10.1038/nn1150

Sütterlin, B., Brunner, T. A., & Opwis, K. (2008). Eye-tracking the cancellation and focus model for preference judgments. *Journal of Experimental Social Psychology, 44(3)*, 904-911. doi: http://dx.doi.org/10.1016/j.jesp.2007.09.003

Teigen, K. H. (1983). Studies in subjective probability III: The unimportance of alternatives. *Scandinavian Journal of Psychology, 24*, 97-105.

Teigen, K. H. (1988). When are low-probability events judged to be "probable"? Effects of outcome-set characteristics on verbal probability judgments. *Acta Psychologica, 67*, 157-174.

Teigen, K. H. (2004). Judgments by representativeness. In R. F. Pohl (Ed.), *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory* (pp. 165-182). Hove, UK: Psychology Press.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101*, 547–567.

Velichkovsky, B. M. (1999). From levels of processing to stratification of cognition: Converging evidence from three domains of research. In B. H. Challis & B. M. Velichovsky (Eds.), *Stratification in cognition and consciousness* (pp. 203-226). Philadelphia, PA: John Benjamins Publishing Company.

Acknowledgements

We would like to thank Bruno Laeng and Siri Leknes for valuable help and insights into the world of eye-tracking. Sulutvedt is supported by a grant from the Ministry of Children, Youths and Families through the Unit of Cognitive Developmental Psychology, Department of Psychology, University of Oslo.