Averaging and Adding in Children's Worth Judgements

Anne Schlottmann[1], Rachel M. Harman & Julie Paine

University College London, UK

Under the normative Expected Value (EV) model, multiple outcomes are additive, but in everyday worth judgement intuitive averaging prevails. Young children also use averaging in EV judgements, leading to a disordinal, crossover violation of utility when children average the part worths of simple gambles involving independent events (Schlottmann, 2000). This study explored the origins of this averaging bias in children's worth judgements, assessing whether averaging also appears for riskless judgements and for other types of risky judgements. In Experiment 1, 8-year-olds judged the worth of having either one or two squares of chocolates in two formally equivalent tasks: Children made additive worth judgements when chocolates varied in size, but used averaging when they varied in winning probability. Performance on the EV task was slightly more advanced when risky followed riskless judgements, with some evidence of transfer. In Experiment 2, 5-year-olds gave additive worth judgements when judging variable fractions of chocolate pies, with displays closely parallel to the spinner discs used for the gambles in Experiment 1. In Experiment 3, 5-year-olds gave additive worth judgements of gambles in which to win either one or two prizes, with alternative rather than independent probabilities of winning. Thus the overgeneralisation of averaging processes to EV judgement, while persistent, neither reflects a general difficulty with additive value judgement, nor with displays showing positive and negative information, nor with risky judgement per se. It may come into play because children have difficulty appreciating the implications of independence, apparent also in other domains. Despite such difficulty, children realize that risky game outcomes go beyond what they can see, and so may apply averaging, as default strategy for population judgement, whereas addition might be the default for judging the sample itself.

A core assumption of Expected Value (EV), dictated by the laws of probability, is that multiple outcomes are additive. Young children, however, violate this assumption, averaging rather than adding the part worth of simple gambles (Schlottmann, 2000). The present study explores the origins of this error, and whether it may be reduced.

From the first year of life, infants are sensitive to probability (Teglas, Girotto, Gonzalez & Bonatti, 2007; Xu & Garcia, 2008) and from 4 years children make judgements of probability and EV that conform to the

[1] Correspondence to the first author at the Division of Psychology & Language Sciences, University College London, Gower Street, WC1E 6BT, Email: a.schlottmann@ucl.ac.uk. We thank staff, children and parents at the participating schools for their kind support.

normative expectations (for review, see Schlottmann & Wilkening, 2011). Studies using Functional Measurement (FM, Anderson, 1981, 1982, 1991, 1996) have shown that children's judgements of how easy it is to randomly draw a winner marble vary with the number of winners and losers on the plate, with a barrel-shaped or fan pattern as predicted by the probability ratio model for various designs (Anderson & Schlottmann, 1991; Acredolo, O'Connor, Banks, & Horobin, 1989; Wilkening & Anderson, 1991). Moreover, children's judgements of how good it is to play a gamble for a prize vary with the likelihood of winning and size of the prize, showing a fan-shaped pattern as predicted by the multiplicative EV model (Bayless & Schlottmann, 2010; Schlottmann, 2001; Schlottmann & Anderson, 1994; Schlottmann & Christoforou, 2005; Schlottmann & Tring, 2005). The traditional view (Hoeman & Ross, 1982; Piaget & Inhelder, 1958, 1975) saw children as non-probabilistic reasoners, and it is true that neither children nor adults are good at explicit reasoning about or computation of probabilities, but the FM data show that even young children have good intuitive probability understanding. Children's intuitions cannot be discounted as non-probabilistic because of their good structural fit with formal probability models.

Against this background it is surprising that children should have difficulty with intuitive EV judgement in situations with multiple outcomes. Under the normative model, EVs of each outcome should simply be added, and adding would seem to be easier than multiplication (Anderson & Cuneo, 1982). However, children average rather than add.

Schlottmann (2000) had children judge how much they would like a game in which they could win a prize if a spinner landed on the winning red segment. Children preferred games with a high probability of winning one prize over 2-prize games to which a second spinner with an additional half chance of winning a second prize had been added, despite the fact that they preferred the two-spinner game when the probability of the first prize was low. This crossover interaction is a qualitative violation of additive utility, under which an additional chance must increase EV; under an adding model, if A+B>A then it follows that C+B>C.

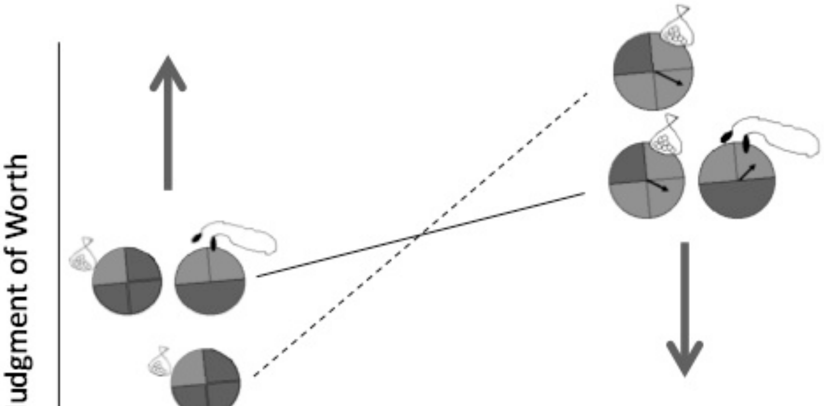

Figure 1. Crossover interaction violates EV additivity. (In Schlottmann (2000), children preferred single spinner games with high probability of

winning the marbles to those with low probability (dashed line). When a second spinner with independent probability of winning the skipping rope was added (solid line), it raised judgement of low probability games, but lowered judgement of high probability games. This is ordinally inconsistent with adding.)

This violation of additivity is not restricted to EV or to children. Butzin and Anderson (1974) found the same crossover, when children judged how much they would like to play with one or two toys of variable attractiveness. Gaeth, Levin, Chakraborty and Levin (1990) found it in adults for evaluation of consumer goods with various tie-in products, Troutman and Shanteau (1976) found it when goods were described by one or two attributes of variable value, and Oden and Anderson (1971) found it in liking for meals. These crossovers are expected under averaging: A medium value pulls up an average of this and a low value, but the same medium value pulls down an average of this and a high value. Hsee's (1998) less-is-better effect may be yet another example. Averaging processes are also pervasive in forming person impressions, which can be seen as a different type of value judgement, both in adults (see review in Anderson, 1996) and children (Hendrick, Franz & Hoving, 1974). Our initial hypothesis was, therefore, that EV averaging appears in overgeneralisation of everyday value judgements.

It is also clear, on the other hand, that even infants have some appreciation of additivity, in the sense that if a quantity/numerosity is added to an existing quantity/numerosity, then infants have a precise expectation of how much more there should be now (Wynn, 1992; Feigenson, Carey & Spelke, 2002). This understanding of additive increases is explicitly linked to the language of numbers from pre-school age (Hughes, 1986). Thus explicit appreciation of additive increases seems to coexist with a tendency to use averaging rather than adding in intuitive judgements.

It might be objected that there is an extensive literature on children's tendency to use additive rather than multiplicative rules for multiplicative concepts, such as area or density (Anderson & Cuneo, 1978; Cuneo, 1982; Wilkening, 1981; Wolf & Algom, 1987; Andrews, Halford, Murphy & Knox, 2009). However, this is a problem of terminology rather than substance: Adding and averaging are both adding-type operations that can be distinguished only under special circumstances, for instance, by comparison of one- and two-value situations, as described above. The studies looking at multiplication concepts have simply not distinguished adding from averaging. Averaging could conceivably be very wide-spread in children's intuitive judgements.

The main question addressed in the present study, accordingly, is whether children can make additive judgements at all, more specifically, whether children would make additive worth judgements at least in riskless situations more similar to situations studied in the literature on pre-schoolers numerical addition (Hughes, 1986), in which the additive increase in the valuable quantity might be more obvious.

To this end, children in Experiment 1 were invited to help a puppet play simple games for one or two chocolate prizes of variable attractiveness. In the riskless size task, children judged how happy the puppet would be when the size of each prize varied. In the risky EV task children judged how happy the puppet would be when the probability of winning each prize varied. The riskless and EV tasks were thus formally identical. If children are capable of additive worth judgements at all, however, we would expect advanced additive performance in the riskless task.

Children made EV judgements either before or after riskless worth judgements. Assuming that children's judgements in the two tasks differ, we can thus also address whether initial experience with additive worth improves later performance in the EV task. Such improvement can only be interpreted, however, if we also know whether initial averaging experience in the EV task impedes subsequent performance in the riskless situation. Transfer only in the normative direction, towards additive EV judgements, would suggest an increase in children's understanding.

EXPERIMENT 1

METHOD

Participants. Thirty-two 8-year-olds (mean age 8 years 9 months, range 8 years 3 months to 9 years 7 months, 14 girls) participated. Children were volunteers of mixed ability from a single year 4 class at a Sussex, UK, primary school (corresponding to US grade 3) with primarily white middle class intake.

Materials. The EV game involved paper spinner discs (13.5 cm diameter) with variable red (win) and blue (lose) segments. There were discs with 7:1 and 1:7 win:lose proportions used during instruction, all red and all blue anchor spinners, and 1:3 2:2 and 3:1 spinners for the experimental stimuli. Discs were placed on a plastic base with a fat grey plastic spinner during instruction, but only the paper discs were used for experimental trials. Pieces of mock chocolate (9 cm square, 1.5 cm deep boxes, covered in silver foil) were the prizes placed by each spinner. The riskless game involved only these mock chocolate squares, with 9 and 2 cm squares used for instruction/anchors, and 3, 5, 7 cm squares used for the experimental stimuli.

The response scale had 17 wooden dowels increasing in 1 cm increments from 2.5 to 18.5 cm height. Children pointed to a stick to indicate how happy a puppet (Lucy Lemur) would be with each game, with bigger sticks for better games. Children have successfully used this scale from 4 years (Anderson & Schlottmann, 1991; Schlottmann, 2001; Schlottmann & Anderson, 1994). Scale usage was elicited in the standard way by instruction with end anchors (Anderson, 1982, chapter 1.)

Design. Each child judged the worth of getting either one or two squares of chocolates in two formally equivalent tasks: In the risky EV task, children judged how happy Lucy would be to play a game for the chocolate(s), with one or two spinners varying in probability of winning it. In the riskless size task, children judged how happy the puppet would to get one or two chocolates varying in size. In each task, children judged two individually randomized replications consisting of a 3x3 within subjects design for the two chocolate games, with small, medium and large probability/size of each prize, interspersed with 3 games involving only a single chocolate (same probabilities/sizes), so there were 24 stimuli in total. Half of the children played the riskless game first, half the EV game.

Procedure. Children were tested individually in a single 20 to 30 minute session at their school. First children met Lucy Lemur who liked chocolate and needed help to get them. If the EV task came first, children first saw a 1:7 red:blue spinner. Children generally knew that it would be easier for this spinner to land on blue, the experimenter (E) confirmed that this was because there was so much blue, and children were shown a spin. The 7:1 spinner was introduced in the same way. Then children were told that "red wins, blue loses" and children were shown the prizes. Seeing one large chocolate they agreed that Lucy would be happy with this, they also agreed she would be even happier with two large chocolates, and when both were removed they thought she would be sad.

To introduce 1- and 2-prize games, children then learned that if a chocolate was placed by a spinner, this meant Lucy would win it if the spinner landed on red. All agreed that Lucy would be happier with a 7:1 spinner game for a chocolate than with the 1:7 game. They were then told that Lucy can sometimes spin one spinner to play for one chocolate prize (one chocolate shown with the 7:1 spinner), but sometimes she gets to spin two spinners at the same time, so that she can win two chocolates (a second chocolate with 1:7 spinner was added). At this stage, 4 children said that the single prize game was better, but were told that 2-prize games are better than 1-prize games, because two prizes are better than one.

The stick scale was introduced, with long sticks for good games, short sticks for bad games, and medium sticks for ok games. Children generally chose the longest stick for Lucy getting two chocolates for sure, a medium stick for a single chocolate, and the shortest stick for no chocolate. Children also generally pointed to the longest stick for a game with 2 all-red discs, and the shortest stick for an all-blue single disc game; all red and all blue anchors were kept beside the corresponding scale ends throughout.

Children next practiced telling Lucy how good each game was. To start, they were reminded that they would see 2-prize games (E points to the long sticks) and 1-prize games (E points to the short sticks). Practice trials consisted of three single prize games presented in order of increasing worth, and three 2-prize games increasing further in worth. If children pointed to the highest stick for the high probability 1-prize game they were reminded that there were also 2-prize games and that two prizes are better than one, so they would run out of sticks if they used the longest stick here. However, children were not shown which stick to use. After the practice, Lucy went for a sleep and children proceeded with the experimental trials, without further feedback. Upon completion of the task, Lucy woke up to admire children's performance, and asked if they could help on a different game. Instruction for the riskless game then proceeded in abbreviated manner

If the riskless game came first, children first saw the different chocolates. All ranked them according to size when asked which Lucy would like best. Children were told that sometimes Lucy could have one chocolate, sometimes two, and sometimes none, and that she would be happy if she won one piece, even happier with two chocolates, but sad if she got none. Children were shown a large chocolate, and then a small chocolate was added; 1 child thought the single chocolate game was better at this stage, and was told again that two chocolates are better than one.

The scale was introduced in the same manner as before, except that 2 large squares of chocolate and an empty chocolate wrapper for no chocolate were used as anchors. Practice and experimental trials followed, as before.

RESULTS AND DISCUSSION

Figure 2 shows children's mean worth judgements for one- and two-prize games in the EV (left panel) and riskless game (right panel). The two rows are for children who saw EV before riskless games (top), and size before EV games (bottom). The data were analysed by mixed model ANOVAs for each task, with chocolate 1 and 2 as within- and task order as between-subjects factor.

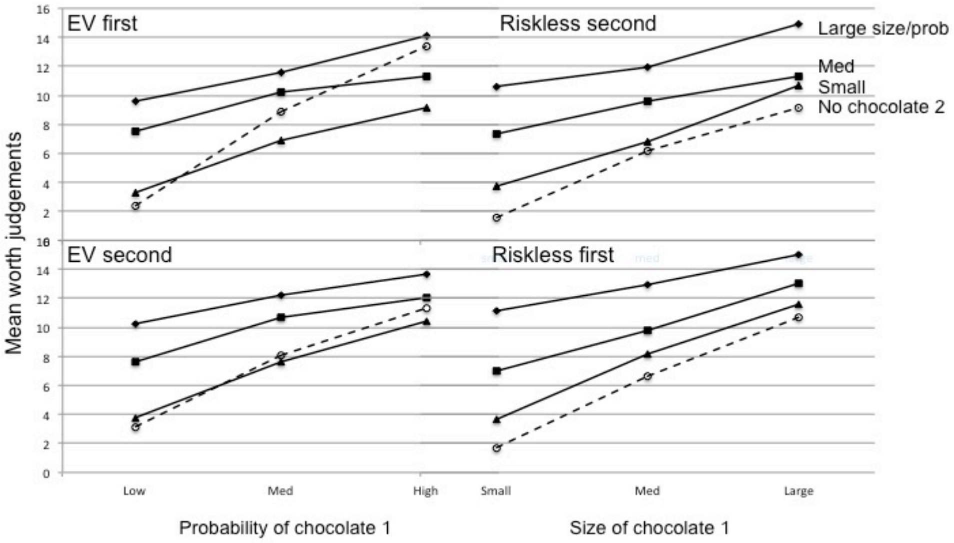

Figure 2. Mean worth judgements for 2-prize games (solid lines) and 1-rize games (dashed line) in the EV (left) and riskless (right) task, as a function of probability/size of chocolate 1 and chocolate 2. (The crossover of the dashed line in the EV task indicates averaging, while in the riskless task the dashed line lies below the solid lines, indicating adding. The averaging crossover is more pronounced in the top panel for children making EV judgements first, but less pronounced when the riskless task came first in the bottom panel, indicating transfer.)

The near parallelism of the solid curves for 2-prize games in all four panels shows that children appropriately considered the worth of both outcomes, with effects for chocolate 1 and 2, F(1,60)>137. (All p<0.05 unless noted.) The slight deviation from parallelism visible in most panels as curves converging towards the right led to significant chocolate 1 x 2 interactions in both tasks, F(4,120)>5.99. This convergence might reflect a slight irregularity in use of the scale, or could indicate minor sub-additivity, but in any event the deviation is small. Task order did not affect judgements for 2-prize games in either EV or riskless task.

Comparing across tasks in Figure 2, both chocolate effects were about 1 point larger in the riskless on the right than in the EV task, F(2,60)>5.30. This is a meaningful task adaptation: The chocolate prize in the EV task has the same size as the large chocolate in the riskless task, but in the EV task it is not won with certainty, only with high probability, leading to slightly lower judgements.

Adding in the Riskless Task. The dashed curve in each panel is for 1-prize games in which only a single chocolate may be won. Normatively, judgements for these games should lie below those for all of the 2-prize games, to which a second chocolate has been added. This pattern appears for the riskless task in the right panels, regardless of whether this task came first or second. Thus, 1-prize games were worth significantly less than 2-prize games with an additional small chance of winning a second chocolate, F(1,30)=61.00, MSE=1.64; interactions with size of the first chocolate and with task order were not significant, F(2,60)=2.39 and F<1.

Averaging in the EV task. In the EV task (left panels), in contrast, the dashed curve crosses over the solid curves, which qualitatively rules out adding and supports averaging. This is best seen the top left panel. Half a chance for the second chocolate raises the judgement when added to a low probability of winning the first chocolate (left point on dashed curve compared to left point on middle solid curve), but lowers the judgement when added to a high chance of winning the first chocolate (right points on the same curves). This is ordinally inconsistent with adding, but predicted by averaging. The crossover was reflected in a significant chocolate 1 x 2 interaction when 1-prize games were compared to 2-prize games with a medium probability of winning the second prize, F(2,60)=43.02, MSE=2.78. These data replicate Schlottmann (2000).

Transfer in the EV task. The 3-way interaction with task order was also significant for the EV task, F(2,60)=4.42, MSE=2.78, as apparent from comparison of the top and bottom left panel. A crossover appears in both panels, but it is more pronounced on the top, when the EV task came first, than on the bottom when it came second. In the bottom left panel, the curve has sunk somewhat, towards its normative lower position, in what may be described as a hybrid pattern between averaging and adding.

This hybrid pattern indicates transfer, with initial experience in the additive riskless task moving subsequent EV judgements slightly closer to the additive model. Note that there was no comparable transfer from EV to riskless task: That children used averaging to make EV judgements first did not move riskless judgements away from additivity, and the 3-way interaction with task order was not significant for the riskless task, F(2,60)=1.29, MSE=1.81. Thus the transfer effect here would seem to indicate learning, not just blind carry-over.

Individual Subjects. We also analysed individual children's data, to check whether the group results were representative of individuals. Of main interest was the question whether the hybrid pattern discussed above appeared for individuals, or whether the group pattern appeared because some children showed an additive, others pure averaging patterns. Because single subject ANOVAs have low power, we used a means-based analysis to classify individuals (Schlottmann, 2001). Children were grouped as showing either averaging (single prize curve reaches/crosses over the medium 2-prize curve at high probability of winning the single prize), or a hybrid pattern (single prize curve crosses over the low, but not medium 2-prize curve at high probability of winning the single prize), or adding (single prize curve no higher than the low 2-prize curve at high probability of winning the single prize). The distribution of individual patterns is in Table 1.

Table 1: Number of children with different response patterns in three experiments

	Averaging	Hybrid	Adding	Other
EXPERIMENT 1				
EV task all	**20**	5	2	5
Riskless task all	4	7	**19**	2
(n=32 8-year-olds per group))				
EV task (first)	**14**	2	-	-
EV task (second)	**6**	3	2	5
Riskless task (first)	3	3	**8**	2
Riskless task (second)	1	4	**11**	-
(n=16 8-year-olds per group)				
EXPERIMENT 2				
Riskless task (mock pies)	1	3	**10**	3
(n=16 5-year-olds)				
EXPERIMENT 3				
EV task (alternative outcomes)	1	-	**7**	2
(n=10 5-year-olds)				

The top rows of Table 1 show that averaging was the majority pattern in the EV task, while in the riskless task addition was most frequent, confirming the group impression. The other categories appeared with similar, lower frequency in each task.

When different task orders are considered, no clear difference appears between the two distributions in the riskless task, but far fewer children averaged when the EV task came second, 88% versus 38%. There was, however, no corresponding clear increase in hybrid or additive patterns, instead the largest increase was in unclassifiable patterns. Thus additive experience had some effect, but was insufficient to induce correct performance. Experience with the additive riskless task may have led to an insight that averaging is not appropriate in the subsequent EV task, rather than to appreciation of the more appropriate additive idea. Learning here may have induced a conflict, rather than provided its resolution.

To conclude, concerning the sources of children's averaging error, the present study clearly shows that averaging is not simply the default strategy reflecting a general difficulty with additive judgements: Children added in the riskless task, yet averaged in the EV task.

It is possible, nevertheless, that addition, while functional in 8-year-olds, is not firmly established at this age, with children falling back on averaging in more difficult circumstances, e.g., when worth judgement is complicated by the presence of uncertainty. In line with this, individuals (when not affected by a preliminary task) almost always used averaging in the EV task, with adding in the riskless task used somewhat less frequently. In light of the possibility that averaging might precede adding developmentally it would seem important to consider younger children's performance.

It is also possible that children's difficulty with the EV task does not relate to EV per se, but rather to the presence of the circular spinner discs used. The spinners present both positive (red wins) and negative (blue loses) information, when the different sized chocolate in the riskless task presented only positive "wins". Averaging might be triggered by the presence of, and perceived need to balance, both positive and negative information, rather than by EV situations.

Experiment 2 therefore tested 5-year-olds in the riskless task only. In contrast to Experiment 1, prizes were not chocolate squares, but round chocolate pies presented in their tins, with some segments visibly missing, which made the pies very similar to the spinner discs used in Experiment 1, presenting both positive and negative information.

EXPERIMENT 2

METHOD

Participants. Sixteen 5-year-olds (mean age 5 years 9 months, range 5 years 5 months to 6 years 3 months, 9 girls) participated. These were year 1 children (corresponding to US Kindergarden) from the same school as in Experiment 1.

Materials. As in Experiment 1, except that prizes here were mock chocolate pies (33.5cm diameter, 1.5cm deep), consisting of silver tins with brown cardboard wedges. Seven pies were used in all, with 7:1, 1:7 (pie pieces:empty slots) used during instruction, with full and empty tins for the anchors, and 1:3, 2:2 and 3:1 pies for the experimental stimuli.

Design and Procedure. Design and procedure were as for the riskless task in Experiment 1, except that Lucy had won a competition in a bakery, so she got to take home any chocolate pie left over, not sold, at the end of the day. The different prizes, scale etc. were introduced as in Experiment 1.

RESULTS AND DISCUSSION

Figure 3 shows children's mean worth judgements of 1 and 2 chocolate prize wins of variable sizes. The solid lines for 2-prize games again show some convergence, i.e., there is possible sub-additivity, but 1-prize games (dashed curve) are judged lower in worth throughout, which is again clear evidence against averaging, for adding.

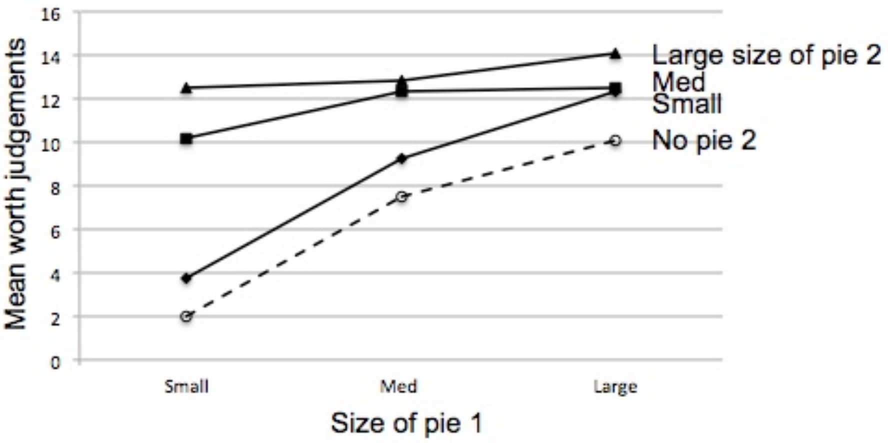

Figure 3. Mean worth judgements for 5-year-olds seeing riskless 2-pie games (solid lines) and 1-pie games (dashed line). (The 1-pie curve lies at the bottom in accord with the additive EV model, but the curve convergence for 2-prize games shows some deviations from additivity.)

In line with this view, the ANOVA for the 2-prize games showed main effects of pie 1 and 2, F(2,30)>104, as well as an interaction, F(4,60)= 22.92, MSE=2.82, reflecting the convergence on the right, possibly due to ceiling effects. When 1-prize games were compared with games in which a small second pie had been added, there were only main effects of pie 1 and 2, F(2,30)>81, with F<1 for the interaction. The group impression was representative of individual children's performance, with the majority (10 of 16 children) showing an adding pattern (Table 1).

In sum, Experiment 2 shows that even young children aged 5 did not average in the riskless task, with judgements clearly reflecting understanding that two prizes are better than one. (These young children might have had a bit more difficulty with use of the response scale, or they may show more subadditivity than the 8-year-olds in Experiment 1, but this was not of major concern here). Averaging is also not linked to some peculiarity of the circular spinner disc stimuli, as adding appeared here with very similar circular stimuli in a riskless task.

EXPERIMENT 3

The question addressed in Experiment 3 is whether averaging appears with any risky EV task. The EV task of Experiment 1 presented two independent events, operationalized with two independent spinners. Thus children could win both, either, or none of the prizes. This, however, is only one possible implementation of a two-outcome situation.

Instead of outcomes being independent, they could be alternatives, i.e., mutually exclusive. This can be operationalized with a single spinner that has 3 different-coloured segments. In 2-prize games, if the spinner lands on one colour, children win one prize, if it lands on the second colour they win a second prize, and if it lands on the 3rd colour they win nothing. If averaging is linked to risky EV judgements generally for children, then we should find the same data pattern here as in Experiment 1.

METHOD

Participants. Ten 5-year-olds (mean age 5 years 7 months) participated. Children were volunteers of mixed ability from a year 1 class at a County Durham, UK, primary school with mainly working class intake.

Materials. As in Experiment 1, except that a small bag of marbles and a skipping rope served as the two prizes. The EV game involved spinner discs with red segments to win the marbles, blue segments to win the rope, and white segments to lose. An all white disc and a half red, half blue disc were used to anchor the scale. Single prize discs with 3:1 and 1:3 red:white and blue:white proportion were also used in the introduction. Experimental 1-prize discs had .125, .25 and .375 probability of the blue outcome, with 1:7, 2:6 and 3:5 (blue:white) proportion (top row in Figure 4). 2-prize discs factorially combined these probabilities with the same probability of the red outcome, yielding discs with 1:1:6, 1:2:5, 1:3:4, 2:1:5, 2:2:4, 2:3:3, 3:1:4, 3:2:3 and 3:3:2 red:blue:white colour proportions (the nine cells in the body of Figure 4)

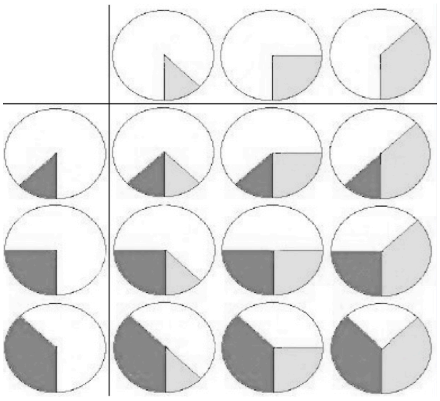

Figure 4. In Experiment 3, 2-prize games with alternative outcomes involved factorial combinations of low, medium or high probability of the red (top row) and blue (left column) outcomes; 1-prize games only had the blue outcome; white was the losing outcome throughout.

Design and Procedure. The design was as for the EV task in Experiment 1 and the procedure was analogous.

RESULTS AND DISCUSSION

Figure 5 shows children's mean judgements in the EV task with alternative outcomes. Children added in this task as well. The data pattern is near parallel for 2-prize games, and the dashed 1-prize curve lies clearly below.

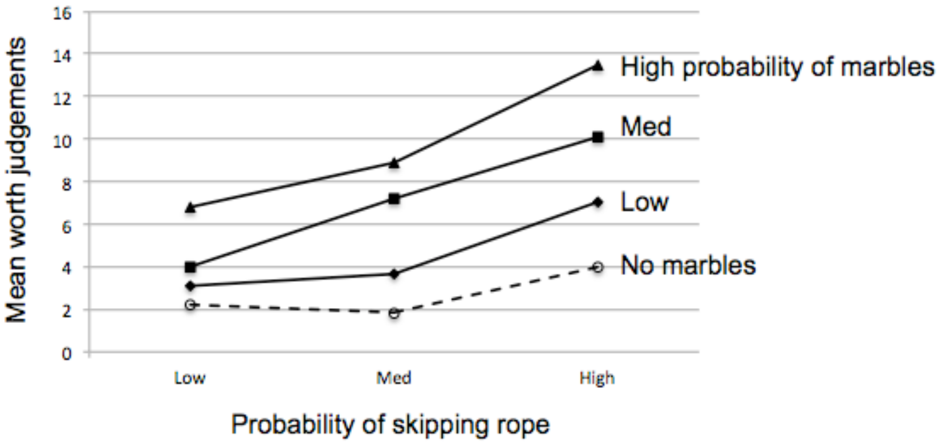

Figure 5. Five-year-olds' mean EV judgements for 2-prize games (solid lines) and 1-prize games (dashed line) in Experiment 3, when outcomes were mutually exclusive, not independent. (Near parallelism of the 4 curves, with the 1-prize curve at the bottom, indicates addition.)

For 2-prize games, the ANOVA showed main effects of the probabilities of both marble and rope outcome, F(2,18)>78. The interaction was significant as well, F(4,36)=5.94, MSE=1.31, however, the curves here show some divergence rather than convergence as in the previous studies. When 1-prize games without the marbles were compared to games to which an additional low chance for the rope outcome had been added, there were only main effects of marble and rope, F>14.69, with F(2,18)=1.77, MSE=3.42, p=.20 for the interaction, reflecting that the dashed 1-prize curve lies clearly below the 2-prize curves, indicating addivity. Table 1 confirms that the majority (7 of 10 children) used adding, so the group data are representative of individuals.

In sum, Experiment 3 shows that averaging is not inevitable in risky judgements either, even for young 5-year-olds.

GENERAL DISCUSSION

In this study, 8-year-olds used averaging to judge the worth of games involving two independent events, replicating Schlottmann (2000). Much younger children, however, made additive worth judgements when chocolate prizes varied in size rather than probability, even with positive-negative displays that were very close to the spinners used in the EV task, and also when two risky outcome were mutually exclusive rather than independent.

Can We Improve EV Additivity in Children? The present study showed that the averaging error is not a passing illusion that is easy to correct: Additive experience improved subsequent EV judgements slightly, but did not make the patterns clearly additive. From the increased variability of performance after additive experience one might speculate that this experience may have brought children to realize that EV averaging was not quite appropriate, but not that addition was the better strategy, highlighting the non-obviousness of the additive idea in the EV task.

That some transfer occurred nevertheless highlights the potential of FM for engineering better understandings. Better transfer tasks might be found with time. The literature on analogical problem solving in young children is rife with examples of children only transferring with surface similarity between teaching and target tasks (Goswami, 1992). From this point of view, more transfer might be expected if the chocolate pie task of Experiment 2, or the alternative probabilities task of Experiment 3 preceded the EV task. Whether increased transfer performance with more similar stimuli would imply increased understanding is, of course, a different question, but perhaps additive experience to dislodge averaging, together with explicit instruction on the relevance of addition might work.

What Is the Origin of EV Averaging? The other main focus here was the possible origin of the averaging error in EV judgement; the present experiments rule out some likely sources. Most importantly, the three studies showed clearly that children are not generally limited to averaging in all their intuitive worth judgements. In two riskless tasks and in the risky task with exclusive outcomes children had no difficulty expressing in their judgements that two prizes are better than one. Importantly, this appeared for children as young as 5 years, so it is unlikely that EV averaging is the residue of a general tendency to average that is pronounced for young children and reduces with age.

The averaging error is also not a specific response to spinner-like stimuli affording an averaging compromise strategy through the presentation of both positive and negative information. This is clear from Experiment 2 with bi-coloured chocolate pies that were visually very similar to the spinners in Experiment 1. A demonstration of averaging with non-spinner displays, e.g., with blind draws from plates of marbles, would further underscore this point, but this is left for future study.

A third possibility was that children have a problem understanding additivity of risky outcomes. However, children added the worth of two alternative risky outcomes in Experiment 3. Note that in Experiment 3 we used two different toy prizes as in Schlottmann (2000), not two identical chocolates as in Experiment 1. This was done to counter the possibility that children might avoid adding component EVs altogether, because children might have recognized the equivalence of the two chocolates and then judged merely EV of a positive outcome of any colour. Children would still have to add the two winning areas, but they would not add component EVs, in other words, instead of

(1) $r = p_r v_r + p_b v_b$,

children's strategy might be better described as

(2) $r = p_{(r+b)}\, v$.

To reduce likelihood of this strategy, Experiment 3 used two different toy prizes, however, a small possibility remains that children ignored these differences and simply judged the probability of winning a prize. A more conclusive test might use two quantitatively very different prizes, making the approach of equation (2) manifestly inappropriate. At present, at any rate, we would tentatively conclude that children seem to have a problem with adding independent risky outcomes, rather than with risky outcomes per se.

Independent risky outcomes might be more difficult for children than alternative risky outcomes, because children need to keep more possibilities in mind than they can see: They may not just win v_1 or v_2, but they could also win both v_1 *and* v_2. Both p_1v_1 and p_2v_2 are visually presented, but p_1p_2 (v_1+v_2) is not. Moreover, since the probabilities of all possible outcomes sum to 1, EV of winning only v_1 is not actually p_1v_1, as visually presented, but $p_1(1-p_2)v_1$, and similarly for v_2. Of course, children need not make the relevant computations, but it is crucial for them to understand that there are three possible non-zero outcomes, not just two, with the third option the intersection of the others.

Difficulty with understanding independent events appears in other research domains as well. In the logical reasoning literature, some argue that children first interpret the connective 'or' as meaning one or the other, but not both, with the inclusive interpretation appearing years later (Braine & Rumain, 1983). The literature on causal reasoning generally shows that children from 2 or 3 years have no difficulty with the idea that if one cause operates, then another does not, while the possibility of two causes operating together seems never considered, with backwards blocking in children's causal attributions (Sobel, Tenenbaum, & Gopnik, 2004), and non-normative causal discounting in conditional reasoning (Ali, Schlottmann, Shaw, Chater & Oaksford, 2010). Crain (2008) argues that the exclusive interpretation of logical 'or' is a pragmatic implicature rather than reflecting a logical limitation, and the same could be true in causal reasoning (Ali et al., 2010), but regardless of this issue, it would seem that thinking about exclusive events may be more familiar or natural for children than thinking about independent events.

That the averaging error here reflects difficulty with the idea of independence could explain why children lack understanding of additivity, but, importantly, does not account for why children misunderstand independent events as involving averaging. More than just lack of understanding, this indicates a bias, and one that is fairly resistant to change. This account therefore still does not go deep enough.

In the introduction, we suggested that EV averaging might be a generalization from everyday value judgement, as demonstrated previously in children and adults, but this hypothesis is not correct: The present experiments clearly show that children can and do make appropriate additive value judgements. So when do children add and when do they average?

In the remainder of this discussion we speculate that adding and averaging originate in two different domains. When judging physical properties, we typically judge what is seen, the sample presented itself, and this may be linked to adding. When judging social/psychological properties in contrast, we typically make a judgement about an underlying property from the sample presented and this may be linked to averaging. Value belongs to neither domain inherently, but can be linked to either, and this may lead to adding or averaging, respectively. To explain how we arrived at this view, we next consider that children not only make averaging errors where adding is appropriate, but also make additive errors where averaging is appropriate.

Interlude: Additive Errors in Intuitive Physics. Physics distinguishes between extensive and intensive properties of a system. Extensive properties depend directly on system size, roughly, concepts of amount, such as volume, mass, number. Intensive properties, in contrast, do not depend on size, and in a homogenous system pertain to every unit of the system; examples are density, speed, temperature, or psychophysical properties such as color or taste intensity. Important for present purposes is that if you combine two systems, then extensive properties combine additively, but intensive properties combine by averaging, e.g., 1 l of cold water and 1 l of hot water gives 2 l (volume is extensive and additive) of warm water (temperature is intensive and combines by averaging).

When material quantities are combined, children make additive judgements of extensive physical properties, shown by the present experiments, or see Anderson and Cuneo (1978, Exp 7). On the other hand, children have great difficulty with intensive properties, documented since Piaget's (1930) studies of flotation and density, in fact, they make additive, extensive errors for intensive properties (Stavy, Strauss, Orpaz & Carmi, 1982; Paik Cho & Go, 2007). Strikingly, up to 10 years children think that when two coloured liquids are combined the resulting colour is darker than the darker of the two, rather than lying in between (Jäger & Wilkening, 2001). Extensive, additive change of physical parameters may be more salient to children than intensive change, because when two quantities are combined, extensive properties always add, with the result "more" relative to either of the source quantities, while intensive change is variable and complex; there could be no change at all (in everyday life one often combines quantities of the same substance, with identical intensive quantities, e.g., filling up a glass of juice), or the change is an increase relative to one quantity and a decrease relative to the other. A child may thus initially equate combination with additive change, and overgeneralize this to intensive properties, leading to what Jäger and Wilkening (2001) call an extensivity bias.

Most important for present purposes, these over-additive errors show that adding processes are as basic in development as averaging. The developmental question of adding versus averaging thus appears not only in worth judgement, but in intuitive judgement more generally. Over-additive errors rule out that EV averaging occurs because the basic intuitive adding-type process is averaging, with adding an idea from numerical cognition that later on also affects intuitive judgement. Rather, both adding and averaging processes appear early, and both are occasionally mis-applied. So again, when do children add, and when do they average?

Physical versus Social Judgement. Our current speculation is that the two intuitions simply arise in two different domains. Adding is a natural response to how-much questions about physical properties, while averaging is a natural response to how-much questions about social/psychological properties. Adding arises, because, as argued above, when material quantities are combined, additive change of extensive properties is highly salient, and much more salient than averaging of intensive quantities.

Averaging, on the other hand, arises when information about social/psychological properties is combined, rather than material quantities. In social/psychological judgement we use the information presented to infer an underlying property, and a good guess at the underlying property is the central tendency, i.e., average, of this information. When combining information in such inferences, the amount of information may change additively, but amount is not what we judge, rather we judge the underlying property, which is typically stable and unchanging, at least within the timeframe of judgement. Our judgements become more extreme with more information, not because the sample amount is changing, nor because the underlying property itself is changing, but because our inference of central tendency becomes more reliable with increased sample size. If, for instance, children judge how likable a person is (Hendrick et al., 1974) or how deserving of Christmas gifts (Schlottmann & Anderson, 1995, 2007), knowing two good things about the person rather than one does not make them twice as good/deserving, but we can be more certain that our judgement is not based on an outlier. Put another way, social judgement is typically a population judgement, an inference from the sample, while physical judgement is typically a judgement of the sample itself.

Value judgement does not belong inherently to either domain and can be constructed as sample or population judgement. In our riskless chocolate case of Experiment 1, what children see is what they get, so they judge how happy they are with the sample itself, the amount of chocolate, producing additive judgements. Note that in the EV case of Experiment 1, children could do the same and additively judge the amount of winning area in the samples, which would produce the correct data pattern. But children understand uncertainty (see review in Schlottmann & Wilkening, 2011), and because they understand that what they see is not what they may get they make an inference, a population judgement, going beyond the information given. They do not make this inference correctly, but the averaging error indicates, paradoxically, that they judge probabilities, not concrete amounts, as also found in other tasks (Schlottmann, 2001). It also shows, importantly, that additivity is not a reflexive response to stimuli with salient extensive properties, but that children can change their frame of reference at will, from judgement of a sample quantity to judgement of the underlying population characteristic.

On face value, our finding of EV additivity with exclusive risky outcomes in Experiment 3 does not fit with this interpretation. However, two arguments we already made reconcile this study with the present view: First, children may have added winning area in Experiment 3, rather than component EVs, i.e., they made extensive judgements. Second, the three possible alternative outcomes were directly given in Experiment 3, in contrast to Experiment 1 with independent events, where the four possible alternative outcome combinations must be inferred. Thus, when alternative outcomes are shown there is less need for inference than in the independent events case. Children might improve with independent events as well, if these were re-presented in terms of the four alternative outcome combinations.

Conclusion. In three experiments, children between 5 and 8 years judged the worth of multiple outcome situations sometimes by addition, sometimes by averaging. Averaging in the independent events situation amounts to a bias that violates the additivity prescription of EV. We speculate that averaging occurs because children construct the judgement as requiring an inference from the sample presented to an underlying population property. Adding occurs, in contrast, when no inference is required and children simply judge the sample itself. Exclusive outcome representations of risky events allow the latter. More generally, the present view predicts that children might be helped to distinguish better between judgements requiring addition or averaging, if intensive physical properties were introduced by reference to 'quasi-intensive' social properties, while additive non-material properties might be introduced by reference to extensive material properties. Many studies have shown that children have functional probability and utility intuitions that structurally correspond to the normative prescriptions (Schlottmann & Wilkening, 2011). The averaging bias in EV judgements highlights one limitation of these intuitions.

REFERENCES

Acredolo, C., O'Connor, J., Banks, L., & Horobin, K. (1989). Children's ability to make probability estimates: Skills revealed through application of Anderson's functional measurement methodology. *Child Development, 60*, 933-945.

Ali, N., Schlottmann, A., Shaw, A., Chater, N., & Oaksford, M. (2010). Causal discounting and conditional reasoning in children. In N. Chater & M. Oaksford (Eds), *Cognition and conditionals: Probability and logic in human thought,* 117-134. Oxford: Oxford University Press.

Anderson, N. H. (1981). *Foundations of information integration theory*. New York: Academic Press.

Anderson, N. H. (1982). *Methods of information integration theory*. New York: Academic Press.

Anderson, N. H. (1991). *Contributions to information integration theory* (Vol. I-III). Hillside, NJ: Erlbaum.

Anderson, N. H. (1996). *A functional theory of cognition*. Mahwah, NJ: Erlbaum.

Anderson, N. H., & Schlottmann, A. (1991). Developmental study of personal probability. In N. H. Anderson (Ed.), *Contributions to information integration theory. Vol. III: Developmental* (pp. 111-134). Hillsdale, NJ: Erlbaum.

Anderson, N. H., & Cuneo, D. O. (1978). The height + width rule in children's judgements of quantity. *Journal of Experimental Psychology: General, 107*, 335-378.

Andrews, G., Halford, G.S., Murphy, K. & Knox, K. (2009). Integration of weight and distance information in young children: The role of relational complexity. *Cognitive Development*, 24(1), 49-60.

Bayless, S. & Schlottmann, A. (2010). Skill-related uncertainty and expected value in 5- and 7-year-olds. *Psicologica. Special Issue on Functional Measurement.* 31(3), 677-687.

Butzin, C. A. & Anderson, N.H. (1974). Functional measurement of children's judgements. *Child Development*, **44**, 529-537.

Braine, M. D. S., and Rumain, B. 1983. Logical reasoning. In J. Flavell and E. Markman (Eds.), *Handbook of child psychology: Vol. 3. Cognitive development*, 261-340. New York: Academic Press.

Crain, S. (2008). The acquisition of disjunction. *Language and Speech*, 51(1&2), 151-169.

Cuneo, D. O. (1982). Children's judgements of numerical quantity: A new view of early quantification. *Cognitive Psychology,* 14, 13-44.

Feigenson, L., Carey, S. & Spelke, E. (2002). Infants' discrimination of number vs. continuous extent. *Cognitive Psychology*, 44(1), 33-66.

Gaeth, G.J., Levin, I.P., Chakraborty, G., & Levin, A.M. (1990). Consumer evaluation of multi-product bundles. *Marketing Letters*, **2**(1), 47-57.

Goswami, U. (1992). *Analogical reasoning in children*. Hove: Lawrence Erlbaum Associates.

Hendrick, C. Franz,C.M., & Hoving, K.L. (1974). How do children form impressions of persons? They average. *Memory and Cognition*, 3(3), 325-328.

Hoemann, H. W., & Ross, B. M. (1982). Children's concepts of chance and probability. In C. J. Brainerd (Ed.), *Children's logical and mathematical cognition* (pp. 93-121). New York: Springer.

Hsee, C.K. (1998). Less is better: When low-value options are valued more highly than high-value options. *Journal of Behavioral Decision Making*, **11**, 107-121.

Hughes, M. (1986). *Children and number: Difficulties in learning mathematics.* Oxford: Blackwell.

Jäger, S. & Wilkening, F. (2001). When light and light make dark: The development of cognitive averaging. *Journal of Experimental Child Psychology*, 70(4), 323-345.

Oden, G. C. & Anderson, N.H. (1971). Differential weighting in integration theory. *Journal of Experimental Psychology,* **89**(1), 152-161.

Paik Cho, S.H. & Go, B.K. (2007) Korean 4- to 11-year-old student conceptions of heat and temperature. *Journal for Research in Science Teaching*, 44: 284–302.

Piaget, J (1930). *The child's conception of physical causality*. M. Gabain, (Transl). London: Routledge & Kegan Paul.

Piaget, J., & Inhelder, B. (1958). *The growth of logical thinking in children and adolescents* (A. Parsons & S. Milgram, Trans.). New York: Basic Books.

Piaget, J., & Inhelder, B. (1975). *The origin of the idea of chance in children* (L. Leake, Jr., P. Burrell & H. Fishbein, Trans.). New York: Norton.

Schlottmann, A. (2000). Children's judgements of gambles: A disordinal violation of additive utility. *Journal of Behavioral Decision Making*, **13**, 77-89.

Schlottmann, A. (2001). Children's probability intuitions: Understanding the expected value of complex gambles. *Child Development, 72*(1), 103-122.

Schlottmann, A., & Anderson, N. H. (1994). Children's judgements of expected value. *Developmental Psychology, 30*(1), 56-66.

Schlottmann, A., & Anderson, N. H. (1995). Belief revision in children: Serial judgement in social cognition and decision-making domains. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **21(5)**, 1349-1364.

Schlottmann, A. & Anderson, N.H. (2007). Belief learning and revision studied with information integration theory. *Teorie & Modelli, Special Issue on Applications of Functional Measurement in Psychology,* **12(1-2),** 63-76.

Schlottmann, A. & Christoforou, M. (2005). *Why are young children so good at expected value judgement?* Symposium on child and adolescent decision-making. Biennial

Meetings of the Society for Research in Child Development. Atlanta, Georgia, USA.

Schlottmann, A., & Tring, J. (2005). How children reason about gains and losses: Framing effects in judgement and choice. *Swiss Journal of Psychology, 64*(3), 153-171.

Schlottmann, A., & Wilkening, F. (2011, in press). Early developments in judgement and decision: Probability and expected value, serial processing, heuristics and biases. In M. Dhami, A. Schlottman & M. Waldmann (Eds.), *Judgement and decision-making as a skill: Learning, development, evolution.* Cambridge: University Press.

Shanteau, J. (1974). Component processes in risky decision making. *Journal of Experimental Psychology*, 103(4), 680-691.

Stavy, R. Strauss, S., Orpaz, N., & Carmi, C. (1982). U-shaped behavioural growth in ratio-comparisons. In S. Strauss and R. Stavy (Eds.), *U-shaped behavioural growth* (pp.11-36). NY: Academic Press.

Sobel, D.M., Tenenbaum, J.S., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science*, *28*, 303-333.

Teglas, E., Girotto, V., Gonzalez, M., & Bonatti, L. L. (2007). Intuitions of probabilities shape expectations about the future at 12 months and beyond. *Proceedings of the National Academy of Sciences of the United States of America, 104*(48), 19156-19159.

Troutman, C.M. & Shanteau, J. (1976). Do consumers evaluate products by adding or averaging attribute information? *Journal of Consumer Reseasrch*, 3(2), 101-106.

Wilkening, F, (1981). Integrating velocity, time and distance information: A developmental study. *Cognitive Psychology*, 13,231-247.

Wilkening, F., & Anderson, N.H. (1991). Representation and diagnosis of knowledge structures in developmental psychology. In N. H. Anderson (Ed.), *Contributions to information integration theory: Vol III Developmental* (pp. 45-80). Hillsdale, NJ: Erlbaum.

Wolf , Y., & Algom, D. (1987). Perceptual and memorial constructs in children's judgements of quantity: A law of across-representation invariance. *Journal of Experimental Psychology: General*, 116(4), 381-397.

Wynn, K. (1992). Addition and subtraction by human infants. *Nature* **358**, 749 – 750.

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences of the United States of America, 105*(13), 5012-5015.

(Manuscript received: XXX; accepted: XXX)