View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Archivio istituzionale della ricerca - Università di Padova

Journal of Cognitive Psychology, 2013
http://dx.doi.org/10.1080/20445911.2013.801485

Routledge
Taylor & Francis Group

The effect of stimulus variability on children's judgements of quantity

Gianmarco Altoè[1] and Franca Agnoli[2]

[1]Department of Pedagogy, Psychology and Philosophy, University of Cagliari, Cagliari, Italy
[2]Department of Developmental Psychology and Socialization, University of Padova, Padova, Italy

This study investigates the effect of stimulus variability on development of the ability to make quantity judgements related to area. Participants were 241 children (aged 4, 5, 6, 8, and 12 years) and 82 university students, who were asked to compare the quantities in 2 sets of 5 chocolate bars of constant width but variable length. Participants indicated which set contained more chocolate or that the amounts of chocolate were equal. Judgement accuracy of 12-year-olds and adults decreased monotonically as the variance of bar lengths increased. In younger children, performance was low when variance was very low or very high, but accuracy was higher for intermediate levels of variance, thus resulting in an inverted U-shaped effect. This pattern was confirmed in a second experiment in which we controlled for a possible age-related response bias against "equal" judgements. Findings suggest that judgements of quantity are based on a mixture of learned heuristics and comparisons of approximate quantity representations, both of which develop throughout childhood.

Keywords: Quantity judgements; Area; Stimulus variability; Cognitive development.

The concept of variability (i.e., dispersion of observed data) has a central role in statistics and in quantitative decisions in everyday life. Understanding variability may improve inductive reasoning, which often involves tasks characterised by the presence of variation in observed data (Nisbett, Krantz, Jepson, & Kunda, 1983). Despite the widespread belief in the importance of this concept, surprisingly little work has focused on how reasoning about variability develops (Garfield & Ben-Zvi, 2005). The current study investigates development of the ability to make quantity judgements in the presence of variability. Specifically, we ask children and adults to compare the quantities represented by 2 sets of 5 vertical bars, while keeping the variability of bar heights constant in 1 set and manipulating variability of the bar heights in the other set.

Research in this area has developed along 2 relatively separate lines of inquiry in developmental psychology and in statistics education. In the field of developmental psychology, there is a long history of research investigating children's ability to make quantity judgements, beginning with Piaget's seminal work on centration (Piaget, 1941/1952) and conservation (Piaget & Inhelder, 1948/1967). According to Piaget, conservation emerges between ages 7 and 8 years, when children reach the concrete operational stage of development. Before this stage, children have

Correspondence should be addressed to Gianmarco Altoè, Dipartimento di Pedagogia, Psicologia e Filosofia, Università di Cagliari, Via Is Mirrionis, 109123 Cagliari, Italy. E-mail: giamma.alto@gmail.com
This article was originally published with errors. This version has been corrected. Please see erratum/corrigendum (http://dx.doi.org/10.1080/02687038.2013.813765).

© 2013 Taylor & Francis

difficulty focusing on more than 1 dimension of a stimulus because their thought is centred. Piaget proposed that centration is a consequence of the absence of mobility that characterises early thought. Later work demonstrated that children younger than 7–8 years are able to solve tasks involving conservation (Field, 1981, 1987), and that the ability to make quantity judgements is more complex and develops more gradually than was previously believed (e.g., Gigerenzer & Richter, 1990). Children's judgements of the quantities represented by areas have been extensively studied for different sized rectangular shapes (e.g., Anderson & Cuneo, 1978; Cuneo, 1980; Gigerenzer & Richter, 1990). Anderson and Cuneo (1978) and Cuneo (1980) found that 3-, 4-, and 5-year-old children's judgements of a rectangle's area were determined by the sum of its length and width, whereas older children's judgements were correctly dependent on the product of its length and width. A decade later, Gigerenzer and Richter (1990) proposed an alternative theoretical approach called Perceptual Constancy (Brunswik, 1934), in which the whole perceived area is a basic theoretical concept. In a series of experiments with children aged between 4 and 9 years and adults, the authors showed that young children's use of the height + width rule was not supported. Based on their findings, the authors proposed a 3-step process model: (1) centring, in which children focus on the longer side of the rectangle (i.e., children pay attention to both dimensions of the rectangle, but tend to overestimate the effect of the longer dimension to formulate their judgements); (2) limited perceptual constancy, in which children's area perception is influenced by shape, but this dependency is related both to the rectangle under judgement and to the series of rectangles presented during the experiment; and (3) perceptual constancy, in which area perception matches physical area, independent of shape. Findings indicated that centring and limited perceptual constancy strongly overlapped in children aged between 4 and 9 years, whereas perceptual constancy was typically observable in adults.

More recently, researchers have demonstrated that humans share a system for approximating the number of objects in a scene (Halberda & Feigenson, 2008; for reviews, see also Cantlon, Platt, & Brannon, 2009; Dehaene, 2009). This sense of number is supported by an internal approximate number system (ANS) that produces a primitive sense of number rapidly and automatically. A central psychophysical characteristic of the ANS is that it produces behaviour following Weber's law: People's ability to discriminate 2 approximate number representations does not depend on the total number of items or the absolute difference between them, but on the ratio between the 2 quantities (Feigenson, Dehaene, & Spelke, 2004). There is increasing evidence that area, like number, relies on an approximate representation system (Lourenco, Bonny, Fernandez, & Rao, 2012), and that the representational precision of this system may improve with age (Cantlon et al., 2009). Lourenco and Longo (2010) provided direct evidence for the presence of shared representations for space, number, and time in preverbal infants, who were able to transfer associative learning across magnitude dimensions. More recently, 3-year-olds have been shown to learn the meaning of the word ''more'' in context of both approximating number and approximating area, suggesting an underlying similarity between these 2 dimensions (Odic, Pietroski, Hunter, Lidz, & Halberda, 2013). In a related study, Odic, Libertus, Feigenson, and Halberda (2012) tested forty 3- to 6-year-old children and adults in both a number and an area discrimination task in which participants selected the greater of 2 quantities across a range of ratios. The authors found that, like number acuity, area acuity steadily improved during childhood.

In the field of statistics education, students' reasoning about variability at different ages has been investigated using a variety of contexts (Garfield & Ben-Zvi, 2008). For example, Lehrer and Schauble (2007) assessed elementary school children's reasoning about variability in 2 contrasting contexts. In the measurement context, participants were asked to measure the heights of several objects. Results showed that children were able to create rudimentary statistical indices of central tendency and spread. In the natural or ''biological'' context (i.e., growth of plants), however, subjects manifested difficulties in handling sources of natural variation and related statistics. Interestingly, the authors found that students' understanding of variability could be improved by specific activities involving explorations of sampling and comparing distributions. Similar findings were reported by Bakker (2004) in a study of adolescents attending Grades 7 and 8. By proposing a number of instructional activities to foster participants' reasoning about key statistical concepts including variability, Bakker found that activities in which students were encouraged to

reason about sampling and shape of data were the most effective. Although students learn relatively easily how to compute formal measures of variability, they rarely understand what these summary statistics represent, either numerically or graphically (Garfield & Ben-Zvi, 2008). Furthermore, students have difficulty recognising the importance of these formal measures and in connecting them with other summary statistics. For instance, DelMas and Liu (2007) demonstrated that even trained college students have strong difficulties relating the concept of variability to the concept of centre and comparing the degree of variability across groups. From an inferential perspective, the understanding of the relationships between variability and the concepts of distribution, estimation, and sampling is one of the most problematic issues in students' statistical reasoning (Cobb, McClain, & Gravemeijer, 2003). Indeed, Obrecht, Chapman, and Suàrez (2010) demonstrated that laypeople appropriately use information about sample variance if it is embedded in a supportive context, but they fail to do so if the context implies inconsistent variance information.

At least 3 critical issues can be identified in the preceding literature review. First, many studies have focused on variability embedded in complex contexts and in connection with other statistical concepts, rather than on variability per se. Much effort has been devoted to understanding variability from an inferential perspective (e.g., sampling; see Reading & Shaughnessy, 2004), whereas the descriptive perspective has been largely neglected despite being the key prerequisite for reasoning correctly about variation. Second, most of the studies concern students' performance and the development of effective training programmes, whereas little is known about the cognitive mechanisms underlying statistical reasoning. An increased integration of the developmental and statistics educational literature is necessary to improve our knowledge about the cognitive processes pertaining to the development of reasoning about variability. Finally, research on young children's first intuitions concerning variation is scarce. A developmental approach may shed light on the origins of the concept of variability as well as its course across different ages.

The current study draws on both the developmental and the statistics education literature to investigate the influence of variability on quantity judgements of children aged between 4 and 12 years and adults. We focus on this age range because prior work suggests that children as young as 4 years are able to perform area quantity judgements through the application of simple rules (Anderson & Cuneo, 1978), and that the ability to use more complex strategies consolidates around the age of 11–12 years (Gigerenzer & Richter, 1990). Given that judgement of area plays an essential role in the development of mathematical concepts in children (e.g., Halberda & Feigenson, 2008), and because researchers in statistics education commonly use tasks involving graphical

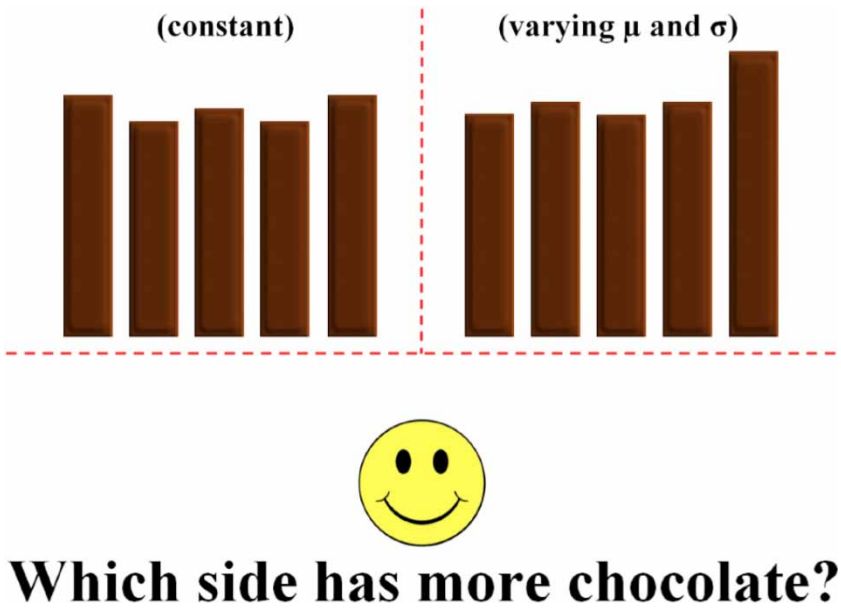

Figure 1. Example stimulus with the constant set on the left and comparison set on the right.

representations of areas, such as histograms (e.g., DelMas & Liu, 2007), we employed an area judgement task to assess participants' performance.

In the experiment, participants are asked to compare the quantities represented by 2 sets of 5 vertical bars such as those shown in Figure 1, while keeping the mean and variance of bar heights constant in 1 set (the constant set) and manipulating the mean and variance of the bar heights in the other set (the comparison set). These 2 sets of bars appear similar to histograms, and the quantity represented by a set of bars is equivalent to the number of observations represented by a histogram. The bars are not histograms, however, because their horizontal positions are irrelevant. Indeed, the vertical bars were described as chocolate, and participants were asked to say which set contained the most chocolate. We use 5 bars in each set because using fewer bars (e.g., 2 or 3) may increase the risk that participants concentrate on a single bar rather than on the overall variability of bars. In addition, results of a pilot study indicated that the proportion of correct responses among the youngest children was significantly greater than chance using 5 bars.

Children should be able to perform this task easily if all the chocolate bars in each set were equally long; they would simply select the set containing longer bars. The task is more difficult when bar lengths are variable; some bars may be longer and some shorter than the bars in the same position in the other set. Indeed, when the variability of bar lengths is very high in the comparison set, it may contain both bars longer and shorter than any in the other set, making the comparison between sets difficult. We predict, therefore, that the accuracy of quantity judgements will decrease as the variability of bar lengths in a comparison set increases. We also predict that quantity judgements will improve with age as children develop the capability to attend selectively only to the bars of unequal length and to combine the results of multiple potentially inconsistent comparisons.

EXPERIMENT 1

Method

Participants. Participants included 241 children and 82 adults residing in Northern Italy. The children (38 4-year-olds, 33 5-year-olds, 39 6-years-olds, 67 8-year-olds, and 64 12-year-olds) were recruited at 2 kindergartens, 4 primary schools, and a middle school. Adults ($M = 23.71$ years, $SD = 2.48$) were undergraduate students majoring in psychology at the University of Padova. We obtained written informed consent from the school principals and both parents as well as verbal assent from all the children.

Materials and design. Materials were constructed for 3 training trials and 15 experimental trials. The stimulus for each trial consisted of 2 sets of 5 bars, as shown in Figure 1. The bar widths were all 2 cm. In the constant set the mean bar length μ was 7.50 cm and the standard deviation σ was .36 cm. The 15 comparison sets were constructed by factorially combining 3 mean bar lengths and 5 standard deviations. The mean bar length in the comparison set was 7.86 cm (i.e., more chocolate), 7.14 cm (i.e., less chocolate), or 7.50 cm (i.e., the same amount of chocolate).

A comparison set was configured by hand for each condition with the added constraint that all bars have lengths between 4.5 and 10.5 cm. Figure 2 shows all 15 stimuli with the constant set on the left and the comparison set on the right. The quantity represented by the comparison set is larger in the first column, smaller in the second column, and equal to the constant set in the third column. The standard deviation of bar lengths in the comparison set was 0 (all equal height) in the first row, .36 cm (the same as the constant set) in the second row, .72 cm (twice the standard deviation of the constant set) in the third row, 1.08 cm (3 times the standard deviation of the constant set) in the fourth row, or 1.55 cm (more than 4 times the standard deviation of the constant set) in the fifth row. The stimuli in the fifth row contain outlier bars, using Tukey's (1977) definition of an outlier as a point that falls more than 1.5 times the interquartile range above the third quartile or below the first quartile of data. Specifically, in Stimulus 13 the comparison set contains a very short bar, in Stimulus 14 a very high bar, and in Stimulus 15 both a very short and a very long bar.

Procedure. Children participated individually during school hours in a quiet room. The children were seated in front of a computer at a comfortable reading distance (approximately 60 cm). They were instructed that they would be comparing the quantity of 2 sets of chocolate bars in a series of graphical representations. Specifically,

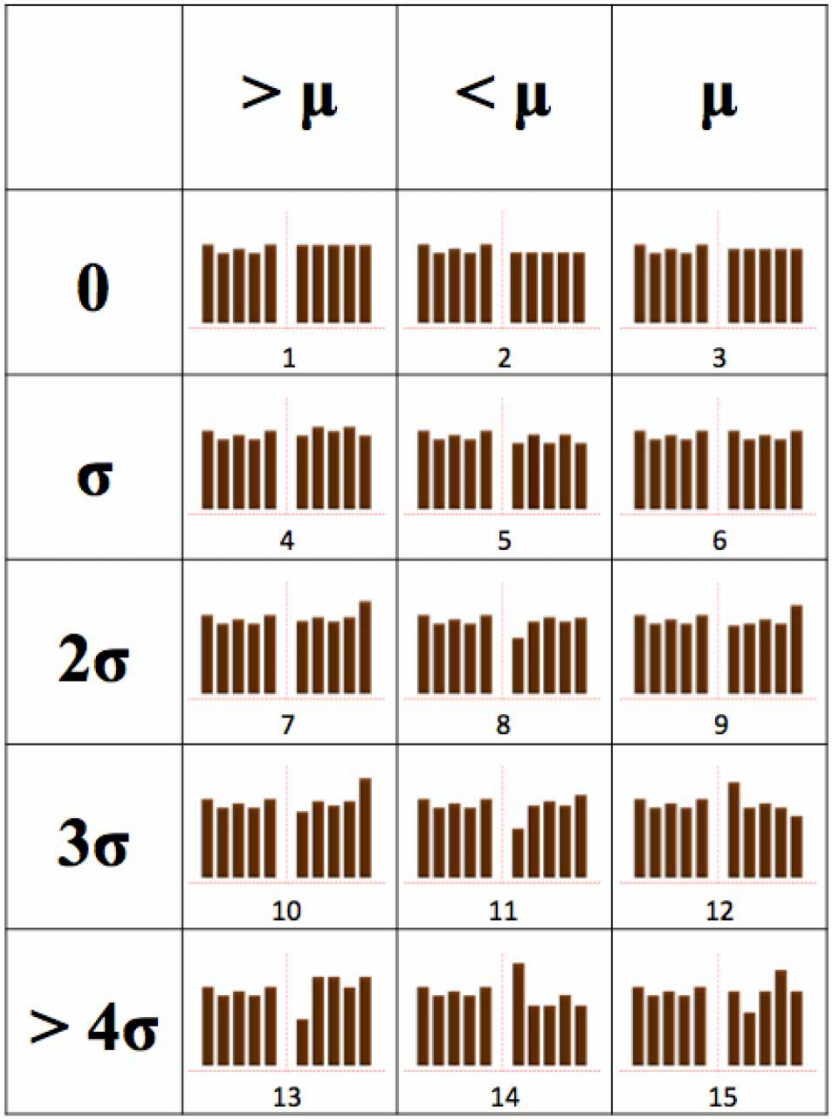

Figure 2. The 15 stimuli with the comparison set on the right. The mean bar length of the constant set was $\mu = 7.50$ cm. The mean bar length of the comparison set was greater than μ in the first column, less than μ in the second column, and equal to μ in the third column. The standard deviation of the constant set was $\sigma = .36$ cm. The standard deviation of the comparison set increases from 0 in the first row to more than 4σ in the fifth row. [To view this figure in colour, please visit the online version of this Journal.]

children were asked "Which side has more chocolate? This side, that side, or are they the same?" Participants were instructed to answer spontaneously, taking as much time as they needed to look at each stimulus. They were also asked not to linger on a specific stimulus to prevent them from giving answers based on the application of complex computing strategies. The children responded orally and by pointing, and their responses were audiotaped and recorded by the experimenter using paper-and-pencil. During the test trials, the experimenter gave only neutral-positive feedback unconnected to the participant's performance. To ensure that participants understood the task, they completed 3 training trials in which the correct responses were left, right, and equal before proceeding to the 15 test trials. During these trials, the experimenter provided indications with regard to response time to discourage responding too rapidly (e.g., without thinking) or too slowly (e.g., trying to measure and remember the height of every single bar). The side of appearance of the constant set and the order of the 15 experimental trials were counterbalanced across participants.

Adults participated in a classroom setting. Each adult was given a booklet containing 3 training trials and 15 experimental trials, 1 per

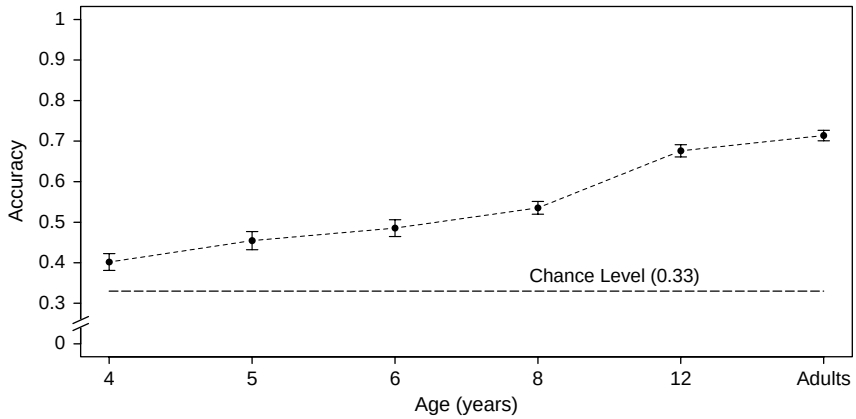

Figure 3. Mean proportion of correct responses by age for Experiment 1. Error bars represent standard errors of the mean ($N = 323$).

page. Four versions of the booklet were created combining 2 different presentation orders of the 15 stimuli and counterbalancing the side on which the constant set appeared. Adults were instructed to solve the 3 training and 15 experimental trials 1 at a time without turning back to prior pages, drawing in the booklet, or using any other strategy to determine the answer. They were asked to simply look at the pictures and to provide an answer without imposing any time limit for each trial. Subjects were also asked not to linger on a specific stimulus to prevent them from giving answers based on the application of complex computing strategies. On each page adults judged which side contained the most chocolate and responded by checking 1 of 3 responses: left side, equal, or right side.

Results

Because the data were repeated measurements of a categorical response, a logistic mixed-effects model approach was used (Baayen, 2008; Jaeger, 2008; Pinheiro & Bates, 2000). In this model the dependent variable was accuracy (correct or incorrect response). The fixed effects were age group (6 levels: 4-, 5-, 6-, 8-, 12-year-olds, and adults), mean chocolate bar lengths in the comparison set (3 levels: more, less, or the same as the constant set), and standard deviation of the chocolate bar lengths in the comparison set (quantitative variable with 5 increasing values). To evaluate the potential quadratic effect of bar length variability on performance, the model included the quadratic term for the standard deviations of chocolate bar lengths. To evaluate whether the effects of the manipulated experimental variables were constant across age groups, all 2-way interactions including age were tested. Subjects were treated as a random effect. To assess the significance of both fixed and random effects, we carried out a series of likelihood ratio tests (LRT) for nested models[1] based on the chi-square distribution (Pinheiro & Bates, 2000). As suggested by Wagenmakers (2007), we also considered the Bayesian Information Criterion (BIC). Because both approaches always yielded the same results, only the results of the LRT are presented. Detailed information on the model, including estimated parameters and odds ratios as a measure of effect size, is reported in Table A1 (available via the supplementary tab on the article's online page at http://dx.doi.org/10.1080/20445911.2013.801485). All analyses were performed using R software (R Core Development Team, 2009).

The mean proportion of correct responses (shown in Figure 3) increased monotonically with age (.40 for 4-year-olds, .45 for 5-year-olds, .49 for 6-year-olds, .54 for 8-year-olds, .68 for 12-year-olds, and .71 for adults) resulting in a significant main effect of age on participants' performance, $\chi^2(4) = 214.09$, $p < .001$. A planned comparison analysis was performed comparing each age group with the adjacent age group (i.e., 4-year-olds vs. 5-year-olds, 5-year-olds vs. 6-year-olds, and so on). Proportion correct was significantly greater for 12-year-olds than 8-year-olds

[1] Each test was performed according to the principle of marginality (Type II tests), i.e., testing each term after all others, except ignoring the term's higher order relatives (for details see Fox, 1997).

($p < .01$) and significantly greater for 8-year-olds than 6-year-olds ($p < .05$).

We also found a significant main effect of mean amount of chocolate, $\chi^2(2) = 713.54$, $p < .001$, with participants performing better when the amount of chocolate in the constant and comparison sets was unequal. The proportion correct was .73 ($SE = .01$) when the comparison set was larger and .68 ($SE = .01$) when the comparison set was smaller. In contrast, the proportion correct was only .33 ($SE = .01$) when the comparison and constant sets contained the same amount of chocolate. As shown in Figure 4, when the 2 sets contained equal amounts of chocolate, errors were strikingly greater for 4-, 5-, 6-, and 8-year-olds compared to 12-year-olds and adults, resulting in a significant interaction of mean chocolate bar length and age, $\chi^2(2) = 191.54$, $p < .001$. Judgements of every age group were much less accurate when the 2 sets represented equal quantities. The mean performance of the 4 youngest groups of children did not achieve 20% correct when the quantities were equal, and 4-year-old children achieved only 2% correct. The overall superior performance of 12-year-olds and adults was primarily due to their judgements of sets representing equal quantities.

As expected, the standard deviation of chocolate bar lengths affected performance (see Figure 5), but the relationship was complex. Both the linear, $\chi^2(1) = 118.77$, $p < .001$, and quadratic, $\chi^2(1) = 35.06$, $p < .001$, terms for the standard deviation were statistically significant. These effects were moderated, however, by age. As predicted, the performance of 12-year-olds and adults decreased as the standard deviation of chocolate bar lengths increased. The performance of 4-, 5-, 6-, and 8-year-old children, however, was highest for intermediate levels of variability, resulting in a surprisingly inverted U-shaped effect of bar length variability on performance. These observations are supported by a significant interaction between both the linear term and age, $\chi^2(5) = 29.23$, $p < .001$, and the quadratic term and age, $\chi^2(5) = 16.37$, $p < .01$.

Stimulus feature analysis. Comparing the quantities represented by 2 sets of bars is difficult. The overall mean percentage correct was only 55%, and even the adults only achieved 71% correct. We know that children and adults can reliably compare the quantities represented by 2 rectangles of equal width and different heights. The current experiment requires comparing 2 sets of 5 rectangles. This comparison could be accomplished by mentally summing or averaging the quantities of all 5 rectangles in each set and then comparing these 2 results. In this case, differences in performance would reflect abilities to integrate the bar lengths and compare the results. Judgements may, however, rely on comparisons of features of the stimuli, such as bars of equal length, exceptionally long bars, and exceptionally short bars. In this case, differences in performance would reflect differences in the features employed in these comparisons. Table 1 presents the responses of each age group to each of the 15 stimuli. We examined the stimuli to assess the performance that would result from applying simple feature comparison strategies and compared it with the observed performance of each age group.

Stimuli 7, 9, 10, 12, 14, and 15 all have 1 bar that is substantially longer than the others. The majority of children in the 4 youngest age groups

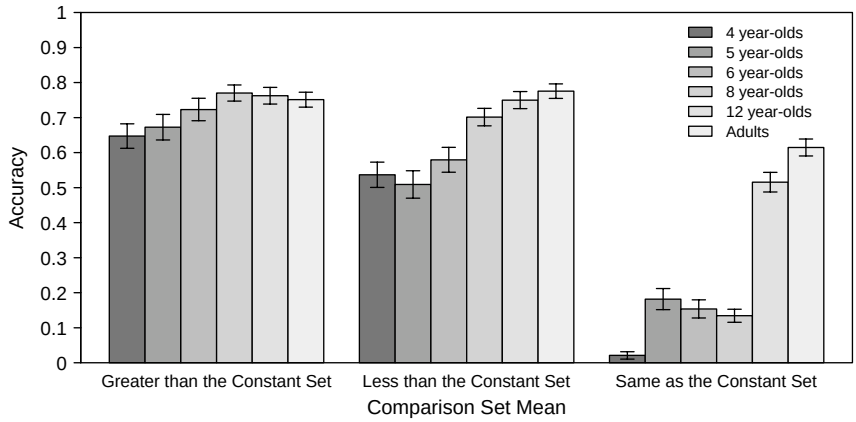

Figure 4. Mean proportion of correct responses by age and mean amount of chocolate in the comparison set for Experiment 1. Error bars represents standard errors of the mean ($N = 323$).

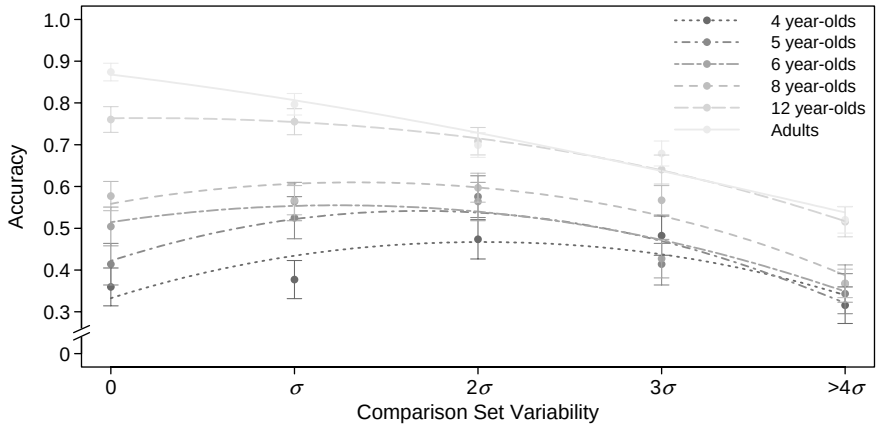

Figure 5. Mean proportion of correct responses by age and stimulus variability for Experiment 1. Error bars represent standard errors of the mean. Lines represent estimated effects of the model ($N = 323$).

consistently selected the set containing this longer bar. Selecting the set with the longest bar yields the correct answer for Stimuli 7 and 10, and the percentage of children who selected this answer was 83% for Stimulus 7 and 77% for Stimulus 10. For the remaining 4 stimuli of this group the set containing the long bar is the incorrect answer, but the percentage of children (excluding the 12-year-olds) who selected it was 67% for Stimulus 14, 72% for Stimulus 9, 74% for Stimulus 12, and 64% for Stimulus 15. There were no systematic age differences in their responses to these stimuli. It appears that many children adopted a strategy of selecting the side with the long bar when there was a single exceptionally long bar. The 12-year-old children and adults, in contrast, were much less likely to base their judgements only on the longest bar. The percentages of 12-year-old children who incorrectly selected the longest bar for Stimuli 9, 12, 14, and 15 were 55%, 52%, 34%, and 30%, respectively. The percentages of adults who incorrectly selected the longest bar for these 4 stimuli were 46%, 40%, 10%, and 18%, respectively.

Of the remaining stimuli, 8, 11, and 13 all have 1 bar that is substantially shorter than the other bars. A possible strategy is to avoid the set with

TABLE 1
Percentage of left (L), equal (E), and right (R) responses for each stimulus and each age group in Experiment 1

		Comparison greater than constant							*Comparison less than constant*							*Comparison equals constant*						
			Age							*Age*							*Age*					
SD		S#	4	5	6	8	12	A	S#	4	5	6	8	12	A	S#	4	5	6	8	12	A
0	L	1	42	36	20	15	5	1	2	*50*	*58*	*74*	*87*	*80*	*83*	3	52	46	44	64	16	10
	E		3	15	18	7	15	6		0	24	13	7	19	16		*3*	*18*	*15*	*9*	*68*	*86*
	R		*55*	*49*	*62*	*78*	*80*	*93*		50	18	13	6	1	1		45	36	41	27	16	4
σ	L	4	50	15	5	8	5	2	5	*68*	*61*	*69*	*73*	*89*	*88*	6	53	27	46	36	16	11
	E		5	21	31	28	25	20		8	18	18	19	9	12		*0*	*33*	*36*	*33*	*67*	*73*
	R		*45*	*64*	*64*	*64*	*70*	*78*		24	21	13	8	2	0		47	40	18	31	17	16
2σ	L	7	21	6	5	12	2	1	8	*58*	*79*	*79*	*87*	*89*	*84*	9	34	12	23	15	8	10
	E		0	15	8	3	12	17		10	15	13	3	9	16		*5*	*15*	*3*	*7*	*37*	*44*
	R		*79*	*79*	*87*	*85*	*86*	*82*		32	6	8	10	2	0		61	73	74	78	55	46
3σ	L	10	21	9	21	13	5	9	11	*63*	*39*	*44*	*79*	*74*	*83*	12	29	15	13	15	12	12
	E		0	15	10	3	14	18		0	15	18	5	13	15		*3*	*9*	*15*	*7*	*36*	*48*
	R		*79*	*76*	*69*	*84*	*81*	*73*		37	46	38	16	13	2		68	76	72	78	52	40
>4σ	L	13	34	12	10	22	16	12	14	*29*	*18*	*23*	*25*	*42*	*50*	15	39	20	21	30	22	26
	E		0	18	10	3	20	38		3	12	10	12	24	40		*0*	*15*	*8*	*10*	*48*	*56*
	R		*66*	*70*	*80*	*75*	*64*	*50*		68	70	67	63	34	10		61	64	72	60	30	18

The standard deviation (*SD*) increases from 0 in the top row to more than 4σ in the bottom row. Correct responses are italicised. For each stimulus, the association between response and age was tested using chi-square. All chi-squares were significant at $p < .01$. The effect size for each test, measured as Cramer's phi, ranged between .19 and .51 (mean = .32; median = .31).

the shortest bar, and performance for Stimuli 8 and 11 suggests that this strategy developed relatively late. For these 2 stimuli, the short bar is in the comparison set and the correct response is the constant set. Averaging over these 2 stimuli, the percentages of correct responses for 4-, 5-, 6-, 8-, and 12-year-old children were 61%, 59%, 62%, 83%, and 82%, respectively, and 84% of adults responded correctly. About 60% of children in the 3 youngest age groups responded correctly, and about 83% of 8-year-olds, 12-year-olds, and adults responded correctly, suggesting that many of the oldest children and adults adopted a similar strategy.

Stimulus 13 is interesting because the comparison set contains a short bar but also contains the most chocolate, so the strategy of avoiding the set with the shortest bar yields the wrong answer. The percentage of correct responses for 4-, 5-, 6-, 8-, and 12-year-old children were 66%, 70%, 80%, 75%, and 64%, respectively, but only 50% of adults responded correctly, resulting in a significant chi-square, $\chi^2(5) = 51.21$, $p < .001$, Cramer's phi $= .28$. This is the only stimulus for which younger children were more accurate than older children and adults. Apparently many adults ruled out the comparison set because of its short bar.

The strongest evidence of age differences was found, surprisingly, in Stimuli 1, 2, and 3, in which all bars of the comparison set were equally long. Averaging across Stimuli 1 and 2, the percentage of correct responses for 4-, 5-, 6-, 8-, and 12-year-olds was 53%, 53%, 68%, 82%, and 80%, respectively. In striking contrast, for Stimulus 3, in which the 2 sets contain the same amount of chocolate, the mean accuracy for the 4 youngest groups was 3%, 18%, 15%, and 9%, respectively. Children in these 4 age groups were unable to recognise reliably that both sets of bars represented equal quantity, whereas 68% of 12-year-old children and 86% of adults responded correctly. Indeed, 4-year-old children almost never selected "equal" as a response; they responded correctly to only 2.1% of the 5 equal stimulus sets. The oldest children, in contrast, responded correctly to 52% of the equal stimulus sets.

EXPERIMENT 2

As expected, the performance of 12-year-old children and adults decreased as stimulus variability increased. Surprisingly, the performance of children ranging in age from 4 to 8 years achieved its maximum at intermediate levels of variability. Their performance was lower for stimuli with both low and high variability. Another striking difference between the performance of younger and older participants was the frequency of their responses that the stimuli were equal. Younger children rarely gave this response even when the stimuli were, in fact, equal. The proportion of correct responses for stimuli with equal quantities was only 12% for children from 4 to 8 years of age but 52% for 12-year-old children and 61% for adults. Apparently children had a response bias against selecting the equal response. Could the observed age differences in the effects of stimulus variability on performance be due to age differences in response bias? There is evidence that the concept "same" develops in childhood, influencing children's understanding of quantity (Cantlon, Fink, Safford, & Brannon, 2007; Cowan, 1991). Hence, the effect of variability on performance observed in our data may partly be influenced by this decision bias. To control for this potentially confounding effect, we ran an additional experiment in which (1) the 5 stimuli with sets having the same quantity were eliminated, and (2) the response option "the same" was omitted. Furthermore, we decided to focus our attention on 6- and 8-year-olds because in the former group, the reversed U-shaped effect of variability was still present, whereas in the latter group this effect tended to disappear.

Method

Participants were 64 children (30 6-year olds and 34 8-year olds) recruited in a primary school following the same informed consent procedure as in Experiment 1. The materials and procedure were the same as in Experiment 1 except that Stimuli 3, 6, 9, 12, and 15 were eliminated and participants were required to indicate which stimulus was larger on each trial.

Results

Because the data were repeated measurements of a binary response, we used a logistic mixed-effects model. The dependent variable was accuracy. The fixed effects were age group (6- and 8-year olds), mean chocolate bar lengths in the comparison set (more or less than the constant set), and standard deviation of chocolate bar lengths in the

comparison set. To evaluate the quadratic effect of bar length variability, the model included the quadratic term for the standard deviations of chocolate bar lengths. To test our main hypothesis that the effect of variability depends on age we evaluated the interactions of age with both the linear and quadratic term of the standard deviation of chocolate bar lengths. Detailed information on the model, including estimated parameters and odds ratios as a measure of effect size, is reported in Table A2 (available via the supplementary tab on the article's online page at http://dx.doi.org/10.1080/20445911.2013.801485).

Mean proportions correct, presented in Figure 6 as a function of age and stimulus variability, were greater than those observed in Experiment 1 for these 2 age groups. This increase in performance is not surprising because in this experiment there were only 2 response alternatives. The 8-year-old children achieved significantly higher performance than the 6-year-olds, $\chi^2(1) = 12.14$, $p < .001$. Performance was also significantly better, $\chi^2(1) = 16.10$, $p < .001$, when the comparison set was larger ($M = .81$, $SE = .02$) than when the constant set was larger ($M = .67$, $SE = .03$). As in Experiment 1, both the linear, $\chi^2(1) = 23.07$, $p < .001$, and quadratic, $\chi^2(1) = 29.63$, $p < .001$, terms of the standard deviation of chocolate bar lengths were significant. The effect of stimulus variability was, however, moderated by age (see Figure 6), as evidenced by a significant interaction of age and the linear term, $\chi^2(1) = 8.89$, $p < .001$.

These results confirm the nonmonotonic effect of stimulus variability on younger children's performance. As in Experiment 1, the strongest evidence of a reversed U-shape effect was found for the youngest children, and this effect attenuates with age.

GENERAL DISCUSSION

Judging the relative sizes of 2 quantities in the presence of variability is a fundamental problem in all the sciences. We rely on the mathematics of statistics and probability in the sciences to guide these judgements, but in both the sciences and in everyday life people may view data representations such as histograms and form opinions about magnitudes.

We investigated development of the ability to make quantity judgements in the presence of variability. Comparing the quantities represented by 2 sets of 5 bars is similar to but simpler than comparing 2 histograms because the positions of the bars are irrelevant in quantity judgements. These judgements were surprisingly difficult even for adults, who responded incorrectly to 29% of the stimuli. Recognising that 2 sets represented equal quantities proved especially difficult. The mean percentage correct when quantities were equal was only 12% for children and 61% for adults. There was a strong bias for finding 1 set to be larger than the other, a bias that is especially strong in young children and slowly weakens with age. If a similar bias occurs with histograms, it could cause people to perceive differences that do not exist.

We expected to find that quantity judgement performance would increase with age and decrease as the variance of bar lengths increases. As expected, we found that mean performance increases monotonically from 4 years to early adulthood; in addition, we found marked inter-individual differences. Some young children performed this task more accurately than some adults. Also as expected, adults' performance decreases as variability of bar lengths increases,

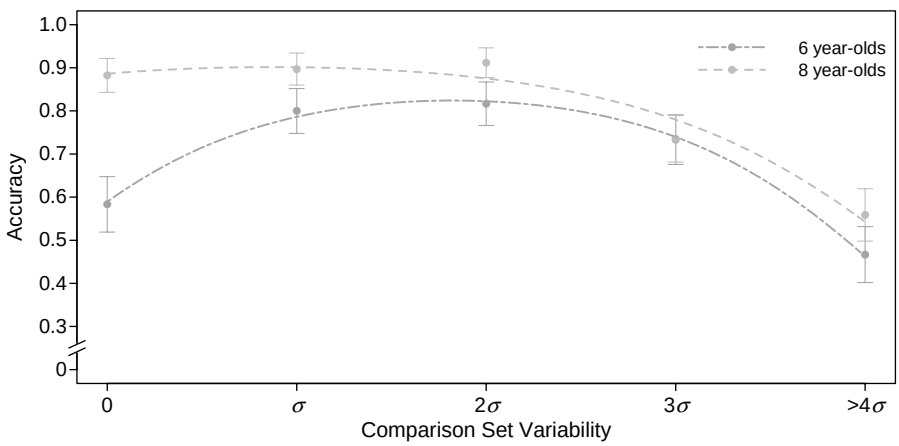

Figure 6. Mean proportion of correct responses by age and stimulus variability for Experiment 2. Error bars represent standard errors of the mean. Lines represent estimated effects of the model ($N = 64$).

but surprisingly, the youngest age groups achieve highest performance for intermediate variability. It was also surprising to find that children of all ages performed better than adults for 1 stimulus (Stimulus 13).

This task requires integrating and comparing the information represented by the lengths of 5 bars. How children and adults perform the task apparently depends on characteristics of the stimuli related to variability. As Figures 5 and 6 show, the effect of age on performance is strongest when all bars are of equal length in the comparison set. For these stimuli the task requires generating and comparing an approximate internal representation of the quantities in the 2 sets. The youngest age groups are able to compare these internal approximations only slightly better than chance, whereas adults achieve about 90% accuracy.

When the lengths of bars in the comparison set vary, the stimuli acquire features such as bars that are noticeably longer or shorter than all others. The results suggest that children and adults employ heuristics based on such features when judging the quantities in some stimuli, and the heuristics employed vary with age. One such heuristic is to choose the comparison set when it contains a bar longer than any in the constant set. Most children aged 4 to 8 years responded to the 6 stimuli containing a long bar in accordance with this heuristic, resulting in a correct response for 2 stimuli and an incorrect response for 4 stimuli. Most of the 12-year-old children and adults successfully avoided the errors that this heuristic produces.

Note that this heuristic also explains the inverted U-shape of performance by the youngest age groups in Figures 5 and 6. Their performance is low when comparison set variability is zero because their ability to approximate and compare these quantities has not fully developed. Their performance increases as variability increases because this heuristic yields the correct answer for 1 of the 2 stimuli with variability of 2 sigma and 3 sigma. For the stimuli with variability greater than 4 sigma this heuristic yields the wrong answer for 1 of the 2 stimuli, depressing performance and yielding the inverted U-shape performance curve.

The role and development of heuristics have been observed in other areas of judgement and decision making in children and adolescents, and in some situations older children and adults who have learned heuristics perform worse than younger children (Furlan, Agnoli, & Reyna, 2012; Jacobs & Potenza, 1991; Reyna & Ellis, 1994). Some older children and adults apparently learned another heuristic applicable in this task: The quantity in a set of bars is smaller when 1 of the bars is shorter than any others. This heuristic yielded a wrong answer for 1 of the 2 stimuli containing a short bar (Stimulus 13), and children of all ages outperformed adults on that stimulus.

These results suggest that the effect of variability on quantity judgement is complex because judgements are based on a mixture of learned heuristics and comparisons of approximate quantity representations, both of which develop throughout childhood. Through experience with various quantities, children learn to approximate and compare quantities more accurately, and they also learn heuristics based on features of quantity sets that emerge as variability increases. Consequently, performance depends on the specific sets presented. Manipulating stimulus characteristics (such as the number of bars and bar lengths) and measuring response times provide avenues for exploring learned heuristics in future research on the effects of variability.

Original manuscript received July 2012
Revised manuscript received March 2013
First published online June 2013

REFERENCES

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

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Bakker, A. (2004). Reasoning about shape as a pattern in variability. *Statistics Education Research Journal*, *3*, 64–83.

Brunswik, E. (1934). *Wahrnehmung und Gegenstandswelt* [*Perception and the world of objects*]. Leipzig: Deuticke.

Cantlon, J., Fink, R., Safford, K., & Brannon, E. M. (2007). Heterogeneity impairs numerical matching but not numerical ordering in preschool children. *Developmental Science*, *10*, 431–440.

Cantlon, J. F., Platt, M. L., & Brannon, E. M. (2009). Beyond the number domain. *Trends in Cognitive Sciences*, *13*, 83–91.

Cobb, P., McClain, K., & Gravemeijer, K. P. E. (2003). Learning about statistical covariation. *Cognition and Instruction*, *21*, 1–78.

Cowan, R. (1991). The same number. In D. Durkin & B. Shire (Eds.), *Language in mathematical education: Research and practice* (pp. 445–464). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cuneo, D. O. (1980). A general strategy for quantity judgments: The height + width rule. *Child Development*, *51*, 299–301.

Dehaene, S. (2009). Origins of mathematical intuitions: The case of arithmetic. *Annals of the New York Academy of Sciences*, *1156*, 232–259.

DelMas, R. C., & Liu, Y. (2007). Students' conceptual understanding of the standard deviation. In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 87–116). New York, NY: Lawrence Erlbaum Associates.

Feigenson, L., Dehaene, S., & Spelke, E. S. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*, 307–314.

Field, D. (1981). Can preschool children really learn to conserve? *Child Development*, *52*, 326–334.

Field, D. (1987). A review of preschool conservation training: An analysis of analyses. *Developmental Review*, *7*, 210–251.

Fox, J. (1997). *Applied regression analysis, linear models, and related methods.* Thousand Oaks, CA: Sage.

Furlan, S., Agnoli, F., & Reyna, V. F. (2012). Children's competence or adults' incompetence: Different developmental trajectories in different tasks. *Developmental Psychology.* Advance online publication. doi:10.1037/a0030509

Garfield, J., & Ben-Zvi, D. (2005). A framework for teaching and assessing reasoning about variability. *Statistics Education Research Journal*, *4*, 92–99.

Garfield, J., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice.* Dordrecht: Springer.

Gigerenzer, G., & Richter, H. R. (1990). Context effects and their interaction with development: Area judgments. *Cognitive Development*, *5*, 235–264.

Halberda, J., & Feigenson, L. (2008). Developmental change in the acuity of the ''number sense'': The approximate number system in 3-, 4-, 5-, and 6-year-olds and adults. *Developmental Psychology*, *44*, 1457–1465.

Jacobs, J. E., & Potenza, M. (1991). The use of judgment heuristics to make social and object decisions: A developmental perspective. *Child Development*, *62*, 166–178.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446.

Lehrer, R., & Schauble, L. (2007). Contrasting emerging conceptions of distribution in contexts of error and natural variation. In M. C. Lovett & P. Shah (Eds.), *Thinking with data* (pp. 149–176). Mahwah, NJ: Lawrence Erlbaum Associates.

Lourenco, S. F., Bonny, J. V., Fernandez, E. P., & Rao, S. (2012). Nonsymbolic number and cumulative area representations contribute shared and unique variance to symbolic math competence. *Proceedings of the National Academy for Science*, *109*, 18737–18742.

Lourenco, S. F., & Longo, M. R. (2010). General magnitude representation in human infants. *Psychological Science*, *21*, 873–881.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, *90*, 339–363.

Obrecht, N. A., Chapman, G. B., & Suàrez, M. T. (2010). Laypeople do use sample variance: The effect of embedding data in a variance-implying story. *Thinking and Reasoning*, *16*, 26–44.

Odic, D., Libertus, M. E., Feigenson, L., & Halberda, J. (2012). Developmental change in the acuity of approximate number and area representations. *Developmental Psychology.* Advance online publication. doi:10.1037/a0029472

Odic, D., Pietroski, P., Hunter, T., Lidz, J., & Halberda, J. (2013). Young children's understanding of ''more'' and discrimination of number and surface area. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 451–461.

Piaget, J. (1952). *The child's conception of number.* London: Routledge & Kegan Paul. (Original work published 1941)

Piaget, J., & Inhelder, B. (1967). *The child's conception of space.* New York, NY: W. W. Norton. (Original work published 1948)

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS.* New York, NY: Springer.

R Development Core Team. (2009). *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing. ISBN 3-900051-07-0. Available from http://www.Rproject.org

Reading, C., & Shaughnessy, M. (2004). Reasoning about variation. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 201–226). Dordrecht: Kluwer Academic.

Reyna, V. F., & Ellis, S. C. (1994). Fuzzy-Trace Theory and framing effects in children's risky decision making. *Psychological Science*, *5*, 275–279.

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison-Wesley.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin and Review*, *14*, 779–804.

APPENDIX A

TABLE A1
Results of the logistic mixed-effects model with accuracy as dependent variable for Experiment 1

Fixed effects	*B*	*SE*	*Odds ratio*	$\chi^2(df)$
Age				214.09*** (5)
5-year-olds	.50	.71	1.64	
6-year-olds	1.45	.70	4.29*	
8-year-olds	1.51	.64	4.53*	
12-year-olds	2.35	.63	10.48***	
Adults	3.40	.64	30.01***	
Comparison set mean				713.54*** (2)
Less than the constant set	−.47	.21	.62*	
Same as the constant set	−4.55	.53	.01***	
Comparison set standard deviation (linear)	1.21	.38	3.36**	118.77*** (1)
Comparison set standard deviation (quadratic)	−.20	.06	.82**	35.06*** (1)
Age × Mean in the comparison set				191.54*** (8)
5-year-olds × Less than the constant set	−.24	.32	.79	
6-year-olds × Less than the constant set	−.20	.31	.82	
8-year-olds × Less than the constant set	.10	.28	1.11	
12-year-olds × Less than the constant set	.41	.28	1.50	
Adults × Less than the constant set	.62	.27	1.87*	
5-year-olds × Same as the constant set	2.25	.60	9.26***	
6-year-olds × Same as the constant set	1.77	.60	5.88**	
8-year-olds × Same as the constant set	1.31	.58	3.71*	
12-year-olds × Same as the constant set	3.38	.56	29.5***	
Adults × Same as the constant set	3.85	.56	47.0***	
Age × Comparison set standard deviation (linear)				29.234*** (5)
5-year-olds × Standard deviation (linear)	−.10	.53	.90	
6-year-olds × Standard deviation (linear)	−.57	.52	.57	
8-year-olds × Standard deviation (linear)	−.33	.48	.72	
12-year-olds × Standard deviation (linear)	−1.02	.47	.36*	
Adults × Standard deviation (linear)	−1.69	.47	.18***	
Age × Comparison set standard deviation (quadratic)				16.37** (5)
5-year-olds × Standard deviation (quadratic)	−.01	.09	.99	
6-year-olds × Standard deviation (quadratic)	.06	.09	1.06	
8-year-olds × Standard deviation (quadratic)	.01	.08	1.01	
12-year-olds × Standard deviation (quadratic)	.12	.08	1.13	
Adults × Standard deviation (quadratic)	.21	.07	1.23**	

Baseline category for age was ''4-year-olds''. Baseline category for comparison set mean was ''greater than the constant set''. For comparison set standard deviation, the degree of the estimated term (linear or quadratic) is reported in parentheses. Random effect was subject. Number of observations = 4845. Number of subjects = 323. *$p < .05$, **$p < .01$, ***$p < .001$.

TABLE A2
Results of the logistic mixed-effects model with accuracy as dependent variable for Experiment 2

Fixed effects	*B*	*SE*	*Odds ratio*	$\chi^2(1)$	*p-value*
Age (8-year-olds)	2.86	.98	17.41	12.14	<.001
Comparison set mean (less than the constant set)	−.81	.20	.45	16.10	<.001
Comparison set standard deviation (linear)	2.08	.48	8.00	23.07	<.001
Comparison set standard deviation (quadratic)	−.37	.08	.69	29.63	<.001
Age × Comparison set standard deviation (linear)	−1.27	.73	.28	8.89	<.01
Age × Comparison set standard deviation (quadratic)	.15	.12	1.17	1.68	.194

Baseline category for age was ''6-year-olds''. Baseline category for comparison set mean was ''greater than the constant set''. For comparison set standard deviation, the degree of the estimated term (linear or quadratic) is reported in parentheses. Random effect was subject. Number of observations = 640. Number of subjects = 64.