Empirical Paper

ijbd

International Journal of Behavioral Development
2023, Vol. 47(3) 253–264

DOI: 10.1177/01650254231160126
journals.sagepub.com/home/jbd
SAGE

Too easy, too hard, or just right: Lifespan age differences and gender effects on task difficulty choices

Sabine Schaefer[1], Michaela Riediger[2], Shu-Chen Li[3] and Ulman Lindenberger[4]

Abstract

In everyday life, individuals often need to make choices about the difficulty level of tasks they wish to perform. Here, we investigate age- and gender-related differences in the monitoring of discrepancies between the difficulty of a given task and one's own performance level, and in the likelihood to select task difficulties that match one's performance level. Male and female children, teenagers, younger adults, and older adults (total $N = 160$) were asked to play a modified version of the BINGO game. Task difficulty was operationalized as the number of cards played simultaneously. We expected that (a) discrepancies between individuals' self-selected difficulty levels and their objectively assessed maximum manageable task difficulty (MMTD) would be lowest in early adulthood; (b) children and teenagers, on average, would select relatively difficult task difficulties; and (c) males would overestimate their performance levels, on average, to a greater extent than females. As predicted, younger adults selected task difficulties closest to their MMTD. All other age groups, including older adults, chose task difficulties above their MMTD. The expected gender differences were restricted to children, with boys showing more pronounced performance overestimations than girls. Children and teenagers fluctuated more in their difficulty choices than adults, and many of them, especially boys, occasionally chose difficulty levels far beyond their performance capabilities. We conclude that task-difficulty choices are an interesting topic for lifespan studies. Future research should systematically vary the physical risk involved in a task, and also include the presence of peers.

Keywords

Children, teenagers, young adults, old adults, task-difficulty, gender differences, risk taking

Task engagement in daily life often comes with the need to perceive and evaluate the relation between how difficult a given task is relative to one's own ability level. Being able to judge how well one can perform a given task and to choose a task difficulty level that corresponds to one's abilities generally leads to better performance outcomes. One example is the selection of a proficiency level for skiing lessons. Choosing a class that is too advanced may result in failure and injuries, but choosing a class that is too easy may result in lack of progress and boredom. Performance and learning levels will be highest if the difficulty level of the class is just right. Another example is the selection of a suitable height in high jump. If the chosen height is too high, the athlete will fail. If the height is too low, mastering it will not reveal the actual performance potential of the athlete, due to a ceiling effect (for an example for ceiling effects in an educational context, see Staus et al., 2021).

To choose a suitable level of task-difficulty, one needs to know one's actual performance level in a given task. Metacognition improves with age throughout childhood and adolescence (e.g., Destan & Roebers, 2015; Forsberg et al., 2021), and children become increasingly able to select suitable task-difficulty levels (Niebaum & Munakata, 2020) or adjust their study times in cognitive tasks (Dufresne & Kobasigawa, 1989). Depending on the context, older adults have shown overconfidence (Crawford & Stankov, 1996; Shing et al., 2009), but also underconfidence (Hertzog & Touron, 2011) in cognitive tasks.

In most research paradigms on metacognition, subjects judge their learning or predict the outcome of their cognitive operations, but they do not choose the difficulty level of a task themselves. Therefore, individual or developmental differences in the discrepancy of performance and task-difficulty selection cannot be detected. We argue that over- or underestimations of one's performance potential and resulting miscalibrations of task-difficulty choices are an interesting research question. Choosing

[1]Saarland University, Germany
[2]Friedrich Schiller University Jena, Germany
[3]TU Dresden, Germany
[4]Max Planck Institute for Human Development, Germany

Corresponding author:
Sabine Schaefer, Institute of Sport Sciences, Saarland University, Campus Building B 8.1, 66123 Saarbrücken, Germany.
Email: sabine.schaefer@uni-saarland.de

unsuitable task-difficulty levels is likely to result in suboptimal performance outcomes. Overestimations lead to failure, and underestimations keep individuals from showing their full performance potential. In addition, actual performance of the task was rarely measured repeatedly in previous research, such that a trial-by-trial tuning of metajudgements could not be addressed (for an example, see Forsberg et al., 2021).

In the current study, we investigated age differences in metacognition from a lifespan perspective by asking participants to actively choose a suitable level of task-difficulty. Specifically, we examined differences in (a) *monitoring* the discrepancies between task difficulty and own task performance and (b) *selecting* levels of task difficulty that correspond to one's task performance. We compared participants of different ages and genders in this respect. Informed by the work of Brim (1992), we proposed the concept of *selection margins* to characterize age-graded differences in difficulty monitoring and difficulty selection (see also Riediger et al., 2006). Selection margins are defined as the discrepancy between the task difficulty that an individual can maximally manage (*maximum manageable task difficulty* [MMTD]) given his or her currently available processing resources, and the task difficulty he or she actually chooses to work on. Selection margins can be *progressive, neutral, and conservative*, depending on whether individuals select task conditions that overtax, match, or undertax their current performance levels, respectively. We constructed a task in which neutral selection margins are a requirement for optimal performance outcomes.

For two reasons, we assumed that children and older adults experience greater difficulties in monitoring discrepancies between demands and performance than adolescents and younger adults. First, cognitive resources are less stable in childhood and late adulthood than in early adulthood. Relative to younger adults, children and older adults undergo greater age-graded changes in several broad abilities of cognitive mechanics (cf. abilities of fluid intelligence), such as information processing speed, executive control, working memory, and episodic memory (e.g., Li et al., 2004). In addition, these abilities vary more from moment to moment, reflecting greater process fluctuations in these age groups (Li et al., 2004; Lindenberger & van Oertzen, 2006; Mella et al., 2016; Papenberg et al., 2013; Rutter et al., 2020). Second, prior research indicates that cognitive monitoring mechanisms operate less efficiently in childhood and old age than in early adulthood (e.g., Clawson et al., 2017; Dodson et al., 2007; Gajewski et al., 2018; Hämmerer et al., 2011; Kray et al., 2021; Schneider, 2008). Taken together, children and older adults show larger performance fluctuations from trial to trial, and are less precise in monitoring their performances over time. Therefore, we expected that children and older adults, on average, would monitor the association between performance and difficulty less accurately than adolescents and younger adults. We also expected that the task choices of children and older adults would deviate to a greater extent from their current task proficiency than the choices of younger adults. It is possible that the direction of these discrepancies reflects a general long-term ontogenetic trend. For fluid-cognitive abilities, for example, which are relevant for the task that we used in our study, children are on a growth, and older adults, on a decline trajectory. Recent experiences of changes in performance levels could make children susceptible to *overestimating* their abilities in relation to task difficulty, resulting in progressive selection margins, whereas older adults could tend to *underestimate* their own abilities, resulting in conservative selection margins (see also Hertzog & Touron, 2011).

At the same time, we noted that research on individual differences in overconfidence among college students would suggest a different pattern of findings, with lower performance being associated with greater overconfidence, regardless of age (Kruger & Dunning, 1999; Serra & DeMarree, 2016). In their seminal paper, Kruger and Dunning (1999) found that participants scoring in the lowest quartile on tests of either humor, grammar, or logic considerably overestimated their test performance and ability in the respective domain. According to this line of reasoning, children, teenagers, and older adults could be expected to show more *progressive* selection margins than younger adults, given their comparatively more limited cognitive resources.

To test the selection margins concept and our hypotheses, a task was needed that fulfills the following requirements: (a) It should be interesting and motivating for participants from different age groups varying in age from childhood to older adulthood. (b) Task performance should involve fluid-cognitive abilities, for which age-related differences informing our hypotheses exist. (c) Highest task performance should be achieved only when task difficulty is matched with one's performance potential. Choosing higher or lower task difficulty levels should result in considerably lower task performance. (d) The task should lend itself to repeated assessments, thus allowing measurement of participants' maximum manageable task-difficulty. Pilot testing determined that a task derived from the well-known BINGO game fulfills all of these requirements. Participants are presented with cards containing four numbers. Their task is to search all BINGO cards for a called-out target number, and mark it when present, within a fixed time-window. The task difficulty varies depending on the number of BINGO cards, and participants selected the number of to-be-played BINGO cards in our study. Task performance was determined as the difference between the number of correctly marked BINGO cards minus the number of BINGO cards that could not be completed. Thus, both underestimations (i.e., choosing too few cards) and overestimations (i.e., choosing too many cards) led to suboptimal performances. Our version of the BINGO task therefore required that participants are optimally calibrated to their actual performance level when choosing a suitable task-difficulty level.

Two fluid-cognitive abilities that are related to BINGO task performance are cognitive speed and reasoning, since participants have to choose a suitable number of cards to play with, and find the target numbers on their cards as quickly as possible. Cognitive speed and reasoning fall into the broad fluid domain of cognitive abilities (Horn, 1989), which increase during childhood and adolescence, and decline in the course of normal aging. To assess the influence of cognitive resources on these decisions, we included measures of speed and reasoning as fluid-cognitive covariates in our study. We also included knowledge of vocabulary as a covariate related to crystallized abilities, which is expected to show increases well into later adulthood (Li et al., 2004).

In addition, we repeatedly asked our participants to report their motivation and emotional reactions to the BINGO game. If age groups differ in these dimensions, this may contribute to their task-difficulty choices, in addition to age differences in cognitive and metacognitive abilities.

When planning the current study, we intended to also look at gender differences in selection margins. In general, males tend to

be more risk-taking than females. A meta-analysis of 150 studies by Byrnes et al. (1999) investigated gender differences in risk-taking for different types of tasks (e.g., self-reported versus observed behaviors), different task contents (e.g., smoking, driving, drinking and drugs), and at different ages. For most domains, males showed higher levels of risk taking than females (see also Figner & Weber, 2011; Fisk, 2018). Biological theories refer to genetic and hormonal influences when explaining these differences, arguing that competitiveness and "a taste for risky confrontation" will help males to fight for resources (Wilson & Daly, 1985). The theory by Arnett (1992) argues that personal characteristics (e.g., sensation seeking) and cultural values and expectations both influence risk-taking.

Empirical studies also investigated gender differences in performance estimations. Gasser and Tan (2005) asked young adults to throw darts at targets, with the dartboard being hidden behind a screen, such that no performance feedback was provided. Subjects were asked to estimate their performance (distance from the target) after each throw. Males were more confident, but not better calibrated, than females. De Pater et al. (2009) provided university students with the opportunity to choose challenging tasks in a diagnostic situation claimed to assess their management potential. Women chose fewer challenging tasks than men, although both genders were aware that choosing a challenging task would be more valuable to assess somebody's management potential. The meta-analysis by Byrnes et al. (1999) indicates that gender differences in risk-taking are larger in children and teenagers than in young adults (see also Abbott-Chapman et al., 2008; Katzir et al., 2018; Little, 2006).

Concerning developmental changes over childhood and adolescence, the dual systems perspective predicts that risk-tasking and sensation seeking behaviors reach a peak in adolescence. This is due to an early-maturing socioemotional-incentive processing system, in combination with gradual increases of cognitive control abilities into the early 20s (for a review, see Shulman et al., 2016), based on the maturation of underlying brain circuits (Casey et al., 2011). In a study by Paulsen et al. (2011), children were even less risk-averse than adolescents. In addition, males report higher levels of sensation-seeking and risk-taking than females, and a more protracted development over adolescence (Shulman et al., 2015). In the context of this study, we expected that males, and especially boys, would show more progressive selection margins than females.

To summarize, this study tested the following set of hypotheses. We expected that (a) younger adults would be more likely to minimize the mismatch between task performance (operationalized as the self-selected difficulty levels) and available resources (operationalized as the objectively assessed MMTD) than children and older adults; (b) self-selected difficulty levels in younger adults would fluctuate less within individuals than the self-selected difficulty levels of children and older adults, reflecting a lifespan peak in monitoring efficiency in early adulthood; and (c) children would make more progressive selection decisions. For older adults, both types of deviations were considered possible: More conservative selection margins may be seen if older adults base their decisions on the experience of previous ontogenetic resource declines, but more progressive selection margins may reflect overestimations related to poorer overall performance levels. We decided to test 9-year-old children and teenagers (13–14 years), as well as young and old adults, since these groups have repeatedly shown clear age differences in cognition and task-monitoring mechanisms (e.g., Li et al., 2004), and risk-taking in adolescence has attracted increased research attention in the past years (Shulman et al., 2016). In addition, we included gender as a factor in the analyses. We developed a multi-tasking paradigm to test these predictions. The paradigm permits participants to simultaneously play and monitor any number from two up to twenty BINGO cards. To maximize their game scores, participants need to select an appropriate (that is, just manageable) number of cards to play. MMTDs were empirically determined at the beginning and end of the study, and selection margins were calculated in relation to the performance level of each individual.

Method

Participants

For the statistical power analysis, we focused on the expected differences between age groups in selection margins. In a study by Schaefer et al. (2022), the effect size (ES) for the difference in selection margins between young and older adults varied between Cohen's $d = 1.1$ (study 1) and Cohen's $d = 2.3$ (study 2). Both effects are considered to be very large. However, findings may differ by the exact type of task used (Hildenbrand & Sanchez, 2022). Using GPower 3.1 (Faul et al., 2007) for an analysis of variance (ANOVA) on age differences in selection margins in 4 age groups, with an alpha = .05 and power = 0.80, the projected sample size needed for a medium to large ES of $f = 0.3$ is approximately $N = 128$. Thus, our sample size of $N = 160$ should be more than adequate for the main objective of this study.

We tested 9-year-olds ($M = 9.5$, $SD = 0.4$), 13- to 14-year-olds ($M = 14.1$, $SD = 0.6$), 20- to 25-year-olds ($M = 23.4$, $SD = 1.4$), and 70- to 75-year-olds ($M = 72.7$, $SD = 1.5$), with 40 participants in each age group. The sample was stratified by sex, with almost equal numbers of males and females in each group (children: 21 males/19 females; teenagers: 21/19; young adults: 20/20; old adults: 20/20). Participants were drawn from the participant pool of the Max Planck Institute for Human Development (MPI) in Berlin. The 9-year-olds went to local elementary schools. For the teenagers, 29 out of 40 participants went to a higher tier school type which prepares them for university entrance (= grammar school). Eleven teenagers went to other types of secondary schools. Out of the 40 young adults, 31 had finished grammar school, and 9 had finished another (usually lower) secondary school track. Except for 2 participants with self-employed professions, the older adults were all retired. Out of 40 older adults, 18 had finished a university degree, 3 had finished grammar school, 13 had finished middle school, and 6 had finished lower school tracks. Participants were tested in age-homogeneous group sessions with up to five participants per group, with each session lasting between one and one and a half hours, and they received 60 Euro for their participation. The present study was conducted in the context of "Research Unit 448, Binding: Functional Architecture, Neural Correlates, and Ontogeny," which was funded by the Deutsche Forschungsgemeinschaft (DFG, project number 5468744).

Table 1. Cognitive Covariates and Maximum Manageable Task Difficulties (MMTDs) for Sessions 2 and 5 and Selection Margin Scores Based on These Values.

	Sample			
	Children	Teenagers	Younger adults	Older adults
N	40	40	40	40
Digit symbol score (correct items)				
M	33.2	50.6	58.5	42.8
SD	6	10.1	9.5	10.1
Range	22–49	29–74	43–79	23–65
Mehrfachwahl-Wortschatztest (MWT-A) (correct items)				
M	13.9	24.9	30.6	32.8
SD	2.9	4.2	3.5	2.5
Range	Sep–20	17–33	20–36	21–36
Figural analogies (correct solutions)				
M	12.1	13.7	16.6	10
SD	4.1	4.2	3.2	4.6
Range	0–18	0–21	Oct-21	0–18
Maximum manageable task-difficulty score (MMTD), Session 2				
M	4.2	6.65	7.75	4.85
SD	0.96	1.01	1.36	0.96
Range	2.4–6.0	5.2–9.0	5.0–10.2	3.0–7.2
Selection margin score, MMTD from Session 2				
M	0.77	0.55	0.32	0.83
SD	1.44	1.31	0.20	1.04
Range	−2.1 to 5.2	−3.3 to 3.5	-2.7 to 2.6	-1.0 to 3.2
MMTD score, Session 5				
M	4.43	7.33	8.28	5.14
SD	1.03	1.15	1.25	0.84
Range	3.0–7.0	5.8–9.4	5.6–10.2	3.0–7.0
Selection margin score, MMTD from Session 5				
M	0.54	−0.13	-0.21	0.55
SD	1.47	1.19	0.17	0.98
Range	−2.1 to 4.6	−3.1 to 2.5	-2.7 to 2.3	-1.6 to 3.1
Prior BINGO experience				
% no	58	80	58	13
% yes	32	20	42	87
% missing	10			

Note. SD = standard deviation. Cognitive covariates included cognitive speed (Digit Symbol), vocabulary (MWT-A), and reasoning (Figural Analogies). Maximum manageable task-difficulty scores (MMTD) represent the average number of cards played to achieve the five highest scores of the session. Selection margin scores refer to the deviation of the number of cards played from one's MMTD score. Positive values indicate progressive selection margins, and negative values conservative selection margins.

The Ethics Committee of the Max Planck Institute for Human Development approved of the study.

Several standardized cognitive tests were administered to describe the samples. To measure cognitive speed, the "Digit Symbol Substitution Test" of the Wechsler Intelligence Scales (D. Wechsler, 1981) was administered. Vocabulary was assessed with the MWT-A, for which participants have to find the words among a selection of pseudo-words (Lehrl et al., 1991). Reasoning was tested with the Figural Analogies test (Thorndike et al., 1954). Table 1 presents the results of these tests. Consistent with the developmental literature, cognitive speed and reasoning improve during childhood and adolescence and show declines in late adulthood, whereas knowledge of vocabulary increases into old age (e.g., Li et al., 2004). We also asked participants whether they had ever played the Bingo before.

Apparatus and Experimental Task

BINGO cards were presented on WACOM touch screens connected to Macintosh G5 computers (touch screen size 34 cm × 27 cm). Each BINGO card was 3 cm x 3 cm in size. Up to eight cards were displayed in each row of the display. Participants used a special pen for touching the screen, and the angle of the monitor was adjusted individually to maximize comfort.

Four numbers were presented on each of the cards (see Figure 1). The number of cards to be played could differ between 2 and 20, depending on the phase of the study. In each trial, 20 to-be-searched for numbers were presented auditorily one after the other through earphones and visually on the screen, 10 of which were target numbers that were shown on at least one of the cards. The interstimulus interval was 7,500 ms, and the response

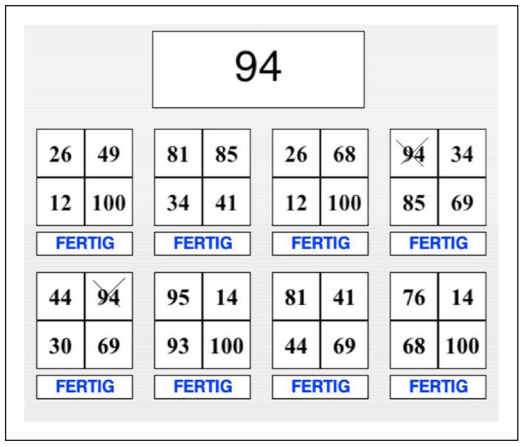

Figure 1. Example of the BINGO Task With 8 Cards.
Note. The number shown on top needs to be currently searched.

interval was 5,500 ms. The end of the response interval was signaled by a tone. The participants' task was to touch the called numbers as fast as possible on all the cards that contain the number, and to touch the "Done" button when all numbers on a given card had been crossed out. Participants were instructed to search for the number by going through all the cards in a left-to-right fashion, like in reading. After touching a specific number, the number was crossed out. When all the numbers of a card had been crossed and the "Done" button had been touched, the entire card changed its color from white to light gray. A point was given only when all numbers on a card had been crossed, and the "Done" button had been touched. For all the cards that were not completed, either because participants had missed at least one of the numbers on the cards, or because they forgot to touch the "Done" button, one point was deducted from the overall performance score. Throughout a trial, all numbers displayed on the cards were actually presented, such that the score for perfect performance was equal to the number of cards played. Participants were instructed to optimize their score while playing the BINGO game, but they did not receive any additional incentives (like money or gifts) depending on their BINGO performances.

Procedure

Each participant took part in five sessions. In Session 1, a battery of cognitive covariates and several computerized questionnaires were administered. The following sessions assessed performance on the BINGO task. Sessions 2 and 5 used fixed difficulty levels to estimate the MMTD of each participant. In Sessions 3 and 4, participants self-selected the number of cards to play. They could choose any number from 2 to 20 cards.

Session 2 started with the instruction how to play BINGO and a practice trial with three cards. Participants played 21 trials of the BINGO game with fixed difficulty levels in a pseudo-randomized order. Pilot experiments revealed pronounced age differences in the number of cards that can be played successfully. Therefore, children and older adults played 2 to 8 cards, whereas younger adults and teenagers played 5 to 11 cards, with 3 trials per condition.

The dynamic phase of the study took place in Session 3 and 4. In each of the sessions, participants played eight blocks of BINGO, each block consisting of two trials. Participants chose the to-be-played number of cards for each block, with the instruction to maximize their score. The score could only be optimized if the chosen task-difficulty level was neither too low nor too high. To assess the accuracy of participants' performance monitoring in the dynamic phase, participants reported the number of played/completed/uncompleted cards after each trial (for details, see supplement 1). One of the sessions in the dynamic assessment phase (i.e., Session 3 or 4 in counterbalanced order) was administered with veridical performance feedback, and the other one without feedback. In conditions without feedback, participants did not receive any information about their performance. In conditions with feedback, participants received auditory feedback on whether or not they had found all numbers after each response interval, and how many cards had been played/completed/not completed and the score for each trial. Throughout the study, participants were also asked questions about their emotional reaction to the BINGO game (see supplement 2 for details).

The assessment in Session 5 was identical to the one in Session 2.

Overview of Analyses

Fluctuations in the number of cards chosen during the dynamic phase of the study are analyzed with a mixed-design ANOVA with age group (4) and gender (2) as between subjects-factors. MMTDs are analyzed with a mixed-design ANOVA with age group (4) as between-subjects factor and session (2) as within-subjects factor. Selection margin scores are analyzed with an ANOVA with age group and gender as between-subjects factors, for the overall selection margins as outcome variable, as well as for the selection margins assessed in the session with and without feedback. X^2 tests are run to compare the incidence of choosing far too many cards across age groups and gender. Regressions address whether selection margin scores can be predicted by ability level, age, gender, or a combination thereof.

For all ANOVAs, F values and partial Eta square values for ESs are reported (small effect $\eta^2p = .01$, medium effect $\eta^2p = .06$, large effect $\eta^2p = .14$). To follow up significant interactions in the ANOVAs, independent samples t-tests were used. The alpha level used to interpret statistical significance was $p < .05$. Analyses were run with SPSS version 26.

Results

BINGO Performance: Fluctuations in Number of Cards Played

Table 2 presents descriptive information on the BINGO performance in the various sessions, age groups, and gender.[1] An ANOVA with age group (4) and gender (2) as between subjects-factors revealed significant differences between age groups in the intraindividual standard deviations of the number of cards chosen, $F(3, 152) = 34.18$, $p < .001$, $\eta^2p = .403$, with a strong linear trend of decreasing standard deviations with increasing age, $p < .001$ (children: $M = 3.2$, $SD = 1.5$; teenagers: $M = 2.1$, $SD = 1.3$; younger adults: $M = 1.3$, $SD = 0.9$; older adults: $M = 0.9$, $SD = 0.6$). There was also a significant effect of gender, $F(1, 152) = 4.93$, $p = .028$, $\eta^2p = .031$, and a significant interaction of age group and gender, $F(3, 152) = 4.48$, $p = .011$, $\eta^2p = .071$, due to 9-year-old

Table 2. BINGO Performances as a Function of Session, Age Group and Gender.

Session	Measure	Sample							
		Children (*n* = 40)		Teenagers (*n* = 40)		Younger Adults (*n* = 40)		Older Adults (*n* = 40)	
		Male (*n* = 21)	Female (*n* = 19)	Male (*n* = 21)	Female (*n* = 19)	Male (*n* = 20)	Female (*n* = 20)	Male (*n* = 20)	Female (*n* = 20)
2	Number of cards played								
	M	4.9	4.9	7.8	7.8	7.8	7.8	4.9	4.9
	SD	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Filled cards without "Done" (%)								
	M	3.9	3.2	4.1	2.8	2.5	1.7	5.6	6.9
	SD	4.1	2.4	2.3	2.2	2.6	1.1	3.3	5.3
	Range	0–20	0–8	0–8	1–8	0–5	0–10	2–23	1– 16
	Score								
	M	−0.2	0.4	0.2	1.3	3.8	2.4	1.3	1.3
	SD	1.6	1.4	2.4	2.4	1.9	2.9	1.3	1.3
	Range	−4 to 2	−3 to 3	−4 to 5	−3 to 5	−1 to 7	−5 to 7	−2 to 3	−1 to 4
3	Number of cards played								
	M	5.9	4.8	7.1	7.5	8.1	7.7	5.9	5.7
	SD	1.6	1.5	1.6	1.5	1.1	1.8	1.1	1.2
	Range	3–7	3–9	3–10	5–11	3–10	5–11	4–8	3–8
	SD of cards played								
	M	4.5	2.5	2.3	2.0	1.2	1.2	1.1	0.8
	SD	1.9	1.8	1.8	1.5	0.6	1.1	1.0	0.4
	Range	1–6	0–6	0–6	1–6	0–3	0–5	0–5	0–2
	Filled cards without "Done" (%)								
	M	0.8	1.5	2.3	1.4	1.3	1.6	4.6	6.1
	SD	0.9	2.0	2.3	1.6	1.5	1.7	3.3	5.9
	Range	0–5	0–8	0–8	0–4	0–3	0–7	0–18	1–20
	Score								
	M	−0.8	0.8	1.9	3.3	5.4	4.6	2.0	1.7
	SD	2.4	1.3	2.2	1.4	1.8	1.9	1.5	1.8
	Range	−4 to 4	−7 to 4	−2 to 6	1 to 6	0 to 9	0 to 9	−1 to 4	−4 to 5
4	Number of cards played								
	M	4.7	4.3	6.6	7.6	8.4	8.1	5.5	5.6
	SD	1.6	1.5	1.2	1.4	1.4	1.9	1.0	1.4
	Range	2–7	2–8	4–9	5–11	4–7	3–8	4–7	3–8
	SD of cards played								
	M	2.5	1.4	1.3	1.7	1.2	1.0	0.5	0.7
	SD	1.8	1.6	1.4	1.3	1.1	0.9	0.4	0.4
	Range	0–6	0–6	0–6	0–5	0–4	0–4	0–1	0–2
	Filled cards without "Done"(%)								
	M	2.3	1.5	2.3	1.1	1.2	1.2	3.2	7.7
	SD	2.9	1.8	2.3	1.2	1.4	1.4	3.4	17.8
	Range	0–6	0–9	0–6	0–9	0–3	0–5	0–81	0–15
	Score								
	M	0.2	1.4	3.0	3.4	5.1	5.0	2.9	2.4
	SD	1.9	1.8	2.1	1.9	1.9	1.8	1.0	2.2
	Range	−3 to 2	−5 to 5	−1 to 6	−1 to 7	1 to 8	1 to 9	1 to 5	−4 to 6
5	Number of cards played								
	M	5.0	5.0	8.0	8.0	8.0	8.0	5.0	5.0
	SD	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Filled cards without "Done" (%)								
	M	2.8	3.2	4.6	2.2	1.3	2.1	3.7	4.1
	SD	2.1	2.7	3.6	2.0	1.2	2.5	3.2	2.8
	Range	0–5	1–12	0–11	0–13	0–4	0–10	0–12	0–9
	Score								
	M	−0.2	0.3	0.5	2.9	4.8	3.7	1.8	1.9
	SD	1.6	1.8	3.3	2.8	2.0	2.6	1.3	1.3
	Range	−3 to 3	−3 to 4	−6 to 6	−3 to 6	−2 to 7	−2 to 8	−1 to 4	−1 to 4

Note. SD = standard deviation. Sessions 2 and 5 used predetermined difficulty levels. Participants chose the number of cards in Sessions 3 and 4. "Filled cards without "Done" (%)" refers to the percentage of completed cards for which participants did not hit the "Done" button in time. "Score" is the number of completed cards minus cards that could not be completed.

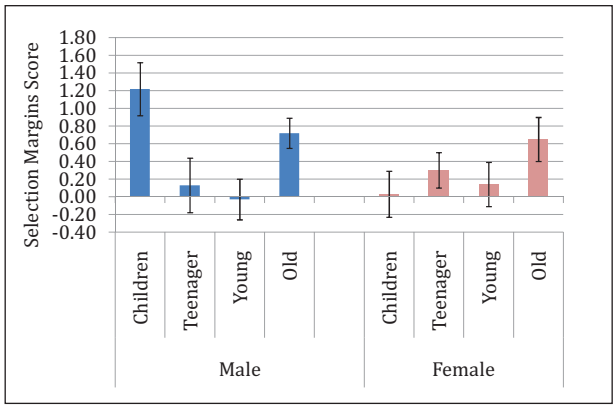

Figure 2. Selection Margin Scores by Age Group and Gender.
Note. Selection margins are calculated by comparing the number of cards played to the maximum manageable task-difficulty of the individual. Positive values represent overestimations (progressive selection margins). Males are shown in blue, and females in pink. The range of selection margins scores in the current study was between -3.3 and 5.2 (see Table 1 for details). The following sample sizes apply: male children ($n = 21$), male teenagers ($n = 21$), male young adults ($n = 20$), male old adults ($n = 20$), female children ($n = 19$), female teenagers ($n = 19$), female young adults ($n = 20$), and female old adults ($n = 20$). Error bars = *SE* mean.

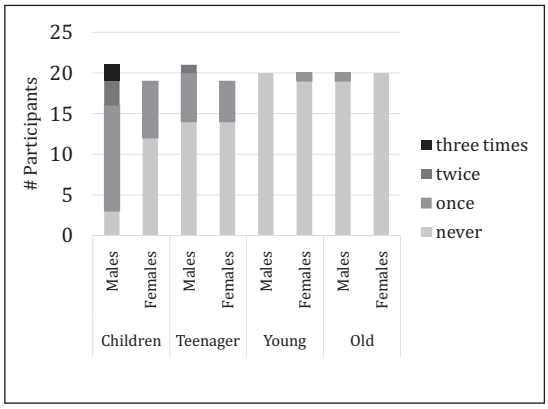

Figure 3. Number of Participants Choosing a Too-High Difficulty Level (8 or More Cards) by Age Group and Gender.
Note. The following sample sizes apply: male children ($n = 21$), male teenagers ($n = 21$), male young adults ($n = 20$), male old adults ($n = 20$), female children ($n = 19$), female teenagers ($n = 19$), female young adults ($n = 20$), and female old adults ($n = 20$).

boys ($M = 3.88$, $SD = 1.21$) fluctuating more in the number of cards played than 9-year-old girls ($M = 2.51$, $SD = 1.49$), $t(38) = 3.16$, $p = .003$, $d = 1.0$, but comparable fluctuations across males and females in all the other age groups. Concerning developmental trends, our hypothesis received only partial support. Whereas children and teenagers showed higher standard deviations than younger adults, older adults actually fluctuated less than younger adults in the choice of their task difficulties, $t(68.7) = 2.35$, $p = .022$, $d = .52$.

Supplement 4 reports additional block-by-block analyses on the number of cards played over the course of sessions 3 and 4.

Maximum Manageable Task Difficulty (MMTD) Scores

For Sessions 2 and 5, the trials of every participant were rank-ordered by the score that had been achieved, and the average number of cards played for the five best scores of the session was calculated. Table 2 presents the values by age group and session, aggregated over gender, since males and females did not differ in their MMTD scores, and there was also no interaction of gender and age group. A mixed-design ANOVA with age group as between-subjects factor and session as within-subjects factor revealed significant differences in MMTD scores between age groups, $F(3, 156) = 119.69$, $p < .001$, $\eta^2p = .697$, with younger adults showing higher performances than teenagers, who had higher scores than older adults and children. Participants achieved higher MMTDs in Session 5 compared to Session 2, $F(1, 155) = 40.74$, $p < .001$, $\eta^2p = .207$, and there was no interaction of session and age group ($p = .067$).

Selection Margin Scores

Since MMTD scores increased over the course of the study, two selection margin scores were calculated for each individual, one in relation to the MMTD score of Session 2, and the other one in relation to the MMTD score of Session 5 (see Table 2). Figure 2 presents the selection margins scores for males and females in the four age groups, averaged across the MMTDs from Sessions 2 and 5. Positive values indicate progressive selection margins, whereas negative values conservative selection margins.

An ANOVA with age group and gender as between-subjects factors revealed significant differences in the selection margin scores of the four age groups, $F(3, 152) = 3.01$, $p = .032$, $\eta^2p = .056$, and the quadratic trend for this effect reached significance ($p = .004$). As expected, younger adults' selection margins were very accurate, while children's selection margins were progressive. Older adults also showed progressive selection margins. The main effect of gender did not reach significance ($p = .197$), but there was a significant interaction of age group and gender, $F(3, 152) = 3.32$, $p = .021$, $\eta^2p = .062$, reflecting pronounced gender differences in selection margin scores among children, with boys ($M = 1.22$, $SD = 1.36$) showing higher scores than girls ($M = .03$, $SD = 1.15$), $t(38) = 2.98$, $p = .005$, $d = .73$. In the other three age groups, differences between males and females were not reliable.

To investigate whether receiving objective performance feedback during the dynamic phase of the study influenced selection margin decisions, we conducted a mixed-design ANOVA on selection margin scores with age group and gender as between-subjects factors and feedback condition (2) as a within-subjects factor. As for the overall selection margin scores, the main effect of age group and the interaction of age group and gender were significant. There was no main effect of feedback condition, $F(3, 152) = .01$, $p = .91$, $\eta^2p = .000$, and there was also no interaction of this effect with age group ($p = .940$) or gender ($p = .549$), and no three-way interaction of age group, gender, and feedback ($p = .774$). Objective performance feedback failed to influence selection margin decisions.

To obtain a better picture of influences on task-difficulty choices, we additionally investigated the incidence of extreme choices, that is, how many times participants chose a number of cards to play with that was 8 or more cards higher than one's MMTD. Figure 3 presents these data by age group and gender.

Age group differences in extreme choices were highly reliable, $X^2(3, N=160)=53.0$, $p<.001$. In the sample of 9-year-olds, boys chose far too many cards more often than girls, $X^2(1, N=40)=10.17$, $p<.01$. In all other age groups, gender differences did not reach significance.

Children and older adults showed lower performance levels than teenagers and younger adults, and were more progressive in their selection margin scores. These findings can be related to studies reporting a greater overconfidence in people who tend to perform poorly (Kruger & Dunning, 1999). To follow this up, we conducted a hierarchical regression investigating the extent to which selection margin scores could be predicted by MMTDs, by the age contrasts comparing children and older adults to teenagers and younger adults (contrast 1), children to older adults (contrast 2), and teenagers to younger adults (contrast 3), or by gender, and whether the interaction of these factors improved prediction quality. Throughout all the models that were run, performance in the BINGO game (MMTDs) was the only predictor that consistently showed significant β-values (e.g., $\beta=-.31$, $p<.01$, $R^2=.098$ for the first model), supporting the findings by Kruger and Dunning (1999) that people with poor performances tend to overestimate their abilities. The inclusion of the age contrasts, or gender, or the interactions of the predictors did not result in significant β-values that were consistent across models.

To shed further light on potential influence on task-difficulty choices in the BINGO task, supplement 2 reports age differences in emotional reactions to the BINGO game over time, and supplement 3 reports the correlations of cognitive covariates, BINGO performances and selection margins.

Discussion

This study investigated age- and gender-related differences in monitoring discrepancies between task difficulty and task performance, and in selecting task difficulties that optimize performance outcomes. An experimental paradigm modeled after the popular BINGO game was designed to address these goals. Since participants decided on suitable task-difficulty levels repeatedly over the course of several sessions, we could assess their strategic choices in a fine-grained manner, and also investigate trial-by-trial tuning of metajudgements (see also Forsberg et al., 2021, and Supplement 4). As expected, younger adults were more successful in playing the game than teenagers, older adults and children.

In line with our predictions, children fluctuated more in their task-difficulty choices. However, contrary to expectations, older adults, rather than younger adults, were the age group that fluctuated the least. Age group differences in this measure were strongly influenced by some individuals occasionally choosing far too many cards, a tendency that was absent among adults, but present in children and teenagers. This may be related to higher tendencies of risk-taking and sensation seeking in younger age groups (Paulsen et al., 2011; Shulman et al., 2015). Apparently, older adults acquired a stable perception of how many cards they should be playing, and made their choices accordingly. Future research should add measures on personality traits to the selection margins paradigm.

We also expected that younger adults would be most accurate in choosing task difficulties that correspond to their proficiency level. The data were consistent with this prediction. In fact, on average, younger adults' selection margins did not deviate significantly from zero. They skillfully calibrated their choices in relation to their task proficiency. The average selection margins of teenagers were less progressive than those of children, suggesting a developmental progression in the ability of choosing a difficulty level that is in line with one's task proficiency. Self-perceptions of competence in different domains of functioning become more realistic as children get older (Destan & Roebers, 2015; Forsberg et al., 2021; Jacobs et al., 2002). Supplement 3 addresses whether cognitive performances are correlated with BINGO performances and the selection margins decisions. Fluid intellectual abilities (i.e., cognitive speed and reasoning) correlate with BINGO MMTDs, but the relationships between cognitive covariates and selection margins are less consistent. This suggests that the miscalibration observed in children and older adults cannot be entirely explained by differences in cognitive or metacognitive abilities.

Supplement 2 adds to this by showing that age groups differ in some of their emotional reactions to the BINGO task. For example, older adults consistently report to be more motivated to play the BINGO game, as compared to the other age groups. This may have influenced their "progressive" tendency in the BINGO task. Future research should investigate the influence of emotional and motivational factors on task-difficulty choices in more detail, and it should also consider using different sets of cognitive tasks to address the generalizability of our findings.

Two recent studies have used the selection margins approach in the context of physical tasks. Schaefer and colleagues (2021) asked participants who were in the age range from childhood to young adulthood to perform rope skipping, soccer dribbling, and manual tracing tasks. Instead of choosing a specific task-difficulty level, participants were asked to predict their performance of the upcoming trial. To succeed, participants needed to deliver accurate performance predictions. Overestimations were discouraged, since participants lost all their points of the respective trial if they failed to reach their predicted performance. Across all tasks, there was a consistent decrease of overestimations with increasing age: Selection margins became less progressive, and participants overestimated their performance in a lower number of trials.

Another study based on the paradigm (Schaefer et al., 2022) used two motor tasks—carrying a tray with cube-towers, and stepping over a cross-bar—in younger and older adults. Participants were instructed to choose a suitable level of task-difficulty, by adjusting the height of the cube-tower and the bar. Again, overestimations (cube-towers collapsing or the bar falling down) led to the loss of all points of the respective trial. Task strategies were influenced by the physical risk of the respective task: For the tray-carrying task, older adults were more risk-tolerant in their task-difficulty choices. When stepping over the crossbar, older adults left a larger "safety-buffer" than young adults. This may be an adaptive strategy, because in real life, stepping over on obstacle involves a larger risk of physical harm than carrying an object. Wearing an age simulation suit, which mimics the sensory-motor declines of old adulthood, made young adults adopt a more careful strategy in the stepping-over task, similar to old adults.

Older adults of the current study showed a progressive selection bias, choosing too many cards on average. In fact, the average overestimation of about half a card across all trials was

comparable for 9-year-olds and older adults. However, the type of choices associated with progressive selection margins differed markedly between age groups. Children and teenagers were much more likely than adults to occasionally make choices that exceeded their proficiency level by far, whereas older adults consistently chose difficulty levels in slight excess of their MMTD. Over- or underestimations also depend on physical risk (Schaefer et al., 2022). In the current study, playing a computerized BINGO game did not involve any risk of physical harm. Although playing unsuitable difficulty-levels resulted in suboptimal scores, this loss of points may not have been perceived as very frustrating by our participants. Increasing the salience and importance of the task by offering rewards for successful performances should be tested in future research with the paradigm. However, the emotional reactions to the BINGO game presented in supplement 2 indicate that all age groups were rather motivated to play the game. Interestingly, children consistently reported that their scores were lower than expected, and this tendency did not change throughout the study.

One limitation of the current study is that changes in performance cannot be investigated over time in Sessions 3 and 4. Choosing too many cards to play with puts participants under a pronounced time-pressure, and some numbers will be missed. This means that 1 point for each uncompleted card is subtracted from their score. "Filled cards" or "points" cannot be interpreted as meaningful performance measures under these circumstances. However, Supplement 4 presents a block-by-block analysis of the number of cards played over the course of Sessions 3 and 4. Age differences in the number of cards played became more pronounced over time in the current study. It is possible that the decision to expose participants to different difficulty levels in Sessions 2 and 5 (depending on the to-be-expected performance range of each age group) may have led to anchoring effects in some of our participants. Future research with the paradigm should use tasks with more readily interpretable performance outcomes for each trial (like in the study by Schaefer et al., 2022), and consider confronting each participant with all possible task-difficulty levels.

In the current study, we found gender differences in the selection margins of the 9-year-olds, with boys occasionally showing strong overestimations. When making such extreme choices, boys may have approached the task in a more playful way and simply wanted to "see what happens." Risk-taking and sensation seeking show a peak in adolescence for computerized tasks (see Shulman et al., 2016, for a review). In addition, there is evidence for rather high levels of risk-taking during childhood in situations involving a risk for physical harm (Little, 2006; Morrongiello & Dawber, 2004). Although most of the studies on children's physical risk-taking focus on possibilities to prevent them from taking risks, it also has been argued that risk engagement is an important resource through which children learn from their own mistakes (Christensen & Mikkelsen, 2008). In the cognitive domain, a study by Shin et al. (2007) suggested that more progressive choices are adaptive in childhood. The authors had Kindergarten, first- and third-grade children work on a multi-trial sort-recall task, and asked them to predict how many items they would recall prior to each trial. Children who overestimated their recall more strongly showed greater gains in recall over the lists than children who overestimated their performance less drastically, supporting the adaptivity of children's overestimation of their cognitive abilities (see also Bjorklund, 1997). On the other hand, a recent study by Destan and Roebers (2015) showed that 6-year-olds who underestimated their performances in a memory task (as compared to "realists" and "over estimators") were better in allocating study times, showed a more adequate control of incorrectly recognized items and more accurate confidence judgments.

In the present study, performance improvements over the course of the study (as reflected as changes in MMTD scores from Session 2 to Session 5) were not reliably associated with selection margins, neither for the entire sample (r=.04) nor for any of the age groups (children: r=-.04; teenagers: r=.15; younger adults: r=.19; older adults: r=.08). Partial correlations controlling for the influence of gender revealed the same pattern of findings. We additionally separated the children and teenagers into two groups, depending on whether the individual had chosen far too many cards at least once ("risk-takers," n=38) or never ("risk-avoiders," n=42). There were no significant differences in performance improvements between the two groups (risk-takers: M=0.32, SD=0.91; risk-avoiders: M=0.57, SD=0.81), $t(78)=1.27$, p=.210. Nevertheless, boys' tendency to occasionally choose a number of cards far beyond their MMTD may have long-term benefits, as it exposes them to a larger range of experience.

When planning the study, we were unsure whether older adults would show progressive or conservative selection margins. On average, they chose difficulty levels that were slightly higher than their MMTD in the current study. These findings are in line with other empirical studies on the ability to accurately judge one's performance, in which older adults overestimated rather than underestimated their performance (Crawford & Stankov, 1996; Dodson et al., 2007; Shaw & Craik, 1989; K. Wechsler et al., 2018). Older adults participating in the present study may have had an optimistic attitude toward their own present and future performance potential. They also evaluated their own current functioning as being better than that of their typical age peers (Riediger et al., 2014), and they reported to be highly motivated to play the game (see Supplement 2). Future research should also include middle-aged adults.

Hierarchical regression analyses revealed that task proficiency predicted selection margin scores. This is consistent with the observation that less well performing individuals tend to overestimate their performance (Kruger & Dunning, 1999). But why do people ever choose too many cards if their estimation of how many cards were completed on previous trials is rather accurate (see supplement 1)? A phenomenon called "planning fallacy" might have influenced selection margin decisions. As described by Buehler et al. (1994), people tend to underestimate their task completion times, probably because they focus on plan-based scenarios rather than on relevant past experiences when generating their predictions. In the context of the BINGO game, this could be reflected in an overly optimistic attitude ("This time, I will pay attention to all the numbers, and I will not miss a single one.").

Children were the only group in which males were more risk-taking than females. Genetic and/or hormonal differences between boys and girls might underlie these behaviors (Slutske et al., 2011), in addition to gender-role specific cultural expectations (Cárdenas et al., 2012; Guiso et al., 2008; Morrongiello et al., 2010). In the motor domain, Mondschein et al. (2000) had

mothers of 11-month-olds judge how steep a slope could be that their child would be able to crawl down. Mothers of girls underestimated their performance, and mothers of boys overestimated their performance, although girls' and boys' actual performance levels did not differ. With the current data set, we cannot disentangle the influences of genes, hormones, and socialization on gender differences in our BINGO game. In any case, the influence of these factors seems to have diminished in the teenagers, who are generally more realistic about their performance potential, independent of gender. Future research should consider taking the pubertal status of the adolescents into account.

Summary and Outlook

When given the opportunity to choose the number of to-be-played cards, younger adults were most accurate in choosing difficulty levels that optimized their scores, whereas participants in the other age groups showed progressive selection margins by choosing too many cards, especially 9-year-old boys. Furthermore, children and teenagers occasionally selected far too many cards, thereby showing high fluctuations in selection margins. In contrast, older adults fluctuated the least. Gender differences were only reliable in children, with higher levels of risk taking in boys than in girls. Older adults' tendency to overestimate their performance, on average, might have been influenced by motivational factors, or by the lack of danger for physical harm when taking risks in the current task. Future research should investigate how task-difficulty choices can contribute to positive and negative developmental outcomes. When does being overly optimistic about one's performance potential lead to beneficial consequences for future performance (e.g., when learning a new skill as a child or teenager), and when is being careful and risk-avoidant a better strategy (e.g., when crossing a slippery street intersection as an old pedestrian)? Since the current study cannot clearly disentangle the influence of meta-cognition and risk-taking factors as mechanisms underlying our findings, future research should address this issue. In addition, performing in front of an audience may influence task strategies, since levels of risk-tasking in adolescence have been shown to increase when peers are present (Albert et al., 2013; Silva et al., 2015). Investigating such influences across the lifespan would be an interesting avenue for future research.

Acknowledgements

We thank Natalie Ebner, Jutta Heckhausen, Mike Martin, and Yee Lee Shing for helpful discussions. We also want to thank Rozalina Angelova, Dulce Erdt, Axinja Hachfeld, Viola Jucksch, Sebastian Mohnke, and Elisabeth Wenger for their help with data collection, and Jan Steinkraus and Elmar Tampe for programming the experimental tasks.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was funded by the Max Planck Institute for Human Development in Berlin.

ORCID iD

Sabine Schaefer https://orcid.org/0000-0001-7890-9132

Supplemental Material

Supplemental material for this article is available online.

Note

1. Since the difficulty levels for Sessions 2 and 5 of the study were predetermined, the number of cards played is identical for children and older adults (two to eight cards) and teenagers and younger adults (five to eleven cards). Session 2 includes an additional practice trial with three cards.

References

Abbott-Chapman, J., Denholm, C., & Wyld, C. (2008). Gender differences in adolescent risk taking: Are they diminishing? *Youth & Society*, *40*, 131–154. https://doi.org/10.1177/0044118X07309206

Albert, D., Chein, J., & Steinberg, L. (2013). Peer influences on adolescent decision making. *Current Directions in Psychological Science*, *22*(2), 114–120. https://doi.org/10.1177/0963721412471347

Arnett, J. (1992). Reckless behavior in adolescence: A developmental perspective. *Developmental Review*, *12*, 339–373. https://doi.org/10.1016/0273-2297(92)90013-R

Bjorklund, D. F. (1997). The role of immaturity in human development. *Psychological Bulletin*, *122*, 153–169. https://doi.org/10.1037/0033-2909.122.2.153

Brim, G. (1992). *Ambition: How we manage success and failure throughout our lives*. Basic Books.

Buehler, R., Griffin, D. W., & Ross, M. (1994). Exploring the "planning fallacy": Why people underestimate their task completion times. *Journal of Personality and Social Psychology*, *67*, 366–381. https://doi.org/10.1037/0022-3514.67.3.366

Byrnes, J. P., Miller, D. C., & Schafer, W. D. (1999). Gender differences in risk taking: A meta-analysis. *Psychological Bulletin*, *125*, 367–383. https://doi.org/10.1037/0033-2909.125.3.367

Cárdenas, J.-C., Dreber, A., von Essen, E., & Ranehill, E. (2012). Gender differences in competitiveness and risk taking: Comparing children in Colombia and Sweden. *Journal of Economic Behavior & Organization*, *83*, 11–23. https://doi.org/101016/j.jebo.2011.06.008

Casey, B. J., Jones, R. M., & Somerville, L. H. (2011). Braking and accelerating of the adolescent brain. *Journal of Research on Adolescence*, *21*(1), 21–33. https://doi.org/j.1532-7795.2010.00712.x

Christensen, P., & Mikkelsen, M. R. (2008). Jumping off and being careful: Children's strategies of risk management in everyday life. *Sociology of Health and Illness*, *30*, 112–130. https://doi.org/10.1111/j.1467-9566.2007.01046.x

Clawson, A., Clayson, P. E., Keith, C. M., Catron, C., & Larson, M. J. (2017). Conflict and performance monitoring throughout the lifespan: An event-related potential (ERP) and temporospatial component analysis. *Biological Psychology*, *124*, 87–99. https://doi.org/0.1016/j.biopsycho.2017.01.012

Crawford, C., & Stankov, L. (1996). Age differences in the realism of confidence judgements: A calibration study using tests of fluid and crystallized intelligence. *Learning and Individual*

Differences, *8*, 83–103. https://doi.org/10.1016/S1041-6080(96)90027-8

De Pater, I. E., Van Vianen, A. E. M., Fischer, A. H., & Van Ginkel, W. P. (2009). Challenging experiences: Gender differences in task choices. *Journal of Managerial Psychology*, *24*, 4–28. https://doi.org/10.1108/02683940910922519

Destan, N., & Roebers, C. M. (2015). What are the metacognitive costs of young children's overconfidence? *Metacognition and Learning*, *10*, 347–374. https://doi.org/10.1007/s11409-014-9133-z

Dodson, C. S., Bawa, S., & Krueger, L. E. (2007). Aging, metamemory, and high-confidence errors: A misrecollection account. *Psychology and Aging*, *22*, 122–133. https://doi.org/10.1037/0882-7974.22.1.122

Dufresne, A., & Kobasigawa, A. (1989). Children's spontaneous allocation of study time: Differential and sufficient aspects. *Journal of Experimental Child Psychology*, *47*, 274–296. https://doi.org/10.1016/0022-0965(89)90033-7

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Figner, B., & Weber, E. U. (2011). Who takes risks when and why? Determinants of risk taking. *Current Directions in Psychological Science*, *20*, 211–216. https://doi.org/10.1177/0963721411415790

Fisk, S. R. (2018). Who's on top? Gender differences in risk taking produce unequal outcomes for high-ability men and women. *Social Psychology Quarterly*, *81*(3), 185–206. https://doi.org/10.1177/0190272518796512

Forsberg, A., Blume, C. L., & Cowan, N. (2021). The development of metacognitive accuracy in working memory across childhood. *Developmental Psychology*, *57*, 1297–1317. https://doi.org/10.1037/dev0001213

Gajewski, P. D., Ferdinand, N. K., Kray, J., & Falkenstein, M. (2018). Understanding sources of adult age differences in task switching: Evidence from behavioral and ERP studies. *Neuroscience & Biobehavioral Reviews*, *92*, 255–275. https://doi.org/10.1016/j.neubiorev.2018.05.029

Gasser, M., & Tan, R. (2005). Performance estimates and confidence calibration for a perceptual-motor task. *North American Journal of Psychology*, *7*, 457–468.

Guiso, L., Monte, F., Sapienza, P., & Zingales, L. (2008). Culture, gender, and math. *Science*, *320*, 1164–1165. https://doi.org/10.1126/science.1154094

Hämmerer, D., Li, S.-C., Müller, V., & Lindenberger, U. (2011). Lifespan differences in electrophysiological correlates of monitoring gains and losses during reinforcement learning. *Journal of Cognitive Neuroscience*, *23*, 579–592. https://doi.org/10.1162/jocn.2010.21475

Hertzog, C., & Touron, D. T. (2011). Age differences in memory retrieval shift: Governed by feeling-of-knowing? *Psychology and Aging*, *26*, 647–660. https://doi.org/10.1037/a0021875

Hildenbrand, L., & Sanchez, C. A. (2022). Metacognitive accuracy across cognitive and physical task domains. *Psychonomic Bulletin & Review*, *29*, 1524–1530. https://doi.org/10.3758/s13423-022-02066-4

Horn, J. L. (1989). Models of intelligence. In R. L. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 29–73). University of Illinois Press.

Jacobs, J. E., Lanza, S., Osgood, D. W., Eccles, J. S., & Wigfield, A. (2002). Changes in children's self-competence and values: Gender and domain differences across grades one through twelve. *Child Development*, *73*, 509–527. https://doi.org/10.1111/1467-8624.00421

Katzir, T., Kim, Y.-S. G., & Dotan, S. (2018). Reading self-concept and reading anxiety in second grade children: The roles of word reading, emergent literacy skills, working memory and gender. *Frontiers in Psychology*, *9*, Article 1180. https://doi.org/10.3389/fpsyg.2018.01180

Kray, J., Kreis, B. K., & Lorenz, C. (2021). Age differences in decision making under known risk: The role of working memory and impulsivity. *Developmental Psychology*, *57*(2), 241–252. https://doi.org/10.1037/dev0001132

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*, 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Lehrl, S., Merz, J., Burkhard, G., & Fischer, S. (1991). *Manual zum MWT-A*. Perimed Fachbuch Verlag.

Li, S.-C., Lindenberger, U., Hommel, B., Aschersleben, G., Prinz, W., & Baltes, P. B. (2004). Transformations in the couplings among intellectual abilities and constituent cognitive processes across the life span. *Psychological Science*, *15*, 155–163. https://doi.org/10.1111/j.0956-7976.2004.01503003.x

Lindenberger, U., & von Oertzen, T. (2006). Variability in cognitive aging: From taxonomy to theory. In F. I. M. Craik & E. Bialystok (Eds.), *Lifespan cognition: Mechanisms of change* (pp. 297–314). Oxford University Press.

Little, H. (2006). Children's risk-taking behaviour: Implications for early childhood policy and practice. *International Journal of Early Years Education*, *14*, 141–154. https://doi.org/10.1080/09669760600661427

Mella, N., Fagot, D., & de Ribaupierre, A. (2016). Dispersion in cognitive functioning: Age differences over the lifespan. *Journal of Clinical and Experimental Neuropsychology*, 38(1), 111–126. https://doi.org/10.1080/13803395.2015.1089979

Mondschein, E. R., Adolph, K. E., & Tamis-LeMonda, C. S. (2000). Gender bias in mothers' expectations about infant crawling. *Journal of Experimental Child Psychology*, *77*, 304–316. https://doi.org/10.1006/jecp.2000.2597

Morrongiello, B. A., & Dawber, T. (2004). Identifying factors that relate to children's risk-taking decisions. *Canadian Journal of Behavioral Science*, *36*, 255–266. https://doi.org/10.1037/h0087235

Morrongiello, B. A., Zdzieborski, D., & Normand, J. (2010). Understanding gender differences in children's risk taking and injury: A comparison of mothers' and fathers' reactions to sons and daughters misbehaving in ways that lead to injury. *Journal of Applied Developmental Psychology*, *31*, 322–329. https://doi.org/10.1016/j.appdev.2010.05.004

Niebaum, J., & Munakata, Y. (2020). Deciding what to do: Development in children's spontaneous monitoring of cognitive demands. *Child Development Perspectives*, *14*(4), 202–207. https://doi.org/10.1111/cdep.12383

Papenberg, G., Hämmerer, D., Müller, V., Lindenberger, U., & Li, S.-C. (2013). Lower theta inter-trial phase coherence during performance monitoring is related to higher reaction time variability: A lifespan study. *NeuroImage*, *83*, 912–920. https://doi.org/10.1016/j.neuroimage.2013.07.032

Paulsen, D. J., Platt, M. L., Huettel, S. A., & Brannon, E. M. (2011). Decision-making under risk in children, adolescents, and young adults. *Frontiers in Psychology*, *2*, Article 72. https://doi.org/10.3389/fpsyg.2011.00072

Riediger, M., Li, S.-C., & Lindenberger, U. (2006). Selection, optimization, and compensation as developmental mechanisms of adaptive resource allocation: Review and preview. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (6th ed., pp. 289–313). Elsevier.

Riediger, M., Voelkle, M., Schaefer, S., & Lindenberger, U. (2014). Charting the life course: Age differences and validity of beliefs about lifespan development. *Psychology and Aging*, *29*, 503–520. https://doi.org/10.1037/a0036228

Rutter, L. A., Vahia, I. V., Forester, B. P., Ressler, K. J., & Germine, L. (2020). Heterogeneous indicators of cognitive performance and performance variability across the lifespan. *Frontiers in Aging Neuroscience*, *12*, Article 62. https://doi.org/10.3389/fnagi.2020.00062

Schaefer, S., Bill, D., Hoor, M., & Vieweg, J. (2022). The influence of age and age simulation on task-difficulty choices in motor tasks. *Aging, Neuropsychology, and Cognition*. Advance online publication. https://doi.org/10.1080/13825585.2022.2043232

Schaefer, S., Ohlinger, C., & Frisch, N. (2021). Choosing an optimal motor-task difficulty is not trivial: The influence of age and expertise. *Psychology of Sport and Exercise*, *57*, 102031. https://doi.org/10.1016/j.psychsport.2021.102031

Schneider, W. (2008). The development of metacognitive knowledge in children and adolescents: Major trends and implications for education. *Mind, Brain, and Education*, *2*, 114–121. https://doi.org/10.1111/j.1751-228X.2008.00041.x

Serra, M. J., & DeMarree, K. D. (2016). Unskilled and unaware in the classroom: College students' desired grades predict their biased grade predictions. *Memory & Cognition*, *44*, 1127–1137. https://doi.org/10.3758/s13421-016-0624-9

Shaw, R. J., & Craik, F. I. M. (1989). Age differences in predictions and performance on a cued recall task. *Psychology and Aging*, *4*, 131–135. https://doi.org/10.1037/0882-7974.4.2.131

Shin, H., Bjorklund, D. F., & Beck, E. F. (2007). The adaptive nature of children's overestimation in a strategic memory task. *Cognitive Development*, *22*, 197–212. https://doi.org/10.1016/j.cogdev.2006.10.001

Shing, Y. L., Werkle-Bergner, M., Li, S.-C., & Lindenberger, U. (2009). Committing memory errors with high confidence: Older adults do but children don't. *Memory*, *17*, 169–179. https://doi.org/10.1080/09658210802190596

Shulman, E. P., Harden, K. P., Chein, J. M., & Steinberg, L. (2015). Sex differences in the developmental trajectories of impulse control and sensation-seeking from early adolescence to early adulthood. *Journal of Youth and Adolescence*, *44*, 1–17. https://doi.org/10.1007/s10964-014-0116-9

Shulman, E. P., Smith, A. R., Silva, K., Icenogle, G., Duell, N., Chein, J., & Steinberg, L. (2016). The dual systems model: Review, reappraisal, and reaffirmation. *Developmental Cognitive Neuroscience*, *17*, 103–117. https://doi.org/10.1016/j.dcn.2015.12.010

Silva, K., Shulman, E. P., Chein, J., & Steinberg, L. (2015). Peers increase late adolescents' exploratory behavior and sensitivity to positive and negative feedback. *Journal of Research on Adolescence*, *26*(4), 696–705. https://doi.org/10.1111/jora.12219

Slutske, W. S., Bascom, E. N., Meier, M. H., Medland, S. E., & Martin, N. G. (2011). Sensation seeking in females from opposite- versus same-sex twin pairs: Hormone transfer or sibling imitation? *Behavioral Genetics*, *14*, 533–542. https://doi.org/10.1007/s10519-010-9416-3

Staus, N. L., O'Connell, K., & Storksdieck, M. (2021). Addressing the ceiling effect when assessing STEM out-of-school time experiences. *Frontiers in Education*, *6*, Article 690431. https://doi.org/10.3389/feduc.2021.690431

Thorndike, R. L., Hagen, E., & Lorge, I. (1954). *Cognitive Abilities Test*. Houghton Mifflin Harcourt.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale—Revised (WAIS-R)*. Psychological Corporation.

Wechsler, K., Drescher, U., Janouch, C., Haeger, M., Voelcker-Rehage, C., & Bock, O. (2018). Multitasking during simulated car driving: A comparison of young and older persons. *Frontiers in Psychology*, *9*, Article 910. https://doi.org/10.3389/fpsyg.2018.00910

Wilson, M., & Daly, M. (1985). Competitiveness, risk taking, and violence: The young male syndrome. *Ethology and Sociobiology*, *6*, 59–73. https://doi.org/10.1016/0162-3095(85)90041-X