Metacognition and Learning (2025) 20:12
https://doi.org/10.1007/s11409-025-09416-2

Performance judgment in mathematical and reading competence in adults

Maximilian Seitz[1] · **Anna Hawrot**[1] · **Kathrin Lockl**[1]

Received: 11 September 2024 / Accepted: 14 March 2025

Abstract
Past research has revealed that people often hold overestimated perceptions of their performance in various domains. While there is a debate over how overestimation might facilitate motivation to a certain degree, intentional learning processes typically benefit from accurate judgments. With the focus on lifelong learning, accurate performance judgments should be important in all age groups. However, research on older age groups who do not participate in formal schooling (e.g., middle-aged adults) has primarily investigated performance judgments using laboratory tasks that cannot be equated with complex academic skills. Therefore, the current study investigated age-related differences in global performance judgments (postdictions) and their inaccuracy (or: bias) in adulthood using cross-sectional data from a large-scale, representative German panel study (adults aged 25–67). We investigated judgments of mathematical and reading competence using parametric and nonparametric models to capture the patterns in the data, controlling for educational attainment. In both mathematical and reading competence, we found that while performance judgments decreased, the observed bias increased in older age groups. In addition, there was evidence for small gender differences in performance judgment but not in the bias. The findings provide a comprehensive insight into age-related differences in performance judgments and their accuracy in adulthood.

Keywords Performance judgement · Procedural metacognition · Mathematical competence · Reading competence · Adulthood

Past research has revealed that people often hold inaccurate, usually overestimated, perceptions of their skills, character, and performance (see e.g., Dunning et al., 2004 for a review). Such inaccuracies concern global self-assessments (e.g., self-concept) as well as task-specific judgments (e.g., estimated test performance), and may carry manifold negative consequences at school, at work, and in private life (e.g., Dunlosky & Rawson, 2012; Dunning et al., 2004). The topic of self-assessments, their accuracy, and changes over time have been popular in various strands of research in psychology and educational studies due to their well-documented links with school learning and academic achievement (see Marsh

✉ Maximilian Seitz
maximilian.seitz@lifbi.de

[1] Leibniz Institute for Educational Trajectories, Wilhelmsplatz 3, 96047 Bamberg, Germany

Published online: 08 April 2025

et al., 2022 for self-concept research and Metcalfe, 2009 for metacognition research). However, the focus on learning and achievement generally went hand in hand with a focus on children, adolescents, and young adults because they participate in formal schooling in which achievement constitutes a key outcome. Research on older age groups has primarily investigated memory performance in various laboratory tasks (for a review, see Dodson, 2017), which is difficult to equate with learning complex skills. Moreover, such studies give little information on the pattern of age-related differences because they often rely on comparing older and younger adults lumped into broad age groups.

The quickly evolving world, globalization, rapid technological advances, and lengthened activity in the labor market, in connection with a longer life span, necessitate intentional learning in various contexts, including the workplace, home, or local community, throughout adulthood (e.g., Hager, 2020). Self-assessments may play a central role in lifelong learning. Limited knowledge of learning-related self-assessments, their accuracy, and changes over time limits the understanding of factors affecting skill and knowledge acquisition in adulthood in both formal and non-formal contexts. Moreover, it precludes an adequate design of learning environments for adult learning. Therefore, the current study investigated age-related differences in global performance judgments and their accuracy in adulthood. We used data on performance judgments in two standardized tests on mathematical and reading competence and the accuracy of these judgments collected from a large representative cross-sectional sample of German adults aged 25 to 67. Analytically, we employed and compared both parametric and nonparametric modeling approaches. Moreover, to better understand individual differences in performance judgments, we examined the role of gender.

Theoretical framework

In research on metacognition, performance judgments are seen as indicators of metacognitive processes – also referred to as procedural metacognitive skills (e.g., W. Schneider et al., 2022). Metacognitive processes differ depending on the stage of learning (acquisition, retention, retrieval) and include self-monitoring (a bottom-up process) and self-regulation (a top-down process). Self-monitoring refers to keeping track of where one is on the way toward understanding or remembering, which serves as the basis for self-regulation and self-initiated learning behavior (Nelson, 1990; Nelson & Narens, 1994). Thus, metacognitive monitoring reflects individuals' ability to evaluate their learning progress. Since performance judgments serve as the basis for self-regulation, the accuracy of these judgments is central. It can be argued that goal-directed self-regulation is greatly facilitated if performance judgments are realistic. Accurate judgments allow for choosing appropriate learning strategies and allocating adequate learning time so that learning is efficient and goal-directed (Dunlosky & Rawson, 2012).

In studies that examine metacognitive monitoring skills, participants are typically asked to judge their performance before, during, or after working on a memory task or a comprehension test. That is, they provide predictions or postdictions of their performance, respectively (for more information see Flavell et al., 2002; Roebers, 2002; Schneider & Lockl, 2008). Hence, performance judgments differ depending on the time point of assessment. They also differ depending on whether they relate to individual items (local judgments) or the entire test (global judgments; Hacker et al., 2008).

Performance judgments can be related to test performance. This aspect is also referred to as absolute accuracy or "calibration", describing the extent to which predicted or postdicted and actual performance tally up. Several studies have shown that postdictions are more accurate than predictions (e.g., Hacker et al., 2000; Maki & Serra, 1992), probably because more information about the test items is available after completing the test (Maki & Serra, 1992). At the same time, high performance is linked to higher and more accurate judgments (e.g., Hacker et al., 2008). Regarding interindividual differences, it has been shown that measures of absolute accuracy or calibration are more stable in comparison to relative judgments and that it is more likely to detect reliable interindividual differences when they exist (Kelemen et al., 2000) using these measures.

Age differences in performance judgments and their accuracy

Previous research on performance judgments and their accuracy in adults has shown varied results, depending on the task and type of judgment (pre- or postdictive, global or local). Most studies used laboratory tasks involving memory and perception (e.g., Dodson, 2017; McWilliams et al., 2023; Overhoff et al., 2021), or sometimes reasoning and general knowledge (e.g., Burns et al., 2016; Dahl et al., 2010). These studies generally indicate that older adults are less accurate than younger adults when making local retrospective judgments of performance in memory and perceptual tasks (see Dodson, 2017 for a review; Overhoff et al., 2021). However, the accuracy of retrospective judgments may remain intact with age for tasks relying on general knowledge (Dahl et al., 2010; see Dodson, 2017 for a review), although some studies report a decline (e.g., Burns et al., 2016). Moreover, accuracy may remain intact or even improve with age for prospective performance judgments – although this has been mostly shown with memory tests (Daniels et al., 2009; Eakin et al., 2014).

The existing research on the development of global retrospective judgments and their accuracy in adults suggests that older age groups tend to give lower global postdictions of their performance than younger adults (Devolder, 1993; Touron et al., 2010) but the accuracy of their postdictions remains stable with age. Devolder (1993), Devolder et al. (1990), Hertzog et al. (1994), and Kidder et al. (1997) reported that younger and older adults did not differ in the accuracy of global postdictions in various retrospective and prospective memory tasks and practical problem-solving skills. However, in a study by Touron et al. (2010) older adults had a lower accuracy of postdicted performance in a working memory task. However, studies using general knowledge tests have been rare.

Nevertheless, the above-mentioned studies, although informative, share important limitations. First, most of them used laboratory tasks of memory performance that are not representative of typical learning tasks or domains. Second, they grouped older and younger adults into broad age categories, which could result in age differences in accuracy being undetectable. Third, age effects could be confounded by other aspects that are not always considered, such as educational background. Past studies have included mainly university students as young adults (e.g., Hertzog et al., 1994; Touron et al., 2010), making comparisons to a broader group of older adults difficult. Furthermore, due to a reliance on group comparisons, these studies do not allow drawing detailed inferences on how people of various ages differ in their judgments and accuracy.

At the same time, research on cognitive aging has documented a significant decline in various cognitive abilities, for instance, processing speed, working memory, fluid intelligence, or executive functions over adulthood (e.g., Hartshorne & Germine, 2015; Lacreuse et al., 2020). This suggests that metacognitive skills, as cognitive in nature, may decline

as well. Executive functions, which decline over adulthood (Lacreuse et al., 2020), have been suggested to contribute to metacognition, although with mixed empirical support (e.g., Filippi et al., 2020; Pennequin et al., 2010; Perrotin et al., 2008; Roebers & Feurer, 2016; Wu & Was, 2023). However, the studies used manifold measures of metacognitive skills, which makes integrating their results challenging. Nevertheless, given that metacognition and executive functions share certain functional areas in the brain (e.g., Roebers & Feurer, 2016) that show aging-related changes (e.g., Fleming & Dolan, 2012; Stuss, 2011), it seems likely that both will decline with age.

Additionally, research on adult populations has revealed that people commonly believe that various cognitive skills, including verbal and numerical ability, decline with age (Vaportzis & Gow, 2018). Drawing on stereotype threat and self-categorization theory, it can be assumed that such deficit expectations may also affect test performance (Haslam et al., 2012). Therefore, older adults' performance in certain ability domains may be lower due to cognitive aging and the decline may be further amplified by their stereotypical expectation of poorer performance. As a result, the decline in performance may be higher than the decline in expectations, leading to an increased overestimation (lower accuracy) in older adults.

Gender differences in performance judgments

An indication that performance judgments and their accuracy may differ between genders can be found in theoretical models that explain how the information available during the judgment process affects the judgment itself. These models list self-beliefs among various task- and person-related factors (e.g., Ehrlinger & Dunning, 2003; Koriat et al., 2008; Zhao & Linderholm, 2008; for a discussion, see Händel et al., 2020). At the same time, gender differences in domain-specific academic self-beliefs, for instance, self-efficacy and self-concept, are well documented in student populations (e.g., Huang, 2013; Mejía-Rodríguez et al., 2021; OECD, 2019). Although data on such self-beliefs in adulthood are, understandably, a lot more limited and fragmented, with results for adults usually pooled across broad age groups, they still provide relevant cues. In the metanalysis by Huang (2013), men aged 23 or older had a higher self-efficacy in mathematics than women from the same age group. Moreover, adult women from Western countries more often report anxiety about performing calculations or have higher mathematics anxiety than men (Hart & Ganley, 2019; OECD & Statistics Canada, 2011), which suggests their lower mathematics self-concept as well. However, data on literacy or reading self-concept and self-efficacy in adulthood is largely lacking. Nevertheless, the language domain and reading are still considered stereotypically feminine (e.g., Bonnot & Jost, 2014; Espinoza & Strasser, 2020; Steffens & Jelenec, 2011), which suggests potential higher self-perceptions in the reading domain in adult women in comparison to men.

The current study

Taken together, research on global performance judgments and their accuracy throughout adulthood is limited. Little is known about fine-grained age differences and potential gender differences, which limits the understanding of factors influencing skill and knowledge acquisition in adulthood in both formal and non-formal contexts. Moreover, it precludes

the design of learning environments in a way that caters to the specific needs of adult learners. Therefore, the current study investigated age-related differences in global performance judgments in adulthood. In addition, we focused on potential gender differences. To get a broader picture of the phenomenon, we included judgments in mathematics and reading. We expected performance judgments to be lower in older adults in both domains (H1, H2). Due to the likely influence of gender stereotypes on self-perceptions, we expected performance judgments to be lower in women in mathematics (H3) and lower in men in reading (H4). Concerning the accuracy of performance judgments, we hypothesized a higher bias (inaccuracy) in older adults in both domains (H5, H6). Moreover, we expected that self-perceptions, possibly influenced by gender stereotypes, would reduce such overestimation, resulting in a lower bias in women in mathematics (H7) and a lower bias in men in reading (H8).

Method

Sample

The current study used data from the adult cohort of the German National Educational Panel Study (NEPS SC6; Blossfeld & Roßbach, 2019). NEPS SC6 has an annual data collection plan and draws on adults born between 1944 and 1986 living in Germany. The panel study was conceptualized as a representative cohort in Germany regarding gender, educational attainment, birth year, place of living, and municipality size (Hammon et al., 2016). Because panel attrition was selective regarding participants' age, marital status, residency, and household size (Stöckinger et al., 2018; Zinn et al., 2020), design weights are provided in the official Scientific Use File (Hammon et al., 2016). The current study focused on the third survey wave (data collection: October 2010 to May 2011), in which the data was collected in the participants' homes (i.e., self-report measures and competence assessment). Overall, $n = 7396$ participants were interviewed at home in this survey wave (Aust et al., 2012). The analytical sample in the current study drew on $n = 6890$ cases that participated in the competence tests (50.35% female; 16.78% migration background; 93.15% German household language; $M_{age} = 47.24$, $SD = 11.30$, $Min = 25$, $Max = 67$). The mean position on the International Socio-Economic Index of Occupational Status Index (ISEI-08, Ganzeboom et al., 1992) was slightly above average: $M_{ISEI} = 50.02$, $SD = 21.42$, $Min = 11.56$, $Max = 88.96$. However, it should be noted that not all respondents participated in or finished both competence tests. Overall, we analyzed data from subsamples of $n = 5015$ for mathematical competence and $n = 5057$ for reading competence (overlap: 46.18%). Systematic differences between these subsamples due to, for example, sociodemographic aspects were not expected because the test assignment was randomized in the study design.

Competence tests

In addition to the survey questionnaires, the participants completed a mathematics and reading test specifically developed for NEPS. They measured mathematical and reading competencies necessary for successful participation in modern society, and therefore they

did not directly follow the school curriculum. Both tests were administered by a trained interviewer at the beginning of a face-to-face interview in the participants' homes in a paper–pencil format (the test time was 28 min each). Due to survey constraints, not all participants of NEPS SC6 were given all competence tests; those who participated in the respective tests were given the same items in the same sequence. For all analyses, we use percentage scores; a descriptive overview is provided in Table 1.

The mathematics test included 21 items that covered five content areas (quantity, change and relationships, shape and space, data and chance) and, at the same time, required six cognitive components for successful task completion (applying technical skills, modeling, arguing, communication, representing, problem solving). The items had a multiple-choice format with one correct answer, a yes–no format that required a decision whether a statement was correct or not considering the information provided in the task (e.g., in a table), or a short open-ended format that required typing in a number or a single word. A detailed description of the test's theoretical framework is available in Neumann et al. (2013), whereas example items are provided by Schnittjer and Duchhardt (2015). Marginal reliability equaled 0.780 and the test had good psychometric properties (see Jordan & Duchhardt, 2013 for details).

The reading test included 30 items covering five text types (information, commenting or augmenting, literary, instruction, and advertising) and, simultaneously, three cognitive requirements necessary for successful task completion (finding information, drawing text-related conclusions, and reflecting and assessing). The texts sought to require only minimal prior knowledge for adequate comprehension. The items had a multiple-choice format with one correct answer, a yes–no format that required a decision whether a statement was correct or not considering information provided in the text, or a matching format that required assigning subheadings to different text sections. A detailed description of the test's theoretical framework is available in Gehrer et al. (2013), whereas example items can be found in Gehrer et al. (2012). Marginal reliability equaled 0.717 and the test had good psychometric properties (see Hardt et al., 2013 for details).

Table 1 Descriptive overview of performance judgment indicators and competence scores

	Total *M* (SD)	Men *M* (SD)	Women *M* (SD)
Mathematical competence: Performance judgment (Min = 0; Max = 1)	0.62 (0.23)	0.67 (0.22)	0.57 (0.18)
Mathematical competence: Performance judgment bias (Min = −1; Max = 1)	0.18 (0.18)	0.18 (0.18)	0.19 (0.18)
Reading competence: Performance judgment (Min = 0; Max = 1)	0.64 (0.22)	0.63 (0.22)	0.66 (0.22)
Reading competence: Performance judgment bias (Min = −1; Max = 1)	0.06 (0.17)	0.06 (0.16)	0.06 (0.17)
Mathematical competence (Min = 0; Max = 1)	0.45 (0.23)	0.50 (0.23)	0.38 (0.21)
Reading competence (Min = 0; Max = 1)	0.58 (0.20)	0.56 (0.20)	0.60 (0.19)

The term "bias" refers to the inaccuracy of performance judgments when compared to the test performance; mathematical competence (men $n = 2471$; women $n = 2544$); reading competence (men $n = 2515$; women $n = 2542$); all values are weighted

Performance judgments and their accuracy

Immediately after each test, the participants were asked to estimate the number of items they presumably answered correctly. Therefore, the estimates were so-called *postdictions* or *retrospective judgments of performance accuracy* (Hardt et al., 2013; Schraw, 2009) widely used measures of metacognitive monitoring. Since they referred to the whole test, they were global judgments. To create an indicator of performance judgment, the number of items participants thought they solved correctly was divided by the number of items in the test. In addition, we used "bias" (inaccuracy) as an easily interpretable indicator of accuracy. For this, the proportion of items solved correctly was subtracted from the item proportion the participants thought they solved correctly. Zero indicates perfect accuracy, whereas values below and above zero indicate under- or overestimation, respectively. We used four indicators: two indicators of performance judgment in mathematics and reading and two respective bias indicators. All indicators were available in the dataset.

Control variables

The analyses controlled for educational attainment because, in older cohorts, access to education was more limited (e.g., Gebel & Pfeiffer, 2010). At the same time, more educated people may give higher performance judgments and be more accurate due to their higher competencies associated with prior experience with similar tasks (see Hacker et al., 2008). Without controlling for education, age differences in performance judgments and accuracy may reflect between-cohort differences in educational attainment. For the same reason, we controlled for education while examining the role of gender. Women, especially in older generations, had more limited access to educational opportunities than men (e.g., Gebel & Pfeiffer, 2010), and, as higher education is linked to higher competencies, this might also be a source of confounding (see Hacker et al., 2008). In other words, without adequate control, gender differences may reflect differences in educational attainment. In the current study, educational attainment was used as a self-reported categorical indicator of endowed or acquired social and individual resources. We used the International Standard Classification of Education (ISCED-97; UNESCO, 2006), adapted for Germany (S. L. Schneider, 2008).

Data analysis

Although there are some findings on performance judgment in adults based on laboratory experiments in the literature (e.g., Devolder, 1993; Devolder et al., 1990; Hertzog et al., 1994; Touron et al., 2010), there is no clear age-related pattern, especially regarding domain-specific competencies. Therefore, we tested both parametric and nonparametric approaches to model the functional form of age on performance judgment as well as the accuracy of performance judgment. More specifically, we used linear models with and without a quadratic term to model the relationship between participants' age and performance judgment or performance judgment bias, respectively. These models are compared using explained variance and information criteria. For a nonparametric approach, we used a generalized additive model to capture potentially non-linear patterns in the data, which was possible due to the large sample size. Generalized additive models allow for a data-driven investigation of the relationship between the criterion and predictor variables. They

are often used for exploratory modeling because they capture linear and nonlinear relationship patterns. Such models are more flexible in capturing the peculiarities of the data and use smooth functions to visualize the relationship between variables (Wood, 2017). For evaluating the models, we drew on the explained variance and deviance (i.e., a goodness-of-fit measure on how well the model explains variability in the data relative to a saturated model). In all models, educational level and gender were included as linear control variables. Due to model complexity and the risk of overfitting, these variables were added as linear factors in the nonparametric models. Data preparation and calculating the parametric models were done in STATA 17 (StataCorp, 2021), while the generalized additive models were calculated in R 4.4.0 (R Core Team, 2023) using the package "mgcv" 1.9–1 (Wood, 2011). To account for potential selectivity due to longitudinal attrition, we used the design weights provided in the Scientific Use File (Hammon et al., 2016).

Results

Table 1 provides a descriptive overview of the performance indicators in mathematical and reading competence. These indicators refer to how many items the participants thought to have solved correctly (performance judgment) and how biased their judgment was when compared to their actual test performance. Mathematical competence ($r = -0.31$, $p < 0.01$) and reading competence ($r = -0.23$, $p < 0.01$) were negatively correlated with participants' age. Men reported higher performance judgment in mathematical competence compared to women ($t(5013) = 21.66$, $p < 0.01$, $d = 0.61$), while women reported higher performance judgment in reading competence compared to men ($t(5055) = 3.88$, $p < 0.01$, $d = 0.11$). When adjusting for their test performance, both men and women had a similar bias in both content domains. On average, men's and women's bias indicators were positive. That is, men and women stated that they solved more items correctly than they really did. It should be mentioned that level differences between the two tests cannot be directly interpreted as the tests had a different solving probability (Table 1). Information on calibration as an alternative indicator of absolute accuracy (e.g., Fleming & Lau, 2014) can be found in the supplement (Table 1S). An overview of all bivariate correlations can also be found in the supplement (Table 2S).

Performance judgment in mathematical competence

In the linear model predicting performance judgment in mathematical competence while controlling for education level (Fig. 1), there was a small negative effect of age, suggesting that older age groups had lower performance judgment ratings ($\beta = -0.10$, $p < .01$). In addition, women had lower performance judgment ratings than men ($\beta = -0.21$, $p < .01$). The model explained a small amount of variance ($R^2 = 0.15$). In the quadratic model, the quadratic term for age was significant ($\beta = -0.35$, $p < 0.01$). However, the information criteria and lack of an increase in explained variance indicated that the quadratic model had limited incremental value compared to the linear model (Table 2).

Regarding the nonparametric approach, performance judgment in mathematical competence in different age groups showed a small but significant negative trend (Fig. 1S); effective degrees of freedom (EDF) = 8.99, $p < 0.01$ (REML = 326.76). Descriptively, this may be due to the rather pronounced difference between people aged 50–60 years. That is, participants over the age of about 50 were more likely to state having solved fewer

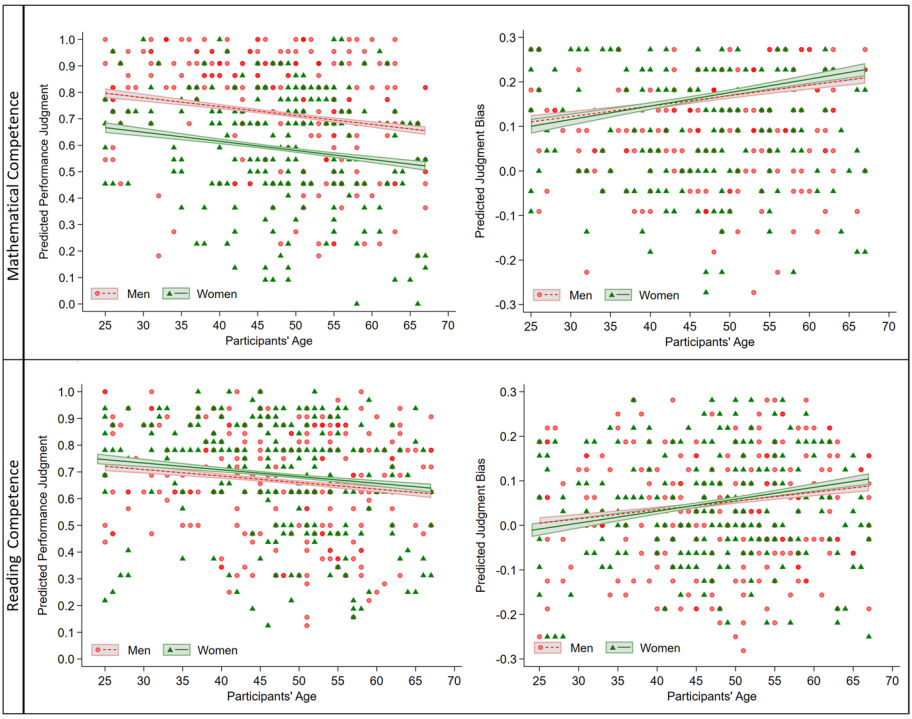

Fig. 1 Linear models. *Note*. Men in red (mathematics: $n = 2470$; reading: $n = 2515$); women in green (mathematics: $n = 2544$; reading: $n = 2542$); with 95% confidence intervals; weighted regression models controlled for education level; for readability, the scatter plots show a random subsample of 10% of cases

mathematics problems correctly than younger participants. As in the linear model, gender had a significant effect, suggesting that women's average level of performance judgment was lower than men's (unstandardized coefficient = −0.10, 95% CI [−0.11; −0.08], $p < 0.01$). The explained variance of the model was small and comparable to the parametric models (Adj. $R^2 = 0.15$; deviance explained = 15.90%). Due to parsimony, we conclude that the linear model fits the data best.

Regarding performance judgment bias in mathematical competence, there was a significant age effect in the linear model (Fig. 1), suggesting that older age groups overestimated their mathematical competence ($\beta = 0.14$, $p < 0.01$). There was no significant gender effect ($\beta = -0.01$, $p = 0.72$) and the model explained a small amount of variance ($R^2 = 0.06$). In the quadratic model, the quadratic age term was also significant ($\beta = 0.36$, $p < 0.01$). However, the explained variance and information criteria suggest that the incremental value of the quadratic model is limited (Table 2). Similarly, the generalized additive model revealed that the bias in performance judgment was higher in older age groups (Fig. 1S); EDF = 8.99, $p < 0.01$ (REML = 381.61). As in the parametric models, there were no differences between men and women in the overall level of performance judgment bias in mathematical competence (unstandardized coefficient = 0.00, 95% CI [−0.01; 0.01], $p = 0.74$). The explained variance of the model was small and comparable to the parametric models (Adj. $R^2 = 0.05$; deviance explained = 6.02%). Therefore, we again conclude that the linear model fits the data best.

Table 2 Model comparison

		R^2		AIC		BIC	
		Linear	Quadratic	Linear	Quadratic	Linear	Quadratic
Mathematical Competence	Performance Judgment	0.15	0.15	−1252.29	−1260.18	−1169.19	−1170.68
	Performance Judgment Bias	0.06	0.06	−2692.21	−2699.72	−2609.11	−2610.23
Reading Competence	Performance Judgment	0.10	0.11	−1381.03	−1390.54	−1297.89	−1301.01
	Performance Judgment Bias	0.04	0.05	−3504.02	−3522.83	−3420.88	−3433.30

The models differ only in how the predictor "participants' age" was included

As the models suggested a negative effect of age on performance judgment and a positive effect on performance judgment bias, we additionally ran a sensitivity analysis to check whether this effect was significant when ability level was controlled for. Therefore, we calculated a linear regression analysis predicting the performance judgment ratings with participants' age, only including cases within 1/8 of a standard deviation around the mean of the mathematical competence test scores (subsample $n = 420$; all age groups adequately represented). In both models predicting performance judgment and performance judgment bias, age was not a significant predictor (Table 3S, 4S). Therefore, when only looking at a subsample with comparable test scores, age was not associated with performance judgment or the accuracy of performance judgment in mathematical competence.

Finally, we included an exploratory interaction term between age and gender in the reported models predicting performance judgments as well as performance judgment bias in mathematical competence. While the interaction term was not significant in the linear and quadratic models, it was significant in the nonparametric models (i.e., performance judgment and performance judgment bias). However, it resulted in no improvement of the adjusted R^2, higher REML values, and lower explained deviance scores, suggesting overfitting or a more complex relationship than the data support (Babyak, 2004). Therefore, we do not further interpret these exploratory results of the generalized additive models.

Performance judgment in reading competence

Regarding the performance judgment in reading competence, there was a negative age effect in the linear model (Fig. 1), suggesting that older age groups had lower performance judgment ratings ($\beta = -0.07$, $p < 0.01$). In addition, women had higher ratings than men ($\beta = 0.09$, $p < 0.01$). The model explained a small amount of variance ($R^2 = 0.10$). In the quadratic model, the quadratic term for age was significant ($\beta = -0.39$, $p = 0.02$). However, the information criteria and small additional effect on explained variance indicate that the quadratic model has limited incremental value when compared to the linear model (Table 2). Regarding the nonparametric approach, there was also a significant decrease in older age groups (Fig. 1S); EDF = 8.99, $p <$. 01 (REML = 277.62). As in the linear model, gender had a significant effect, suggesting that women's average level of performance judgment in reading competence was higher than men's (unstandardized coefficient = 0.04, 95% CI [0.03; 0.05], $p < 0.01$). The explained variance of the model was small (Adj. $R^2 = 0.11$; deviance explained = 11.40%) and comparable to the parametric models. Due to parsimony, we conclude that the linear model fits the data best.

Finally, the linear model predicting performance judgment bias in reading competence indicated a positive age effect ($\beta = 0.11$, $p < 0.01$), namely that older age groups showed more overestimation. There was no significant gender effect ($\beta = -0.01$, $p = 0.77$) and the model explained a small amount of variance ($R^2 = 0.04$). In the quadratic model, the quadratic age term was significant ($\beta = -0.54$, $p < 0.01$). However, the information criteria and small additional effect on explained variance indicated that the quadratic model had little incremental value when compared to the linear model (Table 2). Regarding the nonparametric approach, there was a similar positive effect of age on performance judgment bias in reading competence (Fig. 1S); EDF = 8.99, $p < 0.01$ (REML = −774.46). As in the linear model, there were no differences between men and women (unstandardized coefficient = 0.00, 95% CI [−0.01; 0.01], $p = 0.85$). The explained variance of the model was small and comparable to the parametric models (Adj. $R^2 = 0.05$; deviance explained = 5.19%). Again, we conclude that the linear model fits the data best.

Additionally, we ran a sensitivity analysis to check whether the effect of age was significant when the ability level was controlled for. We calculated a linear regression analysis predicting the performance judgment ratings with participants' age, while only including cases within 1/8 of a standard deviation around the mean of the reading test scores (subsample $n = 521$; all age groups adequately represented). In both models predicting performance judgment and performance judgment bias, age was not a significant predictor (Table 5S; Table 6S). Therefore, when only looking at a subsample with comparable test scores, age was not associated with performance judgment or the accuracy of performance judgment in reading competence.

Finally, we included an exploratory interaction term between age and gender in all reported models predicting performance judgments as well as performance judgment bias in reading competence. The interaction term only had a significant effect in the generalized additive model predicting performance judgment. However, this resulted in a negative partial R^2, higher REML values, and lower explained deviance scores, suggesting overfitting or a more complex relationship than the data can support (Babyak, 2004). Therefore, we do not further interpret this exploratory result.

Discussion

The current study investigated performance judgments in domain-specific competencies in different age groups using data from a representatively drawn German study. We chose both parametric and nonparametric approaches to examine differential patterns between age groups, and men and women while controlling for educational attainment. Due to parsimony, the linear models showed the best fit when compared to the quadratic and generalized additive models. However, it should be mentioned that the parametric and the nonparametric models cannot be directly compared. Our findings suggest that older age groups differ from younger ones in giving lower judgments in both mathematical and reading competence. In addition, our modeling approach indicates distinct differences by gender, even when controlling for educational attainment. Men reported higher performance judgment in mathematical competence, while women reported higher performance judgment in reading competence, in line with hypotheses H1-H4. Regarding the accuracy of performance judgments, the results indicated that older age groups were generally more inaccurate (overestimation). However, it should be mentioned that this conclusion was not supported by the sensitivity analyses that included only individuals with average test performance. Therefore, our results do not allow a definitive conclusion regarding H5 and H6. Furthermore, men and women did not differ in bias (hypotheses H7 and H8 were not supported).

Theoretical implications

In line with previous research (e.g., Dunning et al., 2004), our findings suggest a rather general pattern of overestimation in all age groups. While it is typically found that the accuracy for some aspects of metacognitive monitoring remains stable (e.g., Devolder et al., 1990; Hertzog et al., 1994), our analyses of the full NEPS SC6 sample indicated that older age groups show more overestimation compared to younger age groups (similarly, Touron et al., 2010). This might be because previous studies have mostly used laboratory memory tasks and included potentially selective samples grouped into broad age categories. The

latter did not allow for examining the pattern of age differences. In contrast, we used tests of complex competencies and included a large and representative sample that allowed treating age as a predictor. Still, it should be noted that age differences were small (i.e., less than one standard deviation covering the whole age range). Furthermore, it is important to point out that one should be cautious about interpreting differences in judgment bias when there are differences in the criterion performance (for a discussion, see Dunlosky et al., 2015). Given that our results showed that mathematical and reading literacy were negatively associated with age, we conducted additional sensitivity analyses, including a subsample with comparable average performance levels. These analyses showed that the age effects on judgment accuracy were no longer significant, which may indicate that differences in judgment bias across the full sample may be, at least in part, an artifact of differences in performance. Unfortunately, it was impossible to disentangle age effects in judgment bias and performance level for the full sample. Thus, the question remains whether the observed differences in judgment bias reflect actual monitoring ability or are, at least partly, the result of different performance levels. Additionally, we cannot comment on patterns of overestimation in adults over 67, who might be more affected by memory decline influencing metacognitive monitoring (Hertzog & Shing, 2011) due to a lack of data.

Regarding gender differences, women judged their performance lower than men in mathematics but higher than men in reading, concurring with other studies on global performance judgments in these domains in adults (for reading, see Golke et al., 2022; for mathematics, see Händel et al., 2020). They also tally up with past research on global self-evaluations in adulthood (e.g., mathematics self-concept, self-efficacy, Hart & Ganley, 2019; Huang, 2013) and align with gender-stereotypical connotations of mathematics and reading (e.g., Bonnot & Jost, 2014; Espinoza & Strasser, 2020), However, this should not be generalized, as the results typically differ depending on a specific task or domain and the type of judgment (see e.g., Händel et al., 2020; Pallier, 2003; Rivers et al., 2021).

Contrary to our hypotheses, we found no gender differences in bias. This was unexpected because we argued that gender stereotypes and gender differences in self-perceptions (e.g., Hart & Ganley, 2019; Huang, 2013) should reduce overestimation and make judgments in mathematics more accurate in women and judgments in reading more accurate in men. Moreover, in a study by Händel et al. (2020), female undergraduate students had a lower bias in their global performance judgments in a mathematics test. The pattern of gender differences in bias, just as in the case of performance judgments, seems highly dependent on the task or domain and the type of judgment (Pallier, 2003; Rivers et al., 2021). Overall, the topic requires further research, tapping into domain-related stereotypes and gender identity (for a discussion, see Golke et al., 2022).

This study included mathematics and reading. Although it is tempting to compare the results in the two domains to contribute to the discussion on domain-general versus domain-specific aspects of metacognitive monitoring (see e.g., Fleming et al., 2014; Gutierrez et al., 2016; Schraw et al., 1995), such direct comparisons are not informative. On one hand, the two competence domains tap into distinct types of knowledge, skills, and mental processes. On the other hand, the two tests differed in their solving probability, which likely affected both performance judgments and their accuracy (bias), making them incomparable. In other words, a more difficult test likely results in lower performance judgments. Moreover, since higher skills are linked with lower bias (e.g., Hacker et al., 2008), the performance judgments in a more difficult test are likely to be more biased.

Practical implications

Research on performance judgments often focuses on younger participants, typically in a (semi-)structured or institutionalized context. However, lifelong learning in various settings has increased in many societies, highlighting the importance of self-regulated learning and metacognitive abilities in older age groups. Our study indicates that adult learners may face increasing challenges in assessing their performance. While younger learners also overestimate their abilities, we provide some indication that older learners' accuracy decreases or, given the results of the sensitivity analyses, at least remains at the same level of overestimation. This should be relevant for teachers and educators involved in adult education. On the one hand, poor accuracy may lead to inadequate learning behavior and hinder performance (Dunlosky & Rawson, 2012), so educators/teachers should aim at improving monitoring abilities (e.g., Mikkilä-Erdmann & Iiskala, 2020). On the other hand, making their judgments more accurate, by effectively grounding their overly positive self-views, may diminish motivation, leading to poorer learning behavior (e.g., decreased learning time). Therefore, it might be important to also focus on motivational aspects and realistic learning goals when teaching adults (for a conceptual framework, see Sheffler et al., 2021).

Limitations and future directions

Our study focused on global performance judgments, so the results should not be generalized to other types of judgments. Performance judgments may differ in how fine-grained they are (global versus local) and when they are made (pre- versus postdictions; Händel et al., 2013; Schraw, 2009). They depend on task- and non-task-specific factors to a varying degree, and therefore carry different information and represent manifold aspects of metacognitive monitoring (e.g., Rutherford, 2017; Schraw, 2009). As a result, they may show a different pattern of age-, gender-, and education-related differences. Moreover, global performance judgments have been criticized because they do not entertain the possibility of differentiating between judgment bias and metacognitive sensitivity (Fleming & Lau, 2014). Future studies should also include other types of judgments because it could help in understanding the internal representations of learning processes better (Filippi et al., 2020).

More specifically, a central shortcoming of the current study is that we could only compare different age groups using cross-sectional data. In other words, we cannot comment on potential mechanisms that drive changes in how test performance is evaluated or what factors support stability in the accuracy of such judgments. Although empirical evidence on generational differences in, for example, personality or self-concept is mixed (for an overview, see Twenge et al., 2015), we cannot rule out that generational aspects, especially in combination with differences in educational attainment and gender inclusiveness, might influence how performance judgments are made and how they develop over time.

Overall, the models only explained a modest amount of variance for performance judgment, and even less so for the accuracy of performance judgment. We focused on participants' age, gender, and education, but there should be cognitive, emotional, and motivational factors that contribute in a more direct way to how such evaluations are made. Theoretically, we drew on, among others, the accessibility model that suggests that we use various sources of information for evaluating our performance in cognitive ability tests (Koriat et al., 2008). Regarding cognitive factors, familiarity, and working memory capacity have been identified as processes relevant to evaluating test characteristics (e.g., Touron

et al., 2010). Regarding emotional factors, test anxiety typically negatively influences performance judgments (for a student population, see Silaj et al., 2021), which might be especially relevant for the mathematics test (Hart & Ganley, 2019). Finally, motivational factors should be central to participants' engagement (i.e., achievement motivation) and what they think about their performance in general (i.e., self-efficacy, Jiang & Kleitman, 2015).

Conclusion

While the current study has several limitations, it is important to note that we used a large representative sample of adults of a broad age range living in Germany, whereas most studies on performance judgments typically focus on a selective sample of young adults (i.e., college students). Although further longitudinal data are needed, our findings highlight that older age groups might show lower global performance judgments in mathematical and reading competence. In general, individuals of all ages, both men and women tend to overestimate their performance. We consider this study a step further toward understanding changes in the self-assessment of cognitive abilities in adulthood.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11409-025-09416-2.

Acknowledgements The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

This paper uses data from the National Educational Panel Study (NEPS; see Blossfeld & Roßbach, 2019) Starting Cohort Adults (NEPS Network, 2023). From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS has been carried out by the Leibniz Institute for Educational Trajectories (LIfBi, Bamberg) in cooperation with a nationwide network.

Author's Contribution Maximilian Seitz (Conceptualization, Data curation, Methodology, Writing – Original draft preparation); Anna Hawrot (Conceptualization, Methodology, Writing – Original draft preparation, Writing – Review & Editing); Kathrin Lockl (Conceptualization, Methodology, Writing – Review & Editing).

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability All data are available for researchers; the analytic code is available upon request to the authors: https://doi.org/10.5157/NEPS:SC6:13.0.0

References

Aust, V., Gilberg, R., Hess, D., Kersting, A., Kleudgen, M., & Steinwede, A. (2012). *Methodenbericht: NEPS Startkohorte 6 Haupterhebung 2010/2011 B67 [NEPS Starting Cohort 6 methods report*

2010/2011 B67]. infas Institut für angewandte Sozialwissenschaft GmbH. https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC6/3-0-0/Methodenbericht_SC6_W3_B67.pdf

Babyak, M. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine, 66*(3), 411–421. https://doi.org/10.1097/00006842-200405000-00021

Blossfeld, H.-P., spsampsps Roßbach, H.-G. (2019). *Education as a lifelong process: The German National Educational Panel Study (NEPS)*. Springer VShttps://doi.org/10.1007/978-3-658-23162-0

Bonnot, V., & Jost, J. T. (2014). Divergent effects of system justification salience on the academic self-assessments of men and women. *Group Processes & Intergroup Relations, 17*(4), 453–464. https://doi.org/10.1177/1368430213512008

Burns, K. M., Burns, N. R., & Ward, L. (2016). Confidence—More a personality or ability yrait? It depends on How it is measured: A comparison of young and older adults. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00518

Dahl, M., Allwood, C. M., Rennemark, M., & Hagberg, B. (2010). The relation between personality and the realism in confidence judgements in older adults. *European Journal of Ageing, 7*(4), 283–291. https://doi.org/10.1007/s10433-010-0164-2

Daniels, K. A., Toth, J. P., & Hertzog, C. (2009). Aging and recollection in the accuracy of judgments of learning. *Psychology and Aging, 24*(2), 494–500. https://doi.org/10.1037/a0015269

Devolder, P. A. (1993). Adult age differences in monitoring of practical problem-solving performance. *Experimental Aging Research, 19*(2), 129–146. https://doi.org/10.1080/03610739308253927

Devolder, P. A., Brigham, M. C., & Pressley, M. (1990). Memory performance awareness in younger and older adults. *Psychology and Aging, 5*(2), 291–303. https://doi.org/10.1037/0882-7974.5.2.291

Dodson, C. S. (2017). Aging and memory. In *Learning and memory: A comprehensive reference* (pp. 403–421). Elsevier. https://doi.org/10.1016/B978-0-12-809324-5.21053-5

Dunlosky, J., Mueller, M. L., spsampsps Thiede, K. W. (2015). Methodology for investigating human metamemory: Problems and pitfalls. In J. Dunlosky spsampsps S. K. Tauber (Eds.), *The Oxford handbook of metamemory* (pp. 23–38). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199336746.001.0001

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction, 22*(4), 271–280. https://doi.org/10.1016/j.learninstruc.2011.08.003

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5*(3), 69–106. https://doi.org/10.1111/j.1529-1006.2004.00018.x

Eakin, D. K., Hertzog, C., & Harris, W. (2014). Age invariance in semantic and episodic metamemory: Both younger and older adults provide accurate feeling-of-knowing for names of faces. *Aging, Neuropsychology, and Cognition, 21*(1), 27–51. https://doi.org/10.1080/13825585.2013.775217

Ehrlinger, J., & Dunning, D. (2003). How chronic self-views influence (and potentially mislead) estimates of performance. *Journal of Personality and Social Psychology, 84*(1), 5–17. https://doi.org/10.1037/0022-3514.84.1.5

Espinoza, A. M., & Strasser, K. (2020). Is reading a feminine domain? The role of gender identity and stereotypes in reading motivation in Chile. *Social Psychology of Education, 23*(4), 861–890. https://doi.org/10.1007/s11218-020-09571-1

Filippi, R., Ceccolini, A., Periche-Tomas, E., & Bright, P. (2020). Developmental trajectories of metacognitive processing and executive function from childhood to older age. *Quarterly Journal of Experimental Psychology, 73*(11), 1757–1773. https://doi.org/10.1177/1747021820931096

Flavell, J. H., Miller, P. H., & Miller, S. A. (2002). *Cognitive development* (4th ed.). Pearson.

Fleming, S. M., Ryu, J., Golfinos, J. G., & Blackmon, K. E. (2014). Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain*, *137*(10), 2811–2822. https://doi.org/10.1093/brain/awu221

Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society b: Biological Sciences, 367*(1594), 1338–1349. https://doi.org/10.1098/rstb.2011.0417

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience, 8*(443), 1–9. https://doi.org/10.3389/fnhum.2014.00443

Ganzeboom, H. B., De Graaf, P. M., & Treiman, D. J. (1992). A standard international socio-economic index of occupational status. *Social Science Research*, *21*(1), 1–56. https://doi.org/10.1016/0049-089X(92)90017-B

Gebel, M., & Pfeiffer, F. (2010). Educational expansion and its heterogeneous returns for wage workers. *Journal of Contextual Economics-Schmollers Jahrbuch, 130*(1), 19–42. https://doi.org/10.3790/schm.130.1.19

Gehrer, K., Zimmermann, S., Artelt, C., & Weinert, S. (2012). *The assessment of reading competence (including sample items for Grade 5 and 9)* [Scientific Use File 2012, Version 1.0.0]. University of Bamberg, National Educational Panel Study.

Gehrer, K., Zimmermann, S., Artelt, C., & Weinert, S. (2013). NEPS framework for assessing reading competence and results from an adult pilot study. *Journal for Educational Research Online*, *5*(2), 50–79. https://doi.org/10.25656/01:8424

Golke, S., Steininger, T., & Wittwer, J. (2022). What makes learners overestimate their text comprehension? The impact of learner characteristics on judgment bias. *Educational Psychology Review, 34*(4), 2405–2450. https://doi.org/10.1007/s10648-022-09687-0

Gutierrez, A. P., Schraw, G., Kuch, F., & Richmond, A. S. (2016). A two-process model of metacognitive monitoring: Evidence for general accuracy and error factors. *Learning and Instruction*, 44, 1–10. https://doi.org/10.1016/j.learninstruc.2016.02.006

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, 92*(1), 160–170. https://doi.org/10.1037/0022-0663.92.1.160

Hacker, D. J., Bol, L., & Kneer, M. C. (2008). Metacognition in education: A focus on calibration. In J. Dunlosky & R. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 429–455). Psychology Press.

Hager, P. J. (2020). Concepts and definitions of lifelong learning. In M. London (Ed.), *The Oxford Handbook of Lifelong Learning, Second Edition*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197506707.013.2

Hammon, A., Zinn, S., Aßmann, C., & Würbach, A. (2016). *Samples, weights, and nonresponse: The adult cohort of the National Educational Panel Study (wave 2 to 6)*. Leibniz Institute for Educational Trajectories. https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC6/9-0-0/SC6_6-0-0_W.pdf

Händel, M., Artelt, C., & Weinert, S. (2013). Assessing metacognitive knowledge: Development and evaluation of a test instrument. *Journal of Educational Research Online, 5*(2), 162–188. https://doi.org/10.25656/01:8429

Händel, M., de Bruin, A. B. H., & Dresel, M. (2020). Individual differences in local and global metacognitive judgments. *Metacognition and Learning, 15*(1), 51–75. https://doi.org/10.1007/s11409-020-09220-0

Hardt, K., Pohl, S., Haberkorn, K., & Wiegand, E. (2013). *NEPS technical report for reading – Scaling results of Starting Cohort 6 for adults in main study 2010/11* (NEPS Working Paper 25). Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://www.lifbi.de/Portals/2/Working%20Papers/WP_XXV.pdf

Hart, S. A., & Ganley, C. M. (2019). The nature of math anxiety in adults: Prevalence and correlates. *Journal of Numerical Cognition, 5*(2), 122–139. https://doi.org/10.5964/jnc.v5i2.195

Hartshorne, J. K., & Germine, L. T. (2015). When does cognitive functioning peak? The asynchronous rise and fall of different cognitive abilities across the life span. *Psychological Science, 26*(4), 433–443. https://doi.org/10.1177/0956797614567339

Haslam, C., Morton, T. A., Haslam, S. A., Varnes, L., Graham, R., & Gamaz, L. (2012). "When the age is in, the wit is out": Age-related self-categorization and deficit expectations reduce performance on clinical tests used in dementia assessment. *Psychology and Aging, 27*(3), 778–784. https://doi.org/10.1037/a0027754

Hertzog, C., Saylor, L. L., Fleece, A. M., & Dixon, R. A. (1994). Metamemory and aging: Relations between predicted, actual and perceived memory task performance. *Aging, Neuropsychology, and Cognition, 1*(3), 203–237. https://doi.org/10.1080/13825589408256577

Hertzog, C., & Shing, Y. L. (2011). Memory development across the life span. In K. L. Fingerman, C. A. Berg, J. Smith, & T. C. Antonucci, *Handbook of life-span development* (pp. 299–330). Springer New York.

Huang, C. (2013). Gender differences in academic self-efficacy: A meta-analysis. *European Journal of Psychology of Education, 28*(1), 1–35. https://doi.org/10.1007/s10212-011-0097-y

Jiang, Y., & Kleitman, S. (2015). Metacognition and motivation: Links between confidence, self-protection and self-enhancement. *Learning and Individual Differences, 37*, 222–230. https://doi.org/10.1016/j.lindif.2014.11.025

Jordan, A.-K., & Duchhardt, C. (2013). *NEPS technical report for mathematics—Scaling results of Starting Cohort 6–Adults* (NEPS Working Paper 32). Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://www.lifbi.de/Portals/2/Working%20Papers/WP_XXXII.pdf

Kelemen, W. L., Frost, P. J., & Weaver, C. A. (2000). Individual differences in metacognition: Evidence against a general metacognitive ability. *Memory & Cognition, 28*(1), 92–107. https://doi.org/10.3758/BF03211579

Kidder, D. P., Park, D. C., Hertzog, C., & Morrell, R. W. (1997). Prospective memory and aging: The effects of working memory and prospective memory task load. *Aging, Neuropsychology, and Cognition, 4*(2), 93–112. https://doi.org/10.1080/13825589708256639

Koriat, A., Nussinson, R., Bless, H., spsampsps Shaked, N. (2008). Information-based and experience-based metacognitive judgments. In J. Dunlosky spsampsps R. Bjork (Eds.), *Handbook of Metamemory and Memory* (pp. 117–135). Psychology Press. https://doi.org/10.4324/9780203805503.ch7

Lacreuse, A., Raz, N., Schmidtke, D., Hopkins, W. D., & Herndon, J. G. (2020). Age-related decline in executive function as a hallmark of cognitive ageing in primates: An overview of cognitive and neurobiological studies. *Philosophical Transactions of the Royal Society b: Biological Sciences, 375*(1811), 20190618. https://doi.org/10.1098/rstb.2019.0618

Maki, R. H., & Serra, M. (1992). The basis of test predictions for text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(1), 116–126. https://doi.org/10.1037/0278-7393.18.1.116

Marsh, H. W., Pekrun, R., & Lüdtke, O. (2022). Directional ordering of self-concept, school grades, and standardized tests over five years: New tripartite models juxtaposing within- and between-person perspectives. *Educational Psychology Review, 34*, 2697–2744. https://doi.org/10.1007/s10648-022-09662-9

McWilliams, A., Bibby, H., Steinbeis, N., David, A. S., & Fleming, S. M. (2023). Age-related decreases in global metacognition are independent of local metacognition and task performance. *Cognition, 235*, 105389. https://doi.org/10.1016/j.cognition.2023.105389

Mejía-Rodríguez, A. M., Luyten, H., & Meelissen, M. R. M. (2021). Gender differences in mathematics self-concept across the world: An exploration of student and parent data of TIMSS 2015. *International Journal of Science and Mathematics Education, 19*(6), 1229–1250. https://doi.org/10.1007/s10763-020-10100-x

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science, 18*(3), 159–163. https://doi.org/10.1111/j.1467-8721.2009.01628.x

Mikkilä-Erdmann, M., & Iiskala, T. (2020). Developing learning and teaching practices for adults. In E. K. Kallio (Ed.), *Development of adult thinking* (pp. 123–140). Routledge.

Nelson, T. O. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. 26, pp. 125–173). Elsevier. https://doi.org/10.1016/S0079-7421(08)60053-5

Nelson, T. O., spsampsps Narens, L. (1994). Why investigate metacognition? In J. Metcalfe spsampsps A. P. Shimamura (Eds.), *Metacognition* (pp. 1–26). The MIT Press. https://doi.org/10.7551/mitpress/4561.003.0003

Neumann, I., Duchhardt, C., Grüßing, M., Heinze, A., Knopp, E., & Ehmke, T. (2013). Modeling and assessing mathematical competence over the lifespan. *Journal for Educational Research Online*, *5*(2), 80–109. https://doi.org/10.25656/01:8426

Nguyen, C., Leanos, S., Natsuaki, M. N., Rebok, G. W., & Wu, R. (2018). Adaptation for growth via learning new skills as a means to long-term functional independence in older adulthood: Insights from emerging adulthood. *The Gerontologist*https://doi.org/10.1093/geront/gny128

OECD. (2019). *PISA 2018 results (Volume II): Where all students can succeed*. PISA, OECD Publishing-https://doi.org/10.1787/5f07c754-en

OECD spsampsps Statistics Canada. (2011). *Literacy for life: Further results from the adult literacy and life skills survey*. OECDhttps://doi.org/10.1787/9789264091269-en

Overhoff, H., Ko, Y. H., Feuerriegel, D., Fink, G. R., Stahl, J., Weiss, P. H., Bode, S., & Niessen, E. (2021). Neural correlates of metacognition across the adult lifespan. *Neurobiology of Aging, 108*, 34–46. https://doi.org/10.1016/j.neurobiolaging.2021.08.001

Pallier, G. (2003). Gender differences in the self-assessment of accuracy on cognitive tasks. *Sex Roles, 48*(5/6), 256–276. https://doi.org/10.1023/A:1022877405718

Pennequin, V., Sorel, O., & Mainguy, M. (2010). Metacognition, Executive Functions and Aging: The Effect of Training in the Use of Metacognitive Skills to Solve Mathematical Word Problems. *Journal of Adult Development, 17*(3), 168–176. https://doi.org/10.1007/s10804-010-9098-3

Perrotin, A., Tournelle, L., & Isingrini, M. (2008). Executive functioning and memory as potential mediators of the episodic feeling-of-knowing accuracy. *Brain and Cognition, 67*(1), 76–87. https://doi.org/10.1016/j.bandc.2007.11.006

R Core Team. (2023). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rivers, M. L., Fitzsimmons, C. J., Fisk, S. R., Dunlosky, J., & Thompson, C. A. (2021). Gender differences in confidence during number-line estimation. *Metacognition and Learning, 16*(1), 157–178. https://doi.org/10.1007/s11409-020-09243-7

Roebers, C. M. (2002). Confidence judgments in children's and adult's event recall and suggestibility. *Developmental Psychology, 38*(6), 1052–1067. https://doi.org/10.1037/0012-1649.38.6.1052

Roebers, C. M., & Feurer, E. (2016). Linking executive functions and procedural metacognition. *Child Development Perspectives, 10*(1), 39–44. https://doi.org/10.1111/cdep.12159

Rutherford, T. (2017). The measurement of calibration in real contexts. *Learning and Instruction, 47*, 33–42. https://doi.org/10.1016/j.learninstruc.2016.10.006

Schneider, S. L. (2008). Applying the ISCED-97 to the German educational qualifications. In S. L. Schneider (Ed.), *The International Standard Classification of Education (ISCED-97). An evaluation of content and criterion validity for 15 European countries* (pp. 76–102). Mannheim Centre for European Social Research.

Schneider, W., & Lockl, K. (2008). Procedural metacognition in children: Evidence for developmental trends. In J. Dunlosky & R. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 391–409). Psychology Press.

Schneider, W., Tibken, C., spsampsps Richter, T. (2022). The development of metacognitive knowledge from childhood to young adulthood: Major trends and educational implications. In *Advances in Child Development and Behavior* (Vol. 63, pp. 273–307). Elsevier. https://doi.org/10.1016/bs.acdb.2022.04.006

Schnittjer, I., & Duchhardt, C. (2015). *Mathematical competence: Framework and exemplary test items*. Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/Kompetenzen/NEPS_com_ma_2015_en.pdf

Schraw, G. (2009). Measuring metacognitive judgments. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 415–429). Routledge.

Schraw, G., Dunkle, M. E., Bendixen, L. D., & Roedel, T. D. (1995). Does a general monitoring skill exist? *Journal of Educational Psychology, 87*(3), 433–444. https://doi.org/10.1037/0022-0663.87.3.433

Sheffler, P., Rodriguez, T. M., Cheung, C. S., & Wu, R. (2021). Cognitive and metacognitive, motivational, and resource considerations for learning new skills across the lifespan. *Wires Cognitive Science, 13*(2), e1585. https://doi.org/10.1002/wcs.1585

Silaj, K. M., Schwartz, S. T., Siegel, A. L. M., & Castel, A. D. (2021). Test anxiety and metacognitive performance in the classroom. *Educational Psychology Review, 33*(4), 1809–1834. https://doi.org/10.1007/s10648-021-09598-6

StataCorp. (2021). *Stata statistical software: Release 17*. StataCorp LLC.

Steffens, M. C., & Jelenec, P. (2011). Separating implicit gender stereotypes regarding math and language: Implicit ability stereotypes are self-serving for boys and Men, but not for girls and women. *Sex Roles, 64*(5–6), Article 5–6. https://doi.org/10.1007/s11199-010-9924-x

Stöckinger, C., Kretschmer, S., & Kleinert, C. (2018). *Panel attrition in NEPS Starting Cohort 6: A description of attrition processes in waves 2 to 7 with regard to nonresponse bias (NEPS Survey Paper No. 35)*. Leibniz Institute for Educational Trajectories. https://doi.org/10.5157/NEPS:SP35:1.0

Stuss, D. T. (2011). Functions of the frontal lobes: Relation to executive functions. *Journal of the International Neuropsychological Society, 17*(05), 759–765. https://doi.org/10.1017/S1355617711000695

Touron, D. R., Oransky, N., Meier, M. E., & Hines, J. C. (2010). Metacognitive monitoring and strategic behaviour in working memory performance. *Quarterly Journal of Experimental Psychology, 63*(8), 1533–1551. https://doi.org/10.1080/17470210903418937

Twenge, J. M., Gentile, B., & Campbell, W. K. (2015). Birth cohort differences in personality. In M. Mikulincer, P. R. Shaver, M. L. Cooper, & R. J. Larsen (Eds.), *APA handbook of personality and social psychology. Volume 4: Personality processes and individual differences* (pp. 535–552). American Psychological Association.

United Nations Educational, Scientific, and Cultural Organization (UNESCO). (2006). *International Standard CLassification of Education ISCED 1997*. UNESCO Institute for Statistics. https://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-1997-en_0.pdf

Vaportzis, E., & Gow, A. J. (2018). People's beliefs and expectations about how cognitive skills change with age: Evidence from a U.K.-wide aging survey. *The American Journal of Geriatric Psychiatry, 26*(7), 797–805. https://doi.org/10.1016/j.jagp.2018.03.016

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semi-iparametric generalized linear models. *Journal of the Royal Statistical Society (b), 73*(1), 3–36. https://doi.org/10.1111/j.1467-9868.2010.00749.x

Wood, S. N. (2017). *Generalized additive models: An introduction with R*. Chapman and Hall/CRChttps://doi.org/10.1201/9781315370279

Wu, M., & Was, C. A. (2023). The relationship between executive functions and metacognition in college students. *Journal of Intelligence, 11*(12), 220. https://doi.org/10.3390/jintelligence11120220

Zhao, Q., & Linderholm, T. (2008). Adult metacomprehension: Judgment processes and accuracy constraints. *Educational Psychology Review, 20*(2), 191–206. https://doi.org/10.1007/s10648-008-9073-8

Zinn, S., Würbach, A., Steinhauer, H. W., & Hammon, A. (2020). Attrition and selectivity of the NEPS starting cohorts: An overview of the past 8 years. *AStA Wirtschafts- und Sozialstatistisches Archiv, 14*(2), 163–206. https://doi.org/10.1007/s11943-020-00268-7