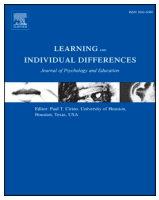
Learning and Individual Differences 122 (2025) 102764

Contents lists available at ScienceDirect

Learning and Individual Differences

journal homepage: www.elsevier.com/locate/lindif



The accuracy of performance judgements and academic achievement: A two-sample two-wave study of German primary and lower secondary school students

Anna Hawrot *, Timo Gnambs, Kathrin Lockl

Leibniz Institute for Educational Trajectories (LIfBi), Bamberg, Germany

ARTICLE INFO

Keywords:
Academic achievement
Performance judgements
Metacognitive monitoring
Procedural metacognition
Response surface analysis

ABSTRACT

In metacognition research, performance judgements and their accuracy are considered pivotal for self-regulated learning and task performance. However, their long-term impact on academic achievement remains under-researched. This study investigated the role of performance judgements and their accuracy for later maths competencies and explored whether this relationship varied with age. We used data on student performance judgements in a maths test, actual performance, and performance in a maths test two years later collected from 5551 German primary and 4780 lower secondary school students. Response surface analyses supported none of the five competing hypotheses that we investigated. They indicated the dominant role of past competencies and a positive, although weaker, effect of judgements, especially at high competence levels. Students in both samples overestimated their performance, with secondary school students being more accurate. The study suggests refining theoretical models to better link past performance, performance judgements, and accuracy to short- and long-term achievement.

Educational relevance statement

The accuracy of students' self-evaluation of their skills and performance is considered crucial, as it influences their learning behaviour. However, the long-term impact of accurate self-evaluations on performance remains uncertain. This study revealed no link between the accuracy of self-evaluated performance in a maths test and performance in a maths test two years later. Instead, initial performance was found to be most important for later performance. We also observed a positive, though weaker, link between the overall level of self-evaluation and performance two years later, especially in students who performed high in the initial test.

1. Introduction

Research on self-evaluations has revealed that people often hold inaccurate, usually overestimated, perceptions of their skills, character, and performance. This puzzling result triggered numerous studies in various sub-disciplines of psychology trying to shed light on the mechanisms leading to such inaccuracies on the one hand and their potential consequences on the other (see e.g., Dunning et al., 2004 for a review). For instance, metacognition research has investigated the role that the accuracy of monitoring one's own learning plays for memory retention (e.g., Dunlosky & Rawson, 2012), educational psychology has explored how the accuracy of self-perceptions of ability, as an aspect of motivation, affect academic achievement (e.g., Paschke et al., 2020), whereas social and personality psychology has researched the link between self-enhancement, a trait-like tendency to under- or overestimate own skills and capabilities, and personal adjustment (e.g., Humberg et al., 2018).

The question of the consequences of flawed self-evaluations seems particularly important in the school context because they may affect student learning behaviour (e.g., Metcalfe, 2009). Many researchers, especially in the field of metacognition, have postulated that accurate or only slightly inflated judgements of performance in learning tasks are optimal for academic achievement because they allow students to recognise whether a learning goal has been met and, therefore, enable proper self-regulation of learning (e.g., Rutherford, 2017a; Stone, 2000; Thiede & Dunlosky, 1999; Winne, 2011). However, although it is well documented that children of primary and secondary school age often overestimate their task performance (e.g., García et al., 2016; Mirandola et al., 2018; Roebers & Spiess, 2017), little is known about how the evaluation of one's performance influences performance in the long run.

* Corresponding author at: Leibniz Institute for Educational Trajectories, Wilhelmsplatz 3, 96047 Bamberg, Germany.
E-mail address: anna.hawrot@lifbi.de (A. Hawrot).

https://doi.org/10.1016/j.lindif.2025.102764
Received 11 November 2024; Received in revised form 12 June 2025; Accepted 18 July 2025
Available online 23 July 2025

Therefore, it remains unclear whether accurate performance judgements are beneficial or not in the long run.

In contrast, research on the accuracy of another type of self-evaluations, that is self-perceptions of ability (e.g., academic self-concept or self-efficacy), has emphasised that not accurate but over-optimistic self-evaluations, as key for maintaining motivation and task persistence, may positively affect achievement (e.g., Bouffard et al., 2011; Taylor & Brown, 1988). However, empirical evidence on the link is mixed (e.g., Chiu & Klassen, 2009, 2010; Gresham et al., 2000; Paschke et al., 2023).

Irrespective of the hypothesized link to achievement, both strands of research have struggled with the methodological problem of relating the accuracy of self-evaluations to actual performance, which may at least partially explain the inconsistent results of previous studies. Typical accuracy measures such as difference and residual scores, although commonly used, are confounded with self-evaluations, making it impossible to differentiate the effect of accuracy from the effect of self-evaluations on potential outcomes (see Humberg et al., 2018 for a detailed discussion of the problem). However, Humberg, Nestler, and Back (2019) have recently proposed response surface analyses (Edwards & Parry, 1993) as a novel method for verifying hypotheses on the link between the accuracy of self-estimates and various outcomes. The method overcomes the limitations of previous methodological and statistical approaches, allowing the detection of the effects of accuracy. In their research, Humberg, Dufner, et al. (2019) focused on self-enhancement and its role for personal adjustment. However, to our knowledge, the method has not yet been used in research on metacognition.

Therefore, this study investigates the role that the accuracy of performance judgements plays for later student academic competencies, while taking into account that the link may change with student age. To this end, it uses data on student judgements of their performance in a maths competence test, actual performance in the test, and performance in a maths competence test two years later collected from two independent nationwide samples of German primary and lower secondary school students. Moreover, to avoid the methodological shortcomings of past studies, it employs the analytical strategy proposed by Humberg, Nestler, and Back (2019).

2. Theoretical framework

In the field of metacognition research, performance judgements are seen as indicators of metacognitive processes—also referred to as procedural metacognitive skills (e.g., 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 you are in your goal of understanding and remembering which serves as the basis for self-regulation and self-initiated learning behaviour (Nelson, 1990; Nelson & Narens, 1994). Thus, metacognitive monitoring reflects an individual's ability to evaluate how well he or she is progressing. Since performance judgements serve as the basis for self-regulation, it is important how accurate these judgements are. Goal-directed self-regulation seems to be possible only if performance judgements are realistic. Accurate judgements allow appropriate strategies to be chosen and learning time to be allocated so that optimal learning outcomes can be achieved (Dunlosky & Rawson, 2012).

In studies that examine metacognitive monitoring skills, students are typically asked to judge their performance before, during or after working on a memory task or 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 judgements differ depending on the time point of assessment but also on whether they relate to individual items or the entire test (Hacker et al., 2008). When it comes to estimating one's overall performance in advance or retrospectively, this aspect is also referred to as absolute accuracy or 'calibration' and it concerns the correspondence between predicted or postdicted and actual overall performance. On the other hand, relative accuracy or 'resolution' describes the accuracy in monitoring the relative recallability or comprehension of different items. The focus of this paper is on absolute accuracy.

Overall, research indicates that students often overestimate their performance, with lower performing students exhibiting more overconfidence than higher performing students (Hacker et al., 2000). Moreover, judgements relating to the entire test (aggregate items) usually lead to less overestimation compared to judgements that refer to individual items (item-by-item judgements), also known as the aggregation effect (Gigerenzer et al., 1991; Griffin & Tversky, 1992). Furthermore, postdictions tend to be more realistic than predictions because predictions are mainly based on expectations of what may happen, whereas postdictions rely more on individuals' experience concerning the content of the test and one's performance (Hacker et al., 2000). Thus, a condition in which aggregate judgements are assessed after the test should result in relatively little overestimation.

2.1. The accuracy of performance judgements and academic achievement

The relationships between performance judgements and academic achievement can be summarized with five competing hypotheses. Two of them are discussed in research on metacognition (the self-knowledge and optimal margin hypotheses), whereas the remaining three are popular mostly in research on self-perceptions (e.g., self-efficacy or self-concept, Humberg, Dufner, et al., 2019; Paschke et al., 2023). Although the latter hypotheses originate from another strand of research, self-perceptions of ability, as subjective assessments of one's performance and capabilities in specific domains (Marsh et al., 2017), bear similarities with performance judgements, making the three hypotheses worth discussing as well. Specifically, similarly to performance judgements, ability self-perceptions require self-monitoring, which belongs to metacognitive experiences (Efklides, 2011). Moreover, they also are supposed to affect learning, and, as a result, academic achievement (e.g., Marsh et al., 2017), although the expected mechanism is motivational rather than metacognitive. However, unlike performance judgements, they represent metacognitive knowledge about the self and own strengths and weaknesses rather than the metacognitive monitoring of learning (Efklides, 2011).

Although we discuss the five hypotheses in detail, due to methodological problems with separating the effect of self-evaluations from the effect of their accuracy (see e.g., Humberg et al., 2018; Humberg, Dufner, et al., 2019), we will refrain from discussing empirical evidence for and against each of them.

2.1.1. Self-knowledge hypothesis

The first hypothesis, called the self-knowledge hypothesis, is dominant in metacognition research. It states that accurate performance judgements are optimal for academic achievement, with both over- and underestimation having a detrimental effect. It is because accurate judgements, as an element of self-monitoring, are considered necessary for recognising whether a learning goal has been achieved. Therefore, they shape learning behaviour, learning effort, and enable proper self-regulation of learning (e.g., Dunlosky & Rawson, 2012; Hacker and Bol, 2019; Rutherford, 2017b; Stone, 2000; Thiede & Dunlosky, 1999; Winne, 2011). Failing to recognise the lack of sufficient understanding may prevent students from adjusting their strategies and allocating resources (e.g., time, effort) to a learning task, as well as diminish their ability to meet future learning goals that build on that material. Similarly, failing to recognise that a learning goal has been met may, for instance, lengthen learning time without improving performance ('the labour-in-vain effect', e.g., Nelson & Leonesio, 1988), or leave limited resources for other learning tasks (e.g., Hacker and Bol, 2019; Winnie, 2011), amounting to poorer achievement as a result.

2.1.2. Optimal margin hypothesis

However, it has been occasionally suggested that slightly inflated performance judgements may be beneficial for achievement, at least in certain conditions, because they allow maintaining motivation (Norman, 2020; Zimmerman & Moylan, 2009). Such a notion aligns with the theories that indicate motivation as an important prerequisite for self-regulated learning, including self-monitoring (e.g., Zimmerman, 2011), which in turn is linked with achievement (e.g., Guo, 2022). It also tallies up with models that integrate and interconnect personal dispositions (including self-perceptions of ability) and metacognition (e.g., the MASRL model, Efklides, 2011). The hypothesis, although occasionally mentioned in metacognition research, originates from the theory of the optimal margin of illusion (Baumeister, 1989), which refers to self-perceptions of ability. The theory postulates that slightly to moderately inflated perceptions of the self and world may yield affective benefits that allow maintaining a healthy level of motivation and effort. In contrast, seeing the self and the world accurately or worse than reality is associated with lowered mood and motivation. Overly inflated perceptions, in turn, lead to negative consequences as well, with judgement errors, self-handicapping behaviours, or maladaptive persistence on unachievable tasks as examples (Baumeister, 1989; Lopez et al., 1998). As a result, low, accurate, and overly positive perceptions are expected to lower achievement.

2.1.3. Beneficial self-evaluation bias hypothesis

The first hypothesis that is not prevalent in metacognition research but popular in studies on the accuracy of self-perceived abilities (self-concept, self-efficacy; e.g., Bouffard et al., 2011; He & Côté, 2023; Paschke et al., 2023) is the beneficial self-evaluation bias hypothesis. First proposed, not without controversy, by Taylor and Brown (1988), it postulates that inflated self-perceptions promote motivation, engagement, and persistence, ultimately leading to better performance. Analogously, underestimated self-perceptions are detrimental because they do not foster such qualities. In the school context, inflated academic self-concept and academic self-efficacy have been discussed as forces promoting interest, persistence, motivation, and confidence in actions that increase the chances of success, as well as overall school adjustment (e.g., Bouffard et al., 2011; Chen, 2003; Gonida & Leondari, 2011; Martin & Debus, 1998), all being factors that translate into achievement. However, the hypothesis has been criticised, among others, for not making a clear distinction between the effects of the accuracy of self-perceptions and self-perceptions themselves. For example, it remains unclear why it is the positive bias and not the positive self-view itself that fuels motivation and effort. Readers interested in this hypothesis are referred to the thorough summary of this discussion in Paschke et al. (2023).

2.1.4. Detrimental self-evaluation bias hypothesis

Occasionally, overestimated self-perceptions have also been discussed as either associated with various negative phenomena (e.g., ego involvement, narcissism; Grijalva & Zhang, 2016; Robins & Beer, 2001) or simply detrimental, especially in the long run (Robins & Beer, 2001). In the academic context, such overestimation has been suggested to contribute to disengagement and decreased personal importance of school and learning when one's performance is consistently lower than expected (Robins & Beer, 2001). Overestimation may also prompt students to use self-protective patterns of causal attributions, with failure attributed to external and uncontrollable causes (e.g., luck), preventing them from identifying the true reasons behind achievement below expectations and further lowering it (Sticca et al., 2017). Additionally, it may lead to unrealistic study patterns (Rohr & Ayers, 1973), for instance, the underestimation of time and effort necessary to perform at a certain level, causing insufficient resource allocation and ultimately poorer achievement (Sticca et al., 2017; Talsma et al., 2019). Underestimation, in turn, may induce anxiety and serve, therefore, as a motivational technique that increases effort (Rohr & Ayers, 1973). In other words, the detrimental self-evaluation bias hypothesis is opposite to the beneficial self-evaluation bias hypothesis.

2.1.4.1. Positive self-evaluation hypothesis. The positive self-evaluation hypothesis states that self-perceptions of ability, for instance, self-concept or self-efficacy, affect performance, regardless of their accuracy or bias (e.g., Schunk & Mullen, 2012). In the academic context, the positive relationship results from students with more positive self-views having stronger motivation and engaging more in behaviours that support achievement, like putting more effort into learning, exhibiting persistence in the face of difficulty (e.g., Doménech-Betoret et al., 2017; Guay et al., 2010; Trautwein et al., 2009). Multiple studies have revealed the positive relationships between school- and learning-related self-perceptions and academic achievement, although they also indicate that the link is not unidirectional but reciprocal (e.g., Talsma et al., 2018; Wu et al., 2021).

2.2. The role of age

In school children and adolescents, academic achievement (e.g., Freund et al., 2021; Rescorla & Rosenthal, 2004; Shin et al., 2013) and metacognitive skills (e.g., Bayard et al., 2021; Schneider et al., 2022) improve with age. As children progress through school, they become more independent and self-directed in learning (Harding et al., 2019; Zimmerman & Martinez-Pons, 1990). This is reflected, among others, in declining parental assistance in homework, particularly in higher grades (e.g., Williams et al., 2002).

Looking at metacognitive skills, research suggests that preschool, kindergarten and young primary school children are particularly optimistic and significantly overestimate their performance (Schneider & Lockl, 2008). For example, in a study by Lipko et al. (2009), five-year-old children were asked to estimate how many pictures they would recall in a memory experiment. The results showed that the children were overconfident about their future memory performance and, interestingly, this overconfidence persisted over several trials even though the children had the experience of recalling considerably fewer pictures than they had predicted. In comparison, older primary school children have been shown to provide more accurate performance judgements than younger primary school children (Pressley et al., 1987). Additionally, age-related improvements in children's monitoring have also been reported for confidence judgements, particularly when it comes to the differentiation between correct and incorrect answers. More specifically, confidence judgements tend to be more accurate in older school children and adolescents because they feel more uncertain when they give incorrect responses compared to younger school children (Roebers, 2002; von der Linden & Roebers, 2006). Overall, considerable evidence across various monitoring indicators suggests that metacognitive monitoring substantially improves during the primary and early secondary school years (Schneider et al., 2022).

Developmental trends have also been observed in the extent to which students use their monitoring skills to regulate their learning behaviour. That is, higher correlations between measures of monitoring and control have been found in children aged nine years or older compared to children aged seven or eight years, suggesting that as students grow older, their learning behaviour is more strongly based on their monitoring judgements (Krebs & Roebers, 2010; Lockl & Schneider, 2003; Metcalfe & Finn, 2013). Thus, age-related improvements occur not only in monitoring accuracy but also in the way monitoring is used to regulate one's own learning behaviour. As demonstrated by O'Leary and Sloutsky (2017), a difficulty for young children seems to consist particularly in their inability to initiate metacognitive monitoring and control on their own. Their study showed that, when provided with feedback and a concrete strategy, children as young as five years old improved their monitoring and exhibited evidence of control, but they did not do so spontaneously.

Although these studies have investigated the short-term consequences of monitoring on students' learning behaviour, it is conceivable that both aspects of children's metacognitive skills (monitoring and monitoring-based control) are important for their learning behaviour in the long term. Young children's optimism when judging their performance may lead them to believe that they do not need to search for more adaptive strategies or allocate more study time while learning. When little adaptive learning behaviour is shown over many occasions, it could accumulate over time and lead to poorer learning outcomes in the long term. As children grow older, their overconfidence typically decreases and their ability to base their learning behaviour on self-monitoring increases. This may prompt them to recognise learning gaps, to seek more adaptive strategies or spend more time studying, which may eventually lead to higher learning growth.

Consequently, improvements in students' metacognitive skills may contribute to their progress in academic achievement and the influence of metacognitive skills may increase as students' progress through the school. To our knowledge, such long-term consequences have not been explicitly tested yet.

2.3. Confounding of the effects of accuracy and self-evaluations

Investigating the role of accuracy of self-evaluations for various outcomes, despite its intuitive simplicity, has proved to be challenging. Past research has typically used two types of accuracy measures (Humberg et al., 2018): difference scores (e.g., Sticca et al., 2017; Talsma et al., 2019) and residual scores (e.g., Bouffard et al., 2011; Gonida & Leondari, 2011; Robins & Beer, 2001). Difference scores are calculated by subtracting self-evaluations (e.g., the estimated number of correctly solved tasks in a test) from a criterion measure (e.g., the actual number of correctly solved tasks). Residual scores are the residuals from the regression of self-evaluations on a criterion measure. Because both measures of accuracy are correlated with self-evaluations, the analyses employing them confound the effect of accuracy with the effect of self-evaluations themselves. This may lead to incorrect conclusions that accuracy is linked with an outcome when, in fact, only self-evaluations are (Humberg et al., 2018). This unfavourable property of both accuracy measures has been extensively discussed in the literature (e.g., Edwards, 1994; Humberg et al., 2018; Krueger et al., 2017) and contributed to the development of statistical approaches that allow disengaging the effects of accuracy and self-perceptions, namely, response surface analysis (e.g., Edwards & Parry, 1993; Humberg et al., 2018, 2022). Thanks to describing the relationships between self-evaluations, a criterion measure, and an outcome in a three-dimensional space without the use of an accuracy indicator as a separate variable, response surface analysis avoids the confounding inherent to the use of difference and residual scores (e.g., Edwards & Parry, 1993; Humberg et al., 2018)

3. The present study

This study aimed at verifying the role that the accuracy of performance judgements plays for later academic achievement, while taking into account that the strength of the relationship might change with students' age. To this end, we used data on student judgements on their performance in a test of maths competencies, actual performance in the test, and performance in a maths competence test two years later collected in two independent samples of primary and lower secondary school students. Moreover, in contrast to past research, we avoided confounding the effect of accuracy with the effect of self-evaluations themselves by employing response surface analysis (Edwards & Parry, 1993; Humberg, Nestler, & Back, 2019) to test competing hypotheses on the discussed link.

We tested five competing hypotheses in total, namely: (H1) the self-knowledge hypothesis stating that the accuracy of performance judgements is positively associated with maths competencies; the more accurate the judgements, the higher the competencies; (H2) the optimal margin hypothesis stating that slightly overestimated performance judgements are associated with higher maths competencies, or, in other words, that a modest positive bias is beneficial; (H3) the beneficial self-evaluation bias hypothesis stating that overestimated performance judgements are associated with higher maths competencies, or, in other words, that overconfidence enhances competencies; (H4) the detrimental self-evaluation bias hypothesis stating that the overestimation of performance is negatively associated with maths competencies; the more overestimated judgements, the lower the competencies; (H5) the positive self-evaluation hypothesis stating that the more positive performance judgements, the higher maths competencies, irrespective of the accuracy of judgements. These hypotheses are graphically summarized in Fig. 1.

Although we tested five different hypotheses, we expected to find empirical support for either the self-knowledge or optimal margin hypothesis because the two had been discussed and accepted in metacognition research.

Based on the existing evidence indicating substantial developmental progression in metacognitive monitoring and control from early primary school to early secondary school (Schneider et al., 2022), the study included two samples: second graders in primary school and fifth graders in secondary school. According to the literature on age-related changes in children's ability to base their learning behaviour on monitoring (e.g., O'Leary & Sloutsky, 2017; Schneider et al., 2022), we expected the link between the accuracy of performance judgements and academic competencies two years later to be stronger in lower secondary than primary school students.

4. Methods

4.1. Samples

This study uses data from the German National Educational Panel Study (NEPS; Blossfeld & Roßbach, 2019). The NEPS is a multi-cohort nationwide research project that follows people of different ages, from newborns to the elderly, to better understand how their educational and occupational trajectories unfold over the life course. We draw on data from two independent samples included in the NEPS comprising children from Grade 2 in primary school (Sample 1) and Grade 5 in lower secondary school (Sample 2). Each sample was drawn using a stratified multistage sampling design (see Aßmann et al., 2019). First, a random sample of schools at the primary or lower secondary level offering education in the relevant grades was selected that was stratified according to the major school types in Germany. Then, in each school, all students from two randomly drawn classes for whom parental consent could be obtained were invited to participate in the respective assessment and a follow-up assessment two years later, that is, in either Grade 4 of primary school (Sample 1) or Grade 7 in lower secondary schools (Sample 2). Further details on the two samples are summarized in Berendes et al. (2019) and Thums et al. (2023).

After removing regression outliers (see Supplement A), Sample 1 included 5551 children (51 % girls) from 359 primary schools participating in Grade 2. The students were on average 7.7 years old (SD = 0.4). About 13.0 % had a migrant background, meaning that, at least one of their parents or they themselves were born outside of Germany. Although the children came from various social backgrounds, as indicated by their parents' position on the *International Socio-Economic Index of Occupational Status Index* (ISEI-08; Ganzeboom, 2010) which ranged from 12 to 89, students with low socio-economic status (SES) were underrepresented. On average, their parents had about 14.9 years of education (SD = 2.4) which typically qualifies for university entrance in Germany. Sample 2 included 4870 students (48 % girls) from Grade 5 with a mean age of 10.9 years (SD = 0.5). The socio-demographic composition of this sample was similar to the sample obtained in Grade 2 as shown in Table 1. Almost half of the students from lower secondary school attended an academic track school.

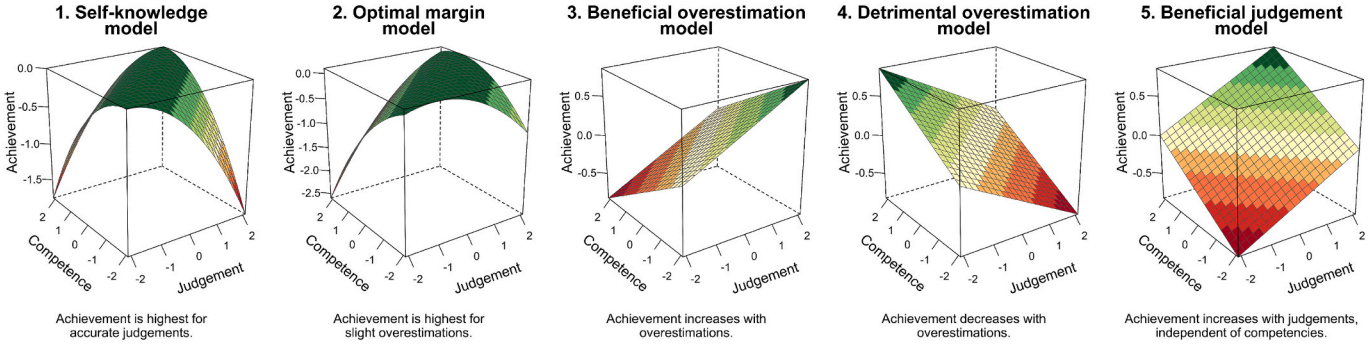


Fig. 1. Response surfaces for hypotheses on the effects of performance judgements on competence development.
Note. The x- and y-axes reflect performance judgements *J* and competencies *C* at the first measurement, while the z-axis reflects achievement *A* (i.e. competencies at the second measurement).

Table 1
Characteristics of samples across measurement occasions.

	Sample 1		Sample 2	
	Grade 2	Grade 4	Grade 5	Grade 7
Sample size	5551	4826	4870	3603
Percentage non-response	0.0 %	13.1 %	0.0 %	26.0 %
Number of schools	359	359	234	191
Percentage female	51.2 %	51.4 %	48.1 %	48.3 %
Mean age (*SD*)	7.7 (0.4)	9.8 (0.4)	10.9 (0.5)	12.9 (0.5)
Percentage migrant background	13.0 %	13.0 %	18.2 %	17.3 %
Percentage academic track	–	–	46.4 %	54.2 %
Mean socio-economic status (*SD*)[a]	58.7 (19.5)	59.0 (19.3)	49.7 (22.3)	50.3 (22.3)
Mean education of parents (*SD*)[b]	14.9 (2.4)	14.9 (2.4)	14.5 (2.4)	14.6 (2.4)
Mean math performance (*SD*)[c]	0.0 (1.0)	0.0 (1.0)	0.0 (1.0)	0.1 (1.0)
Mean performance judgements (*SD*)[c]	0.0 (1.0)	0.0 (1.0)	0.0 (1.0)	0.0 (1.0)

[a] Highest parental international socio-economic index of occupational status (Ganzeboom, 2010).
[b] Parents' highest number of years in education derived from an internationally comparable classification of educational qualifications (Brauns et al., 2003).
[c] *z*-Standardized score obtained in the first assessment.

For each sample, we observed nonresponse rates across measurement occasions that are typical in educational large-scales assessments (Zinn & Gnambs, 2018) ranging between 13 % (Sample 1) and 26 % (Sample 2). The descriptive information in Table 1, however, did not suggest pronounced selection effects. Although migrants and students from lower school tracks had a higher propensity for nonresponse at the follow-up assessment, the respective effects were small. Importantly, nonresponse was only weakly associated with maths competence and performance judgements at the first measurement. Thus, nonresponse did not introduce a substantial bias in the sample compositions across measurement occasions. This was also corroborated by systematic attrition analyses which are summarized in Supplement B.

4.2. Procedure

In the two samples, the students were administered paper-and-pencil tests and questionnaires at school during regular school hours. In primary schools, students who changed schools were followed individually and surveyed by trained interviewers during a home visit. Parents of the participating students were interviewed by phone.

All participants of age and legal guardians of underage participants provided written informed consent before study enrolment. All participants could withdraw from the study at any time. The NEPS study is conducted under the supervision of the German Federal Commissioner for Data Protection and Freedom of Information (BfDI) and in coordination with the German Standing Conference of the Ministers of Education and Cultural Affairs (KMK) and the Educational Ministries of the respective Federal States. All data collection procedures, instruments, and documents were approved by the Data Protection Unit of the Leibniz Institute for Educational Trajectories in line with national ethical and legal regulations.

4.3. Measures

4.3.1. Maths competencies

Students' maths competencies in the two samples were measured with competence tests that were specifically developed for the NEPS. The development of these tests followed a theoretical framework similar to other large-scale assessments (e.g., the Programme for International Student Assessment, OECD, 2017) that adheres to a literacy concept (Weinert et al., 2019). Instead of measuring competencies that are closely tied to a specific school curriculum, the tests aimed to measure maths competencies that are important for successful participation in modern society. The construction rationale adopted for all maths tests (see Neumann et al., 2013) specified four different content areas (i.e., quantity, space and shape, change and relationship, and data and chance) as well as six cognitive components that were required for a successful task solution. The items were constructed in such a way to refer to a specific content area and cognitive component. Each item was accompanied by a multiple-choice (with one correct response option) or short open response format (which typically required a response in the form of a number or a single word). Example items are available in Schnittjer and Duchhardt (2015). The tests were scaled using a one-parametric item response model (Masters, 1982) to provide unidimensional proficiency scores in the form of weighted likelihood estimates (Warm, 1989).

In Sample 1, the two tests that were administered in Grades 2 and 4 included 24 items each. The marginal reliabilities of the tests were 0.79 and 0.73, which indicated a good measurement precision given the brevity of the tests. Psychometric analyses of the tests in the present sample supported a good fit to the item response model, essential unidimensionality, and negligible differential item functioning across different criteria (see Schnittjer et al., 2020; Schnittjer and Gerken, 2018). In Sample 2, the two tests administered in Grades 5 and 7 included 24 and 23 items, respectively. The marginal reliabilities of 0.78 and 0.72 indicated satisfactory measurement precisions of both tests. Several psychometric analyses in the present sample supported the estimation of unidimensional proficiency scores for both tests (see Duchhardt & Gerdes, 2012; Schnittjer & Gerken, 2017).

4.3.2. Performance judgements

After completing the maths competence test, the students were asked to estimate how many items they presumably answered correctly. Their

reports were therefore so called *postdictions* or *retrospective judgements of performance accuracy* (Händel et al., 2013; Schraw, 2009), which are widely used measures of metacognitive monitoring. We used the postdictions given at the first measurement occasion. In primary school (Sample 1), the students answered by indicating either a sad, neutral, or smiley face on a 5-point smiley scale. To familiarize the children with this task, the meaning of all the smilies was explained to them, e.g., that the sad looking face on the left meant that no item was correct and the happy looking face meant that all items were correct. To create an indicator of performance judgements, we transformed the answers into proportions of (presumably) correctly solved items (1 = 0, 2 = 0.25, 3 = 0.5, 4 = 0.75, 5 = 1).[1] In lower secondary school (Sample 2), the students specified the exact number of items they thought they solved correctly. To create an indicator, we divided the number of items that the students thought they solved correctly by the number of items in the test.

4.3.3. Control variables

The analyses controlled for sex (coded 0 for boys and 1 for girls), age (in years), and, in the older sample, school track (coded 0 for non-academic track and 1 for academic track) because these variables showed selection effects in our attrition analyses (see Supplement B). We also controlled for sex because sex differences in maths achievement (e.g., OECD, 2016) and in the accuracy of performance judgements in maths (e.g., Händel et al., 2020) have been documented in some German studies. Within-cohort differences in age, in turn, have been shown to affect achievement due to greater overall maturation, including cognitive maturation, of older students (the relative age effect, see e.g., Cobley et al., 2009; Navarro et al., 2015). At the same time, performance judgements and their accuracy change with age as well (e.g., Bayard et al., 2021; Pressley et al., 1987; Schneider, 2008).

4.4. Statistical analyses

The hypotheses were examined using response surface analyses (Edwards & Parry, 1993; Humberg et al., 2018) in an information-theoretic approach (Burnham and Anderson, 2002). The raw data analysed in this study is available after registration at NEPS Network (2022a, 2022b), while the documented analysis code including the analysis results can be accessed on OSF.

4.4.1. Response surface analyses

The competing hypotheses (see Fig. 1) were tested using response surface analyses (Edwards & Parry, 1993; Humberg et al., 2018) that translated each hypothesis in a second-order polynomial regression with specific constraints.

$$A = b_0 + b_1 J + b_2 C + b_3 J^2 + b_4 JC + b_5 C^2 \quad (1)$$

The full model given by Eq. (1) uses maths achievement at the second measurement occasion as outcome A which is predicted by the linear effects of performance judgement J (b_1) and competence C (b_2) at the first measurement, the respective quadratic effects J^2 (b_3) and C^2 (b_5), and the interaction between judgements and competence JC (b_4). Each hypothesis can be represented as a constrained version of the full model by imposing the parameter constraints given in Supplement C. These constraints give the response surfaces depicted in Fig. 1. The mathematical details of these constraints and how they relate to the different hypotheses are outlined in Humberg et al. (2019).

Since response surface analysis allows describing the relationships between self-evaluations, actual performance, and an outcome in a three-dimensional space, it does not require calculating any indicators of self-evaluation accuracy, and, consequently, avoids their pitfalls. Instead of solely interpreting the regression coefficients, as typically in regression analyses, a response surface pattern, which represents the investigated relationships, is plotted to facilitate interpretation. For instance, in Fig. 1, Model 1 presents a curved response surface where an outcome (student achievement, vertical axis) is the highest when self-evaluations (judgement) agree with actual performance (competence). Therefore, it represents a situation when the accuracy of self-evaluations is pivotal for later competence. For more detail on the interpretation of response surface analyses, see, for instance, Humberg et al. (2018; 2019).

4.4.2. Information-theoretic model evaluation

Instead of examining each hypothesis in isolation, we adopted an information-theoretic approach based on Akaike's (1973) Information Criterion (AIC) and evaluated the empirical evidence for all hypotheses simultaneously (Burnham and Anderson, 2002). For each model, we calculated the Akaike weight which gives a model's likelihood of providing the best explanation for the data in comparison to all other examined models. Thus, it allows ranking all models with regard to their evidentiary values. The Akaike weights not only allow identifying the best model among all examined models, but also highlight potential uncertainties in the model selection process. In case several hypotheses explain the data equally well, this would be reflected in similar Akaike weights. Before calculating the final Akaike weights, we excluded models that were redundant with a nested simpler model, that is, for which the difference in the log-likelihood was smaller than 1 (see Humberg et al., 2019). In such instances the more complex model does not improve the fit of the simpler model despite requiring additional parameters and, thus, should not be used for model selection (Arnold, 2010).

The information-theoretic approach requires specifying a comprehensive set of competing models that can be expected to explain the data (Burnham and Anderson, 2002). Therefore, we included the five models in our analyses that reflect the prevalent hypotheses in prior research (see Fig. 1). Additionally, we considered 15 non-hypothesized candidate models[2] that in principle could serve as plausible alternative explanations for the data. These additional models relaxed some assumptions of the models in Fig. 1 by allowing different curvilinear effects for judgements and competencies. We also considered models without judgement effects (and only effects of competencies) and a model with interaction effects between judgements and competencies. Finally, the model set was completed with a full model that freely estimated all effects in the polynomial regression and a null model that constrained all effects to zero. A description of all considered models is given in Supplement C.

4.4.3. Data transformations

Because the performance judgements and competencies were measured on different scales, the competence scores at the first measurement were transformed into domain percent-correct scores based on

[1] In a pilot study, we compared the two formats of the scale (smiley vs. open format). A total of 606 third-graders were randomly assigned to one of two different groups that were introduced to the different scales. The analyses revealed that the smiley format should be preferred for primary school children because there were significantly fewer missing values in the judgements (1 % in the smiley format versus 8 % in the open format). The correlation between the judgements and the actual test scores was comparable between the groups. However, the level of judgements was somewhat higher in the open format group compared to the smiley format group. That is, children in the open format group on average indicated that they solved 78 % of the items correctly compared to 72 % in the smiley group.

[2] More complex hypotheses, for example, regarding asymmetric or level-dependent congruence effects could be addressed with cubic response surface analyses (Humberg et al., 2022). Because we had no theoretical expectations for these models and exploratory analyses reported in Supplement D found no support for any cubic effect, we refrained from including these specifications in our model set.

the item response model used for scaling the test (Bock et al., 1997). As a result, judgement and competence scores were comparably given as percentages. For the response surface analyses, all variables were *z*-standardized. To preserve commensurability of the variable scales, the *z*-standardization of both predictor variables were based on the pooled mean and standard deviation across the two variables.

4.4.4. Model estimation

The response surface analyses were conducted in *R* (R Core Team, 2024) with *lavaan* (Version 0.6–17, Rosseel, 2012) using a maximum likelihood estimator with cluster-robust standard errors (Savalei, 2014) to account for the nesting of students in different schools. Missing values were handled using a full maximum likelihood approach with auxiliary variables (Enders, 2008) with the help of *semTools* (Jorgensen et al., 2023). The Akaike weights for the information-theoretic model comparisons were calculated in *AICcmodavg* (Mazerolle, 2023). The response surface plots were created with *RSA* (Schönbrodt & Humberg, 2023).

5. Results

An inspection of the joint distributions of the performance judgements and competence scores at the first measurement showed that about 67 % (Sample 1) to 66 % (Sample 2) of the respondents reported performance judgements that were at least half a standard deviation higher than their competence scores (see Table 2). Thus, a substantial percentage of students in both samples overestimated their competencies. These results were mirrored by the univariate density distributions of the performance judgement scores which were highly right-skewed, particularly in primary school (see Fig. 2). Underestimations, on the other hand, were rare and observed for only 6 % (Sample 1) and 7 % (Sample 2) of the respondents. Because about a third of the respondents showed performance judgements that were in agreement with their competencies, we explored the congruence hypotheses further.

Competencies and performance judgements were cross-sectionally correlated in Sample 1 at $r = 0.12$ ($p < .001$) and in Sample 2 at $r = 0.27$ ($p < .001$). The longitudinal correlations between competencies were large and ranged from $r = 0.66$ to 0.75 ($p <. 001$) in the two samples. In contrast, performance judgements exhibited substantially lower correlations with the competence measurements in the follow-up assessment, falling at $r = 0.04$ ($p = .043$) in Sample 1 and $r = 0.25$ ($p < .001$) in Sample 2. As expected, the respective correlation was significantly smaller in primary school than in lower secondary school $z = 10.76$ ($p < .001$). The full descriptive results for all variables are given in Supplement E.

The results of the model evaluations in Table 3 include all models with cumulative Akaike weights exceeding 95 %. Following prevalent recommendations (Burnham & Anderson, 2002), models with a likelihood of being the best among the 20 models included in the candidate set that fell below 5 % were excluded from further interpretations. In primary school (Sample 1), only the full model was supported by the data and, thus, did not substantiate the hypothesized effects of performance judgements depicted in Fig. 1. The competence scores at the second measurement point were significantly ($p < .05$) predicted by linear effects of judgements, $b_1 = 0.02$ ($p < .001$), and competences, $b_2 = 0.76$ ($p < .001$), the quadratic effects of judgements, $b_3 = 0.05$ ($p < .001$), and competencies, $b_5 = 0.06$ ($p < .001$), and the respective interaction, $b_4 = 0.06$ ($p < .001$). As shown in Fig. 3 (left plot), the effect of performance judgement on later achievement depended on the competence level. Judgement effects were stronger at higher competencies and weaker at lower competencies. Together, the two predictors explained 45 % in the variance of achievement.

In Sample 2 from lower secondary school, the confidence set included two models, that is, the interaction model with a likelihood of 78 % and the full model with a likelihood of 22 % (see weights in Table 3). In the interaction model, the competence scores at the second measurement point were significantly ($p < .05$) predicted by linear effects of judgements, $b_1 = 0.10$ ($p < .001$), and competences, $b_2 = 0.63$ ($p < .001$), and the respective interaction, $b_4 = 0.06$ ($p < .001$). As shown in Table 3, the regression coefficients in the full model are very similar. Both models imply a stronger effect of performance judgements on competence development when students have larger initial competencies (see middle plot in Fig. 3). The only difference between the two is a slight curvilinear effect of competencies which is absent from the interaction model. The impact of this effect, however, is rather small as indicated by similar proportions of explained variance in the two models (see R^2 in Table 3).

Taken together, the analyses in the two samples of primary and lower secondary school students did not support the self-knowledge or optimal margin hypotheses that are often discussed in metacognition research. Rather, the response surfaces for the full models in the three samples (see Fig. 3) concordantly suggested that performance judgements and competencies show an interactive effect on achievement that does not reflect a congruence or overestimation effect. Instead, effects of performance judgements on later achievement seem to depend on the competence level of the respondents.

Table 2
Agreement of competencies and judgements.

	Sample size	Percentage	Mean judgement	Mean competence
Sample 1: Grade 2				
Judgements lower than competencies	342	6.16	0.55	0.76
Judgements comparable to competencies	1499	27.00	0.72	0.71
Judgements higher than competencies	3710	66.84	0.87	0.50
Sample 2: Grade 5				
Judgements lower than competencies	345	7.08	0.46	0.70
Judgements comparable to competencies	1331	27.33	0.75	0.73
Judgements higher than competencies	3194	65.59	0.83	0.50

Note. Judgement and competence scores were *z*-standardized.

6. Discussion

This study investigated the role of the accuracy of performance judgements for performance in a test of maths competencies two years later while taking into account that the relationship might change with student age. To this end, we tested a set of competing hypotheses on the above-mentioned relationships in two independent samples of primary and lower secondary school students. Response surface analysis allowed us to correctly relate performance judgements and their accuracy to actual performance.

The analyses aligned with previous findings showing (e.g., Pressley et al., 1987; Roebers, 2002) that primary and lower secondary school children tend to overestimate their performance. Moreover, the correlations between performance judgements and actual performance were lower in Grade 2 than in Grade 5 students, indicating lower accuracy in younger students, which is consistent with past research as well (see e.g., Schneider, 2008 for a review). However, contrary to the expectations, we did not find support for neither the self-knowledge nor optimal margin hypotheses, nor any other tested hypothesis on the link between (the accuracy of) performance judgements and later performance.

6.1. The role of past competencies

The analyses in the two samples revealed the strongest effect of past competencies. One of the probable reasons behind the result is the high

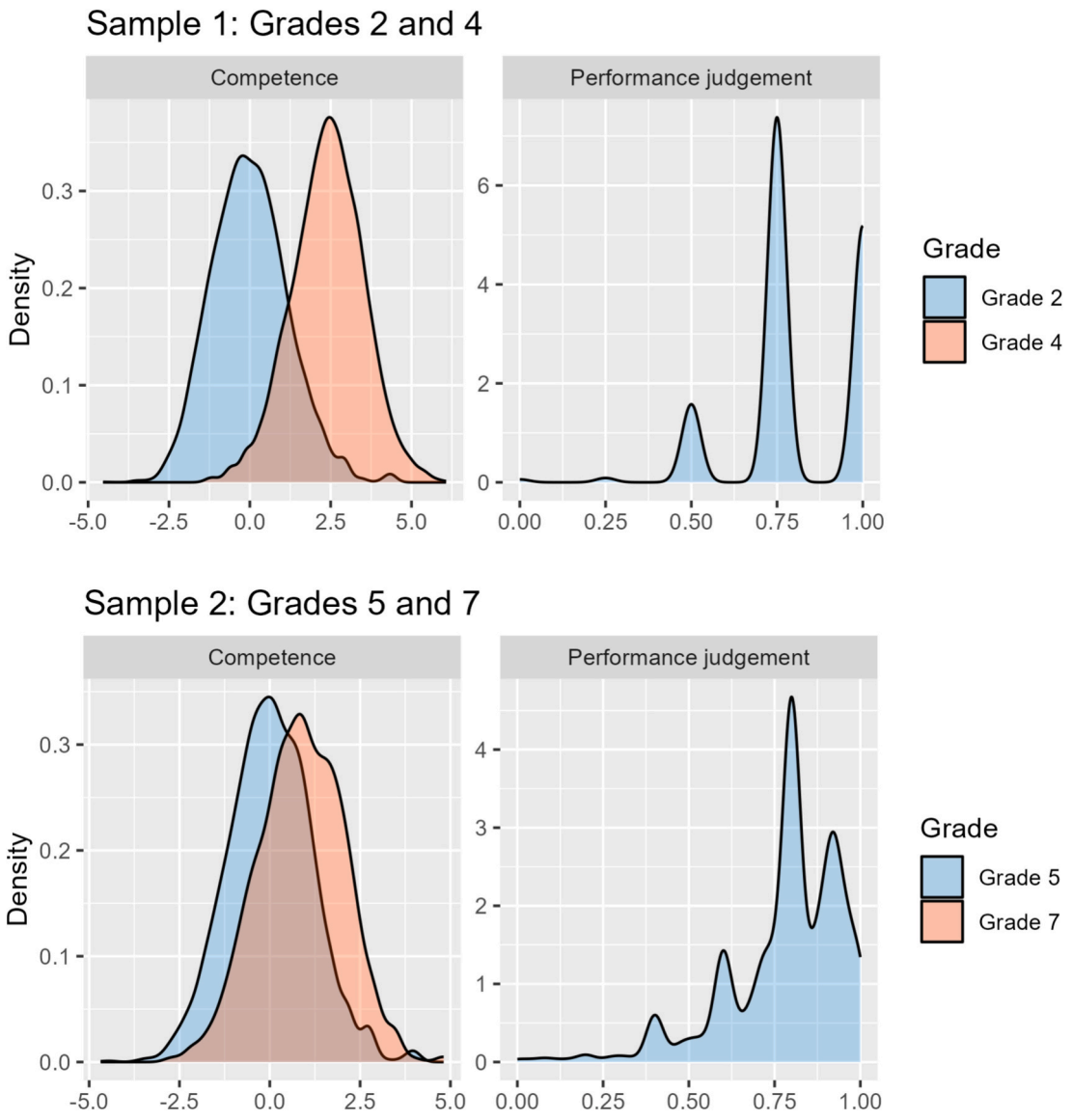


Fig. 2. Density distributions for performance judgements and competencies.

Table 3
Results of information-theoretic model evaluations.

Model	b_1	b_2	b_3	b_4	b_5	K	LL	AIC	Weight	R^2
Sample 1: Grade 2 to 4										
Model 19: Full model	0.022***	0.757***	0.049***	0.062**	0.060***	5	86,443	173,069	1.000	0.445
Sample 2: Grade 5 to 7										
Model 17: Interaction model	0.099***	0.628***	0.000[a]	0.062***	0.000[a]	3	72,730	145,668	0.781	0.307
Model 19: Full model	0.097***	0.633***	0.012	0.056***	0.003	5	72,729	145,671	0.218	0.306

Note. For each sample, the 95 % confidence set of models is reported. Regression coefficients b_1 to b_5 refer to the full polynomial model $A = b_0 + b_1J + b_2C + b_3J^2 + b_4JC + b_5C^2$ with A as maths achievement at the second measurement occasion, J as the performance judgements at the first measurement, and C as the competence at the first measurement. K = Number of estimated regression coefficients (excluding intercept); LL = Log-likelihood of model; AIC = Akaike information criterion; Weight = Akaike weight; R^2 = Variance explained by judgement and competence scores (net of covariate effects). Full results are provided in Supplement F.

[a] Fixed parameter.

** $p < .01$.

*** $p < .001$.

stability of maths competencies, as indicated by the correlations between their repeated measurements (from $r = 0.66$ to 0.75). On the one hand, such a result cannot be considered a surprise due to the abundance of research documenting the role of past performance for future performance. For instance, in a meta-analysis by Scherrer et al. (2025), a two-year stability of test results equalled $r = 0.72$. On the other hand, although no researcher in the field of metacognition would deny the importance of past performance, neither of the tested hypotheses,

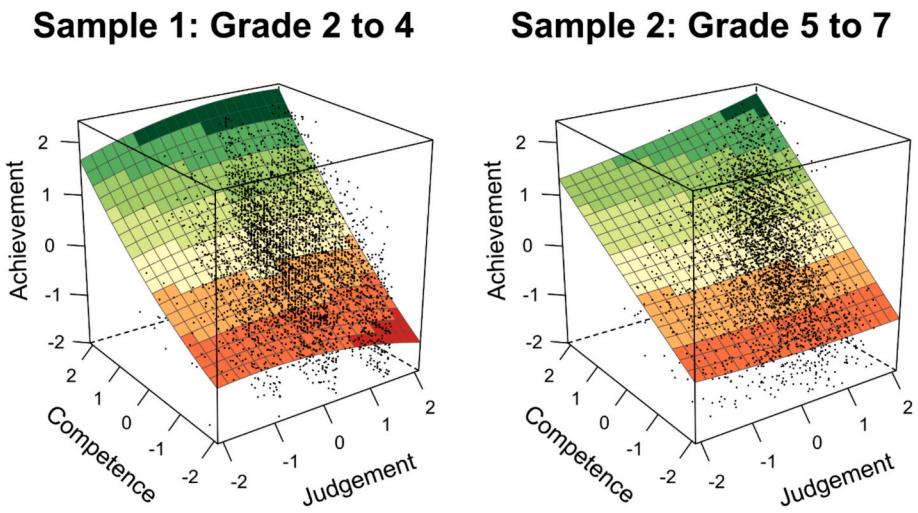


Fig. 3. Response surfaces for full models in the two samples.
Note. The x- and y-axes reflect performance judgements *J* and competencies *C* at the first measurement, while the z-axis reflects achievement *A* (i.e. competencies at the second measurement).

including the models discussed in metacognition research, took past performance into account. However, theoretical models in metacognition research focus on the role of momentary and task-specific processes of metacognitive monitoring for learning behaviour and task performance (e.g., Nelson & Narens, 1994; Thiede & Dunlosky, 1999), not on the accumulation of skills over time resulting from such processes taking place repeatedly at various occasions over extended periods, as we did in this study. Moreover, unlike in this study, studies on metacognitive monitoring often use laboratory tasks that do not rely on prior knowledge and, therefore, do not require taking past performance into account. As a consequence, the models of metacognitive monitoring usually do not consider past performance. Moreover, they place a lot of importance on the accuracy of performance judgements because accurate judgements serve as a base for proper self-regulation of learning (e.g., Rutherford, 2017a; Stone, 2000; Thiede & Dunlosky, 1999; Winne, 2011). As a result, although accuracy is considered pivotal for performance, performance itself is often taken into account indirectly, by looking at the discrepancy (or lack of thereof) between judgement and performance. Our results suggest that it may be beneficial for the field to integrate past performance into its theoretical models. This would require, among others, clarifying whether and how past performance, performance judgements, and their accuracy interplay, and how they contribute to shaping future performance, both in the short and long run.

6.2. The role of performance judgements

Besides the dominant effect of past competencies, the analyses revealed a positive but relatively weak effect of performance judgements, especially in students with high prior competencies. The result suggests that students, especially high achievers, may benefit from judging their own performance as high. As the beneficial self-evaluation hypothesis (e.g., Schunk & Mullen, 2012) proposes, high performance judgements may help maintain effort and motivation (e.g., Doménech-Betoret et al., 2017; Guay et al., 2010; Trautwein et al., 2009). It is conceivable that positive judgements may be particularly beneficial when the student sees the positive effects of their work or effort—an outcome more probable in high achievers. The effect of performance judgements on performance in a maths competence test two years later was weaker in primary than secondary school students, which suggests that performance judgements may become more important in older students. Hence, although these effects were quite small, judging one's performance as high seemed to be slightly more important for learning gains among secondary school students. A very tentative explanation could again refer to the potential motivational role of performance judgements. Academic motivation declines during the late years of primary school and lower secondary school (e.g., Scherrer & Preckel, 2019) whereas children in the early years of primary school are still intrinsically motivated. Therefore, judging one's performance as high could be particularly beneficial at an age when motivation usually decreases. In other words, high performance judgements might counteract a loss of motivation in secondary school. However, it cannot be excluded that the result was due to a different way of measuring performance judgements in the younger sample.

6.3. The role of accuracy

We expected either the self-knowledge or optimal margin hypothesis, both of which highlight the pivotal role of accuracy, to gain empirical support in this study. However, the results did not support the positive role of the accuracy of performance judgements for performance in a maths competence test two years later. In other words, the accuracy of performance judgement did not play a role for later maths competence, and this was true for both primary and secondary school children. Based on literature regarding age-related changes in children's ability to base their learning behaviour on monitoring (e.g., O'Leary & Sloutsky, 2017; Schneider et al., 2022), we had expected a stronger link between accuracy and achievement in secondary school students. However, our findings do not support this assumption. Instead, there was no evidence that accuracy played a role for either age group. Metacognitive monitoring skills develop over primary and lower secondary school, as reflected by improved accuracy of performance judgements (e.g., Bayard et al., 2021; Pressley et al., 1987). However, despite these improvements, the accuracy of performance judgements remains rather low, as indicated by the high percentage of children in both age groups who overestimated their performance. This may be one reason why the accuracy of the judgements is not relevant for competencies in the long run.

The result is, to some extent, unexpected from a theoretical point of view. On the one hand, theoretical models in metacognition research usually focus on the role of metacognitive processes, including metacognitive monitoring, for learning during a specific learning task rather than for performance in other or future tasks. In other words, such models consider metacognitive processes as momentary and largely task-specific (e.g., Nelson & Narens, 1994; Thiede & Dunlosky, 1999), which suggests that perhaps such models should not be used to explain

performance in the long run. On the other hand, since metacognitive monitoring is considered inferential in nature because it relies on various cues (e.g., Koriat et al., 2008), it is supposed to be trainable. Learners can be taught strategies that allow them to gather information helpful in evaluating learning progress, with documented effects of such trainings (see Gutierrez De Blume, 2022 for a meta-analysis). This, in turn, implies that metacognitive monitoring shows at least some task generality, and therefore the accuracy of performance judgements should also affect future performance. However, as mentioned before, other factors than such accuracy may play a prominent role for performance when it is measured with as much delay as in this study (with past performance as potentially the most important factor). Theoretical models in metacognition research may benefit from further clarifying to what extent metacognitive monitoring or which aspects of it are task-specific versus task-general (or even domain-specific versus general) and transferable to other tasks and domains.

Additionally, the MASRL model (Efklides, 2011) may help interpret the results. The model postulates reciprocal relationships between personal dispositions (self-perceptions of ability, motivation, etc. at the Person level) and the self-regulation of learning. High task performance judgements, even if inaccurate (and therefore considered metacognitive failures at the Task x Person level of the model), may feed back to the Person level, increasing the perceptions of ability and motivation. Those perceptions, in turn, can positively affect task engagement and self-regulated learning, which results in higher achievement. Such a feedback loop may explain why performance judgements (irrespective of their accuracy) positively relate to later academic achievement. However, we investigated neither the reciprocal relationships between the two types of self-evaluations, nor their relative contributions or causal paths to achievement. Therefore, as speculative, this interpretation should be treated with caution and requires further research.

6.4. Limitations & future research directions

This study, although run with the utmost rigour, has important limitations that should be taken into account when interpreting its results. First, the study is far from exhaustive with respect to the types and the number of performance judgements. Since various judgements differ in how fine-grained they are (global and local) and when they are made (pre- and postdictions), they depend on task- and non-task specific factors to a varying degree, and therefore carry different information and represent manifold aspects of monitoring (e.g., Rutherford, 2017a; Schraw, 2009). As a result, the role of their accuracy for achievement may differ too. Therefore, future studies should also take other types of performance judgements into account and include multiple measurements to increase measurement accuracy.

Second, we used a different (simpler) scale to measure metacognitive judgements in the younger sample. Although such a scale was more age-appropriate and is widely used in metacognition research with younger children (e.g., Roderer & Roebers, 2010), it introduced an additional factor that could affect the results and made comparisons between age groups more difficult. However, it is important to note that a direct comparison of the level of performance judgements in the two age groups was not our priority in this study. Additionally, the use of five response categories, although age-appropriate, could have reduced variance in student self-evaluations and introduced measurement error, resulting the diminished congruence in self-evaluated and actual performance, and a weaker relationship of both to later maths competencies.

Third, performance judgements and their accuracy change with age (e.g., Bayard et al., 2021; Pressley et al., 1987). Therefore, the time lag between judging one's own performance and filling out a competence measure may play a significant role. Future studies should use designs with shorter time lags.

Fourth, we tested the hypotheses in only one domain (maths), which limits the generalizability of the results. Future studies should replicate the analyses in other domains, for instance, reading or science.

Finally, this was an observational study, which limits causal interpretations. However, in an attempt to remediate the problem, we included adequate control variables.

7. Conclusions

To summarise, the study supported none of the five hypotheses on link between performance judgements and future achievement. Instead, it highlighted the role of prior achievement and showed that judging one's own performance as high might be more beneficial than having accurate judgements. Moreover, high performance judgements become more important as children get older. This suggests that in addition to giving performance feedback and instructions for self-regulated learning, for instance, knowledge about strategies and practicing strategies, it is important to foster student motivation, especially in secondary school, when motivation typically declines (Scherrer & Preckel, 2019).

Moreover, the study, as an attempt to bridge two separate strands of research—on metacognition and competence development—provides an important insight into directions in which theoretical models in metacognition research could be refined in the future.

CRediT authorship contribution statement

Anna Hawrot: Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Timo Gnambs:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis. **Kathrin Lockl:** Writing – review & editing, Writing – original draft, Conceptualization.

Compliance with ethical standards

The NEPS study is conducted under the supervision of the German Federal Commissioner for Data Protection and Freedom of Information (BfDI) and in coordination with the German Standing Conference of the Ministers of Education and Cultural Affairs (KMK) and – in the case of surveys at schools – the Educational Ministries of the respective Federal States. All data collection procedures, instruments, and documents were checked by the data protection unit of the Leibniz Institute for Educational Trajectories (LIfBi). The necessary steps are taken to protect participants' confidentiality according to national and international regulations of data security. All participants of age and legal guardians of underage participants provided written informed consent prior to study enrolment. All participants could withdraw from the study at any time. The analyses are secondary analyses of data published previously.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This paper uses data from the National Educational Panel Study (NEPS; see Blossfeld & Roßbach, 2019). The NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi, Germany) in cooperation with a nationwide network.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.lindif.2025.102764.

Data availability statement

The data that support the findings of this study are available from the Leibniz Institute for Educational Trajectories at doi:https://doi.org/10.5157/NEPS:SC2:10.0.0 and doi:https://doi.org/10.5157/NEPS:SC3:12.1.0. Restrictions apply to the availability of these data, which is the reason why they cannot be provided by the authors of the study. Survey questionnaires are available on the NEPS study website (https://www.neps-data.de). The analysis code is available on OSF.

References

Akaike, H., Petrov, B. N., & Csaki, F. (1973). Information theory as an extension of the maximum likelihood principle. In *Second international symposium on information theory* (pp. 267–281). Akademiai Kiado.

Arnold, T. W. (2010). Uninformative parameters and model selection using Akaike's information criterion. *Journal of Wildlife Management, 74*(6), 1175–1178. https://doi.org/10.2193/2009-367

Aßmann, C., Steinhauer, H. W., Würbach, A., Zinn, S., Hammon, A., Kiesl, H., … Blossfeld, H.-P. (2019). Sampling designs of the National Educational Panel Study: Setup and panel development. In H.-P. Blossfeld, & H.-G. Roßbach (Eds.), *Education as a Lifelong Process* (pp. 35–55). Springer VS.

Baumeister, R. F. (1989). The optimal margin of illusion. *Journal of Social and Clinical Psychology, 8*(2), 176–189. https://doi.org/10.1521/jscp.1989.8.2.176

Bayard, N. S., van Loon, M. H., Steiner, M., & Roebers, C. M. (2021). Developmental improvements and persisting difficulties in children's metacognitive monitoring and control skills: Cross-sectional and longitudinal perspectives. *Child Development, 92*(3), 1118–1136. https://doi.org/10.1111/cdev.13486

Berendes, K., Linberg, T., Müller, D., Wenz, S. E., Roßbach, H.-G., Schneider, T., & Weinert, S. (2019). Kindergarten and elementary school: Starting cohort 2 of the National Educational Panel Study. Blossfeld, H.-P. & Roßbach, H.-G. (Eds.), Education as a lifelong process (2nd ed., Vol. 3, pp. 215–230). Springer.

Education as a lifelong process. In Blossfeld, H.-P., & Roßbach, H.-G. (Eds.), *The German National Educational Panel Study (NEPS). Edition ZfE* (2nd ed.), (2019). Springer VS. https://doi.org/10.1007/978-3-658-23162-0.

Bock, R. D., Thissen, D., & Zimowski, M. F. (1997). IRT estimation of domain scores. *Journal of Educational Measurement, 34*(3), 197–211. https://doi.org/10.1111/j.1745-3984.1997.tb00515.x

Bouffard, T., Vezeau, C., Roy, M., & Lengelé, A. (2011). Stability of biases in self-evaluation and relations to well-being among elementary school children. *International Journal of Educational Research, 50*(4), 221–229. https://doi.org/10.1016/j.ijer.2011.08.003

Brauns, H., Scherer, S., & Steinmann, S. (2003). The CASMIN educational classification in international comparative research. In J. H. P. Hoffmeyer-Zlotnik, & C. Wolf (Eds.), *Advances in cross-National Comparison: A European working book for demographic and socio-economic variables* (pp. 221–244). Springer US. https://doi.org/10.1007/978-1-4419-9186-7_11.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). Springer.

Chen, P. P. (2003). Exploring the accuracy and predictability of the self-efficacy beliefs of seventh-grade mathematics students. *Learning and Individual Differences, 14*(1), 77–90. https://doi.org/10.1016/j.lindif.2003.08.003

Chiu, M. M., & Klassen, R. M. (2009). Calibration of reading self-concept and reading achievement among 15-year-olds: Cultural differences in 34 countries. *Learning and Individual Differences, 19*(3), 372–386. https://doi.org/10.1016/j.lindif.2008.10.004

Chiu, M. M., & Klassen, R. M. (2010). Relations of mathematics self-concept and its calibration with mathematics achievement: Cultural differences among fifteen-year-olds in 34 countries. *Learning and Instruction, 20*(1), 2–17. https://doi.org/10.1016/j.learninstruc.2008.11.002

Cobley, S., McKenna, J., Baker, J., & Wattie, N. (2009). How pervasive are relative age effects in secondary school education? *Journal of Educational Psychology, 101*(2), 520–528. https://doi.org/10.1037/a0013845

Doménech-Betoret, F., Abellán-Roselló, L., & Gómez-Artiga, A. (2017). Self-efficacy, satisfaction, and academic achievement: The mediator role of students' expectancy-value beliefs. *Frontiers in Psychology, 8*, 1193. https://doi.org/10.3389/fpsyg.2017.01193

Duchhardt, C., & Gerdes, A. (2012). *NEPS technical report for mathematics – Scaling results of Starting Cohort 3 in fifth Grade (NEPS working paper no. 19)*. Nationales Bildungspanel: Otto-Friedrich- Universität.

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

Edwards, J. R. (1994). The study of congruence in organizational behavior research: Critique and a proposed alternative. *Organizational Behavior and Human Decision Processes, 58*(1), 51–100. https://doi.org/10.1006/obhd.1994.1029

Edwards, J. R., & Parry, M. E. (1993). On the use of polynomial regression equations as an alternative to difference scores in organizational research. *Academy of Management Journal, 36*(6), 1577–1613. https://doi.org/10.2307/256822

Efklides, A. (2011). Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL model. *Educational Psychologist, 46*(1), 6–25. https://doi.org/10.1080/00461520.2011.538645

Enders, C. K. (2008). A note on the use of missing auxiliary variables in full information maximum likelihood-based structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 15*(3), 434–448. https://doi.org/10.1080/10705510802154307

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

Freund, M.-J., Wolter, I., Lockl, K., & Gnambs, T. (2021). Determinants of profiles of competence development in mathematics and reading in upper secondary education in Germany. *PLoS One, 16*(10), Article e0258152. https://doi.org/10.1371/journal.pone.0258152

Ganzeboom, H. (2010, May). A new international socio-economic index (ISEI) of occupational status for the international standard classification of occupation 2008 (ISCO-08) constructed with data from the ISSP 2002–2007. In *Presentation at the annual conference of the International Social Survey Programme, Lisbon, Portugal*.

García, T., Rodríguez, C., González-Castro, P., González-Pienda, J. A., & Torrance, M. (2016). Elementary students' metacognitive processes and post-performance calibration on mathematical problem-solving tasks. *Metacognition and Learning, 11*(2), 139–170. https://doi.org/10.1007/s11409-015-9139-1

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*(4), 506–528. https://doi.org/10.1037/0033-295X.98.4.506

Gonida, E. N., & Leondari, A. (2011). Patterns of motivation among adolescents with biased and accurate self-efficacy beliefs. *International Journal of Educational Research, 50*(4), 209–220. https://doi.org/10.1016/j.ijer.2011.08.002

Gresham, F. M., Lane, K. L., MacMillan, D. L., Bocian, K. M., & Ward, S. L. (2000). Effects of positive and negative illusory biases. *Journal of School Psychology, 38*(2), 151–175. https://doi.org/10.1016/S0022-4405(99)00042-4

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology, 24*(3), 411–435. https://doi.org/10.1016/0010-0285(92)90013-R

Grijalva, E., & Zhang, L. (2016). Narcissism and self-insight: A review and meta-analysis of narcissists' self-enhancement tendencies. *Personality and Social Psychology Bulletin, 42*(1), 3–24. https://doi.org/10.1177/0146167215611636

Guay, F., Ratelle, C. F., Roy, A., & Litalien, D. (2010). Academic self-concept, autonomous academic motivation, and academic achievement: Mediating and additive effects. *Learning and Individual Differences, 20*(6), 644–653. https://doi.org/10.1016/j.lindif.2010.08.001

Guo, L. (2022). The effects of self-monitoring on strategy use and academic performance: A meta-analysis. *International Journal of Educational Research, 112*, Article 101939. https://doi.org/10.1016/j.ijer.2022.101939

Gutierrez De Blume, A. P. (2022). Calibrating calibration: A meta-analysis of learning strategy instruction interventions to improve metacognitive monitoring accuracy. *Journal of Educational Psychology, 114*(4), 681–700. https://doi.org/10.1037/edu0000674

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.

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

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

Harding, S.-M., English, N., Nibali, N., Griffin, P., Graham, L., Alom, B., & Zhang, Z. (2019). Self-regulated learning as a predictor of mathematics and reading performance: A picture of students in grades 5 to 8. *Australian Journal of Education, 63*(1), 74–97. https://doi.org/10.1177/0004944119830153

He, J. C., & Côté, S. (2023). Are empathic people better adjusted? A test of competing models of empathic accuracy and intrapersonal and interpersonal facets of adjustment using self- and peer reports. *Psychological Science, 34*(9), 955–967. https://doi.org/10.1177/09567976231185127

Humberg, S., Dufner, M., Schönbrodt, F. D., Geukes, K., Hutteman, R., Küfner, A. C. P., … Back, M. D. (2019). Is accurate, positive, or inflated self-perception most advantageous for psychological adjustment? A competitive test of key hypotheses. *Journal of Personality and Social Psychology, 116*(5), 835–859. https://doi.org/10.1037/pspp0000204

Humberg, S., Dufner, M., Schönbrodt, F. D., Geukes, K., Hutteman, R., Van Zalk, M. H. W., … Back, M. D. (2018). Enhanced versus simply positive: A new condition-based regression analysis to disentangle effects of self-enhancement from effects of positivity of self-view. *Journal of Personality and Social Psychology, 114*(2), 303–322. https://doi.org/10.1037/pspp0000134

Humberg, S., Nestler, S., & Back, M. D. (2019). Response surface analysis in personality and social psychology: Checklist and clarifications for the case of congruence hypotheses. *Social Psychological and Personality Science, 10*(3), 409–419. https://doi.org/10.1177/1948550618757600

Humberg, S., Schönbrodt, F. D., Back, M. D., & Nestler, S. (2022). Cubic response surface analysis: Investigating asymmetric and level-dependent congruence effects with third-order polynomial models. *Psychological Methods, 27*(4), 622–649. https://doi.org/10.1037/met0000352

Hacker, D. J., & Bol, L. 2019. Calibration and self-regulated learning: Making the connections. In Dunlosky, J. & Rawson, K. A. (Eds.), The Cambridge handbook of cognition and education (1st ed., pp. 647–677). Cambridge University Press. https://doi.org/10.1017/9781108235631.026.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2023). *semTools: Useful tools for structural equation modeling* (Version Version 0.5–6)) [Computer software]. doi:10.32614/CRAN.package.semTools.

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

Krebs, S. S., & Roebers, C. M. (2010). Children's strategic regulation, metacognitive monitoring, and control processes during test taking. *British Journal of Educational Psychology, 80*(3), 325–340. https://doi.org/10.1348/000709910X485719

Krueger, J. I., Heck, P. R., & Asendorpf, J. B. (2017). Self-enhancement: Conceptualization and assessment. *Collabra: Psychology, 3*(1), Article 28. https://doi.org/10.1525/collabra.91

von der Linden, N., & Roebers, C. M. (2006). Developmental changes in uncertainty monitoring during an event recall task. *Metacognition and Learning, 1*(3), 213–228. https://doi.org/10.1007/s11409-006-9001-6

Lipko, A. R., Dunlosky, J., & Merriman, W. E. (2009). Persistent overconfidence despite practice: The role of task experience in preschoolers' recall predictions. *Journal of Experimental Child Psychology, 103*(2), 152–166. https://doi.org/10.1016/j.jecp.2008.10.002

Lockl, K., & Schneider, W. (2003). Metakognitive Überwachungs- und Selbstkontrollprozesse bei der Lernzeiteinteilung von Kindern [metacognitive monitoring and self-control processes for children's allocation of study time]. *Zeitschrift für Pädagogische Psychologie, 17*(3/4), 173–183. https://doi.org/10.1024//1010-0652.17.34.173

Lopez, D. F., Little, T. D., Oettingen, G., & Baltes, P. B. (1998). Self-regulation and school performance: Is there optimal level of action-control? *Journal of Experimental Child Psychology, 70*(1), 54–74. https://doi.org/10.1006/jecp.1998.2446

Marsh, H. W., Martin, A. J., Yeung, A., & Craven, R. (2017). Competence self-perceptions. In A. J. Elliot, C. Dweck, & D. Yeage (Eds.), *Handbook of competence and motivation*. Guilford Press.

Martin, A. J., & Debus, R. L. (1998). Self-reports of mathematics self-concept and educational outcomes: The roles of ego-dimensions and self-consciousness. *British Journal of Educational Psychology, 68*(4), 517–535. https://doi.org/10.1111/j.2044-8279.1998.tb01309.x

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174. https://doi.org/10.1007/BF02296272

Mazerolle, M. J. (2023). *AICcmodavg: Model selection and multimodel inference based on (Q)AIC(c)* (Version Version 2.3.3) [Computer software]. doi:10.32614/CRAN.package.AICcmodavg.

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

Metcalfe, J., & Finn, B. (2013). Metacognition and control of study choice in children. *Metacognition and Learning, 8*(1), 19–46. https://doi.org/10.1007/s11409-013-9094-7

Mirandola, C., Ciriello, A., Gigli, M., & Cornoldi, C. (2018). Metacognitive monitoring of text comprehension: An investigation on postdictive judgments in typically developing children and children with reading comprehension difficulties. *Frontiers in Psychology, 9*, Article 2253. https://doi.org/10.3389/fpsyg.2018.02253

Navarro, J.-J., García-Rubio, J., & Olivares, P. R. (2015). The relative age effect and its influence on academic performance. *PLoS One, 10*(10), Article e0141895. https://doi.org/10.1371/journal.pone.0141895

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

Nelson, T. O., & Leonesio, J. (1988). Allocation of self-paced study time and the 'labor-in-vain effect'. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(4), 676–686. https://doi.org/10.1037/0278-7393.14.4.676

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

NEPS Network. (2022a). *National Educational Panel Study, scientific use file of starting cohort grade 5* (version 12.1.0) [dataset]. *LIfBi Leibniz Institute for Educational Trajectories., doi.* https://doi.org/10.5157/NEPS:SC3:12.1.0

NEPS Network. (2022b). *National Educational Panel Study, scientific use file of starting cohort kindergarten* [dataset]. *Leibniz Institute for Educational Trajectories (LIfBi)*. https://doi.org/10.5157/NEPS:SC2:10.0.0

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

Norman, E. (2020). Why metacognition is not always helpful. *Frontiers in Psychology, 11*, Article 1537. https://doi.org/10.3389/fpsyg.2020.01537

OECD. (2016). PISA 2015 results (volume I): Excellence and equity in education. *PISA, OECD Publishing*. https://doi.org/10.1787/9789264266490-en

OECD. (2017). *PISA 2015 assessment and analytical framework: Science, reading, mathematic, financial literacy and collaborative problem solving*. OECD Publishing. https://doi.org/10.1787/9789264281820-en

O'Leary, A. P., & Sloutsky, V. M. (2017). Carving metacognition at its joints: Protracted development of component processes. *Child Development, 88*(3), 1015–1032. https://doi.org/10.1111/cdev.12644

Paschke, P., Weidinger, A. F., & Steinmayr, R. (2020). Separating the effects of self-evaluation bias and self-view on grades. *Learning and Individual Differences, 83*, Article 101940. https://doi.org/10.1016/j.lindif.2020.101940

Paschke, P., Weidinger, A. F., & Steinmayr, R. (2023). Linear and nonlinear relationships between self-evaluation and self-evaluation bias with grades. *Learning and Individual Differences, 102*, Article 102266. https://doi.org/10.1016/j.lindif.2023.102266

Pressley, M., Levin, J. R., Ghatala, E. S., & Ahmad, M. (1987). Test monitoring in young grade school children. *Journal of Experimental Child Psychology, 43*(1), 96–111. https://doi.org/10.1016/0022-0965(87)90053-1

R Core Team. (2024). *R: A Language and Environment for Statistical Computing (Version Version 4.3.3) [Computer software]*. R Foundation for Statistical Computing. https://www.R-project.org.

Rescorla, L., & Rosenthal, A. S. (2004). Growth in standardized ability and achievement test scores from 3rd to 10th grade. *Journal of Educational Psychology, 96*(1), 85–96. https://doi.org/10.1037/0022-0663.96.1.85

Robins, R. W., & Beer, J. S. (2001). Positive illusions about the self: Short-term benefits and long-term costs. *Journal of Personality and Social Psychology, 80*(2), 340–352. https://doi.org/10.1037/0022-3514.80.2.340

Roderer, T., & Roebers, C. M. (2010). Explicit and implicit confidence judgments and developmental differences in metamemory: An eye-tracking approach. *Metacognition and Learning, 5*(3), 229–250. https://doi.org/10.1007/s11409-010-9059-z

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., & Spiess, M. (2017). The development of metacognitive monitoring and control in second graders: A short-term longitudinal study. *Journal of Cognition and Development, 18*(1), 110–128. https://doi.org/10.1080/15248372.2016.1157079

Rohr, M. E., & Ayers, J. B. (1973). Relationship of student grade expectations, selected characteristics, and academic performance. *The Journal of Experimental Education, 41*(3), 58–62. https://doi.org/10.1080/00220973.1973.11011410

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36.

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

Rutherford, T. (2017b). Within and between person associations of calibration and achievement. *Contemporary Educational Psychology, 49*, 226–237. https://doi.org/10.1016/j.cedpsych.2017.03.001

Savalei, V. (2014). Understanding robust corrections in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(1), 149–160. https://doi.org/10.1080/10705511.2013.824793

Scherrer, V., Breit, M., & Preckel, F. (2025). The stability of students' academic achievement in school: A meta-analysis of longitudinal studies. *Educational Research Review.* , Article 100687. https://doi.org/10.1016/j.edurev.2025.100687

Scherrer, V., & Preckel, F. (2019). Development of motivational variables and self-esteem during the school career: A meta-analysis of longitudinal studies. *Review of Educational Research, 89*(2), 211–258. https://doi.org/10.3102/0034654318819127

Schneider, W. (2008). The development of metacognitive knowledge in children and adolescents: Major trends and implications for education. *Mind, Brain, and Education, 2*(3), 114–121. https://doi.org/10.1111/j.1751-228X.2008.00041.x

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., & Richter, T. (2022). The development of metacognitive knowledge from childhood to young adulthood: Major trends and educational implications. In Advances (Ed.), *63. in child development and behavior* (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*. National Educational Panel Study: Leibniz Institute for Educational Trajectories. https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/Kompetenzen/NEPS_com_ma_2015_en.pdf.

Schnittjer, I., & Gerken, A.-L. (2017). *NEPS technical report for mathematics: Scaling results of Starting Cohort 3 in Grade 7 (NEPS survey paper no. 16)*. Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://doi.org/10.5157/NEPS:SP16:1.0

Schnittjer, I., Gerken, A.-L., & Petersen, L. A. (2020). *NEPS technical report for mathematics: Scaling results of Starting Cohort 2 in Grade 4 (NEPS survey papers no. 69)*. National Educational Panel Study: Leibniz Institute for Educational Trajectories. https://doi.org/10.5157/NEPS:SP69:1.0

Schnittjer, I., & Gerken, A.-L. (2018). NEPS technical report for mathematics*: Scaling results of starting cohort 2 for grade 2* (NEPS survey paper no. 47). Leibniz Institute for Educational Trajectories, National Educational Panel Study. doi:https://doi.org/10.5157/NEPS:SP47:1.0.

Schönbrodt, F., & Humberg, S. (2023). *RSA: An R package for response surface analysis* (Version Version 0.10.6) [Computer software]. doi:10.32614/CRAN.package.RSA.

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

Schunk, D. H., & Mullen, C. A. (2012). Self-efficacy as an engaged learner. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 219–235). Springer US. https://doi.org/10.1007/978-1-4614-2018-7_10.

Shin, T., Davison, M. L., Long, J. D., Chan, C.-K., & Heistad, D. (2013). Exploring gains in reading and mathematics achievement among regular and exceptional students using growth curve modeling. *Learning and Individual Differences, 23*, 92–100. https://doi.org/10.1016/j.lindif.2012.10.002

Sticca, F., Goetz, T., Nett, U. E., Hubbard, K., & Haag, L. (2017). Short- and long-term effects of over-reporting of grades on academic self-concept and achievement. *Journal of Educational Psychology, 109*(6), 842–854. https://doi.org/10.1037/edu0000174

Stone, N. J. (2000). Exploring the relationship between calibration and self-regulated learning. *Educational Psychology Review, 12*(4), 437–475. https://doi.org/10.1023/A:1009084430926

Talsma, K., Schüz, B., & Norris, K. (2019). Miscalibration of self-efficacy and academic performance: Self-efficacy ≠ self-fulfilling prophecy. *Learning and Individual Differences, 69*, 182–195. https://doi.org/10.1016/j.lindif.2018.11.002

Talsma, K., Schüz, B., Schwarzer, R., & Norris, K. (2018). I believe, therefore I achieve (and vice versa): A meta-analytic cross-lagged panel analysis of self-efficacy and academic performance. *Learning and Individual Differences, 61*, 136–150. https://doi.org/10.1016/j.lindif.2017.11.015

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103*(2), 193–210. https://doi.org/10.1037/0033-2909.103.2.193

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(4), 1024–1037. https://doi.org/10.1037/0278-7393.25.4.1024

Thums, K., Gehrer, K., Gnambs, T., Lockl, K., & Nusser, L. (2023). Data from the National Educational Panel Study (NEPS) in Germany: Educational pathways of students in grade 5 and higher. *Journal of Open Psychology Data, 11*, 3. https://doi.org/10.5334/jopd.79

Trautwein, U., Lüdtke, O., Roberts, B. W., Schnyder, I., & Niggli, A. (2009). Different forces, same consequence: Conscientiousness and competence beliefs are independent predictors of academic effort and achievement. *Journal of Personality and Social Psychology, 97*(6), 1115–1128. https://doi.org/10.1037/a0017048

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427–450. https://doi.org/10.1007/BF02294627

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., Carstensen, C. H., & Lockl, K. (2019). Development of competencies across the life course. In Blossfeld, H.-P. & Roßbach, H.-G. (Eds.), Education as a lifelong process—The German National Educational Panel Study (NEPS) (2nd ed., Vol. 3, pp. 57–82). Springer VS.

Williams, B., Williams, J., & Ullman, A. (2002). *Parental involvement in education (No. 332; BMRB Social Research, Issue 332)*. Department for Education and Skills.

Winne, P. H. (2011). A cognitive and metacognitive analysis of self-regulated learning. In B. Zimmerman, & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 15–32). Routledge/Taylor & Francis Group.

Winnie, P. H. (2011). A cognitive and metacognitive analysis of self-regulated learning. In B. Zimmerman, & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 15–32). Taylor & Francis Group.

Wu, H., Guo, Y., Yang, Y., Zhao, L., & Guo, C. (2021). A meta-analysis of the longitudinal relationship between academic self-concept and academic achievement. *Educational Psychology Review, 33*(4), 1749–1778. https://doi.org/10.1007/s10648-021-09600-1

Zimmerman, B. (2011). Motivational sources and outcomes of self-regulated learning and performance. In B. Zimmerman, & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 49–64). Taylor & Francis Group.

Zimmerman, B., & Martinez-Pons, M. (1990). Student differences in self-regulated learning: Relating grade, sex, and giftedness to self-efficacy and strategy use. *Journal of Educational Psychology, 82*(1), 51–59. https://doi.org/10.1037/0022-0663.82.1.51

Zimmerman, B., & Moylan, A. R. (2009). Self-regulation: Where metacognition and motivation intersect. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 299–315). Routledge/Taylor & Francis Group.

Zinn, S., & Gnambs, T. (2018). Modeling competence development in the presence of selection bias. *Behavior Research Methods, 50*(6), 2426–2441. https://doi.org/10.3758/s13428-018-1021-z