The development of self-assessment accuracy in mathematical word problems: A study of primary school learners

Irena Smetáčková[1], Anna Páchová[1,*]

[1] Faculty of Education, Charles University, Magdalény Rettigová 4, Prague, Czech Republic; Anna.pachova@pedf.cuni.cz

Self-assessment, as a critical component of self-regulated learning, plays a significant role in improving learning achievement. This study focuses on the accuracy of self-assessment in the context of mathematical word problems among primary school students. The evolution of self-assessment accuracy was examined between the second and third grades. The impact of cognitive abilities, mathematical knowledge, and teacher evaluations was investigated. Data were collected from a sample of 542 primary school pupils. Our findings reveal that students demonstrated greater accuracy in self-assessing their performance in mathematical word problems in Year 2 compared to Year 3. This self-assessment accuracy varied based on cognitive abilities and the pedagogical environments. These findings suggest the importance of creating adequate classroom conditions to enhance self-assessment accuracy.

Key words:
self-assessment, learning, metacognition, self-regulation, primary school, mathematics.

Received 9/2025
Revised 11/2025
Accepted 12/2025

1 Introduction

The learning process is a very complex cognitive process. In an effort to enhance the effectiveness of learning, scientific attention has turned to self-regulation in recent decades (Panadero, 2017). Zimmerman's Cyclical Phases Model of self-regulated learning (2000) comprises three phases: forethought, performance, and self-reflection. In the self-reflection phase, students evaluate their task performance and determine their success or failure. This assessment then influences their approach and performance in future tasks.

From the above, it follows that self-assessment is a crucial component of self-regulated learning. Moreover, it is hypothesised, that the relationship between self-regulated learning, metacognition and self-efficacy is mediated by self-assessment (Cera et al., 2013; Schraw et al., 2006). Self-assessment also plays a crucial role in supporting students' learning by actively involving them in assessing their own progress and improve their performance (Ross, 2006). However, self-assessment supports learning and contributes to the development of both metacognition and self-efficacy only when students are able to evaluate their performance accurately – judging weak performance negatively and strong performance positively (Barana et al., 2022; Brown & Harris, 2014).

Although numerous studies have examined self-assessment accuracy (Barana et al., 2022; Brown et al., 2015; Bradshaw, 2001), the existing evidence is heavily concentrated at the secondary and tertiary educational levels. Far less is known about how young learners in elementary school monitor their performance, despite the fact that metacognitive and self-regulatory abilities begin to develop early and show substantial variation at this stage. Existing research with younger children is fragmented and provides inconclusive evidence about how accurately they can judge their own performance. In this article, we are particularly interested in investigating self-assessment in the context of solving word problems among primary school pupils, as this is one of the most challenging aspects of the mathematics curriculum.

1.1 Self-assessment and learning process

Self-assessment is the evaluation of one's own performance, knowledge, or skills (Barana et al., 2022). It involves comparing one's actual performance against a standard or set goals. Klenowski (1995) defined self-assessment as "the evaluation or judgment of 'the worth' of one's performance and the identification of one's strengths and weaknesses with a view to improving one's learning outcomes" (p. 146). Andrade (2019) emphasizes the significance of the assessment process and its role as feedback within the framework of formative assessment. Self-assessment is a process focused on evaluating learning outcomes to provide learners with insights into their strengths and weaknesses. The main goal of self-assessment is to promote self-awareness, helping learners gauge their progress, identify areas for development, and set objectives for improving their performance.

Self-assessment is intrinsically linked to self-regulated learning, particularly as a crucial component of the self-reflection phase in Zimmerman's model (2000). With respect to Vygotsky's tradition of mediated learning (Taber, 2025), based on Bandura's social cognitive theory (2014) and experimental paradigms of cognitive psychology (de Bruin & van Gog, 2012; Andrade, 2019), learning is considered to be the result of the interaction of personal, environmental and behavioral factors and thus efficiency of learning increases due to self-regulation. A self-regulated learner is able to set goals and choose appropriate strategies to achieve them, make effort, monitor learning progress, evaluate achievement, and reflect on their emotional reactions (Pintrich, 2000; Siegesmund, 2017; Zimmerman & Schunk, 2011). As shown by Zimmerman (2000) and others, self-regulated learning involves a number of interacting cognitive, metacognitive (Bakar & Ismail, 2020; Georghiades, 2004), and motivational components, including self-efficacy (Bandura, 2014; Liu-Ambrose et al., 2010).

There are many theories exploring the relationships between these concepts. There is undoubtedly a strong connection between self-assessment and metacognition. Self-assessment involves monitoring and reflecting on one's own work, which are essentially metacognitive processes (Yan, 2020). From this perspective, self-assessment can be considered as part of metacognition. If students have well-developed metacognition and self-assessment, they are able to monitor and evaluate their performance effectively. However, causality can also operate in the opposite direction, where the source of influence determines the relationship to ability or performance. For example, Siegesmund (2017) argues that self-assessment can strengthen students' metacognition and positively influence their self-regulation. At the same time, if students improve the accuracy of their self-assessment, this contributes to the development of metacognition and self-regulation. Thus, self-assessment serves as both a prerequisite and a consequence.

Other studies have also highlighted the relationship between self-assessment and self-efficacy. One explanation lies in the connection through self-regulated learning models (Pintrich, 2000; Zimmerman, 2000). Accurate self-assessment helps learners evaluate their performance, leading to better self-regulation and higher self-efficacy through successful learning experiences. Panadero et al. (2017) demonstrated this by analyzing 19 studies involving 2 305 students, revealing the importance of self-assessment interventions for the development of self-efficacy and self-esteem.

From the discussion above, it is evident that self-assessment is a crucial concept that is strongly connected with self-regulated learning. Moreover, improving self-assessment can enhance related concepts such as metacognition and self-efficacy.

1.2 Accuracy[1] of self-assessment

However, self-assessment aids the learning process and the only if students can accurately evaluate their performance. This means that students should recognize poor performance as such and assess it negatively, and similarly, recognize and assess strong performance positively (Barana et al., 2022).

Numerous studies have shown that learners often make inaccurate self-assessments, often overestimating or underestimating their own abilities. Psychologists refer to this as the Dunning-Kruger effect (Kruger & Dunning, 1999). Moreover, this effect is often more pronounced in lower-achieving students who have difficulty recognizing their poor performance, its causes, and how to improve (Dunning et al., 2004). Inaccurate self-assessment acts as a barrier to self-regulated learning and contributes to poor academic performance.

The accuracy of self-assessment may be influenced by cognitive, motivational, emotional factors, individual characteristics, item characteristics and classroom environment characteristics (Brown et al., 2015). Reduced accuracy in self-assessment can arise from learners' tendencies to be unrealistically optimistic about their abilities, deficiencies in skills and abilities, unclear assessment criteria, and classroom norms that encourage overestimation (Dunning et al., 2004). Highly difficult tasks introduce greater uncertainty into self-assessment, whereas easier tasks allow pupils to judge their performance with higher confidence (Bradshaw, 2001). This means that the difficulty of the task correlates with the accuracy of self-assessment, regardless of individual metacognitive competence (Barana, 2022). In the case of word problems, assessing one's own performance is even more complex, as accuracy depends on the combination of mathematical competence and reading comprehension, which interact in non-trivial ways (Schleppegrell, 2007).

Accuracy in self-assessment shows clear developmental patterns (Bradshaw, 2001). Brown and Harris (2014) found that accuracy varies with both age and performance level, with younger and less proficient elementary pupils displaying lower accuracy. Age therefore plays a central role. Young children's cognitive abilities—particularly working memory, inhibition, and self-regulation—are still developing (Zelazo & Carlson, 2012), which limits their capacity to evaluate their performance accurately. They also have less mature emotional regulation and awareness, making it more difficult for them to reflect honestly, whereas adolescents, with more advanced cognitive and emotional capacities, tend to produce more accurate self-assessments (Harter, 2011).

According to Harter (2011), average levels of self-evaluation, and consequently self-assessment, tend to decline from early childhood into adolescence as children develop social comparison skills and adopt

[1] Some authors use the accuracy of self-assessment, others the consistency (Andrade, 2019). We stay in the term accuracy because we compare whether the student's performance corresponds with his/her self-report on it. The consistency is usualy used in the wider context of comparison with teacher's assesments.

a more critical view of themselves. Piagetian theory further suggests that younger children's moral and evaluative reasoning is heteronomous—anchored in authority and rules—which leads them to perceive self-assessment more as obedience to teacher expectations than as autonomous reflection. Children also rely heavily on teachers' judgments as their primary source of evaluative feedback (Eelder, 2010). With increasing age, pupils develop a more differentiated understanding of their abilities and a stronger capacity for self-regulation, while younger children often display simpler and overly optimistic self-evaluations. However, some longitudinal evidence, such as that from Orth et al. (2021), challenges the notion of a sharp transition in self-efficacy or self-esteem during early adolescence. More longitudinal research—both short- and long-term—is therefore needed.

Bradshaw (2001) also reports associations between self-assessment and cognitive ability, though this relationship is not linear. Accuracy in self-assessment appears to depend on an interplay of cognitive abilities, metacognitive skills, domain-specific knowledge, and task complexity. Importantly, intervention research (Kajamies et al., 2010) shows that training and feedback can substantially improve accuracy, even among individuals with varying cognitive profiles.

Teachers play a critical role in shaping pupils' learning, including their self-assessment practices. They influence students not only through positive expectations and instructional support but also through the nature of learning challenges and classroom interactions. Li and Rubie-Davies (2015) [24] demonstrate that teacher optimism can translate into measurable gains in academic progress. Certain pedagogical practices—especially formative assessment—can enhance pupils' self-assessment and self-regulation (Nicol & Macfarlane-Dick, 2006). Factors supporting the development of self-regulated learners include the use of multiple representations, rich mathematical tasks, productive classroom discourse, scaffolding of strategic behaviour, and adaptive instructional support (Pape et al., 2003). Kramarski and Revach (2009) further argue that teachers need to be capable of self-regulation themselves in order to foster student-centred learning, a condition that has been linked to improved self-regulation among pupils (Perry et al., 2006).

1.3 The relationship between self-assessment and performance in mathematics

In this section, we examine the relationship between self-assessment and performance among elementary school pupils. Self-assessment has received growing attention in both learning theory and educational practice, as it represents a reflective process through which pupils evaluate their work against explicit criteria and identify opportunities for improvement. As a formative practice, it has been shown to enhance learning, support pupils in managing academic demands, and contribute to school success while reducing the risk of school failure (Broadfoot, 2021; Andrade, 2019). Because empirical studies focusing specifically on younger learners remain relatively scarce, we also draw on findings from research with older students.

Mathematical word problems offer a particularly meaningful context for studying self-assessment. They require pupils to engage in higher-order thinking and complex problem-solving, draw on multidisciplinary knowledge, apply mathematical concepts to real-world situations, and reveal conceptual understanding rather than procedural recall (Rendl et al., 2013; Kajamies et al., 2010; Pape et al., 2003). Because solving word problems places considerable demands on metacognitive monitoring and self-efficacy, self-assessment may serve as an important mechanism supporting pupils' performance in this domain.

Findings from Czech research on young learners provide additional support for this argument. Chytrý, Říčan and Živná (2019) showed that metacognitive skills are significantly associated with mathematical performance across different educational approaches and that pupils with more developed metacognition tend to achieve higher mathematical outcomes. The relationship also appears bidirectional: engaging in word problems can strengthen metacognitive skills and self-efficacy, while these skills subsequently support more effective problem solving. In a qualitative study, Tachie (2019) demonstrated that the use of metacognitive strategies—including task analysis, planning, monitoring, checking, reflection, and self-assessment—helped 8th- and 9th-grade learners solve mathematical problems.

Findings from younger pupils also support this relationship. Research with 311 primary school students showed that those who successfully solved mathematical problems demonstrated higher levels of prediction and self-assessment accuracy than less successful pupils (Nováková, 2024). Consistent with this, the study by Chytrý et al. (2019) indicates that metacognitive characteristics can differentiate mathematical performance even among relatively young learners, suggesting that metacognition is not merely an advanced skill but a central component of early mathematical development.

Although Nováková's (2024) study points to a positive link between metacognition and performance, the broader literature reveals considerable variability in the strength of this relationship across studies and age groups. In their meta-analysis, Brown and Harris (2013) found that relationships between self-assessment and performance varied widely, with effect sizes ranging from 0.04 to 1.62 (Cohen's d). This variability was not determined by the type of self-assessment but by its complexity. One possible explanation is the formative nature of assessment. For example, Black, Harrison, Lee, Marshall, and Wiliam

(2004) demonstrated a strong relationship between formative assessment (including self-assessment) and achievement among 11- to 15-year-olds.

Feedback may also influence the strength of the relationship. Sitzmann, Ely, Brown and Bauer (2010) found that correlations between self-assessment and learning were stronger in courses that included feedback than in those without it. Although their meta-analysis focused on adults, evidence suggests that similar mechanisms operate in children. Elder (2010), for instance, showed that both younger (1st grade) and older (4th/5th grade) pupils rely on evaluations provided by others, especially teachers, though older pupils draw on a broader range of sources when forming self-assessments.

One of the few intervention studies targeting elementary pupils (grades 5–6) is the work by Ross et al. (2002), who examined the effects of self-evaluation training on mathematics achievement. Their 12-week program, which included a range of activities aimed at improving self-assessment, led to improved mathematical problem-solving in the experimental group compared to the control group.

Taken together, these findings suggest that comprehensive self-assessment training supported by feedback can lead to more accurate self-evaluations and improved problem-solving performance. However, further research is needed to examine the sources of self-assessment accuracy in younger children in lower grades. Only then will it be possible to develop effective self-assessment interventions tailored to this age group.

As demonstrated by prior research, self-assessment is closely linked to a range of psychological factors that play a pivotal role in school learning. Nevertheless, establishing self-assessment practices in the classroom may be challenging. Individual variability in cognitive abilities and developmental maturity must be taken into account. In light of these considerations, the present study focuses on the accuracy of self-assessment in mathematics across two developmental periods. Mathematical word problems were selected because of their recognized complexity within the curriculum (Rendl et al., 2013). These problems not only reflect real-life applications but also have a documented connection to self-assessment, as discussed earlier. The study seeks to explore changes in self-assessment accuracy in the context of mathematical word problems and examine its relationship with cognitive ability.

The research aims to investigate whether pupils can accurately evaluate their performance in test items, particularly in the challenging domain of mathematical word problems. Our study is guided by the following research questions: (1) How does self-assessment accuracy in solving mathematical word problems develop from Year 2 to Year 3 among elementary school pupils? (2) Is there a relationship between self-assessment accuracy and cognitive ability in solving mathematical word problems?

2 Methodology

This study forms an integral part of two larger research projects that synergistically complement one another. The first project is dedicated to enhancing pupils' problem-solving strategies in mathematical word problems, an area perceived as challenging by elementary school pupils. Regrettably, no longitudinal, comprehensive data from primary school pupils were available for this research. The second project concentrates on strategies employed by teachers to mitigate the risk of school failure. This study encompasses data from 29 primary school classes, gathered through a variety of methods, including tests, standardized psychological instruments, and teacher questionnaires. In this paper, we present the results of our study, which utilizes data collected in the second research project and analyses it within the framework of the objectives established in the first research project.

2.1 Measurements

Mathematical test and word problems

The main tool used in the study was a test on mathematics. They have been developed in several steps following Downing's recommendations (Downing, 2006). The content was based on the Framework Educational Programme for Basic Education (FEP BE) that sets out the expected learning outcomes. The structure of the test, including the format, types of questions, and scoring criteria, was reviewed by a group of educational experts. Test items are systematically generated, reflecting the content and cognitive skills to be assessed. The items were pretested to enhance their quality, considering clarity, relevance, and fairness. The same test specification was used for in Year 2 and Year 3 with respect to increasing knowledge and cognitive skills. Thus, both test results can be compared. The psychometric parameters of both tests were good. Pupils solved tests individually, but in the presence of other classmates, their own teacher, and a researcher whom they already knew. The situation was relatively familiar to them and thus it was not a source of great stress. The tests had no time limit. Due to the COVID-19 situation, the administration of the first test was delayed, resulting in a reduced level of difficulty compared to the second test.

The mathematics test contained 7 tasks with a total score of 100 points. Pupils' answers in tests were recorded in electronic form and then evaluated. The sum of the points for the individual tasks resulted in the total score. Each task was generated to measure a certain educational outcome requested by the Framework Educational Programme for Basic Education (FEP BE). The type of task was either typical or untypical for school lessons in researched classes, however, for the purpose of the study, only typical items were used. The maximum score for each task was set based both on the levels in which the task was typical for school lesson and cognitive challenging. The analysis of self-assessment accuracy across the entire test was presented in a previous article (Švamberk Šauerová & Smetáčková, 2022). However, due to limited comparability between the tests, a closer examination of individual items with similar content was necessary.

The tests in Year 2 and Year 3 contained of one school-typical mathematical word problem (M2 in Year 2 and M3 in Year 3) with similar structures. In both cases, it was complex multi-step problems in which pupils needed to solve individual phases through addition and subtraction. The difference between the tasks was both in the number of steps and in the size of the numbers with which the mathematical operations were performed. The wording of both mathematical word problems is given below.

The task M2 for Year 2 was: *There are 11 girls in the class, 4 more boys than girls. There are ____boys. There are ____fewer girls than boys. There are ____children in the class.* The task diagnoses the pupil's ability to connect two situations described in the addition task, which requires reading comprehension, and the pupil's ability to add two small numbers: $11 + 4$ and $11 + 15$. It is a typical word problem with the comparison operator ("4 more than"). The first subtask was deliberately easier, as it was intended to allow pupils to experience success and motivate them to solve further problems. The context of the task is familiar to the pupils and is part of their life experience. The text of the task is simple, the words 'more' and 'total' correspond to the operation of addition which leads to the correct result. The number range corresponds to the beginning of Year 2 when the number range starts to expand from 20 to 100. The two numbers in the problem (11 and 4) express numbers and are easily modeled using some manipulatives. The problem is compound – the calculation of the first part of the problem is needed for the second part. Between the two parts is sandwiched a statement for the pupil to complete, which tests whether the pupil is aware of the link between the relations more and less.

The task M3 for Year 3 was: *Zdenka got 403 points in the computer game, Tereza got 118 points more than Zdenka, David scored 20 points less than Tereza. Tereza scored ____points. Zdenka and Tereza scored together ____points. David scored ____more points than Zdenka. David scored ____points.* The task, like the previous one, diagnoses the pupil's ability to connect the situations described in the task with additive operations. The key is understanding the text, that is, the discovery of the links between the number of points of each child, and the ability to perform additive operations in the domain of natural numbers up to one thousand. The task context belongs to pupils' lives, but not every pupil needs to have experience with computer games. The number domain corresponds to Year 3. The numbers in the problem represent a number but are too large to use manipulatives. The text of the problem is rather short but contains a lot of information. The problem is complex and requires chaining of thought operations. Completing the first two statements is easy. For the first, the comparison operator (118 more) is added to the state (403 points). The second requires adding two forms, one of which is in the problem statement, and the other is written in the first statement. The completion of these first two statements is identical in idea to that in Problem 1 of the Year 2 test, only the number field is adapted to Year 3. Completing the third statement is more challenging as it requires finding the comparison operator. The pupil can arrive at it in two ways, the first of which is easier. Either the pupil completes the fourth statement first (the comparison operator "20 points less" subtracts from the state "521 points") and then calculates the comparison operator from the two known states (501 and 403). Or the pupil works only with comparison operators (Tereza 118 points more than Zdenka, David 20 points less than Tereza). The number field corresponds to Year 3. The numbers in the problem express a number, but they are too large to use manipulatives.

For each item, after solving it, the pupils were asked the question: *How do you think you managed to solve the problem?* Pupils indicated using circling one of three options: thumbs up (good performance), thumbs horizontally (uncertain or average performance), or thumbs down (poor performance). For each task, including the mathematical complex word problem an accuracy index of the pupil's self-assessment was created, and their sum resulted in an accuracy index for the word problem item and for the math test overall.

Cognitive abilities

Besides tests on mathematics, pupils in Year 2 and Year 3, also completed an independent standardized test measuring cognitive abilities. Nonverbal reasoning was tested by Coloured Progressive Matrices (CPM) in Year 2, respectively by Standard Progressive Matrices (SPM) in Year 3. In each year, the cognitive test was assigned several weeks before the test. The cognitive test was used in a one-off manner and thus may have been influenced by situational factors.

Teachers evaluations of individual pupils

Therefore, we used the teacher's assessment of cognitive abilities as a supplementary contextual data source based on long-term experience with the pupils in school tasks, to provide additional background information for Research Question 2 rather than as a separate research question.

The measure used was a questionnaire in which teachers were asked to make expert judgments about each pupil on 32 items relating to different aspects of school achievement. The following five items related to the assessment of cognitive ability: *He/she can concentrate for 10–15 minutes on one activity; He/she has a good memory, remembers easily; He/she is a logical thinker; He/she is inquisitive, likes to learn new things; He/she immediately understands the task and the teacher's instructions.* Each statement was rated on a scale of 1 (statement about the child is true) to 3 (statement about the child is not true).

The analytical unit was the pupil. For each pupil, characteristics were analyzed as assessed by two tests and within them for one complex mathematical word problem as well as his/her achievements were ascertained with two independent standardized achievement measures.

2.2 Sample

The study included 657 pupils, of whom 49% were boys and 51% were girls. Complete longitudinal data were obtained from 542 pupils. The pupils belonged to 29 primary school classes. The classes were selected at the end of Year 1 and subsequently tested in Year 2 and Year 3. Classes were included based on the following criteria: a) an average-size school and class with a typical proportion of pupils with special educational needs according to national statistics; b) no specific or alternative educational programme; c) agreement from teachers, parents, and pupils to participate in a three-year research project involving extensive data collection through psychological tests, interviews, and classroom observations; d) an experienced teacher, whose practice was consistently rated as effective by school leadership, colleagues, parents, and the Czech School Inspectorate. Although these indicators offer a reasonable proxy for effective teaching, they do not encompass all dimensions of teacher quality, which is a limitation of the study. Class sizes ranged from 15 to 29 pupils, with an average of 23, which corresponds to the national mean in the Czech Republic.

3 Results

Descriptive statistical results of mathematical word problem-solving during Year 2 and Year 3 are presented in Table 1. This comprehensive analysis encapsulates crucial metrics encompassing the mean success rate, average self-assessment scores, and the accuracy of these self-assessments.

Table 1: Scores and self-assessment in mathematical word problems in Year 2 (M2) and Year 3 (M3)

	N	Minimum	Maximum	Mean	Std. Deviation
M2 score	624	0	13	11.18	3.05
M3 score	608	0	12	8.16	4.20
M2 self-assessment	599	1	3	1.25	0.49
M3 self-assessment	591	1	3	1.67	0.63

Both mathematical word problems were proficiently addressed, with an average score of 11.2 out of a maximum of 13 points in M2, and 8.2 out of 12 points in M3. While 62.7% of Year 2 pupils accurately solved the M2 problem, a substantial 27.9% responded partially correctly, and only 9.4% answered incorrectly. However, in Year 3, the scenario underwent a significant transformation, with only 31.3% solving Problem M3 correctly, 48.8% responding partially correctly, and 20% offering incorrect solutions. Notably, despite an overall dominance of high performance, there was a substantial decline in the success rate between Year 2 and Year 3 ($p < 0.001$). The M2 test was less challenging due to its delayed administration caused by the COVID-19 situation. Therefore, the poorer performance on the

M3 test does not necessarily indicate a decline in children's abilities. These differences in performance prompted further consideration of the structural characteristics of the two tasks.

It is therefore important to note that the apparent decline in pupils' accuracy between M2 and M3 may partly reflect methodological characteristics of the two tests. The third-grade task (M3) was not only more cognitively demanding but also included larger numbers and a more complex wording structure. Therefore, the difference in performance could result from the task's higher difficulty rather than a true decrease in self-assessment accuracy.

Since performance and self-assessment accuracy may follow different developmental trajectories, both outcomes were examined separately. The accuracy decreased as well. In M2, 58% of pupils estimated their achievement adequately, while in M3 it was only 32%. The paired t-test shows that this result is statistically significant ($p < 0.001$) and it is a high effect as measured[2] by Cohen's $d = 0.761$. We would reach the same conclusion in the case of the self-assessment for these two tests, with $p < 0.001$. However, this is no longer a large effect, as Cohen's $d = 0.018$. This divergence highlights that changes in performance and changes in the accuracy of self-assessment cannot be interpreted as parallel phenomena.

Based on the first research question, the analysis concerned whether the accuracy of self-assessment evolves over time. Given the observed differences in both performance levels and accuracy between M2 and M3, it was necessary to examine self-assessment patterns in a more controlled way. To ensure that the analysis is not burdened by different success rates in Task 2 and Task 3, we looked at accuracy separately for the group of pupils who gave a completely correct or incorrect answer in each task. The results are presented in Table 2. The table shows that in M2, almost twice as many pupils considered their solution correct in a situation where they solved the problem incorrectly in comparison with M3 (63,3% vs. 36.4%). In contrast only slightly more than half of the successful pupils in M3 considered their solution to be correct.

Table 2: Comparison of the self-assessment accuracy and the achievement

		M2		M3	
Achievement in Task	Self-Assessment	Frequency	Percent	Frequency	Percent
0 point (wrong answer)	Correct answer	136	63.3%	147	36.4%
	Not sure or average performance	69	32.1%	211	52.2%
	Wrong answer	10	4.7%	46	11.4%
	Total	215	100%	404	100%
1 point (correct answer)	Correct answer	330	85.9%	103	55.1%
	Not sure or average performance	49	12.8%	77	41.2%
	Wrong answer	5	1.3%	7	3.7%
	Total	384	100%	187	100%

In addition, we analysed an independent subgroup of pupils who solved both tasks correctly ($N = 143$). The average accuracy of the self-assessment (on a scale of -1; 1) for item M2 was -0.063 ($SD = 0.20$) and for item M3 -0.21 ($SD = 0.27$). This difference was significant ($p < 0.001$) and showed that the accuracy of self-assessment declined over time among same children who are correct solvers. Pupils are more rigorous in assessing their own performance in test tasks and there is an increasing tendency towards self-deprecation. This pattern can be explained methodologically: M2 was generally easier, allowing most pupils to succeed and thus to correctly evaluate their success. In contrast, M3's complexity may have increased uncertainty, leading to more mismatches between actual and perceived performance.

According to the comparison of the accuracy of self-assessment in Year 2 and Year 3, we divided the pupils into several groups. The accuracy of the self-assessment measured by the Bias procedure takes values from -1 to $+1$, where negative values indicate underestimation (the pupil's estimate was worse than the actual performance), positive values indicate overestimation (the estimate was better than the actual performance), and a value of 0 indicates agreement between the estimate and the actual performance. We compared whether the pupil's self-assessment was adequate (accurate), overestimating or underestimating on both the first and second measures. The proportions of each variation in the accuracy of self-assessment are shown in Table 3.

The results show that the accuracy of pupil self-assessment is quite low. Those who are stably capable of accurate self-assessment make up only 1/5. If we also include the subset of pupils who have moved towards accurate self-assessment in at least the second measurement, the proportion rises to just under 1/3. A similar proportion of pupils tend to overestimate their performance and a slightly higher proportion tend to underestimate their performance.

[2] 0.2–0.5: small effect; 0.5–0.8: medium effect; 0.8 or more: large effect (Cohen, 1988)

Table 3: Number of pupils according to the self-assessment accuracy over time

Accuracy of Self-Assessment		Number of pupils	Percent
M2	M3		
Accurate	Accurate	121	22.4%
Overestimation	Overestimation	57	10.5%
Underestimation	Underestimation	48	8.9%
Accurate or Underestimation	Overestimation	115	21.3%
Accurate or Overestimation	Underestimation	152	28.1%
Overestimation or underestimation	Accurate	48	8.9%

For further analyses, we included only three groups of pupils whose self-assessment accuracy was stable across both measures. These groups were (1) pupils who stably underestimated themselves (their self-assessment was below their real performance; Underestimation/ Underestimation in Table 3), (2) pupils whose self-assessment was stably accurate (their self-assessment matched their performance on both tasks; Accurate/Accurate in Table 3), and (3) pupils who overestimated themselves (their self-assessment overestimated their real performance; Overestimation/Overestimation in the Table 3).

The groups were compared in the cognitive ability measured by Coloured Progressive Matrices (CPM) in Year 2, respectively by Standard Progressive Matrices (SPM) in Year 3. The highest score in the CPM might be 36 and in the SPM 60. An additional variable was used in the teacher's evaluation of pupils'cognitive skills. The score for the dimension "Cognitive skills" was computed from five items and presented as an average on a scale from 1 to 3. The means in all three variables for groups according to the development of self-assessment accuracy are shown in Table 4.

Table 4: Cognitive abilities of three groups based on the self-assessment accuracy

	Underestimated pupils ($N = 48$)		Acccuratly estimated pupils ($N = 121$)		Overestimated pupils ($N = 57$)	
	Mean	*SD*	Mean	*SD*	Mean	*SD*
Teacher's Evaluation	1.64	0.60	1.32	0.46	1.64	0.56
CPM (Year 2)	27.06	5.63	29.90	3.60	26.91	4.39
SPM (Year 3)	37.68	8.20	42.11	6.88	35.50	8.72

Within the comparison of individual groups, a one-factor analysis of variance was used.[3] The determined values are in Table 5.

Table 5: ANOVA – Cognitive abilities of three groups based on the self-assessment accuracy

Areas monitored	ANOVA	Effect size – ω^2
Teacher's Evaluation	$F = 10.964$; $p < 0.001$***	$\omega^2 = 0.042$
Post-hoc analysis	$1 > 5$ ($p = 0.001$)***,	$\omega^2_{\text{Lower Index}} = 0.001$
	$5 < 9$ ($p < 0.001$)***	$\omega^2_{\text{Upper Index}} = 0.071$
Raven 1	$F = 12.818$; $p < 0.001$***	$\omega^2 = 0.058$
Post-hoc analysis	$1 < 5$ ($p < 0.001$)***,	$\omega^2_{\text{Lower Index}} = 0.009$
	$5 > 9$ ($p < 0.001$)***	$\omega^2_{\text{Upper Index}} = 0.091$
Raven 2	$F = 15.729$; $p < 0.001$***	$\omega^2 = 0.057$
Post-hoc analysis	$1 < 5$ ($p < 0.001$)***,	$\omega^2_{\text{Lower Index}} = 0.008$
	$5 > 9$($p < 0.001$)***	$\omega^2_{\text{Upper Index}} = 0.089$

Explanations: *($p < 0.1$); **($p < 0.05$); ***($p < 0.001$)

The analysis of variance shows that statistically significant differences at the one percent significance level are found in all areas. The post-hoc analysis conducted through the Scheffeny test shows that the 5 groups always differ from the same groups. While in terms of the teacher's perspective, the values are significantly lower than the other two groups, in terms of the Raven's test the values are significantly higher for both year groups. In terms of substantive significance, these are not very significant effects.

Pupils who were consistently accurate in word problems had significantly higher cognitive ability (as measured by the standardized test and as judged by the teacher) than both other groups. In contrast, the cognitive abilities of pupils who overestimated themselves were the weakest.

[3] Data were assessed for normality using the Shapiro-Wilk test and equality of variances using Brown-Forsythe test. When necessary data were transform to fit assumptions of normality and homogeneity of variance prior the analysis.

In addition to cognitive ability, which is an individual characteristic, it is necessary to ask to what extent the accuracy of self-assessment varies from class to class. In our research, 29 classes of the same grade were included. Comparison of the proportion of each subgroup of pupils according to the accuracy of self-assessment confirmed strong differences between classes. The number of children with consistently accurate self-assessments in word problems varied across classes from 7% to 39%. Some classes had a strong group of underestimating pupils (up to 22%), while others had a group of overestimating pupils (up to 28%). In some classes, the accuracy of self-assessment was unstable, with up to 1/3 of pupils moving from overestimating to underestimating or in opposite direction between Year 2 and Year 3.

4 Discussion

Our study examines patterns of self-assessment accuracy in mathematical word problem-solving among elementary school pupils. We focus on changes in accuracy across grade levels, the influence of cognitive ability, and the role of classroom environments. The research addresses two questions: (1) How does self-assessment accuracy in solving mathematical word problems develop from Year 2 to Year 3? (2) Is there a relationship between self-assessment accuracy and cognitive ability in this domain? Because the Year 2 test was less challenging due to COVID-19 delays, children performed better in Year 2 than in Year 3. This decline in performance was accompanied by a noticeable drop in self-assessment accuracy. Although average accuracy remained relatively stable, older pupils in Year 3 were more likely to believe they had solved the task incorrectly, suggesting a reduced proportion of children capable of accurately judging their performance. Methodologically, these differences may be driven not only by development but also by variation in task difficulty and the limited number of items. With only one complex word problem per grade, random factors (e.g., text comprehension, familiarity with context) could disproportionately influence accuracy. Future studies should therefore employ multiple tasks of varying difficulty levels.

Although our observations cover only two school years, the results align with Harter's (2011) assertion that self-assessment tends to decline as children move from early to middle childhood. This shift often reflects a more critical self-view, which may lead to performance underestimation. At the same time, it is important to consider potential teacher influences—third-grade teachers may apply stricter evaluation criteria than second-grade teachers—an issue explored later in this article.

We also examined the relationship between self-assessment accuracy and cognitive ability. Using standardized tests and teacher evaluations, we found substantial differences in cognitive skills across accuracy groups. This finding supports conclusions by Bradshaw (2001) and Dunning, Heath, and Suls (2004), who identified cognitive ability as a key factor underlying reduced accuracy in self-assessment.

Our study revealed variation in self-assessment accuracy across classrooms. Beyond age and cognitive ability, teaching practices and classroom climate played central roles. Interestingly, the frequency of self-assessment activities did not predict accuracy. This suggests that not only the presence but also the form and quality of self-assessment practices matter. Teachers' communication, instructional strategies, and assessment methods were identified as influential factors, consistent with findings by Brookhart et al. (2004) and Andrade and Valtcheva (2009). More targeted research is needed to understand these dynamics.

Self-assessment is an essential component of self-regulated learning, which is increasingly recognized as foundational for academic success and lifelong learning. Accurate self-assessment enables pupils to reflect on performance, set realistic goals, and assume responsibility for their learning. Our contribution lies in showing that self-assessment accuracy does not automatically improve with age or exposure to self-assessment activities. In fact, accuracy may decline when cognitive demands increase or when classroom conditions become more complex. This challenges the assumption that developmental progression alone enhances metacognitive skills and underscores the need for intentional pedagogical strategies.

Our findings illustrate the dynamic nature of self-assessment and its variation across pupils with different cognitive abilities. Teachers should be mindful of these differences and adjust their approaches accordingly. Simply expecting children to develop more precise self-assessments as they mature is insufficient. Developmental shifts may introduce new challenges, such as heightened sensitivity to self-esteem threats or increased peer comparison, which can impede accuracy. Developing precise self-assessment therefore requires carefully structured practices tailored to pupils' cognitive and metacognitive abilities (Siegesmund, 2017). Pupils with weaker abilities may require targeted interventions to strengthen their capacity for evaluating their performance.

Even in classrooms where teachers implement self-assessment tools, our study highlights inconsistencies in the proportion of pupils who assess themselves accurately. These discrepancies may stem from uniform, one-size-fits-all approaches using a single self-report scale. Such approaches fail to account for pupils' individual tendencies to overestimate or underestimate. Our findings emphasize the need for more nuanced and individualized self-assessment methods. Teachers should adapt strategies and tools to pupils' characteristic self-evaluation patterns, which may be linked to personality and self-esteem. While our study does not provide an exhaustive set of effective techniques, it clearly signals the need for further research in this direction.

Several limitations should be considered when interpreting the findings. The comparability of the two assessment points was influenced by a delay in administering the first test, which resulted in a slightly lower difficulty level than intended. Although tasks M2 and M3 had comparable structures, the broader number range and greater linguistic complexity of M3 may have contributed to differences in performance and accuracy unrelated to development. The small number of test items and participating classes also reduces the generalizability of the findings. Although the same children were tested at both time points, the design does not constitute a fully controlled longitudinal study, as differences in task difficulty may influence interpretations of developmental change. Testing in familiar classrooms reduced stress but may have introduced subtle social influences. Finally, the absence of standardized measures of metacognition and the limited range of tasks likely constrained our analyses. Future research would benefit from longitudinal designs, use of multiple task types, and more controlled conditions.

Future work should employ a more differentiated set of assessment tasks varying in linguistic and numerical complexity. Multiple items at each measurement point would allow for more reliable estimates of accuracy and help distinguish task-specific effects from genuine developmental trends. Incorporating standardized metacognition instruments would also strengthen analyses. A longitudinal design following the same cohort over several years would further clarify how cognitive ability, task difficulty, and instructional context shape the accuracy of pupils' self-evaluations over time.

5 Conclusions

In conclusion, our research aligns with contemporary educational paradigms that emphasize learner agency, self-regulation, and the development of accurate self-assessment. Our findings contribute to the growing body of evidence that self-assessment is not a static skill but one requiring intentional support and differentiation. The study reveals that pupils' self-assessment does not inevitably improve with age; instead, it may decline. This poses a significant educational challenge and highlights the need for pedagogical approaches designed to counteract this decline and increase the proportion of pupils capable of accurate self-evaluation.

Acknowledgment

This study was financially supported by the Ministry of Education, Youth and Sports of the Czech Republic through the grant *"Teachers' understanding of the causes of school failure and the effectiveness of educational interventions"* (Project No. CZ.02.3.68/0.0/0.0/19_076/0016390). The research was further supported by the Technology Agency of the Czech Republic within the project *"Supporting the integration of mathematical, reading and language literacy in primary school pupils"* (Project No. TL03000469), carried out at the Faculty of Education, Charles University.

References

Andrade, H. L. (2019). A critical review of research on student self-assessment. *Frontiers in Education*, *4*, 87. https://doi.org/10.3389/feduc.2019.00087

Andrade, H., & Valtcheva, A. (2009). Promoting learning and achievement through self-assessment. *Theory into Practice*, *48*(1), 12–19. https://doi.org/10.1080/00405840802577544

Bakar, M. A. A., & Ismail, N. (2020). Metacognitive learning strategies in mathematics classroom intervention: A review of implementation and operational design aspect. *International Electronic Journal of Mathematics Education*, *15*(1), em0555. https://doi.org/10.29333/iejme/5937

Bandura, A. (2014). Social-cognitive theory. In R. B. Ewen, *An introduction to theories of personality* (pp. 341–360). Psychology Press.

Barana, A., Boetti, G., & Marchisio, M. (2022). Self-assessment in the development of mathematical problem-solving skills. *Education Sciences*, *12*(2), 81. https://doi.org/10.3390/educsci12020081

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Research into practice: Formative assessment for learning. In E. F. Redish & M. Vicentini (Eds.), *Research on Physics Education* (pp. 91–102). IOS Press. https://doi.org/10.3254/978-1-61499-012-3-91

Bradshaw, B. K. (2001). Do pupils effectively monitor their comprehension? *Reading Horizons: A Journal of Literacy and Language Arts*, *41*(3), 2. Retrieved from https://scholarworks.wmich.edu/reading_horizons/vol41/iss3/2

Broadfoot, P. (2021). *The sociology of assessment: Comparative and policy perspectives. The selected works of Patricia Broadfoot.* Routledge. https://doi.org/10.4324/9780429262609

Brookhart, S. M., Andolina, M., Zuza, M., & Furman, R. (2004). Minute math: An action research study of student self-assessment. *Educational Studies in Mathematics*, *57*(2), 213–227. https://doi.org/10.1023/B:EDUC.0000049293.55249.d4

Brown, G. T., Andrade, H. L., & Chen, F. (2015). Accuracy in student self-assessment: Directions and cautions for research. *Assessment in Education: Principles, Policy & Practice*, *22*(4), 444–457. https://doi.org/10.1080/0969594X.2014.996523

Brown, G. T., & Harris, L. R. (2013). Student self-assessment. In J. H. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. 367–393). Sage.

Brown, G., & Harris, L. R. (2014). The future of self-assessment in classroom practice: Reframing self-assessment as a core competency. *Frontline Learning Research*, *2*(1), 22–30. https://doi.org/10.14786/flr.v2i1.24

Cera, R., Mancini, M., & Antonietti, A. (2013). Relationships between metacognition, self-efficacy and self-regulation in learning. *Journal of Educational, Cultural and Psychological Studies (ECPS Journal)*, *4*(7), 115–141. https://doi.org/10.7358/ecps-2013-007-cera

Chytrý, V., Říčan, J., & Živná, D. (2019). Matematická výkonnost a metakognice žáků základních škol běžných, základních škol Montessori a žáků vyučovaných podle Hejného metody. *Studia paedagogica*, *24*(1), 107–133. https://doi.org/10.5817/SP2019-1-5

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

De Bruin, A. B., & van Gog, T. (2012). Improving self-monitoring and self-regulation: From cognitive psychology to the classroom. *Learning and Instruction*, *22*(4), 245–252. https://doi.org/10.1016/j.learninstruc.2012.01.003

Downing, S. M. (2006). Twelve steps for effective test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3–25). Lawrence Erlbaum Associates Publishers.

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

Elder, A. D. (2010). Children's self-assessment of their school work in elementary school. *Education 3–13*, *38*(1), 5–11. https://doi.org/10.1080/03004270802602044

Georghiades, P. (2004). From the general to the situated: Three decades of metacognition. *International Journal of Science Education*, *26*(3), 365–383. https://doi.org/10.1080/0950069032000119401

Harter, S. (2011). The development of self-representations. In M. R. Leary & J. P. Tangney (Eds.), *Handbook of self and identity* (2nd ed., pp. 229–245). Guilford Press.

Kajamies, A., Vauras, M., & Kinnunen, R. (2010). Instructing low-achievers in mathematical word problem solving. *Scandinavian Journal of Educational Research*, *54*(4), 335–355. https://doi.org/10.1080/00313831.2010.493341

Klenowski, V. (1995). Student self-evaluation processes in student-centred teaching and learning contexts of Australia and England. *Assessment in Education: Principles, Policy & Practice*, *2*(2), 145–163. https://doi.org/10.1080/0969594950020203

Kramarski, B., & Revach, T. (2009). The challenge of self-regulated learning in mathematics teachers' professional training. *Educational Studies in Mathematics*, *72*, 379–399. https://doi.org/10.1007/s10649-009-9204-2

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Li, Z., & Rubie-Davies, C. M. (2017). Teachers matter: Expectancy effects in Chinese university English-as-a-foreign-language classrooms. *Studies in Higher Education*, *42*(11), 2042–2060. https://doi.org/10.1080/03075079.2015.1130692

Liu-Ambrose, T., Davis, J. C., Nagamatsu, L. S., Hsu, C. L., Katarynych, L. A., & Khan, K. M. (2010). Changes in executive functions and self-efficacy are independently associated with improved usual gait speed in older women. *BMC geriatrics*, *10*(1), 25. https://doi.org/10.1186/1471-2318-10-25

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199–218. https://doi.org/10.1080/03075070600572090

Nováková, E. (2024). Metacognitive skills of pupils in primary mathematics education. *Journal of Elementary Education*, *17*(2), 223–239. https://doi.org/10.18690/rei.3274

Orth, U., Dapp, L. C., Erol, R. Y., Krauss, S., & Luciano, E. C. (2021). Development of domain-specific self-evaluations: A meta-analysis of longitudinal studies. *Journal of Personality and Social Psychology*, *120*(1), 145–172. https://doi.org/10.1037/pspp0000378

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, *8*, 422. https://doi.org/10.3389/fpsyg.2017.00422

Panadero, E., Jonsson, A., & Botella, J. (2017). Effects of self-assessment on self-regulated learning and self-efficacy: Four meta-analyses. *Educational Research Review*, *22*, 74–98. https://doi.org/10.1016/j.edurev.2017.08.004

Pape, S. J., Bell, C. V., & Yetkin, I. E. (2003). Developing mathematical thinking and self-regulated learning: A teaching experiment in a seventh-grade mathematics classroom. *Educational Studies in Mathematics*, *53*(3), 179–202. https://doi.org/10.1023/A:1026062121857

Perry, N. E., Phillips, L., & Hutchinson, L. (2006). Mentoring student teachers to support self-regulated learning. *The Elementary School Journal*, *106*(3), 237–254. https://doi.org/10.1086/501485

Pintrich, P. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 452–502). Academic Press.

Rendl, M., Vondrová, N., Hřibková, L., Jirotková, D., Kloboučková, J., Kvasz, L., & Žalská, J. (2013). *Kritická místa matematiky na základní škole očima učitelů* [Critical places of primary mathematics in the eyes of teachers]. Univerzita Karlova v Praze, Pedagogická fakulta.

Ross, J. A. (2006). The reliability, validity, and utility of self-assessment. *Practical Assessment, Research, and Evaluation*, *11*(1), 10. https://doi.org/10.7275/9wph-vv65

Ross, J. A., Hogaboam-Gray, A., & Rolheiser, C. (2002). Student self-evaluation in grade 5–6 mathematics: Effects on problem-solving achievement. *Educational Assessment*, *8*(1), 43–59. https://doi.org/10.1207/S15326977EA0801_03

Šauerová Švamberk, M., & Smetáčková, I. (2022). Accuracy of pupils'self-assessment. *EduPort*, *6*(2), 13–25. https://doi.org/10.21062/edp.2022.009

Schleppegrell, M. J. (2007). The linguistic challenges of mathematics teaching and learning: A research review. *Reading & Writing Quarterly*, *23*(2), 139–159. https://doi.org/10.1080/10573560601158461

Schraw, G., Crippen, K. J., & Hartley, K. (2006). Promoting self-regulation in science education: Metacognition as part of a broader perspective on learning. *Research in Science Education*, *36*(1–2), 111–139. https://doi.org/10.1007/s11165-005-3917-8

Siegesmund, A. (2017). Using self-assessment to develop metacognition and self-regulated learners. *FEMS Microbiology Letters*, *364*(11), fnx096. https://doi.org/10.1093/femsle/fnx096

Sitzmann, T., Ely, K., Brown, K. G., & Bauer, K. N. (2010). Self-assessment of knowledge: A cognitive learning or affective measure? *Academy of Management Learning & Education*, *9*(2), 169–191. https://doi.org/10.5465/AMLE.2010.51428542

Taber, K. S. (2025). Mediated learning leading development: The social development theory of Lev Vygotsky. In B. Akpan & T. Kennedy (Eds.), *Science education in theory and practice: An introductory guide to learning theory* (2nd ed., pp. 275–292). Springer.

Tachie, S. A. (2019). Meta-cognitive skills and strategies application: How this helps learners in mathematics problem-solving. *Eurasia Journal of Mathematics, Science and Technology Education*, *15*(5), em1702. https://doi.org/10.29333/ejmste/105364

Yan, Z. (2020). Self-assessment in the process of self-regulated learning and its relationship with academic achievement. *Assessment & Evaluation in Higher Education*, *45*(2), 224–238. https://doi.org/10.1080/02602938.2019.1629390

Zelazo, P. D., & Carlson, S. M. (2012). Hot and cool executive function in childhood and adolescence: Development and plasticity. *Child Development Perspectives*, *6*(4), 354–360. https://doi.org/10.1111/j.1750-8606.2012.00246.x

Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology*, *25*(1), 82–91. https://doi.org/10.1006/ceps.1999.1016

Zimmerman, B. J., & Schunk, D. H. (2011). Self-regulated learning and performance: An introduction and an overview. In D. H. Schunk & B. Zimmerman (Eds.), *Handbook of self-regulation of learning and performance* (pp. 15–26). Routledge.