A Tablet-Computer-Based Tool to Facilitate Accurate Self-Assessments in Third- and Fourth-Graders

https://doi.org/10.3991/ijet.v13i10.8876

Denise Villányi(✉), Romain Martin, Philipp Sonnleitner,
Christina Siry, Antoine Fischbach
University of Luxembourg, Esch-sur-Alzette, Luxembourg
denise.villanyi@gmx.de

Abstract—Although student self-assessment is positively related to achievement, skepticism about the accuracy of students' self-assessments remains. A few studies have shown that even elementary school students are able to provide accurate self-assessments when certain conditions are met. We developed an innovative tablet-computer-based tool for capturing self-assessments of mathematics and reading comprehension. This tool integrates the conditions required for accurate self-assessment: (1) a non-competitive setting, (2) items formulated on the task level, and (3) limited reading and no verbalization required. The innovation consists of using illustrations and a language-reduced rating scale. The correlations between students' self-assessment scores and their standardized test scores were moderate to large. Independent of their proficiency level, students' confidence in completing a task decreased as task difficulty increased, but these findings were more consistent in mathematics than in reading comprehension. We conclude that third- and fourth-graders have the ability to provide accurate self-assessments of their competencies, particularly in mathematics, when provided with an adequate self-assessment tool.

Keywords—tablet-computer-based testing, student self-assessment, elementary education

1 Student Self-Assessment

Academic self-assessment refers to the activity of evaluating one's (academic) performances or processes [1]. When students are asked to self-assess, they are challenged to think of and to express their understanding and performance. Furthermore, when using such information, teachers can facilitate, adjust, and improve their teaching and learning activities [2] [3]. Studies have shown that self-assessment ability, defined as the ability to assess one's performance with acceptable accuracy, can be taught [4] [1]. Recent publications have even suggested that student self-assessment should no longer be treated solely as an assessment but rather as a core competency for self-regulation [5].

1.1 Self-assessment benefits

In their integrative review on the relation between self-assessment and academic achievement, Brown and Harris (2013) [1] reported positive median self-assessment effects ranging from $d = 0.40$ to 0.45, indicating that student self-assessment is substantially related to educational success; see also [6] [7] [8] [9].

The relation between self-assessment and academic achievement can be explained through the processes of self-regulation [10] [11] [12] and through self-efficacy, the latter defined as "beliefs in one's capabilities to organize and execute the courses of action required to produce given attainments" [13] p. 3. It has been found that self-regulating students engage in three processes to observe and to interpret their behaviors: self-observations by which students concentrate on specific aspects of their performance relevant to their perception of success, self-judgments by which they assess the extent to which they have met their goals, and self-reactions by which they assess how satisfied they are with the results of their actions [10] [8] [11] [12]. A positive or high self-assessment induces higher self-efficacy and consequently leads to goal-setting that is conducive to learning (mastery learning) and to higher persistence in goal-attainment [8] [14] and hence to improved learning outcomes. However, a negative or low self-assessment does not automatically lead to low self-efficacy if students believe in their ability to learn and they adapt their behavior accordingly through self-regulation [15]. Nevertheless, self-assessment does not automatically induce self-regulation. The positive impact depends on a variety of factors, for example, characteristics of the learner, of the task, and task outcome; characteristics of the assessment feedback given; and the conceptual model and setting in which the (self-)assessment takes place [16]. Regarding the last factor, student self-assessment combined with the feedback from a contextually important evaluator (i.e., teacher, tutor, peer) is more likely to induce self-regulatory activity [16]. Self-assessment training influences one or all of the previously mentioned self-regulatory processes and ideally includes interactions with or feedback from teachers or peers [6] [7] [8] [9].

In line with the explanations for how self-assessment is related to achievement, there is empirical evidence that self-assessment positively affects student motivation [11], engagement [17], self-efficacy [18] [19] [20], persistence (specifically on challenging tasks) [21], student choice of reference norms (ipsative vs. social) [8], and the quality of student-teacher relationship [22].

1.2 Self-assessment accuracy

Despite the growing body of knowledge regarding the benefits of student self-assessment, skepticism about learners' self-assessment accuracy remains [1] [23] [24]—with self-assessment accuracy generally defined and operationalized as the consistency between self-assessment and corresponding external judgements (e.g., test scores, school grades, peers' judgments, parents' judgments) [25] [24] [23] [1] [26]. An accurate validation measure of high psychometric quality is recommended when measuring the accuracy of students' self-assessment [25].

Self-assessment accuracy and student age. Beneficial effects of self-assessment can already be found in elementary school children [6] [7] [8] [9] (see also Section 1.1). However, due to the (meta)cognitive immaturity of such young children—in comparison with adults—doubts about young children's self-assessment accuracy are particularly persistent in the literature [27] [28] [29] [30]. Findings from studies with pre- and elementary school children have suggested that self-assessment accuracy depends more on the conditions under which self-assessment is practiced [31] [32] [33] [34] and the appropriateness of the self-assessment instruments [35] [36] [37] than on learners' age per se. Recent reviews on metacognition [38] and metacognitive development [39] have supported this argument.

Regarding the conditions, young students' self-assessment tends to be more accurate (i.e., more strongly correlated with external competency judgments) when measured in a mastery condition than in a competitive one [31]. In Butler's study (1990) [31], in the mastery condition, children had to self-assess their drawing by comparing it with a standard drawing (mastery goal), whereas in the competitive one, they had to self-assess their drawing by comparing it with their peers' drawings. Whereas the mastery condition induced the students to compare their work with a mastery criterion, the competitive condition induced a social comparison with other classmates. In the mastery condition, the correlations between students' self-assessments and teacher judgments were significant ($p < .05$) and moderate to large [40] at all ages: age 5 ($r = .48$), age 7 ($r = .56$), and age 10 ($r = .58$). In the competitive condition, the correlations between students' self-assessment and teacher judgments were significant ($p < .001$) but smaller at ages 5 ($r = .16$) and 7 ($r = .38$) and larger at age 10 ($r = .83$) in comparison with the mastery condition. Butler (1990) concluded from these results that children at age 10 are able to adopt normative goals and criteria for self-assessment in a competitive condition and mastery goals in the match-the-standard condition [31]. Notwithstanding, students' self-assessments tend to be more accurate in mastery goal structures than in competitive educational contexts, independent of age [41].

Regarding self-assessment instruments, self-assessment accuracy tends to increase if the content and specificity of the self-assessment items closely correspond with criterial performance [1] [42]. Precise items about very specific skills tend to yield higher accuracy with objective performance than items about general competency in one domain [24] [32]: "How confident are you that you can correctly spell all words in a one-page story or composition?" [43] versus "I have always done well in writing" [44]; "How confident are you that you can successfully solve equations containing square roots?" [45] versus "I am quite good at mathematics" [46]. The description of specific competencies offers precise criteria against which students can assess their competency; this avoids the use of subjective, less appropriate criteria (e.g., self-serving criteria; social comparison) when students are asked to self-assess a given competency [24] [1] [42]. In the same vein, with respect to adequate validation strategies, symmetry principles derived from Brunswik's lens model [47] suggest that relationships between variables are strongest when these variables are measured on a similar level of abstraction [48].

Finally, a limitation in young learners' competency to verbalize their auto-perceptions [34] or to read and understand written items [37], which are prerequisites of many

self-assessment operationalization (see e.g., [36] [7] [49] [50]), could be misunderstood and misinterpreted as a limitation in students' self-assessment competency or accuracy.

The self-assessment accuracy of lower performers. The self-assessments of lower performing students tend to be less accurate than those of higher performing students [1]. Claes and Salame (1975) [51] found significant differences ($F = 11.25$, $p < .01$) between high and low performers in their self-assessment accuracy. Possible explanations for these findings are that lower performing students may lack an adequate representation of what is expected from them and might not understand the assessment criteria [52] [51] [53], both of which may lead to inaccurate self-assessment. When reporting their school grades, lower performing students may succumb to social desirability or self-enhancement factors [54]. Similar to self-assessment accuracy in relation to student age (see the *Self-assessment accuracy and student age* section), developmental differences or gaps in metacognitive skills [38] are possible explanations for the observed discrepancy. An interesting finding is that low performers seem to gain the most from self-assessment training—which usually includes the training of metacognitive skills (defining assessment criteria and using them for self-assessment)—in terms of performance gains (ES = 0.58) [52] (see also [55]).

2 The Present Study

In the present study, we aimed to rigorously empirically investigate whether or not third- and fourth-graders in elementary school are able to provide accurate self-assessments of key academic competencies (mathematics and reading comprehension) when equipped with a self-assessment tool that combines all the conditions that are favorable for accurate self-assessment (i.e., a non-competitive setting, task-oriented questions, limited reading, and no verbalization required). In our study, we adopted an innovative approach to self-assessment by introducing a tablet-computer-based self-assessment tool that is rich in illustrations and has a language-reduced rating scale, thus reducing the bias that may come along with poor reading and language skills. Classrooms are becoming increasingly digital [56], and tablet technology [57] [58] may help to facilitate and efficiently integrate self-assessment in the classroom. We intend to feed the self-assessment body of knowledge with new, high-quality empirical data. First, we collected self-assessment data from two independent representative samples of third- and fourth-graders (ages 8 to 9 years) in Luxembourg. Second, we used measures of high psychometric quality, namely, the standardized test scores from the national school monitoring program [59] as measures of validation for students' self-assessment.

2.1 Contextual embedding of the study

The present study was located in Luxembourg, which provides a rather unique learning environment. Specifically, Luxembourg has quite a distinct multilingual educational context (three official languages: Luxembourgish, French, and German; bilingual

education in French and German; literacy acquisition in German) and a very heterogeneous student population. Almost half of all students in public education are foreign nationals, and more than half do not speak Luxembourgish as their native language (e.g., [59]). International large-scale assessments (e.g., the OECD's PISA studies) have repeatedly shown that many educational systems in modern societies—and Luxembourg is by far no exception—struggle with the adequate handling of increasingly diverse student populations [60]. Understanding and learning how to effectively deal with highly heterogeneous groups of learners (i.e., solving the problem of how to provide equal opportunities for success to everybody independent of their socioeconomic, sociocultural, and linguistic background) may be considered the largest educational challenge in Luxembourg today.

Although language learning plays a predominant role in elementary education in Luxembourg and consumes over 40% of the total teaching time, 45% of third-graders do not reach the minimum competence standard in German reading comprehension defined by the national school curriculum for this age group [59]. Because reading skills constitute the basis for all future learning, achievement and learning in all other school subjects are at stake [61]. In 2009, Luxembourg's pre- and elementary schools underwent profound changes. Among other changes, the 2009 education act put particular emphasis on formative assessment, regular feedback, and student self-assessment. Given that fair assessments in mathematics should not be confounded with reading skills and that German is the language in which literacy is acquired in Luxembourg, we created an innovative tool that lowers the impact of language and reading in mathematics self-assessment and provides a way for students to assess their German reading comprehension. Thus, we intended to facilitate and support the goals of the education act in the current study.

Luxembourg's increasing diversity, a logical consequence of the demographic change that comes along with a globalized world, is not exclusively a domestic matter. However, owing to several national specificities (e.g., relatively small size, open borders, situated in the heart of Europe, traditionally multilingual, with an economic model built on and relying on immigration), change might occur more quickly in Luxembourg than in other countries. Accordingly, Luxembourg provides a unique educational and societal learning environment, a living laboratory so to speak, that is prototypical and anticipatory of the demographic changes and the related challenges that its geographical and metaphorical neighbors may very likely face over the next decades.

2.2 An innovative tablet-computer-based self-assessment tool

We decided to use tablet technology to develop the digital self-assessment tool (referred to as "the tool"). On the one hand, tablet computers offer the same technological opportunities as computers [57]. On the other hand, tablet devices allow maximal mobility when used in schools and classrooms and have an intuitive design, a simple interface, touch screen function, and multimedia capabilities that facilitate the user's interaction with the program, particularly for pre- and elementary school children (ibid). On the basis of the principles of the *Cognitive Theory of Multimedia Learning* (CTML) for the design of multimedia instructional messages [62] [63] [64], we developed a tool

that fulfills the following requirements (see also the *Self-assessment accuracy and student age* section): (1) self-assessment in a non-competitive setting, (2) task-oriented self-assessment, and (3) self-assessment that requires only limited reading and no verbalization. The details of how these requirements have been taken into consideration in this tool are elaborated below.

1. In order to avoid competition and social comparison between students [31] [41] [42] [25], we based the items used in the tool on an external reference standard: the national school curriculum. We concentrated on the domains of language (reading comprehension in German) and mathematics (arithmetic operations; geometry and space). The items were developed along three proficiency levels, attuned to the required competency standards of third- and fourth-graders.
2. The self-assessment items are concrete tasks on a sub-competency level and on three proficiency levels. We consciously avoided general questions about competency (e.g., How well are you doing in mathematics?) because they are distal from criterial performance and mastery learning objectives. According to CTML, meaningful learning occurs when the processes of selecting, organizing, and integrating take place for visual and verbal representations. A multimedia instructional message consists of pictures (i.e., animated or static) and words (i.e., spoken or written). It is effective if it helps students hold visual and verbal representations in working memory simultaneously [65] [62]. In our tool, we avoided extraneous material that was not related to the self-assessment tasks, and we aimed to highlight essential material [64] [66]. An important objective for the mathematics tool was the reduction of (written) language. Consequently, we based the items on illustrations and short animations combined with written on-screen text, where necessary, to represent the key sub-competencies in the school curriculum. The language used in the tool is German, the language of instruction in Luxembourg's elementary education. Because the tool was designed for classroom use on an individual basis, we decided not to insert spoken language and sound. The reading comprehension tool consists primarily of written text and does not contain many illustrations.
3. The use of concrete objects [34], pictorial inventories [17] [37], and language-reduced answer scales [67] might be effective for helping young students overcome the verbalization and literacy barriers they experience. Figure 1 provides an example of one of the competencies represented in the tool: the application of arithmetic operations in concrete life situations. In a short video, Paul gets 25 euros and wants to spend them on roller coaster rides. The student is told that 1 ticket costs 6 euros. The student is then asked the actual self-assessment item: How many tickets can Paul buy? The rating is introduced by the sentence: I could solve this problem. We used a language-reduced and pictorial visual analog scale (VAS) as the rating scale for the self-assessments. A VAS is appropriate for capturing subjective perceptions [68] and is considered to be reliable when used with children [67]. Furthermore, a comparative study showed that children with an immigration background preferred the language-reduced VAS over the Likert-type scale [67]. The use of a VAS avoids all bias associated with poor language and reading skills. In our tool, a pictorial nodding head and a pictorial shaking head were placed at the opposite ends of a scale-free

line. Students rated how confident they felt about whether they could solve the presented item by moving a slider along the line.

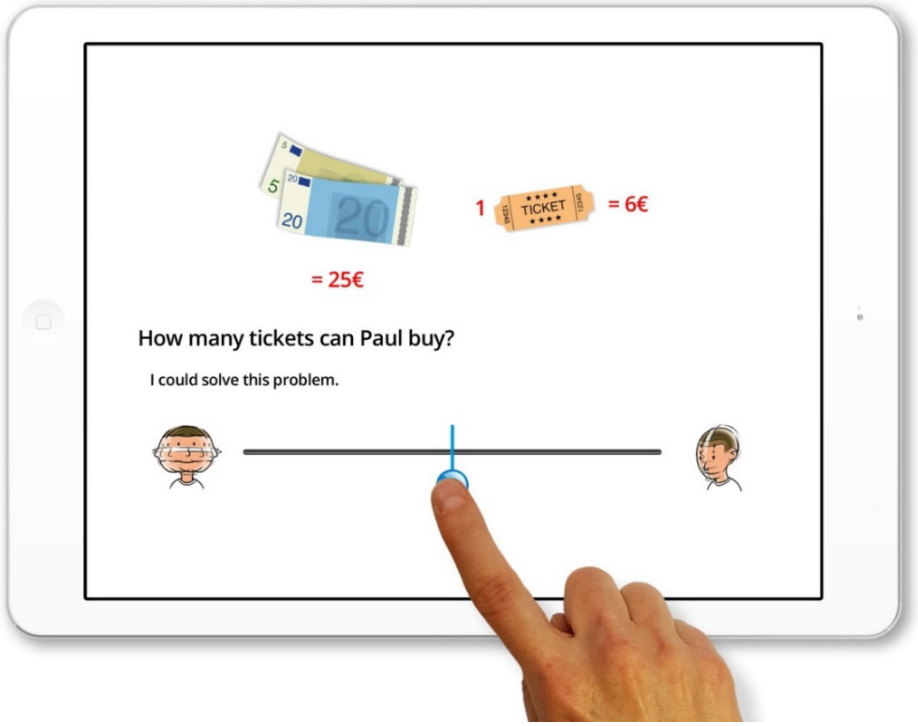

Fig. 1. Example of a competency represented in the self-assessment tool

2.3 Research aims

Skepticism toward elementary school students' self-assessment persists despite the demonstrated benefits of self-assessment in this age group (see Sections 1.1 and 1.2). In the present study, we wanted to investigate whether or not third- and fourth-graders (ages 8 to 9 years) can provide accurate self-assessments of key academic competencies (mathematics, reading comprehension) when equipped with an adequate self-assessment tool.

More concretely, accurate self-assessment requires three things:

1. Consistency between self-assessment and a corresponding external judgment (e.g., test scores, school grades; see the Self-assessment accuracy section). In other words, students' self-assessments should reflect their actual competencies. The strength of the relationship between self-assessment and a test score is typically identified through correlational analysis. If students' self-assessments reflect their actual competencies, medium to large [40] correlations between self-assessment and test scores should emerge. Moreover, when students are organized into proficiency groups on

the basis of their standardized test scores from the national school monitoring program, students in lower proficiency groups should provide lower self-assessments on average in comparison with students in higher proficiency groups. Analysis of variance (ANOVA) can be applied to test for significant differences between these groups.

2. Obtaining self-assessments from third- and fourth-graders requires adapted conditions and instruments (see the Self-assessment accuracy and student age section). Self-assessment accuracy with our tablet-computer-based self-assessment tool (mastery condition, items on the task level) requires that students recognize the inherent difficulty of self-assessment items (see Section 2.2, Point 1, and 2). We can compute self-assessment item mean scores as a function of the (theoretical) difficulty level of the items. ANOVA can be applied to test for significant differences between these mean scores. If students recognize the inherent difficulty of the items, their confidence in solving an item should decrease as item difficulty increases.
3. Lower performers tend to be less accurate in their self-assessments because they do not understand the tasks and assessment criteria or they succumb to social comparison and desirability when asked to report grades (see the Self-assessment accuracy of lower performers section). The tablet-computer-based self-assessment tool has features that can help overcome these obstacles: It is language-reduced and proposes self-assessment on a task level in a non-competitive setting. Consequently, accurate self-assessment with the tablet-computer-based self-assessment tool requires that even lower performers recognize the inherent difficulty of the self-assessment items (see Section 2.2, Point 1, and 2). By organizing the students into proficiency groups on the basis of their standardized test scores from the national school monitoring program, we can compute self-assessment item mean scores for each group as a function of the (theoretical) difficulty level of the items. ANOVA can be applied to test for significant differences between these mean scores within the groups. If lower performers (i.e., students in the lowest proficiency group) recognize the inherent difficulty of the items, their confidence in solving an item should decrease as item difficulty increases.

3 Method

3.1 Sample and procedures

Our study was based on two independent and representative samples from Luxembourg's elementary school population. The samples were chosen randomly from the elementary school districts all over the country. The students were in Grades 3 and 4 and were 8- to 9-years old. Our tool was designed for classroom use, and we facilitated the administration of the tool ourselves by bringing the tablet computers to the schools. Teachers were present during the testing.

The first round of data collection with the tool took place in autumn 2014 in 14 different fourth-grade classes. The final samples consisted of $N = 191$ students (51.31% girls, 48.69% boys) in mathematics and $N = 187$ students (51.87% girls, 48.13% boys)

in reading comprehension. 42.93% of the mathematics and 44.39% of the reading comprehension samples were students who predominantly spoke a language other than Luxembourgish or German at home (*vs.* 53% in the population; see Martin et al., 2015). 54.97% and 56.15% were students with a migration background (first and second generation, respectively); (*vs.* 50% in the population, see Martin et al., 2015). The second round of data collection took place in spring 2015 in 29 different third-grade classes. The final samples consisted of $N = 370$ students (47.03% girls, 52.97% boys) in mathematics and of $N = 340$ students (45.88% girls, 54.12% boys) in reading comprehension. 54.05% of the mathematics and 51.76% of the reading comprehension samples were students who predominantly spoke a language other than Luxembourgish or German at home. 49.19% and 49.41% were students with an immigration background (first and second generation, respectively). We discarded cases from the initial samples with missing data on the performance tests (see the *Standardized tests* section). For Grade 3, we discarded data from $n = 32$ students in the mathematics sample and $n = 25$ students in the reading comprehension sample. For Grade 4, we discarded data from $n = 20$ students in the mathematics sample and $n = 21$ students in the reading comprehension sample. We also discarded cases that were consistently too quick in their self-assessments, with a median time of 2 seconds or less for one rating, and at the same time showed a mean score of 95 or higher on a 0 to 100 visual analog scale (see Section 2.2, Point 3). For Grade 3, we discarded data from $n = 4$ students in the mathematics sample and $n = 29$ students in the reading comprehension sample. For Grade 4, we discarded data from $n = 14$ students in the mathematics sample and $n = 23$ students in the reading comprehension sample. We conducted our study with the approval of the Luxembourg Ministry of Education in accordance with the data protection rules of the National Commission for Data Protection. Parents and students were informed in writing about the scientific background of the study well in advance and were given the opportunity to refuse to participate in the study.

3.2 Measures

Self-assessment. The students' self-assessments in mathematics and German reading comprehension were measured with the tablet-computer-based self-assessment tool. The scales contained 66 items in mathematics and 34 items in reading comprehension. In mathematics, the internal consistency reliability (Cronbach's alpha) was .97 in Grade 3 and .96 in Grade 4. In reading comprehension, the internal consistency reliability (Cronbach's alpha) was .94 in Grade 3 and .95 in Grade 4 (see Table 2). The rating scale we used was a visual analog scale (VAS) that contained 101 hidden positions, allowing students to score from 0 to 100 [0, 100] after each item. We computed the mean self-assessment scores in mathematics and in reading comprehension separately.

Each of the self-assessment items in our tool was assigned one of three possible difficulty levels defined on the basis of an external reference standard: the national school curriculum. These assignments were approved by a group of experts, composed of elementary school teachers and researchers who focused on item and test development and were responsible for the standardized tests used in the Luxembourg school

monitoring program. In the tool, the theoretical item difficulty levels increase from level 1 to level 3. Level 2 items represent the minimum competency standard required for Grade 3; the level 1 items are below and the level 3 items are above this competency standard. In reading comprehension, the number of items per levels 1, 2, and 3 were 9, 13, and 12, respectively. In mathematics, the item distribution was 4, 21, and 41, respectively. Knowing from other studies that students generally tend to self-assess high [18] [43], we deliberately integrated more items on levels 2 and 3 than on level 1 to avoid ceiling effects. Moreover, the number of items ensures a valid representation with regard to the content of the measured competencies. The testing time was 40 minutes for mathematics and 20 minutes for reading comprehension. The introduction to the tool took 10 minutes.

Standardized tests. Luxembourg's school monitoring program [59] consists of yearly standardized tests in mathematics and German reading comprehension in Grade 3. These tests are based on the competency standards of the national school curriculum. We used the standardized test scores from the school years 2013/2014 and 2014/2015 to validate the self-assessment measures because they represent an accurate measure of students' academic competencies with high psychometric quality. On these tests, the person parameters (Warm's Weighted Likelihood Estimator scores; see [69]) for the whole population of third-graders were standardized to $M = 500$ and $SD = 100$. In mathematics, the WLE reliability was .89 in 2013/2014 and .92 in 2014/2015. In reading comprehension, the WLE reliability was .88 in 2013/2014 and .89 in 2014/2015 (see Table 2).

In our study, on the basis of students' standardized test scores, we assigned them to three proficiency groups (see Table 1). The cut scores are theoretically embedded and derived from the national monitoring program [70]. Students in proficiency group 1 performed below the minimum competency standard required for Grade 3; those in proficiency group 2 (>437.57 points in mathematics; >484.90 points in reading comprehension) reached the standard; and those in proficiency group 3 (>520.53 points in mathematics; >543.95 points in reading comprehension) performed above the standard. In both samples, the scores from the standardized tests closely corresponded to the standardized population mean of $M = 500$ and $SD = 100$ (see Table 2).

Table 1. Students by proficiency group in mathematics and reading comprehension in Grades 3 and 4

Grade 3							
Mathematics test				***Reading comprehension test***			
Proficiency group	*n*	*% in sample*	*% in population*	*Proficiency group*	*n*	*% in sample*	*% in population*
1	130	35.14	33.00	1	173	50.88	46.00
2	113	30.54	30.00	2	72	21.18	21.00
3	127	34.32	38.00	3	95	27.94	32.00
Grade 4							
Mathematics test				***Reading comprehension test***			
Proficiency group	*n*	*% in sample*	*% in population*	*Proficiency group*	*n*	*% in sample*	*% in population*
1	35	18.32	26.00	1	67	35.83	42.00
2	66	34.55	31.00	2	50	26.74	24.00
3	90	47.13	43.00	3	70	37.43	34.00

Note. n = number of students per proficiency group. Proficiency group: 1 = below the competency standard for Grade 3; 2 = at the standard; 3 = above the standard. % in sample = percentage in the self-assessment sample. % in population = percentage in the school monitoring population.

Table 2. Intercorrelations between self-assessment scores and test scores in mathematics and reading comprehension in Grades 3 and 4

r	**Grade 3**			
	1	***2***	***3***	***4***
1 Self-assessment mathematics	--			
2 Self-assessment reading comprehension	.49	--		
3 Mathematics test	.58	.38	--	
4 Reading comprehension test	.33	.46	.59	--
Mean	79.95	90.26	478.48	500.08
SD	15.80	12.30	106.94	107.24
Range	4.67 – 100.00	42.53 – 100.00	191.40 – 841.67	167.07 – 869.48
Reliability	.97[a]	.94[a]	.92[b]	.89[b]
r	**Grade 4**			
	1	***2***	***3***	***4***
1 Self-assessment mathematics	--			
2 Self-assessment reading comprehension	.57	--		
3 Mathematics test	.40	.41	--	
4 Reading comprehension test	.28	.45	.61	--
Mean	86.29	91.66	517.08	521.42
SD	12.75	12.86	93.54	99.86
Range	39.50 – 100.00	30.50 – 100.00	182.14 – 788.93	299.89 – 788.57
Reliability	.96[a]	.95[a]	.89[b]	.88[b]

Note. Coefficients in bold are significant at the .01 level (2-tailed).
[a] Cronbach's alpha. [b] WLE reliability

3.3 Statistical analyses

All statistical analyses were computed separately for each sample (Grade 3, Grade 4) and each domain (mathematics, reading comprehension).

We applied correlational analysis (Pearson product moment correlation) between students' mean self-assessment scores and their standardized test scores to check whether students' self-assessments reflected their actual competencies.

After assigning each student to one of the three proficiency groups (see the *Standardized tests* section), we computed a univariate analysis of variance (one-way ANOVA) and post hoc comparisons to test for significant differences between the mean self-assessment scores of these groups. Students' proficiency (three groups) served as the independent variable; the dependent variable was the mean score on the self-assessment scale. We computed another one-way ANOVA and post hoc comparisons to test whether students recognized the inherent difficulty of the self-assessment items. The theoretical item difficulties (three levels) served as the independent variable; the dependent variable was the mean score on the self-assessment scale. Finally, we computed 12 independent one-way ANOVAs and post hoc comparisons, one for each proficiency group in each domain and for each sample (3 x 2 x 2). These analyses allowed us to check whether students recognized the items' inherent difficulty independent of their proficiency. The theoretical item difficulty served as the independent variable; the dependent variable was students' mean score on the self-assessment scale computed separately for each proficiency group.

For all tests, we applied a significance level of $\alpha = .05$. If there was heterogeneity of variance in the independent variable groups, we applied Welch's F test (robust ANOVA) to test for main effects. If there was a main effect, we computed a Games-Howell post hoc test for a pairwise check of the effects between the groups. When homogeneity of variance was confirmed, we applied the ANOVA F test, followed by a Tukey HSD test when there was a significant main effect. Welch's F test, Games-Howell, and Tukey HSD post hoc tests were computed to control the Type I error rate when heterogeneity of variance and nonnormality were present (e.g., [71] [72] [73] [74]). We used Hedge's g, a measure of effect size weighted according to the relative size of each sample.

4 Results

On a general level, self-assessments were high, and there was a tendency for them to be higher for reading comprehension than for mathematics (see Table 2).

4.1 Relation between self-assessment and standardized tests

The correlations between the self-assessment scores and the standardized test scores were moderate to large [40], ranging from $r = .40$ to .58 in mathematics and from $r = .45$ to .46 in reading comprehension (see Table 2). Except for a correlation of $r = .58$ in mathematics in Grade 3, there were no significant differences in the strengths of the relations between the self-assessment and standardized test scores between domains and grade levels.

For mathematics, as student proficiency increased, the mean of the self-assessment scores increased (see Table 3). The ANOVAs showed significant main effects (see Table 3). In Grade 3, post hoc comparisons revealed significant differences between the mean self-assessment scores of all three proficiency groups, with medium to large effect sizes (see Table 3). In Grade 4, there were significant differences between proficiency groups 1 and 3 and groups 2 and 3 with medium to large effect sizes (see Table 3).

For reading comprehension, as student proficiency increased, the mean self-assessment scores increased (see Table 3). The ANOVAs indicated significant main effects (see Table 3). The post hoc comparisons revealed significant differences between the mean self-assessment scores of proficiency groups 1 and 2 and groups 1 and 3, with medium to large effect sizes (see Table 3).

4.2 Self-assessment by item difficulty level

For mathematics, as the theoretical item difficulty increased, students' self-assessment scores decreased (see Table 4). The ANOVAs indicated significant main effects (see Table 4). Post hoc comparisons revealed significant differences between the mean self-assessment scores of item level groups 1 and 3 and levels 2 and 3, with large effect sizes (see Table 4).

For reading comprehension, as the theoretical item difficulty increased, students' self-assessment scores decreased from level 1 to 2, but they increased again from level 2 to 3 (see Table 4). The ANOVAs revealed significant main effects (see Table 4). Post hoc comparisons showed significant differences between the mean self-assessment scores of item level groups 1 and 2, with large effect sizes (see Table 4).

4.3 Self-assessment by item difficulty within proficiency groups

For mathematics, as the theoretical item difficulty increased, students' self-assessment scores decreased in all proficiency groups (see Tables 5 and 6). The ANOVAs revealed significant main effects (see Tables 5 and 6). Post hoc comparisons indicated significant differences between the mean self-assessment scores of item level groups 1 and 3 and levels 2 and 3, with large effect sizes (see Tables 5 and 6).

For reading comprehension, as the theoretical item difficulty increased, students' self-assessment scores decreased from level 1 to 2 but increased again from level 2 to 3 (see Tables 5 and 6). The ANOVAs showed significant main effects (see Tables 5 and 6). In Grade 3, for proficiency group 1, the post hoc comparison reported a significant difference between the mean self-assessment scores of item level groups 1 and 2, with a large effect size (see Table 5). For proficiency group 2, post hoc comparisons indicated significant differences between item level groups 1 and 2 and levels 1 and 3, with large effect sizes (see Table 5). For proficiency group 3, post hoc comparisons did not indicate any effects between groups (see Table 5). In Grade 4, the ANOVA indicated a main effect only for proficiency group 3 (see Table 6). Post hoc comparisons revealed a significant difference between mean self-assessment scores of item level groups 1 and 2, with a large effect size (see Table 6).

Table 3. Self-assessment by proficiency group in mathematics and reading comprehension in Grades 3 and 4

Grade 3							
Self-assessment mathematics							
				Mean differences |*Mi-Mj*|			
Proficiency group	***n***	***M***	***SD***	***1***	***2***	***3***	
1	130	70.52	17.39	--			
2	113	80.83	13.37	**10.32 (0.66)**	--		
3	127	88.83	9.52	**18.31 (1.30)**	**8.00 (0.70)**	--	
Levene's F	***df1***	***df2***	***p***	***Welch's F***	***df1***	***df2***	***p***
16.36	2	367	.000	58.23	2	228.09	.000
Self-assessment reading comprehension							
				Mean differences |*Mi-Mj*|			
Proficiency group	***n***	***M***	***SD***	***1***	***2***	***3***	
1	173	84.71	14.14	--			
2	72	94.84	6.74	**10.13 (0.81)**	--		
3	95	96.89	5.21	**12.17 (1.03)**	2.04	--	
Levene's F	***df1***	***df2***	***p***	***Welch's F***	***df1***	***df2***	***p***
57.3	2	337	.000	51.29	2	195.64	.000
Grade 4							
Self-assessment mathematics							
				Mean differences |*Mi-Mj*|			
Proficiency group	***n***	***M***	***SD***	***1***	***2***	***3***	
1	35	79.21	15.53	--			
2	66	84.40	12.76	5.19	--		
3	90	90.43	9.80	**11.23 (0.96)**	**6.04 (0.54)**	--	
Levene's F	***df1***	***df2***	***p***	***Welch's F***	***df1***	***df2***	***p***
5.24	2	188	.006	10.86	2	79.28	.000
Self-assessment reading comprehension							
				Mean differences |*Mi-Mj*|			
Proficiency group	***n***	***M***	***SD***	***1***	***2***	***3***	
1	67	85.42	15.73	--			
2	50	93.29	10.87	**7.88 (0.57)**	--		
3	70	96.48	7.84	**11.07 (0.90)**	3.19	--	
Levene's F	***df1***	***df2***	***p***	***Welch's F***	***df1***	***df2***	***p***
15.26	2	184	.000	13.5	2	106	.000

Note. n = number of students. Proficiency group: 1 = below the competency standard for Grade 3; 2 = at the standard; 3 = above the standard. M = mean self-assessment score. SD = standard deviation. Games-Howell post hoc tests used for all comparisons. Mean differences in bold are significant at $p < .05$. Effect sizes (Hedge's g) are in bold and in parentheses. Levene's F = F-ratio for equality of variance. Welch's F = robust F-ratio for analysis of variance. ANOVA F = F-ratio for analysis of variance. df_1 = degrees of freedom for the effect of the model. df_2 = degrees of freedom for the residuals of the model. p = probability.

Table 4. Self-assessment by item difficulty level in mathematics and reading comprehension in Grades 3 and 4

Grade 3							
Self-assessment mathematics (N = 370)							
				Mean differences |*Mi-Mj*|			
Item level	n_i	***M***	***SD***	***1***	***2***	***3***	
1	4	93.94	3.33	--			
2	21	89.50	7.14	4.43	--		
3	41	73.70	12.64	**20.24 (1.66)**	**15.80 (1.42)**	--	
Levene's F	df_1	df_2	***p***	***Welch's F***	df_1	df_2	***p***
6.13	2	63	.004	31.31	2	15.46	.001
Self-assessment reading comprehension (N = 340)							
				Mean differences |*Mi-Mj*|			
Item level	n_i	***M***	***SD***	***1***	***2***	***3***	
1	9	93.31	3.35	--			
2	13	88.12	4.37	**5.18 (1.30)**	--		
3	12	90.29	2.95	3.02	2.17	--	
Levene's F	df_1	df_2	***p***	***ANOVA F***	df_1	df_2	***p***
0.57	2	31	.57	5.34	2	31	.010
Grade 4							
Self-assessment mathematics (N = 191)							
				Mean differences |*Mi-Mj*|			
Item level	n_i	***M***	***SD***	***1***	***2***	***3***	
1	4	95.55	2.10	--			
2	21	92.69	4.63	2.86	--		
3	41	82.11	8.63	**13.44 (1.61)**	**10.58 (1.40)**	--	
Levene's F	df_1	df_2	***p***	***Welch's F***	df_1	df_2	***p***
6.53	2	63	.003	31.09	2	16.23	.001
Self-assessment reading comprehension (N = 187)							
				Mean differences |*Mi-Mj*|			
Item level	n_i	***M***	***SD***	***1***	***2***	***3***	
1	9	93.57	2.41	--			
2	13	89.92	3.56	**3.66** (1.16)	--		
3	12	92.13	2.48	1.45	2.21	--	
Levene's F	df_1	df_2	***p***	***ANOVA F***	df_1	df_2	***p***
0.71	2	31	.501	4.37	2	31	.021

Note. N = number of students. ni = number of items per level. Item level: 1 = below the competency standard for Grade 3; 2 = at the standard; 3 = above the standard. M = mean self-assessment score. SD = standard deviation. Games-Howell post hoc tests used for mathematics; Tukey HSD post hoc tests used for reading comprehension. Mean differences in bold are significant at $p < .05$. Effect sizes (Hedge's g) are in bold and in parentheses. Levene's F = F-ratio for equality of variance. Welch's F = robust F-ratio for analysis of variance. ANOVA F = F-ratio for analysis of variance. df1 = degrees of freedom for the effect of the model. df2 = degrees of freedom for the residuals of the model. p = probability.

Table 5. Self-assessment by item difficulty level and proficiency group in mathematics and reading comprehension in Grade 3

Self-assessment mathematics								**Self-assessment reading comprehension**							
Proficiency group 1, n = 130				***Mean differences |Mi-Mj|***				***Proficiency group 1, n = 173***				***Mean differences |Mi-Mj|***			
Item level	n_i	*M*	*SD*	*1*	*2*	*3*		*Item level*	n_i	*M*	*SD*	*1*	*2*	*3*	
1	4	88.50	6.34	--				1	9	88.95	5.45	--			
2	21	81.23	10.42	7.26	--			2	13	81.84	5.61	**7.11 (1.28)**	--		
3	41	63.27	15.16	**25.22 (1.71)**	**17.96 (1.30)**	--		3	12	84.66	4.96	4.29	2.82	--	
Levene's *F*	df_1	df_2	*p*	Welch's *F*	df_1	df_2	*p*	Levene's *F*	df_1	df_2	*p*	ANOVA *F*	df_1	df_2	*p*
3.65	2	63	.032	23.48	2	11.37	.000	.13	2	31	.882	4.7	2	31	.016
Proficiency group 2, n = 113				***Mean differences |Mi-Mj|***				***Proficiency group 2, n = 72***				***Mean differences |Mi-Mj|***			
Item level	n_i	*M*	*SD*	*1*	*2*	*3*		*Item level*	n_i	*M*	*SD*	*1*	*2*	*3*	
1	4	95.11	3.55	--				1	9	97.08	2.24	--			
2	21	91.58	7.28	3.53	--			2	13	93.40	3.95	**3.68 (1.09)**	--		
3	41	73.94	13.78	**21.17 (1.59)**	**17.65 (1.47)**	--		3	12	94.73	1.62	**2.35 (1.23)**	1.34	--	
Levene's *F*	df_1	df_2	*p*	Welch's *F*	df_1	df_2	*p*	Levene's *F*	df_1	df_2	*p*	Welch's *F*	df_1	df_2	*p*
6.23	2	63	.003	30.36	2	15.17	.000	3.65	2	31	.038	4.83	2	18.06	.021
Proficiency group 3, n = 127				***Mean differences |Mi-Mj|***				***Proficiency group 3, n = 95***				***Mean differences |Mi-Mj|***			
Item level	n_i	*M*	*SD*	*1*	*2*	*3*		*Item level*	n_i	*M*	*SD*	*1*	*2*	*3*	
1	4	98.46	1.07	--				1	9	98.39	0.95	--			
2	21	96.12	3.99	2.35	--			2	13	95.58	3.73	2.81	--		
3	41	84.16	9.77	**14.30 (1.52)**	**11.96 (1.44)**	--		3	12	97.18	1.35	1.21	1.61	--	
Levene's *F*	df_1	df_2	*p*	Welch's *F*	df_1	df_2	*p*	Levene's *F*	df_1	df_2	*p*	Welch's *F*	df_1	df_2	*p*
11.43	2	63	.000	38.63	2	31.06	.000	5.01	2	31	.021	5.11	2	20.04	.016

Note. n = number of students per proficiency group. ni = number of items per level. Item level: 1 = below the competency standard for Grade 3; 2 = at the standard; 3 = above the standard. M = mean self-assessment score. SD = standard deviation. Tukey HSD post hoc test used for proficiency group 1 in reading comprehension. Games-Howell post hoc test used for all the other groups. Mean differences in bold are significant at $p < .05$. Effect sizes (Hedge's g) are in bold and in parentheses. Levene's F = F-ratio for equality of variance. Welch's F = robust F-ration for analysis of variance. ANOVA F = F-ratio for analysis of variance. df1 = degrees of freedom for the effect of the model. df2 = degrees of freedom for the residuals of the model. p = probability.

Table 6. Self-assessment by item difficulty level and proficiency group in mathematics and reading comprehension in Grade 4

Self-assessment mathematics								**Self-assessment reading comprehension**							
Proficiency group 1, n = 35				***Mean differences |Mi-Mj|***				***Proficiency group 1, n = 67***				***Mean differences |Mi-Mj|***			
Item level	n_i	***M***	***SD***	***1***	***2***	***3***		***Item level***	n_i	***M***	***SD***	***1***	***2***	***3***	
1	4	92.48	5.52	--				1	9	87.49	4.63	--			
2	21	86.65	8.47	5.82	--			2	13	83.65	4.83	3.84	--		
3	41	74.09	12.29	**18.38 (1.54)**	**12.56 (1.13)**	--		3	12	85.77	4.27	1.72	2.12	--	
Levene's *F*	df_1	df_2	*p*	ANOVA *F*	df_1	df_2	*p*	Levene's *F*	df_1	df_2	*p*	ANOVA *F*	df_1	df_2	*p*
2.35	2	63	.104	12.25	2	63	.000	.12	2	31	.890	1.92	2	31	.163
Proficiency group 2, n = 66				***Mean differences |Mi-Mj|***				***Proficiency group 2, n = 50***				***Mean differences |Mi-Mj|***			
Item level	n_i	***M***	***SD***	***1***	***2***	***3***		***Item level***	n_i	***M***	***SD***	***1***	***2***	***3***	
1	4	94.56	2.09	--				1	9	95.47	2.17	--			
2	21	91.55	5.98	3.02	--			2	13	91.75	4.41	3.73	--		
3	41	79.74	9.47	**14.82 (1.62)**	**11.81 (1.39)**	--		3	12	93.33	4.09	2.14	1.59	--	
Levene's *F*	df_1	df_2	*p*	Welch's *F*	df_1	df_2	*p*	Levene's *F*	df_1	df_2	*p*	ANOVA *F*	df_1	df_2	*p*
5.26	2	63	.008	33.18	2	19.92	.000	1.38	2	31	.267	2.52	2	31	.097
Proficiency group 3, n = 90				***Mean differences |Mi-Mj|***				***Proficiency group 3, n = 95***				***Mean differences |Mi-Mj|***			
Item level	n_i	***M***	***SD***	***1***	***2***	***3***		***Item level***	n_i	***M***	***SD***	***1***	***2***	***3***	
1	4	97.46	1.90	--				1	9	98.04	1.27	--			
2	21	95.87	3.13	1.59	--			2	13	94.61	3.85	**3.44 (1.11)**	--		
3	41	86.96	7.52	**10.50 (1.44)**	**8.91 (1.39)**	--		3	12	97.35	1.28	0.69	2.74	--	
Levene's *F*	df_1	df_2	*p*	Welch's *F*	df_1	df_2	*p*	Levene's *F*	df_1	df_2	*p*	Welch's *F*	df_1	df_2	*p*
8.41	2	63	.001	26.11	2	13.28	.000	6.12	2	31	.006	4.14	2	19.42	.026

Note. n = number of students per proficiency group. ni = number of items per level. Item level: 1 = below the competency standard for Grade 3; 2 = at the standard; 3 = above the standard. M = mean self-assessment score. SD = standard deviation. Tukey HSD post hoc test used for proficiency group 1 in mathematics. Games-Howell post hoc test used for all the other groups. Mean differences in bold are significant at $p < .05$. Effect sizes (Hedge's g) are in bold and in parentheses. Levene's F = F-ratio for equality of variance. Welch's F = robust F-ration for analysis of variance. ANOVA F = F-ratio for analysis of variance. df1 = degrees of freedom for the effect of the model. df2 = degrees of freedom for the residuals of the model. p = probability.

5 Discussion

In the present study, we wanted to investigate whether or not third- and fourth-graders (ages 8 to 9 years) are able to provide accurate self-assessments of their key academic competencies when equipped with an adequate self-assessment tool.

1. The first requirement for an accurate self-assessment was that students' self-assessments reflect their actual academic competencies (see Section 2.3). For both samples and in both domains, students' self-assessment scores had medium to large correlations [40] with their standardized test scores. Moreover, students in lower proficiency groups provided lower self-assessments on average in comparison with students in higher proficiency groups. In other words, students' self-assessments reflected their actual academic competencies.
2. The second requirement for an accurate self-assessment was that students recognize the self-assessment items' inherent difficulty (see Section 2.3). In general, when the (theoretical) item difficulty increased, students' confidence in solving the item decreased. Overall, these findings were less consistent for reading comprehension than for mathematics.
3. The third requirement for an accurate self-assessment was that even lower performers recognize the inherent difficulty of the self-assessment items (see Section 2.3). In general, independent of their affiliation with a performance group, students' confidence in solving the item decreased as item difficulty increased. This finding means that even lower performing students, who have a tendency to provide less accurate self-assessments [1], were able to compare the different items presented in the tool and to recognize the items' inherent difficulty. Overall, and the same as for Point 2 above, these findings were less consistent for reading comprehension than for mathematics.

We conclude from these results that third- and fourth-graders (ages 8 to 9 years) have the ability to provide accurate self-assessments on key academic competencies when provided with an adequate self-assessment tool. Our results were more consistent in the domain of mathematics than in reading comprehension.

5.1 Self-assessment accuracy

Self-assessment accuracy and student age. We deduced from other studies that the accuracy of pre- and elementary school students' self-assessments is strongly influenced by the conditions under which the self-assessment is practiced [31] [32] [33] [34] and the appropriateness of the self-assessment tool [35] [36] [37] rather than students' age per se. Our study offers support for this argument with our main finding that third- and fourth-graders have the ability to provide accurate self-assessments on key academic competencies, particularly in mathematics, when provided with an adequate tool. In other words: under favorable conditions, even young students have the ability to provide accurate self-assessments. In mathematics, our language-reduced and illustration-rich self-assessment tool probably enabled the 8- to 9-year-old students to better

understand the items. In both domains, the language-reduced rating scales allowed students to communicate their auto-perceptions with greater independence from language and reading skills. The items, which were given on a task level instead of a domain level, most likely led to good representations of the skills in question and distracted students from making comparisons with their peers.

Contrary to the general tendency in which students were able to recognize the self-assessment items' inherent difficulty, there was an exception for reading comprehension in both of the samples: Students' self-assessment scores decreased from level 1 to level 2 but increased again for the level 3 items. Most items on level 3 theoretically measure *text interpretation competency*. According to the national school curriculum, *text interpretation competency* is on a higher level of difficulty than the competency *localization and understanding of information in the text*, mostly represented by level 1 and level 2 items in the self-assessment tool. We found the same result pattern in the two samples, showing that students consistently had trouble recognizing an increase in the theoretical difficulty from the level 2 to the level 3 items. This increase in the mean scores from an easier to a more difficult level was not statistically significant (see Tables 4, 5 and 6). Despite this increase, the level 3 item mean scores remained consistently lower than the level 1 item mean scores, but the differences between the two were not statistically significant. Thus, we argue that the differences between level 3 and level 2 as well as between the level 3 and level 1 items are very small and not easily discerned by the students. There might also be a discrepancy between the theoretical and actual difficulties of the level 3 items. Overall, differences between the mean self-assessment scores were less often statistically significant in reading comprehension than in mathematics. This finding shows that the trend in reading comprehension went in the expected direction, but it was not as clear-cut as for mathematics. In item development, it is more difficult to calibrate item difficulty in reading comprehension than in mathematics. For the student, accurate self-assessment is probably more difficult in reading comprehension than in mathematics because assessment criteria and teacher feedback are less clear in reading comprehension compared with mathematics [75].

In self-assessment research with adults, a commonly approved standard for self-assessment accuracy is the strength of the correlation between the self-assessment and some performance measure [24]. Because there is no commonly approved standard for self-assessment accuracy in elementary school, we applied the same standard in our study as for self-assessments with adults. When comparing our results to those of other comparable studies [76] [77] [49] [78] regarding students' age, self-assessments of academic competencies, correlations between self-assessment and a performance measure, group testing on class basis, and sample size, the magnitudes of the correlations in our study tended to be stronger. Even compared with the range of meta-analytic effect sizes listed in Zell and Krizan's (2014) [24] meta-synthesis on the self-assessment of academic ability with adults and adolescents (with mean correlations ranging from r = .21 to .39, with one outstanding value of .63), the magnitudes of the correlations in our study, with 8- and 9-year-olds, tended to be stronger. This finding applies to mathematics as well as reading comprehension. Most likely, our results are due to the language-reduced self-assessment tool, which displayed items on a task level in a non-competitive setting. In addition, the tool fulfills symmetry principles that were derived from

Brunswik's lens model (see the *Self-assessment accuracy and student age* section): Students' self-assessment and their objective performance were measured on similar levels of abstraction.

Self-assessment accuracy and lower performers. Previous studies found that lower performing students tend to be less accurate in their self-assessments than higher performing students [1] [51] (see the *Self-assessment accuracy of lower performers* section). Possible explanations for these findings are a lack of adequate representations of both the expectations and the assessment criteria [52] [51] [53] or the temptation to respond in a socially desirable manner when reporting school grades or comparing their achievements with those of other classmates [54]. In our study, in both domains, self-assessment scores increased as proficiency increased with statistically significant differences between the groups for the majority of comparisons (see Table 3). In mathematics, the less proficient group 1 was as good at recognizing the inherent difficulty of the self-assessment items as the more proficient groups 2 and 3 were (see Tables 5 and 6). In reading comprehension, these findings were less consistent, but there was not a definitive finding that the more proficient groups had better recognition than the less proficient group did (see the *Self-assessment accuracy and student age* section in the Discussion for a possible explanation). Lower performing students, specifically third- and fourth-graders, may lack sufficient competencies in reading comprehension (see the section called *An innovative tablet-computer-based self-assessment tool*). With conventional questionnaires and inventories that depend on written language, these students would have (more) trouble understanding the descriptions of the tasks on which they had to assess themselves. In this sense, a language-reduced self-assessment tool, which displays items on the task level and avoids social comparison, might be particularly beneficial for lower performing students.

5.2 Strengths and limitations

The main objective of our study was to investigate whether or not third- and fourth-graders (ages 8 to 9 years) have the ability to provide accurate self-assessments of key academic competencies when provided with an adequate self-assessment tool. The results discussed in the Discussion section (Points 1, 2, and 3) show that students were able to do so although more consistently in mathematics than in reading comprehension. This outcome can be explained by certain aspects of our tablet-computer-based self-assessment tool; it integrates all the features identified in previous research findings needed to allow elementary school students to provide more accurate self-assessments of their academic competencies. These features are: self-assessment in a non-competitive setting, task-oriented self-assessment, and self-assessment that requires only limited reading and no verbalization (see Section 2.2). The innovation consists of an illustration-rich self-assessment tool with a language-reduced rating scale, thus reducing the bias that may come along with poor reading and language skills. In addition, the tablet-computer's touchscreen offered the students a more intuitive handling of the tool—particularly the slider on the rating scale—compared with a computer keyboard

or mouse. The tool administration and data collection across 43 elementary school classes was successful. Because of these features, too, the tool might once again be particularly beneficial for lower performers.

Our findings are based on two independent and representative samples, randomly chosen out of all the possible elementary school classes in Luxembourg. In general, we found similar patterns in the self-assessments and in the self-assessment accuracy between the two samples, thus confirming the consistency of the measurements with the self-assessment tool (e.g., [79]).

Due to persistent doubts regarding young students' self-assessment abilities, self-assessment research on pre- and elementary school students is still scarce. For this reason, we reviewed findings from different research disciplines and areas (developmental, cognitive, educational, and social psychology; educational science; metacognition research; self-regulation of learning; self-concept and self-efficacy research) to conclude that accurate self-assessment is less a question of age by itself than a question of the conditions under which self-assessment is conducted. This argument allowed us to test the hypothesis that even third- and fourth-graders are able to provide accurate self-assessments of key academic competencies when equipped with an adequate self-assessment tool. We consider this approach to be a strong point of our study, although it implies an oversimplification of the constructs, concepts, and findings of the cited studies. Consequently, we are limited in discussing our results in comparison with the concrete findings from previous studies.

A limitation of our study is the lack of a control group that would have allowed us to compare the accuracy of self-assessment when administered via our innovative tool with self-assessment accuracy in a conventional (e.g., paper pencil) and predominantly language-based setting.

Because the tool was designed for classroom use, testing time was limited (40 plus 20 minutes for the self-assessments in math and German, respectively; 10 minutes for the introduction), and we did not ask students to actually solve the items on the self-assessment tool. Comparing students' actual solutions with their self-assessment of whether they could solve the very same items would have provided an additional analysis for assessing accuracy (e.g., [51] [80]). This might be covered in a future study. Nevertheless, we would like to highlight that the self-assessment tool and its very accurate measure of validation (i.e., standardized tests from Luxembourg's school monitoring program) are based on the same reference standard: the national school curriculum. Items from the standardized tests and the self-assessment tool were problem isomorphs in the majority of cases. The self-assessment tool covers curriculum-relevant competencies, thus allowing high-quality feedback to be provided to teachers and good chances for the tool to be integrated into teaching.

When comparing the self-assessment results between mathematics and reading comprehension, we conclude that the self-assessment tool worked better for the former than the latter. A possible explanation might be that language reduction through illustration—an important feature of the tool—(obviously) did not apply to reading comprehension, except for the rating scale. Another explanation might be that the two factors interacted: In item development, the calibration of item difficulty is more problematic

in reading comprehension than it is in mathematics, but teachers' feedback and assessment criteria are less clear to students in reading comprehension than in mathematics [75], which leads to less accurate self-assessments. The empirical validation of the difficulty levels of the self-assessment items would provide further answers to this question.

6 Conclusions

We conclude from these results that, under favorable conditions, third- and fourth-graders (ages 8 to 9 years) have the ability to provide accurate self-assessments of key academic competencies, but they can do so more consistently in mathematics than in reading comprehension. The favorable conditions are (1) self-assessment in a non-competitive setting, (2) self-assessment items presented on a task level instead of questions about general competency, and (3) the use of a language-reduced and illustration-rich self-assessment tool. The use of tablet technology, specifically a tablet-computer-based self-assessment tool that we developed, was found to be a suitable instrument for providing such conditions, particularly in mathematics.

7 Acknowledgments

The authors thank the *Fonds National de la Recherche (FNR)* for funding the present study. [AFR reference: 6924301].

8 References

[1] G. T. L. Brown and L. R. Harris, "Student self-assessment," in SAGE handbook of research on classroom assessment, J. H. McMillan, Ed. SAGE, 2013, pp. 367–393. https://doi.org/10.4135/9781452218649.n21

[2] D. Black, P., & Wiliam, "Inside the black box : Raising standards through classroom assessment," Phi Delta Kappa, vol. 80, no. 2, pp. 139–148, 1998.

[3] J. Hattie, Visible learning. A synthesis of over 800 meta-analyses relating to achievement. London and New York: Routledge, 2009.

[4] J. A. Ross, "The reliability, validity, and utility of self-assessment," Pract. Assessment, Res. Eval., vol. 11, no. 10, pp. 1–13, 2006.

[5] G. T. L. Brown and L. R. Harris, "The future of self-assessment in classroom practice: Reframing self- assessment as a core competency," Front. Learn. Res., vol. 3, pp. 22–30, 2014.https://doi.org/10.14786/flr.v2i1.24

[6] H. L. Andrade, Y. Du, and X. Wang, "Putting rubrics to the test: The effect of a model, criteria generation, and rubric-referenced self-assessment on elementary school students' writing," Educ. Meas. Issues Pract., vol. 27, no. 2, pp. 3–13, 2008.https://doi.org/10.1111/j.1745-3992.2008.00118.x

[7] D. Fontana and M. Fernandes, "Improvements in mathematics performance as a consequence of self-assessment in Portuguese primary school pupils," Br. Psychol. Soc., vol. 64, pp. 407–417, 1994.https://doi.org/10.1111/j.2044-8279.1994.tb01112.x

[8] J. A. Ross, A. Hogaboam-Gray, and C. Rolheiser, "Student self-evaluation in grade 5-6 mathematics effects on problem-solving achievement," Educ. Assess., vol. 8, no. 1, pp. 43–59, 2002.https://doi.org/10.1207/S15326977EA0801_03

[9] J. P. Gaa, "Effects of individual goal-setting conferences on achievement, attitudes, and goal-setting behavior," J. Exp. Educ., vol. 42, no. 1, pp. 22–28, 1973.https://doi.org/10.1080/00220973.1973.11011437

[10] A. Bandura, "Social-cognitive theory of self-regulation," Organ. Behav. Hum. Decis. Process., vol. 50, pp. 248–287, 1991.https://doi.org/10.1016/0749-5978(91)90022-L

[11] D. H. Schunk, "Goal and self-evaluative influences during children's cognitive skill learning," Am. Educ. Res. J., vol. 33, no. 2, pp. 359–382, 1996. https://doi.org/10.3102/00028312033002359

[12] B. J. Zimmerman, "A social cognitive view of self-regulated academic learning.," J. Educ. Psychol., vol. 81, no. 3, pp. 329–339, 1989.https://doi.org/10.1037/0022-0663.81.3.329

[13] A. Bandura, Social Learning Theory. NJ: Prentice-Hall, Englewood Cliffs, 1977.

[14] M. Komarraju and D. Nadler, "Self-efficacy and academic achievement: Why do implicit beliefs, goals, and effort regulation matter?," Learn. Individ. Differ., vol. 25, pp. 67–72, 2013.https://doi.org/10.1016/j.lindif.2013.01.005

[15] D. H. Schunk and P. A. Ertmer, "Self-regulations and academic learning. Self-efficacy enhancing interventions," in Handbook of self-regulation, B. M., P. R. Pintrich, and M. Zeidner, Eds. 2000, pp. 631–649.

[16] D. L. Dinsmore and H. E. Wilson, "Student participation in assessment: Does it influence self-regulation?," in Handbook of Human and Social Conditions in Assessment, G. T. L. Brown and L. R. Harris, Eds. New York and London: Routledge, 2016, pp. 145–169.

[17] G. Munns and H. Woodward, "Student engagement and student self-assessment: The REAL framework," Assess. Educ. Princ. Policy, Pract., vol. 13, no. 2, pp. 193–213, 2006.

[18] H. L. Andrade, X. Wang, Y. Du, and R. L. Akawi, "Rubric-referenced self-assessment and self-efficacy for writing," J. Educ. Res., vol. 102, no. 4, pp. 287–302, May 2009.https://doi.org/10.3200/JOER.102.4.287-302

[19] C. Glaser, C. Keßler, D. Palm, and J. C. Brunstein, "Improving fourth graders self-regulated writing skills: Specialized and shared effects of process-oriented and outcome-related self-regulation procedures on students' task performance, strategy use, and self-evaluation," Zeitschrift fur Padagogische Psychol., vol. 24, no. 3–4, pp. 177–190, 2010.https://doi.org/10.1024/1010-0652/a000015

[20] Z. Olina and H. J. Sullivan, "Effects of teacher and self-assessment on student performance. Paper presented at the Annual Convention of the American Educational Research Association," 2002.

[21] B. Hughes, H. J. Sullivan, and M. L. Mosley, "External evaluation, task difficulty, and continuing motivation," J. Educ. Res., vol. 78, no. 4, pp. 210–215, 1985.https://doi.org/10.1080/00220671.1985.10885602

[22] M. Griffiths and C. Davies, "Learning to Learn: action research from an equal opportunities perspective in a junior school," Br. Educ. Res. J., vol. 19, no. 1, pp. 43–58, 1993.https://doi.org/10.1080/0141192930190104

[23] M. Ward, L. Gruppen, and G. Regehr, "Measuring self-assessment : Current state of the art," Adv. Heal. Sci. Educ., vol. 7, pp. 63–80, 2002.https://doi.org/10.1023/A:1014585522084

[24] E. Zell and Z. Krizan, "Do People Have Insight Into Their Abilities? A Metasynthesis," Perspect. Psychol. Sci., vol. 9, no. 2, pp. 111–125, 2014.https://doi.org/10.1177/1745691613518075

[25] G. T. L. Brown, H. L. Andrade, and F. Chen, "Accuracy in Student Self-Assessment: Directions and Cautions for Research," Assess. Educ. Princ. Policy Pract., vol. 22, no. 4, pp. 444–457, 2015.https://doi.org/10.1080/0969594X.2014.996523
[26] E. Panadero, G. T. L. Brown, and J. W. Strijbos, "The Future of Student Self-Assessment: a Review of Known Unknowns and Potential Directions," Educ. Psychol. Rev., vol. 28, no. 4, pp. 803–830, 2016.https://doi.org/10.1007/s10648-015-9350-2
[27] J. G. Nicholls, "The development of the concepts of effort and ability, perception of academic attainment, and the understanding that difficult tasks require more ability," Child Dev., vol. 49, no. 3, pp. 800–814, 1978.https://doi.org/10.2307/1128250
[28] J. G. Nicholls, "Achievement motivation: Conceptions of ability, subjective experience, task choice, and performance," Psychol. Rev., vol. 91, no. 3, pp. 328–346, 1984.https://doi.org/10.1037/0033-295X.91.3.328
[29] J. G. Nicholls and A. T. Miller, "The differentiation of the concepts of difficulty and ability," Child Dev., vol. 54, no. 4, pp. 951–959, 1983.https://doi.org/10.2307/1129899
[30] Y. Eshel and Z. Klein, "Development of academic self-concept of lower-class and middle-class primary school children," J. Educ. Psychol., vol. 73, no. 2, pp. 287–293, 1981.https://doi.org/10.1037/0022-0663.73.2.287
[31] R. Butler, "The effects of mastery and competitive conditions on self-assessment at different ages," Child Dev., vol. 61, no. 1, pp. 201–210, Feb. 1990.https://doi.org/10.2307/1131059
[32] D. Stipek and D. Mac Iver, "Review Developmental Change in Children ' s Assessment of Intellectual Competence," Child Dev., vol. 60, pp. 521–538, 1989.https://doi.org/10.2307/1130719
[33] J. Hillyer and T. C. Lye, "Portfolios and second graders' self-assessment of their development as writers," Read. Improv., vol. 33, no. 3, pp. 148–159, 1996.
[34] B. J. Liebovich, "Children's self-assessment," in Issues in early childhood education: Curriculum, teacher education, & dissemination of information, 2000.
[35] V. A. Vo, R. Li, N. Kornell, A. Pouget, and J. F. Cantlon, "Young children bet on their numerical skills: Metacognition in the numerical domain.," Psychol. Sci., vol. 25, no. 9, pp. 1712–1721, 2014.https://doi.org/10.1177/0956797614538458
[36] S. Harter, "The perceived competence scale for children," Child Dev., vol. 53, no. 1, p. 87, Feb. 1982.https://doi.org/10.2307/1129640
[37] S. Harter and R. Pike, "The pictorial scale of perceived competence and social acceptance for young children," Child Dev., vol. 55, no. 6, pp. 1969–1982, Dec. 1984.https://doi.org/10.2307/1129772
[38] M. V. J. Veenman, B. H. A. M. Van Hout-Wolters, and P. Afflerbach, "Metacognition and learning: Conceptual and methodological considerations," Metacognition Learn., vol. 1, no. 1, pp. 3–14, 2006.https://doi.org/10.1007/s11409-006-6893-0
[39] W. Schneider, "The development of metacognitive knowledge in children and adolescents: Major trends and implications for education," Mind, Brain, Educ., vol. 2, no. 3, pp. 114–121, 2008.https://doi.org/10.1111/j.1751-228X.2008.00041.x
[40] J. Cohen, Statistical power analysis for the behavioral sciences, 2nd ed. New York: Academic Press, 1988.
[41] R. Butler, "Are positive illusions about academic competence always adaptive, under all circumstances: New results and future directions," Int. J. Educ. Res., vol. 50, no. 4, pp. 251–256, 2011.https://doi.org/10.1016/j.ijer.2011.08.006
[42] M. Bong and E. M. Skaalvik, "Academic self-concept and self-efficacy: How different are they really ?," Educ. Psychol. Rev., vol. 15, no. 1, 2003.https://doi.org/10.1023/A:1021302408382

[43] F. Pajares, M. D. Miller, and M. J. Johnson, "Gender differences in writing self-beliefs of elementary school students," J. Educ. Psychol., vol. 91, no. 1, pp. 50–61, 1999.https://doi.org/10.1037/0022-0663.91.1.50
[44] H. W. Marsh, Self Description Questionnaire - II: SDQ II. Macarthur, Australia: University of Western Sydney, Self-concept Enhancement and Learning Facilitation Research Centre, 1999.
[45] M. Bong, "Predictive Utility of Subject-, Task-, and problem-specific Self-Efficacy Judgments for Immediate and Delayed Academic Performances," J. Exp. Educ., vol. 70, no. 2, pp. 133–162, 2002.https://doi.org/10.1080/00220970209599503
[46] G. Nagy, H. M. G. Watt, and J. S. Eccles, "The development of students' mathematics self-concept in relation to gender: Different countries, different trajectories?," J. Res. Adolesc., vol. 20, no. 2, pp. 482–506, 2010.https://doi.org/10.1111/j.1532-7795.2010.00644.x
[47] W. W. Wittmann and G. E. Matt, "Aggregation und Symmetrie. Grundlagen einer multivariaten Reliabilitäts- und Validitätstheorie, dargestellt am Beispiel der differentiellen Validität des Berliner Intelligenzstrukturmodells," Diagnostica, vol. 32, no. 4, pp. 309–329, 1986.
[48] D. Wagener and W. W. Wittmann, "Personalarbeit mit dem komplexen Szenario FSYS: Validität und Potential von Verhaltensskalen," Zeitschrift für Pers., vol. 1, no. 2, pp. 80–93, 2002.https://doi.org/10.1026//1617-6391.1.2.80
[49] Y. G. Butler and J. Lee, "On-task versus off-task self-assessments among Korean Elementary School students studying English," Mod. Lang. J., vol. 90, no. 4, pp. 506–518, 2006.https://doi.org/10.1111/j.1540-4781.2006.00463.x
[50] C.-S. Chang and E. Z.-F. Liu, "Exploring the media literacy of Taiwanese elementary school students," Asia-Pacific Educ. Res., vol. 20, no. 3, pp. 604–611, 2011.
[51] M. Claes and R. Salame, "Motivation toward accomplishment and the self-evaluation of performances in relation to achievement," Can. J. Behav. Sci. Can. des Sci. du Comport., vol. 7, no. 4, pp. 398–410, 1975.
[52] J. A. Ross, C. Rolheiser, and A. Hogaboam-Gray, "Effects of self-evaluation training on narrative writing," Assess. Writ., vol. 6, no. 1, pp. 107–132, 1999.https://doi.org/10.1016/S1075-2935(99)00003-3
[53] J. R. Ng and J. K. Earl, "Accuracy in self-assessment: the role of ability, feedback, self-efficacy and goal orientation," Aust. J. Career Dev., vol. 17, no. 3, pp. 39–50, 2008.https://doi.org/10.1177/103841620801700307
[54] N. R. Kuncel, M. Crede, and L. L. Thomas, "The validity of self-reported grade point averages, class ranks, and test scores : A meta-analysis and review of the literature," Rev. Educ. Res., vol. 75, no. 1, pp. 63–82, 2005.https://doi.org/10.3102/00346543075001063
[55] P. M. Sadler and E. Good, "The impact of self- and peer- grading on student learning," Educ. Assess., vol. 11, no. 1, pp. 37–41, 2006.https://doi.org/10.1207/s15326977ea1101_1
[56] D. T. Gordon, The digital classroom. How technology is changing the way we teach and learn. Cambridge, MA: Harvard Education Letter, 2003.
[57] K. Goodwin, "Use of Tablet Technology in the Classroom," Educ. Communities, pp. 1–96, 2012.
[58] O. T. Murray and N. R. Olcese, "Teaching and Learning with iPads, Ready or Not?," TechTrends Link. Res. Pract. to Improv. Learn., vol. 55, no. 6, pp. 42–48, 2011.
[59] R. Martin, S. Ugen, A. Fischbach, and (Eds.), Épreuves Standardisées: Bildungsmonitoring Luxemburg. Nationaler Bericht 2011 bis 2013 [Épreuves Standardisées: School Monitoring for Luxembourg. National report 2011 to 2013]. 2015.
[60] SCRIPT & LUCET, "PISA 2015. Nationaler Bericht Luxemburg [PISA 2015. National report Luxembourg]," Luxembourg, 2016.

[61] P. Engel de Abreu, C. Hornung, and R. Martin, "Wie lernen Kinder Sprachen? Überlegungen zu Spracherwerb und Alphabetisierung in Luxemburg aus der Sicht der Kognitionswissenschaften [How do children learn language? Reflections on language acquisition and literacy from a cognitive science perspective].," in Education report Luxembourg 2015-2: Analyses and findings., T. Lenz and J. Bertemes, Eds. 2015.
[62] R. E. Mayer, "The promise of multimedia learning: using the same instructional design methods across different media," Learn. Instr., vol. 13, no. 2, pp. 125–139, Apr. 2003.https://doi.org/10.1016/S0959-4752(02)00016-6
[63] R. E. Mayer and R. Moreno, "Aids to computer-based multimedia learning," Learn. Instr., vol. 12, no. 1, pp. 107–119, Feb. 2002.https://doi.org/10.1016/S0959-4752(01)00018-4
[64] R. E. Mayer, "Research-based principles for designing multimedia instruction," in Applying Science of Learning in Education: Infusing Psychological Science into the Curriculum, V. A. Benassi, C. E. Overson, and C. M. Hakala, Eds. 2014, pp. 59–71.
[65] R. E. Mayer, Multimedia learning. New York: Cambridge University Press, 2009.https://doi.org/10.1017/CBO9780511811678
[66] K. Hirsh-Pasek, J. M. Zosh, R. M. Golinkoff, J. H. Gray, M. B. Robb, and J. Kaufman, Putting Education in "Educational" Apps: Lessons From the Science of Learning, vol. 16, no. 1. 2015.
[67] H. van Laerhoven, H. J. van der Zaag-Loonen, and B. H. F. Derkx, "A comparison of Likert scale and visual analogue scales as response options in children's questionnaires," Acta Paediatr., vol. 93, no. 6, pp. 830–835, 2004.https://doi.org/10.1111/j.1651-2227.2004.tb03026.x
[68] D. Hasson and B. B. Arnetz, "Validation and findings comparing VAS vs. Likert scales for psychosocial measurements," Int. Electron. J. Health Educ., vol. 8, pp. 178–192, 2005.
[69] T. A. Warm, "Weighted Likelihood Estimation of Ability in Item Response Theory," Psychometrika, vol. 54, pp. 427–450, 1989.https://doi.org/10.1007/BF02294627
[70] A. Fischbach, S. Ugen, and R. Martin, Eds., ÉpStan Technical Report. Luxembourg: University of Luxembourg, 2014.
[71] A. Field, Discovering Statistics Using IBM SPSS, 4th editio. London: Sage, 2013.
[72] D. C. Howell, Statistical Methods for Psychology, 8th editio. USA: Wadsworth, 2013.
[73] P. H. Ramsey, K. Barrera, P. Hachimine-Semprebom, and C.-C. Liu, "Pairwise comparisons of means under realistic nonnormality, unequal variances, outliers and equal simple sizes," J. Stat. Comput. Simul., pp. 125–135, 2009.
[74] J. Algina, T. C. Oshima, and W.-Y. Lin, "Type I error rates for Welch's test and James's second-order test under nonnormality and inequality of variance when there are two groups," J. Educ. Behav. Stat., vol. 19, no. 3, pp. 275–291, 1994.https://doi.org/10.2307/1165297
[75] D. Galloway, E. L. Leo, C. Rogers, and D. Armstrong, "Maladaptive motivational style: the role of domain specific task demand in English and mathematics," Br. J. Educ. Psychol., vol. 66, no. 2, pp. 197–207, 1996.https://doi.org/10.1111/j.2044-8279.1996.tb01189.x
[76] J. P. Connell and B. C. Ilardi, "Self-system concomitants of discrepancies between children's and teachers' evaluations of academic competence," Child Dev., vol. 58, no. 5, pp. 1297–1307, 1987.https://doi.org/10.2307/1130622
[77] D. W. Kwok, D. C. & Lai, "The self-perception of competence by Canadian and Chinese children. Paper presented at the annual convention of the Canadian Psychological Association," 1993.
[78] B. K. Bradshaw, "Do students effectively monitor their comprehension?," Read. Horizons, vol. 41, no. 3, pp. 143–154, 2001.
[79] A. E. R. Association, A. P. Association, and N. C. on M. in Education, Eds., Standards for educational and psychological testing. American Educational Research Association, 2014.

[80] M. H. Van Loon, A. B. H. de Bruin, T. van Gog, and J. J. G. van Merriënboer, "Activation of inaccurate prior knowledge affects primary-school students' metacognitive judgments and calibration," Learn. Instr., vol. 24, no. 1, pp. 15–25, 2013.https://doi.org/10.1016/j.learninstruc.2012.08.005

9 Authors

Denise Villányi has a master's degree in educational science from the Johannes Gutenberg-University of Mainz (Germany). She is currently finishing her doctoral studies in educational psychology at the University of Luxembourg.

Romain Martin is Professor of psychology and empirical educational research at the University of Luxembourg. He specializes in cognitive abilities, educational measurement, and computer-assisted testing. He is currently the vice-rector of academic affairs at the University of Luxembourg.

Philipp Sonnleitner is Dr. in educational psychology from the Free University of Berlin (Germany). His academic research focuses on psychological and educational assessment. He is currently applying his expertise to the Luxembourg school monitoring program.

Christina Siry holds a post of Professor in Learning and Instruction at the University of Luxembourg. She has several active lines of research that focus on the intertwined areas of science learning and learning to teach science, particularly at the primary levels. She focuses on the use of collaborative pedagogies and participatory methodologies as tools for transforming science teacher education and science education.

Antoine Fischbach is Dr. in educational psychology from the University of Trier (Germany). He specializes in large-scale educational assessment, and he is an expert on the Luxembourg school system. He is currently the acting director of the Luxembourg Centre for Educational Testing (LUCET).

Article submitted 20 May 2018. Final acceptance 12 June 2018. Final version published as submitted by the authors.