Self-effective scientific reasoning? Differences between elementary and secondary school students

Kristin Nyberg[1], Susanne Koerber[1], Christopher Osterhaus[2]

[1] University of Education Freiburg, Germany

[2] University of Vechta, Germany

Abstract

Although scientific reasoning is not a formal, independent school subject, it is an increasingly important skill, especially for student learning in science, technology, engineering, and mathematics (STEM) subjects. To promote scientific reasoning effectively, it is important to know its influencing factors. While cognitive influences have been investigated, affective-motivational factors, particularly self-efficacy, have rarely been considered in studies on scientific reasoning. To examine, for the first time, whether self-efficacy can be measured in a task-specific way and whether self-efficacy correlates with students' scientific reasoning performance, the study assessed performance in scientific reasoning and self-efficacy (academic and task-specific) in a sample of 140 fourth graders and 148 eighth graders. As expected, higher correlations emerged for task-specific self-efficacy in both grades. A hierarchical cluster analysis showed that the correlational patterns were not the same across grade levels, with differences in self-estimated performance prevailing between the two grade levels: The largest cluster in Grade 4 (41%) comprised children who significantly overestimated their performance, whereas the largest cluster in Grade 8 (39%) comprised students who gave a realistic estimate of their own performance in scientific reasoning. This cluster was not present in Grade 4. Additional clusters of students who overestimated or underestimated their performance emerged in both grades. The results support the conclusion that self-efficacy expectations are important to consider when fostering scientific reasoning, and the large number of elementary school students who overestimated their performance suggests that not all students might benefit from interventions targeted at increasing self-efficacy.

Corresponding author: Kristin Nyberg, University of Education, Institute of Psychology, Kunzenweg 21, 79117 Freiburg, Germany, kristin.nyberg@ph-freiburg *DOI:* https://doi.org/10.14786/flr.v10i1.955

1.   Introduction

Fostering science, technology, engineering and mathematics (STEM) education in school and integrating technology and engineering into science education is a major challenge in recent years (Bybee, 2010). Nevertheless, STEM education should not be reduced to content knowledge but instead should incorporate the development of scientific attitudes and critical, scientific reasoning or thinking (Osborne, 2013). This broader conceptualization of science skills is mirrored in the conceptualization of scientific literacy in the PISA studies (OECD, 2006, 2015), which encompasses, apart from science content knowledge, scientific reasoning and personal affective-motivational factors. Scientific reasoning can be viewed as a complex construct which includes several components, such as experimentation skills or understanding the nature of science. Previous studies showed evidence for a common conceptual core in scientific reasoning (Koerber et al., 2015) and validated group tests exist to reliably measure scientific reasoning in elementary school and above (e.g., Koerber et al., 2015; Osterhaus et al., 2015).

For fostering scientific reasoning skills, it is important to know factors influencing scientific reasoning. Most research in the area addresses the impact of general cognitive factors like language, intelligence, problem solving, executive functioning, and specific variables (e.g., advanced theory of mind) on the development of scientific reasoning (Koerber et al., 2015; for an overview see Zimmerman, 2007). Research on the influence of affective-motivational factors in scientific reasoning, has been scarce, despite that affective-motivational constructs like self-beliefs and motivation are essential for academic achievement (Cai et al., 2018; Pajares, 1996; Pajares & Valiante, 1997). Among affective-motivational variables, the impact of self-efficacy expectations on academic performance is particularly strong, predicting up to 9% of academic performance in university students (Richardson et al., 2012).

Self-efficacy expectations are classified as competence beliefs based on Bandura's social cognition theory (Bandura, 1986; 1997). Self-efficacy expectations are understood as the belief in one's own ability to cope with a future situation or task. According to Bandura (1997), the precise measurement of self-efficacy should include items that measure beliefs about expectations and performance as close as possible to the identical future tasks or situations, and the more specifically they are assessed, the more accurate the obtained action predictions will be (Bandura, 1997; Schwarzer & Jerusalem, 2002). Students' expectations of self-efficacy are considered to play an important role in the school context. Self-efficacy expectations positively influence academic performance, motivational processes, self-regulation, self-perception, and interest (Bandura & Schunk, 1981; Klassen & Usher, 2010; Pajares & Valiante, 1997; Schunk, 1995). Students with positive self-efficacy expectations spend more time on challenging tasks and situations. Hence, they have the chance to receive more feedback compared to their classmates with lower self-efficacy expectations, who consequently avoid tasks and situations in this domain (Britner & Pajares, 2001; Zeldin & Pajares, 2000). When asking students in Grades 6, 7 and 8 how skilled they would be at solving a math problem, they mainly use their experience with similar math problems to this point, as a basis for their judgment (Usher & Pajares, 2009). Therefore, it is expected that the relation between performance and self-efficacy may be stronger at higher grade levels based on experience.

Researchers agree on the positive correlation between self-efficacy expectations and academic success, mostly referring to typical subjects of the curriculum such as mathematics (Siefer et al., 2020) or writing (Pajares & Valiante, 1997). A meta-analysis by Multon and Brown (1991) with a total sample of roughly 5000 participants reported a correlation of $r = .38$ between academic performance and self-efficacy expectations. This result is backed up years later by a meta-analysis of Honicke and Broadbent (2016). They found a moderate positive correlation between the study performance of university students and their self-efficacy of $r = .33$ across 59 studies. According to Bong (2006), the correlation between self-efficacy expectations and academic performance appears to be higher when measuring self-efficacy more specifically, which is consistent with Bandura's (1997) theory that self-efficacy expectations are context specific. In the following study, the assessment of self-efficacy is based on Bandura's theory of measuring self-efficacy as specifically as possible, i.e. task-specific.

Given this context-dependent requirement of a self-efficacy measure, generalizing results to other disciplines seems difficult, which consequently requires investigation in other domains, subjects, and skills. Research dealing with self-efficacy and skills that are not included as an own subject in the curriculum but have an important relation to other subjects and skills like scientific reasoning is of great interest. Liu et al. (2006), for example, found that science-class self-efficacy expectations of sixth-grade students (e.g., "I am confident I can learn the basic concepts taught in this science class") correlated significantly with understanding science concepts (r =.28). Jansen et al. (2015) investigated performance in scientific literacy and self-efficacy. The sample, which originated from the 2006 PISA survey in Germany, included roughly 5000 secondary school students, most of whom were Grade 9 at the time of the survey. The self-efficacy scale used in the study included eight items that assessed how confident they thought they were in solving a task in scientific literacy (e.g., predict how changes in an environment will affect the survival of certain species). The items that measured performance included real-world science tasks from the field of science, resembling the conceptualization of scientific literacy of the PISA study (Bybee et al., 2010; OECD, 2006). The results of the study suggest that self-efficacy is a significant predictor of scientific literacy (Jansen et al., 2015).

Even though a positive relation between performance and self-efficacy has been found in many domains, it does not necessarily imply that higher self-efficacy is always related to higher performance. Students may over- or underestimate their performance, which can hinder good performance (Bandura, 1986). Students who (strongly) overestimate their performance may not invest in learning the specific task or acquiring metacognitive skills (e.g., strategies to achieve goals), because they are confident to master the situation or task without investing more time and effort (Hadwin & Webster, 2013). Identifying these groups of students is particularly important given that not all students might equally benefit from a direct increase in self-efficacy.

The present study charters new territory: It investigates whether self-efficacy in scientific reasoning can be measured in a task-specific way. While this approach is similar to what has been shown in a specific area of mathematics (Siefer et al., 2020), it is the first time that self-efficacy expectations are assessed in scientific reasoning in two ways of specificity (academic and task-specific). In addition, the study investigates whether there are groups of students with different levels of task-specific self-efficacy and performance in scientific reasoning.

A sample of fourth and eighth graders was selected. Eight graders were included because they had more opportunities to practice scientific reasoning and gain feedback on it, despite being more limited than in mathematics or sports. The fourth graders were chosen because previous studies, examining self-efficacy expectations had predominantly recruited secondary school and university students. Moreover, studies show that perceived academic self-efficacy declines between sixth and eighth Grade (Harter, 1985). The decline can be described by a tendency of elementary school children to have unrealistically high beliefs about their competence (overestimate their competence) whereas as they get older, their self-judgement increasingly matches external evaluations (Nicholls, 1978; Stipek & Hofman, 1980; Pajares & Schunk, 2001). This study intended to shed light on the beginnings of this relation. The study focuses on five key questions: (1) Is our self-efficacy scale suitable to measure task-specific self-efficacy in scientific reasoning for Grades 4 and 8? (2) What is the relation between students' performance in scientific reasoning and their self-efficacy expectations, specifically is this relation already observable in elementary school? (3) Are the correlations higher when measuring task-specific self-efficacy instead of academic self-efficacy? (4) And, are there differences in the strength of the relation between both grade levels? (5) Finally, do diverse clusters exist which categorize students along the task-specific self-efficacy-performance relation (e.g., over- or underestimate and high and low performer)? This last research question pertains to the question whether the relations within the sample are homogeneous or if interindividual differences exist. Since on the one hand, both, self-efficacy and scientific reasoning, can be furthered through intervention or training (Sodian et al., 2002; Margolis & McCabe, 2006), and on the other hand, task-specific self-efficacy measurement provides an accurate insight into self-assessment and related task performance, this allows to specifically focusing on what

needs to be fostered: Self-efficacy or scientific reasoning. Getting better at scientific reasoning skills is especially crucial in relation to the performance in scientific literacy and the important STEM subjects.

2.   Method

2.1   Participants

The sample consisted of $N = 288$ students (155 girls, 133 boys), $n = 140$ fourth graders ($M = 9$ years, 2 months, $SD = 4$ months) and $n = 148$ eighth graders ($M = 13$ years, 3 months, $SD = 6$ months). The students went mostly to 14 middle-class schools close to a mid-sized city in southern Germany. The eighth graders all attended academic track schools (Gymnasium). The data were collected in the first half of the academic year between October 2018 and January 2019. Of the 288 children, 68 (23.6%) spoke at least one language other than German at home (the most frequently reported languages were Russian, English, and French). Student assent and written consent from caretakers were obtained for all participants. Institutional review board (IRB) approval was not sought for the study because the host institution did not have an established IRB.

2.2   Materials

2.2.1   Scientific reasoning

The performance in scientific reasoning was measured with six multiple-select tasks, three tasks testing students' understanding of the nature of science (NOS; Cronbach's α Grade 4 = .53, Grade 8 = .46) and three tasks testing their metaconceptual understanding of experimentation (UNEX Cronbach's α Grade 4 = .52, Grade 8 = .51). Since previous studies revealed that the subcomponents of scientific reasoning skills share a common conceptual core among the subcomponents (Koerber et al., 2015) and there is a correlation between NOS and experimentation skills (Osterhaus et al., 2017), the two scientific reasoning scales were collapsed and used as a single performance measure for scientific reasoning in the subsequent analysis (entire scale Cronbach's α Grade 4 = .49, Grade 8 = .48). Further support for the use of the scale is the result of the study by Osterhaus et al. (2020), who examined a short scale (SPR-I(7)) for validity and reliability and showed similar reliabilities to the six items we used.

NOS. Three NOS tasks (A01 Scientist, A03 Middle Ages and A11 Mistakes in Science) were selected from the Science-P Reasoning Inventory (Koerber et al., 2015; Osterhaus et al., 2020). These three NOS tasks assessed children's understanding of what scientists do and the understanding of the hypothesis-evidence relation (see appendix for a sample item). For each task, three answer options were presented to the students (on a naïve level = 0 points, an intermediate level = 1 point, a scientifically advanced level = 2 points), and they were asked to agree or disagree with each of the answer options. The lowest level answer selected was taken as the final score on the entire item. Thus, the children could obtain a maximum of two points per task and a maximum sum score of six (see Osterhaus et al., 2020 for further coding details).

UNEX. The UNEX tasks were taken from the study of Osterhaus et al. (2015). The students were given three imaginary described experiments (Trees U1, Math textbook U2, Classroom U3). The students were asked to decide whether it was a good or bad experiment to test the different assumptions (see appendix for a sample item). Children were assigned 2 points, when they selected the correct answer and simultaneously rejected the wrong alternatives. One point was given for selecting the correct answer and the intermediate level but rejecting the naïve level (see coding of the NOS tasks).

2.2.2 Self-efficacy expectations

The self-efficacy expectations were measured on two levels: Academic self-efficacy and task-specific self-efficacy expectations.

Academic self-efficacy. The academic self-efficacy expectations were measured with three items, used from Jerusalem and Satow (1999). The scale (WIRKSCHUL) originally comprises 7 items and was validated on 3000 secondary school students. The entire scale was used in our pilot study. From the subsequent interviews and the reliability tests, two items emerged as having poor reliability. In order to achieve a sufficient reliability in both Grade levels for the analysis of the main survey, two additional items had to be excluded. The children were asked to indicate their agreement on a 4-point Likert scale, ranging from "strongly disagree" (1) to "strongly agree" (4) with the following 1) "I can solve even the difficult tasks in class if I exert myself." 2) "Even if I am sick for a longer time, I am still able to perform well" 3) "If a teacher doubts my skills, I am sure that I can still perform well". The scale was applied in German and the description of the items in the text are own translations. Cronbach's α was .61 for Grade 4 and for Grade 8 =.55. The reported reliabilities of > 5 can be interpreted for scales especially with few items (e.g., Nunnally & Bernstein, 1978). The corrected item-total correlation (r_{it}) of .56-58 also point to the reliability of the scale. Criterion validity is provided by the significant relations to task-specific self-efficacy (Grade 8 r= .488 and Grade 4 r= .360), school self-concept (Grade 8 r= .690 and Grade 4 r= .446), and interest (Grade 8 r= .229 and Grade 4 r= .180).

Task-specific self-efficacy. In this study a task-specific self-efficacy scale was used. In contrast to mathematics and other familiar academic areas, scientific reasoning tasks are less familiar to especially fourth graders, thus their respective performance might seem less predictable to them, more so since this competence is usually not explicitly taught in school as an own subject. The results of a pilot study with 57 fourth and eighth graders corroborated this impression. Based on the participants feedback and in accordance with general guidelines on scale construction (Bandura, 2006; Moosbrugger & Kleava, 2012) we designed a very task specific self-efficacy scale (3 items), in which the respective scientific reasoning task and the self-efficacy scale were presented together.

The task-specific self-efficacy were assessed with the following three items together with the scientific reasoning task before the students were asked to solve each of the six scientific reasoning tasks: 1) "I know how to deal with the task" 2) "I am very familiar with such tasks and know how to solve the task" 3) "I would need help to solve the task.". The Likert scale ranged from "strongly disagree" (1) to "strongly agree" (4).

2.2.3 Control variables

With a proficiency test (ELFE 1-6; Lenhard & Schneider, 2006) text comprehension was measured. To measure nonverbal intelligence a subtest of the Cultural Fair Intelligence Test (CFT; Weiß, 2006) was applied.

2.3 Procedure

In both age groups, the testing was conducted as a whole-class testing procedure, with each person working individually on their own booklet. Before answering each scientific reasoning task, the students were asked to fill in the items of the task-specific self-efficacy. In a first step, they were instructed to look at the scientific reasoning task about 25 seconds but not to try to solve the task. The time of 25 seconds appeared in a pilot study to be the best time to give the students enough time to look over the task but not to try to solve it. After they rated the task-specific self-efficacy item, they solved the scientific reasoning task. To avoid confounding effects of reading ability, the items were presented by a PowerPoint presentation and read aloud by an experimenter. The testing took about 60 minutes.

3.   Results

3.1   Core performance and suitability of the task-specific self-efficacy scale

3.1.1   Scientific reasoning

Fourth graders scored an average of 5.23 (*SD* = 3.06) out of 12 points (43.6%), whereas eighth graders performed significantly better, scoring an average of 7.96 (*SD* = 2.46) out of 12 points (66.3%), *t*(277) = -6.03, *p* < .01. According to Cohen (1992), the effect size of *r* = .60 is strong (see Figure 1).

3.1.2   Self-efficacy

Figure 1 shows the mean percent of academic and task-specific self-efficacy expectations transformed to a percentage scale (low to high feelings of self-efficacy). No significant differences in academic self-efficacy were found between grades (*t*(277) = 1.46, *p* = .15), between males and females in Grade 8 (*t*(148) = -.294, *p* = .77) and Grade 4 (*t*(140) = -1.25, *p* = .13). In task-specific self-efficacy, the eighth graders had significantly higher values than the fourth graders, *t*(279) = -3.36, *p* < .01 (see Figure 1). The effect size was *r* = .04, representing a small effect (Cohen, 1992). No significant difference was found between males and females in their task-specific self-efficacy in Grade 8, *t*(146) = 1.92, *p* = .06, and in Grade 4, *t*(139) = -.113, *p* = .91.
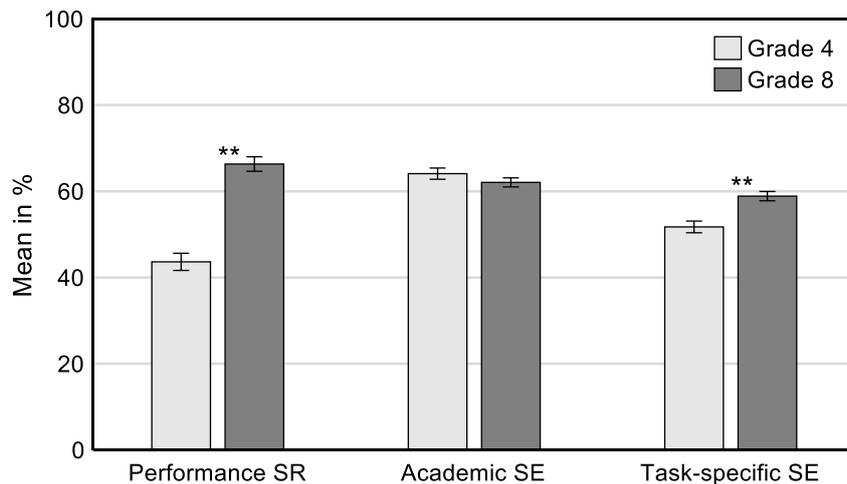
*Figure 1.* Comparison of mean performance in scientific reasoning, task-specific and academic self-efficacy between Grade 4 and 8. **p< .01, SR= scientific reasoning, SE= self-efficacy

The internal consistency in both grades (Cronbach's α Grade 4 = .89, Grade 8 = .88) were high as well as the corrected item-total correlations (r_{it} Grade 4 = .59-77, r_{it} Grade 8= .60-.82) which indicate a good reliability of the scale in both grade levels (Table 1; Bortz & Döring, 2006).

Table 1

Corrected-item-total correlations of the task-specific self-efficacy for Grades 4 and 8

	r_{it-i}																	
	NOS									UNEX								
	1_1	1_2	1_3	2_1	2_2	2_3	3_1	3_2	3_3	4_1	4_2	4_3	5_1	5_2	5_2	6_1	6_2	6_3
G 4	.70	.64	.68	.73	.73	.77	.77	.76	.75	.62	.59	.65	.59	.63	.67	.68	.64	.65
G 8	.62	.60	.61	.75	.81	.78	.68	.73	.76	.74	.76	.77	.65	.60	.82	.79	.80	.75

G=Grade, 1= A01 Scientist, 2= A03 Middle Ages, 3= A11 Mistakes in Science, 4= U1 Trees, 5= U2 Math textbook, 6= U3 Classroom

Content and criterion validity are ensured by the process of scale construction and correlations in line with theoretical findings (Marsh, 2018; Schukajlow et al., 2012). Significant relations were found with academic self-efficacy (Grade 4 r= .360, Grade 8 r= .488), academic self-concept (Grade 4 r= .395, Grade 8 r= .329), and interest (Grade 4 r= .542, Grade 8 r= .467).

Furthermore, we tested whether there was measurement invariance for the task-specific self-efficacy scale across Grades 4 and 8, using R and the lavaan package (Rosseel, 2012). The model shows scalar measurement invariance, following the criterion that maximum change per level in comparative fit index [CFI] and root-mean-square error of approximation [RMSEA] should not exceed a certain threshold (CFI ≤ -.010 and RMSEA ≤ .015; Chen, 2007; for further discussions see Cheung & Rensvold, 2002). Fit indices were as follows: Configural (CFI= .997; RMSEA= .043), metric (CFI= .994; RMSEA= .051) and scalar (CFI= .989; RMSEA= .066). All fit indices of the model accuracy remain in a good range (e.g., Hu & Bentler, 1999), and the maximum-change thresholds for CFI and RMSEA are not exceeded by the scalar model. Scalar invariance implies that the factor structure, factor loadings, and intercepts remain invariant across the two grade levels. Thus, a comparison of items between participants from different groups (here Grade 4 and 8) is possible with regard to the latent variable measured (here task-specific self-efficacy).

3.2   Relation between scientific reasoning and self-efficacy

3.2.1   Academic self-efficacy and scientific reasoning

As shown in Table 2, academic self-efficacy expectations correlated significantly with scientific reasoning in Grade 8 but not in Grade 4. No significant differences in the correlation in Grade 8 between females and males were found (p = .34). The correlations were controlled for intelligence and reading ability.

3.2.2   Task-specific self-efficacy and scientific reasoning

Performance in scientific reasoning and task-specific self-efficacy correlated significantly in both grades (Table 2). Again, no significant differences in the correlations between females and males were observed (Grade 4, p = .45; Grade 8, p =.24). The correlations were controlled for intelligence and reading ability.

Table 2

Correlations between scientific reasoning and academic self-efficacy or task-specific self-efficacy

		Academic SE	Task-specific SE
Grade 4	Scientific reasoning	.076	.194*
Grade 8	Scientific reasoning	.253***	.290***

***p<.001, *p<.05, SE = self-efficacy

3.3 Cluster analysis

The correlation analysis showed a positive relation of task-specific self-efficacy and the performance in scientific reasoning, but the magnitude of the correlation could point to heterogenous patterns within the correlation. A hierarchical cluster analysis was performed to investigate whether there are clusters of students who underestimated or overestimated their performance in relation to their actual performance. Therefore, the values of the performance in scientific reasoning and the task-specific self-efficacy were z-standardized. For the performance, the z-standardized residuals were applied to eliminate the common variance between the performance and task-specific self-efficacy. The squared Euclidean distance was taken as an approximate measure. As a clustering algorithm, the Ward method was selected. Ward's method has been shown to give better results with small clusters (e.g., Everitt et al., 2001). This procedure was similar to the approach of previous studies (e.g., Hallet et al., 2010; Siefer et al., 2020).

The cluster analysis was conducted separately for Grades 4 and 8 because different profiles in the agreement between scientific reasoning and self-efficacy could be expected in each age group. The analysis revealed a four-cluster solution for the fourth graders and a five-cluster solution for the eighth graders. The number of clusters was determined based on dendrograms and cophenetic correlations. Preliminary, on the whole no gender differences were found, except in the very small cluster 1 and 5, in Grade 4 and Grade 8 respectively.

Grade 4. Cluster 1 (underestimators/good performance) included 18 students (14.5%) who showed the highest performance in scientific reasoning across all clusters. Task-specific self-efficacy was also rated high, but it was below their performance. In other words, this group of students slightly underestimated their performance (see Figure 2). The cluster contained significant more females than males, $\chi^2(1, n=17) = 3.56, p < .05$.

Cluster 2 (overestimators/poor performance) was the most frequent cluster ($n = 51$, 41.1%). Students performed below average, but despite their weak performance, they showed high task-specific self-efficacy indicating overestimation of their performance (see Figure 2). No significant difference between the number of female or male students in this cluster was observed.

Cluster 3 (strong underestimators/good performance) ($n = 24$, 19.4%) could best be described as having high-performance and low task-specific self-efficacy. The students seemed to largely underestimate their performance (see Figure 2). Again, no significant difference was found in the number of female and male students.

The last cluster 4 (slight underestimators/poor performance) contained $n = 31$ (25%) students showing poor performance and low task-specific self-efficacy. The task-specific self-efficacy was even poorer than the performance, which showed that students underestimate their performance (see Figure 2). Again, there was no significant difference between the numbers of female and male students.
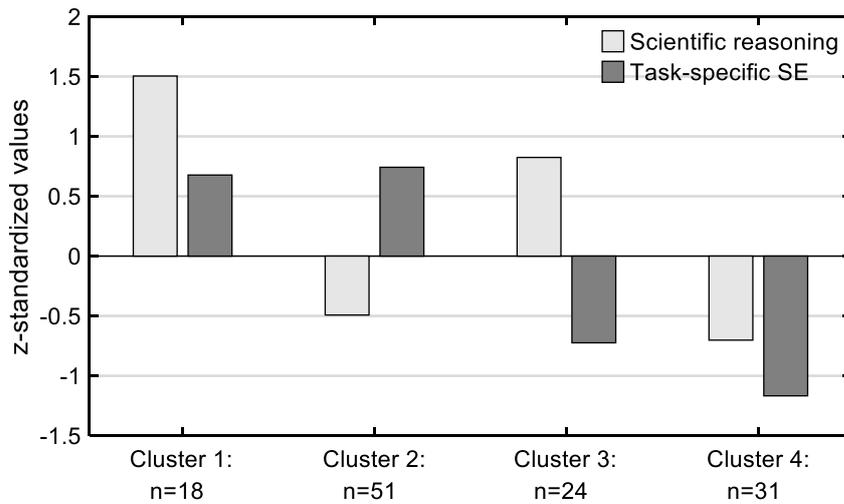
Figure 2. Showing the resulting clusters with performance in scientific reasoning (z-standardized) and task-specific self-efficacy (z-standardized) in Grade 4. Cluster 2 is the most frequent cluster. SE = self-efficacy

A one-way analysis of variance (ANOVA) was performed to statistically confirm the differences between the clusters. In Grade 4, cluster assignment had a significant effect on the performance in scientific reasoning, $F(4,131) = 90.567$, $p < .001$. The performance of the different clusters differed significantly among students across different clusters, with only the performance in cluster 2 ($M = -0.49$, $SD = 0.56$) not being significantly different from the performance in cluster 4 ($M = -0.70$, $SD = 0.50$). Cluster assignment also significantly affected the level of task-specific self-efficacy ($F(4,130) = 91.05$, $p < .001$). All clusters varied significantly in their level of task-specific self-efficacy, except for cluster 1 ($M = -0.74$ $SD = 0.48$) and 2 ($M = -0.68$ $SD = 0.67$), which were not significantly different.

Grade 8. The most frequent cluster in Grade 8 was cluster 1 (realistic estimators/good performance) with $n = 52$ (39%) students. The students showed an above-average performance and a high task-specific self-efficacy, indicating that the students' self-evaluated self-efficacy was close to their actual performance (see Figure 3). No significant difference between the number of female or male students in this cluster was observed.

Cluster 2 (strong overestimators/very poor performance) contained $n = 9$ (6.7%) students with a poor performance and a high task-specific self-efficacy, indicating a high overestimation of their performance (see Figure 3). Again, no significant difference was found in the number of female and male students.

In cluster 3 (strong underestimators/good performance) $n = 21$ (15.8%), students had a high performance and a poor task-specific self-efficacy. The students were high performance underestimators (see Figure 3). There was no significant difference between the numbers of female and male students in this cluster. Students ($n = 25$, 18.8%) in Cluster 4 (overestimators/very poor performance) demonstrated poor performance and average task-specific self-efficacy, indicated below-average overestimation. Also, no significant difference between the number of female or male students in this cluster was observed.

Cluster 5 (underestimators/average performance) with $n = 28$ (19.5%) included students with an average performance and a poor task-specific self-efficacy. The students underestimated their performance. Cluster 5 was the only cluster to have a significant difference between the number of female and male students with significantly more males in the cluster, $\chi^2(1, n = 25) = 3.85$. $p < .05$.

ANOVAs were also conducted for Grade 8 to test for differences between the clusters. Cluster assignment had a significant effect on the performance in scientific reasoning, $F(5,140) = 79.90$, $p < .001$. The performance of the clusters differed significantly, except between cluster 2 ($M = -1.28$, $SD = 0.42$) and cluster 4 ($M = 1.31$, $SD = 0.65$). Cluster assignment also significantly affected the level of task-specific self-efficacy, $F(5,139) = 82.08$, $p < .001$. All clusters varied significantly in their level of task-specific self-efficacy, except between cluster 3 ($M = -0.75$ $SD = 0.53$) and 5 ($M = -0.88$ $SD = -0.40$).
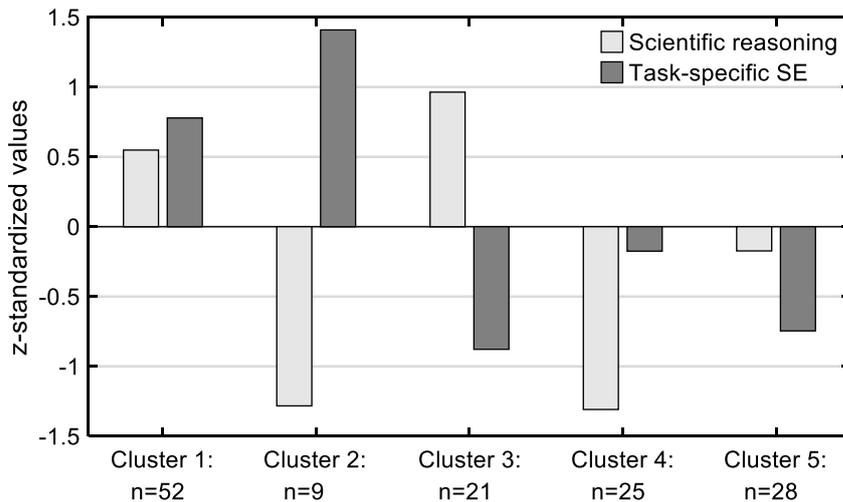

Figure 3. Showing the resulting clusters with performance in scientific reasoning (z-standardized) and task-specific self-efficacy (z-standardized) in Grade 8. Cluster 1 is the most frequent cluster. SE = self-efficacy

4. Discussion

The present study addressed five questions (1) Is our self-efficacy scale suitable to measure task-specific self-efficacy in scientific reasoning for Grades 4 and 8? (2) What is the relation between students' performance in scientific reasoning and their self-efficacy expectations, specifically is this relation already observable in elementary school? (3) Are the correlations higher when measuring task-specific self-efficacy instead of academic self-efficacy? (4) And, are there differences in the strength of the relation between both grade levels? (5) Finally, do diverse cluster exist which categorize students along the task-specific self-efficacy-performance relation (e.g., over- or underestimate and high and low performer)?

4.1 Core Performance and suitability of the used scales

As expected, performance in scientific reasoning was significantly higher in Grade 8 than in Grade 4. Nonetheless, no ceiling effect was observed in the performance of the eighth graders. This finding suggests that performance in scientific reasoning continues to develop through the elementary school years into the secondary school years and may not be completed by the end of the secondary school years, which is in line with previous findings (Bullock et al., 2009; Bullock & Ziegler, 1999; Koerber et al., 2015).

When studying students' self-efficacy, it should be emphasized that self-efficacy could be measured task-specifically for scientific reasoning skills– in elementary and secondary school students. And, in contrast to many other studies, the present study investigated two levels of self-efficacy:

Academic self-efficacy and task-specific self-efficacy. The item characteristics and criterion-related correlations show comparable values for Grade 4 and 8 and indicate that the scales can be applied in both grades. The results related to academic self-efficacy should be interpreted carefully, as the scale does not show high internal consistency, nevertheless (significant) correlations were found, which indeed could have been stronger with a scale having higher internal consistency. Nevertheless, for a replication of the results, especially in relation to the academic self-efficacy scale, the low internal consistency should be included in further considerations. The item characteristics, correlations of the task-specific self-efficacy scale, the high internal consistency and scalar measurement invariance across both grade levels suggest the reliable and valid use of the scale for Grades 4 and 8.

The two grades revealed similar academic self-efficacy with no significant difference (Grade 8: 64% vs. Grade 4: 62%). The task-specific self-efficacy, however, differed significantly between the two grades. Eighth graders reported significantly higher task-specific self-efficacy than fourth graders. This finding is consistent with the development of self-efficacy expectations as described by Bandura (1997) who identified different sources of self-efficacy. Two of the crucial sources for building self-efficacy expectations may be previous mastery experiences and vicarious experiences (Usher, 2009). Mastery experiences occur when students experience feelings of success on a particular task, which engenders the belief that they can succeed in the task again. Vicarious experiences refer to observing someone else perform the task and solve it. Students learn from this observation that they may also succeed at the task. Young students often have little opportunity to benefit from such experiences in building their self-efficacy beliefs in this specific area because they are not so familiar with scientific reasoning tasks yet. However, eighth graders are more likely to have the opportunity in school to work on science problems in physics or NwT (science and technology) than fourth graders. Therefore, it is not surprising that eighth graders rated themselves significantly higher in task-specific self-efficacy.

4.2 Relation between scientific reasoning and self-efficacy

The correlation between academic self-efficacy and scientific reasoning was found only in Grade 8. Task-specific self-efficacy and performance in scientific reasoning correlated significantly across Grades 4 and 8, with a higher correlation emerging for eighth graders. Compared to other studies, we found correlations that tended to be lower. Multon and Brown (1991) reported a mean correlation of $r = .38$ in their meta-analysis, and Honicke and Broadbent (2016) found a similar result of $r = .33$. However, studies with different samples and heterogenous instruments were included in the studies. In a study by Siefer et al. (2020), task-specific self-efficacy correlated with the performance in a specific mathematical area. They reported a correlation of $r = .39$ between task-specific self-efficacy and the students' performance on a test of linear functions in Grade 8 and 9. These studies also would have hinted towards a higher agreement between performance and self-efficacy if self-efficacy was measured specifically. The lower correlations in our study may have resulted from scientific reasoning not being assigned as an independent subject to the curriculum. Receiving feedback and benefiting from mastery experiences in scientific reasoning is more difficult for scientific reasoning tasks (as opposed to mathematics). Consequently, students receive less formal feedback on their performance (e.g., in the form of grades). Fourth-grade students especially have little opportunity to profit from mastery experiences in scientific reasoning because the opportunities to work on science problems, as required in the here used scientific reasoning tests, are missing.

In general, in contrast to self-efficacy measures concerning science content knowledge (e.g., in the 2015 PISA study; Schiepe-Tiska et al., 2016), we observed no gender differences neither in Grade 4 nor in Grade 8 in the (scientific reasoning) task-specific self-efficacy and the correlations. This is consistent with previous findings (e.g., Koerber et al., 2015) who also reported no gender differences in scientific reasoning.

4.3   Differences in the agreement between scientific reasoning and task-specific self-efficacy

The positive correlations between performance and task-specific self-efficacy in Grades 4 and 8 suggested that higher performance is associated with higher task-specific self-efficacy. However, the rather low correlation coefficients could also indicate interindividual differences between students. For that purpose, a hierarchical cluster analysis was conducted. Looking at the found clusters the most noteworthy result was the largest cluster in Grade 8, which contained the students who judged their performance realistically and showed a high performance. This fits with the findings of Siefer et al. (2020), which also showed that the largest group of the sample (33%) realistically assessed themselves in the domain of linear functions in Grades 8 and 9. The cluster with the students, who realistically judge their performance, however, was nonexistent in Grade 4. Many fourth-grade students were assigned to the cluster of overestimators who performed poorly. This in line with previous research findings that elementary school students are more likely to overestimate themselves, in contrast to the beginning of secondary school, where self-efficacy appears to decline (Harter, 1985; Zimmerman, 1995). Students in Grade 4 seemed to mostly overestimate or underestimate their abilities. Being able to realistically judge one's performance requires the skill to realistically judge the required task at an abstract level, which could be difficult for fourth graders. Thus, the result for Grade 4 students could have reflected in a non-realistic judgment of their performance. For example, a study by Kruger and Dunning (1999) suggests the skills needed for a certain task or domain are the same skills needed to evaluate one's performance. A reasonable assumption is that the fourth graders, who are not as good at scientific reasoning as the eighth graders, may not be as skilled at judging their skills and may consequently overestimate or underestimate their scientific reasoning skills. Students become more realistic in their self-efficacy evaluations over time. Nevertheless, clusters are more divergent in Grade 8 in their levels of performance and self-efficacy. It is well established in literature that both great overestimation and underestimation of one's own abilities can be a barrier to performance (e.g., Bandura, 1986; Hattie, 2013). If students are overconfident that they can complete the task or situation well, this may lead them to not take proper strategies or other actions to master the task (Hadwin & Webster, 2013). In contrast, learners with insufficient confidence may be absorbed with cognitive resources that they have to invest more effort in achieving the goal than they should have (Hattie, 2013). The generally high number of underestimator distributed across Grade 4 and 8 in this study could find themselves in a helpless position, which then results in poorer performance in scientific reasoning, which in turn is critical for performance in scientific literacy and the important STEM subjects. This differentiated information is especially relevant for teachers, as it is crucial for them to 1) realize at all, that there might be a substantive heterogeneity in their class with respect to students' estimation of their own performance. 2) This differentiated information helps teacher providing feedback that results in a realistic confidence with at most minimal overestimation, as this allows for the best student performance (e.g., Bandura, 1977) 3) This in turn, might support students in their self-regulation behavior. And finally, 4) this promotional opportunity appears useful in the context that self-efficacy is a measure that can be improved through intervention or training (Sodian et al., 2002; Margolis & McCabe, 2006).

4.4   Limitations and future directions

The cross-sectional design used in the present study allows for a thorough investigation of self-efficacy expectations and their relation to scientific reasoning skills. When interpreting the results, the rather low reliability of the academic self-efficacy scale should be taken into account. To understand and explain interindividual differences in the agreement between performance in scientific reasoning and task-specific self-efficacy, variables that provide information about the use of feedback from teachers and parents would be important. Furthermore, we cannot conclude whether self-efficacy influences scientific reasoning or, conversely, whether scientific reasoning influences self-efficacy. Bandura (1997) argued for a reciprocal relation between skills and self-efficacy. Lawson et al. (2007) found support for the hypothesis that reasoning skills are a good predictor of self-efficacy but not the other way around. The fact that results from specific domains cannot simply be transferred across

domains was shown in the findings from Schöber et al. (2018). For the math domain, the authors found support for a self-enhancement approach (i.e. self-efficacy influences math performance), whereas for writing skills, a skill-development approach was supported (i.e. writing skills influence self-efficacy). Only longitudinal studies can reveal the direction and strength of the relations in the context of scientific reasoning skills. The longitudinal design also allows to examine how self-efficacy and scientific reasoning develop over time. The outcomes could provide relevant implications for a potential intervention. Identifying where to target an intervention is critical, whether to focus on scientific reasoning skills or self-efficacy expectations and to identify the developmental stage that may be most sensitive to intervention.

4.5   Conclusion

The present study showed, for the first time, that self-efficacy can be measured in a task-specific manner in scientific reasoning and found a positive correlation between (task-specific) self-efficacy and performance in scientific reasoning, already in students at the end of elementary school. This suggests that a precise and task-specific measurement of self-efficacy can detect effects already in elementary school children. In addition, our results show substantial interindividual differences and differences in age groups in the agreement between self-efficacy and scientific reasoning. This is an important outcome, especially regarding the influence of scientific reasoning skills on science content knowledge and the relevant STEM subjects. Therefore, this is a possible starting point for promoting this academic discipline.

Keypoints

- Self-efficacy can be measured in a task-specific manner in scientific reasoning- for both elementary and secondary school students
- Stronger correlation between self-efficacy and scientific reasoning with more precisely measured self-efficacy
- Differences in the agreement between scientific reasoning and task-specific self-efficacy among and across grades

References

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory.* Engelwood Cliffs.

Bandura, A. (1997). *Self efficacy: The exercise of control*. Freeman.

Bandura, A (2006). Guide to the construction of self-efficacy scales. In F. Pajares & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (pp. 307-337). Information Age.

Bandura, A., & Schunk, D. H. (1981). Cultivating competence, self-efficacy and intrinsic interest through proximal self-motivation. *Journal of Personality and Social Psychology, 41*, 586–598. https://doi.org/10.1037/0022-3514.41.3.586

Bong, M. (2006). Asking the right question. How confident are you thatyou could successfully perform these tasks? In F. Pajares & T. C. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (pp.287–305). Information Age.

Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* [Research Methods and Evaluation for Social Scientists]. Springer.

Britner, S. L., & Pajares, F. (2001). Self-efficacy beliefs, motivation, race, and gender in middle school science. *Journal of Women and Minorities in Science and Engineering, 7*, 271–285. https://doi.org/10.1615/JWomenMinorScienEng.v7.i4.10

Bullock, M., Sodian, B., & Koerber, S. (2009). Doing experiments and understanding science: Development of scientific reasoning from childhood to adulthood. In W. Schneider & M. Bullock (Eds.). *Human development from early childhood to early adulthood: Findings from a 20 year longitudinal study* (pp.173–197). Psychology Press.

Bullock, M., & Ziegler, A. (1999). Scientific reasoning: Developmental and individual differences. In F. E. Weinert & W. Schneider (Eds.), *Individual Development from 3 to 12: Findings from the Munich Longitudinal Study* (pp. 38–54). Cambridge University Press.

Bybee, R. W. (2010). Advancing STEM education: A 2020 vision. *Technology and Engineering Teacher, 70,* 30–35.

Cai, D., Viljaranta, J., & Georgiou, G. K. (2018). Direct and indirect effects of self-concept of ability on math skills. *Learning and Individual Differences, 61,* 1–38. https://doi.org/10.1016/j.lindif.2017.11.009

Chen, F.F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*, 464-504. https://dx.doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. Structural Equation Modeling, *9*, 233-255. https://doi.org/10.1207/S15328007SEM0902_5

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159. https://doi.org/10.1037/0033-2909.112.1.155

Everitt, B. S., Landau, S., & Leese, M. (2001). *Cluster analysis* (4th ed.). Oxford University Press.

Hadwin, A. F., & Webster, E.A. (2013). Calibration in goal setting: examining the nature of judgments of confidence. *Learning and Instruction*, *24*, 37-47. https://doi.org/10.1016/j.learninstruc.2012.10.001

Hattie, J. (2013). Calibration and confidence. Where to next? Learning and Instruction, *24*, 62-66. https://doi.org/10.1016/j.learninstruc.2012.05.009

Hallet, D., Nunes T., & Bryant, P. (2010). Individual differences in conceptual and procedural knowledge when learning fractions. *Journal of Educational Psychology, 102,* 395-406. https://doi.org/10.1037/a0017486

Harter, S. (1985). *Manual for the Self-Perception Profile for Children*. University of Denver

Honicke, T., & Broadbent, J. (2016). The influence of academic self-efficacy on academic performance: A systematic review. *Educational Research Review, 17,* 63–84. https://doi.org/10.1016/j.edurev.2015.11.002

Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55. https://doi.org/10.1080/10705519909540118

Jansen, M., Scherer, R., & Schroeders, U. (2015). Students' self-concept and self-efficacy in the sciences: Differential relations to antecedents and educational outcomes. *Contemporary Educational Psychology, 41,* 13–24. https://doi.org/10.1016/j.cedpsych.2014.11.002

Jerusalem, M., & Satow, L. (1999). Schulbezogene Selbstwirksamkeitserwartung. In R. Schwarzer & M. Jerusalem (Eds.), *Skalen zur Erfassung von Lehrer- und Schülermerkmalen.* Freie Universität Berlin.

Klassen, R.M., & Usher, E. L. (2010). Self-efficacy in educational settings: Recent research and emerging directions. In T. C. Urdan & S. A. Karabenick (Eds.), *The Decade Ahead: Theoretical Perspectives on Motivation and Achievement* (pp.1–33). Emerald.

Koerber, S., Mayer, D., Osterhaus, C., Schwippert, K., & Sodian, B. (2015). The development of scientific thinking in elementary school: A comprehensive inventory. *Child Development*, *86,* 327–336. https://doi.org/10.1111/cdev.12298

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

Lawson, A.E., Banks, D.L., & Lovgin, M. (2007). Self-efficacy, reasoning ability, and achievement in college biology. *Journal of research in Science Teaching, 44*, 706-724. https://doi.org/10.1002/tea.20172

Lenhard, W., & Schneider, W. (2006). *ELFE 1-6: Ein Leseverständnistest für Erst- bis Sechstklässler.* Hogrefe.

Liu, M., Hsieh, P., Cho, Y., & Schallert, D. L. (2006). Middle school students' self-efficacy, attitudes, and achievement in a computer-enhanced problem-based learning Environment. *Journal of Interactive Learning Research, 17,* 225–242.

Margolis, H., & McCabe, P. P. (2006). Improving self-efficacy and motivation: What to do, what to say. *Intervention In School & Clinic*, *41*, 218–227. https://10.1177/10534512060410040401

Marsh, H.W., Pekrun, R., Parker, P. D., Murayama, K., Guo, J., Dicke, T., & Arens, A.K. (2018). The murky distinction between self-concept and self-efficacy: Beware of lurking jingle-jangle fallacies. Journal of Educational Psychology, *111*, 331-353. https://doi.org/10.1037/edu0000281

Moosbrugger, H., & Kelava, A. (2012). *Testtheorie und Fragebogenkonstruktion* [Test theory and questionnaire construction]. Springer.

Multon, K. D., & Brown, S. D. (1991). Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation. *Journal of Counseling Psychology, 18,* 30–38.

Nagengast, B., Marsh, H. W., Scalas L.F., Xu, M. K., Hau, K. T., & Trautwein, U. (2011). Who took the "x" out of expectancy-value theory? a psychological mystery, a substantive-methodological synergy, and a cross-national generalization. *Psychological Science*, *22,* 1058–1066. https://doi.org/10.1177/0956797611415540

Nicholls, J. G. (1979). Development of perception of own attainment and causal attributions for success and failure in reading. *Journal of Educational Psychology, 71*, 94–99. https://doi.org/10.1037/00220663.71.1.94

Nunnally, J.C. & Bernstein I.H. (1978). *Psychometric Theory*. New York: McGraw-Hill

OECD (2006). *Assessing scientific, reading and mathematical literacy: A framework for PISA 2006*. Paris: OECD.

OECD (2015). *OECD science, technology and industry scoreboard 2015: Innovation for growth and society*. Paris: OECD.

Osborne, J. (2013). The 21st century challenge for science education: Assessing scientific reasoning. *Thinking Skills and Creativity*, *10,* 265–279. https://doi.org/10.1016/j.tsc.2013.07.006

Osterhaus, C., Koerber, S., & Sodian, B. (2020). The Science-P Reasoning Inventory (SPR-I): Measuring emerging scientific reasoning skills in primary school. *International Journal of Science Education*, *42,* 1087-1107. https://doi.org/10.1080/09500693.2020.1748251

Osterhaus, C., Koerber, S., & Sodian, B. (2017). Scientific thinking in elementary school: Children's social cognition and their epistemological understanding promote experimentation skills. *Developmental Psychology, 53*, 450-462. https://doi.org/10.1037/dev0000260

Osterhaus, C., Koerber, S., & Sodian, B. (2015). Children's understanding of experimental contrast and experimental control: an inventory for primary school. *Frontline Learning Research, 3,* 56–94.

Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research*, *66,* 543–578. https://doi.org/10.3102/00346543066004543

Pajares, F., & Schunk, D.H. (2001). The development of academic self-efficacy. In A. Wigfield & J.S. Eccles (Eds.), Development of achievement motivation (pp. 15-31). Academic Press.

Pajares, F., & Valiante, G. (1997). Influence of Self-Efficacy on Elementary Students' Writing. *The Journal of Educational Research, 90,* 353–360. https://doi.org/10.1080/00220671.1997.10544593

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin, 138,* 353–387. https://doi.org/10.1037/a0026838

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*, 1-36. https://doi.org/10.18637/jss.v048.i02

Schiepe-Tiska, A., Simm, I., & Schmidtner, S. (2016). Motivationale Orientierungen, Selbstbilder und Berufserwartungen in den Naturwissenschaften in PISA 2015. In K. Reiss, C. Sälzer, A. Schiepe-Tiska, E. Klieme, & O. Köller (Eds.), *PISA 2015. Eine Studie zwischen Kontinuität und Innovation* (pp. 99–132). Waxmann.

Schöber, C., Schütte, K., Köller, O., McElvany, N., & Gebauer, M. M. (2018). Reciprocal effects between self-efficacy and achievement in mathematics and reading. *Learning and Individual Differences, 63,* 1–11. https://doi.org/10.1016/j.lindif.2018.01.008

Schunk, D. H. (1995). Self-efficacy and education and instruction. In J. E. Maddux (Ed.), *Self efficacy, adaptation, and adjustment: Theory, research, and application* (pp. 281–303). Plenum Press.

Schukajlow, S., Leiss, D., Pekrun, R., Blum, W., Muller, M., & Messner, R. (2012). Teaching methods for modelling problems and students' task-specific enjoyment, value, interest and self-efficacy expectations. *Educational Studies in Mathematics*, *79*, 215–237. https://10.1007/s10649-011-9341-2

Schwarzer, R., & Jerusalem, M. (2002). Das Konzept der Selbstwirksamkeit. *Zeitschrift Für Pädagogik*, *44,* 28–53.

Siefer, K., Leuders, T., & Obersteiner A. (2020). Leistung und Selbstwirksamkeitserwartung als Kompetenzdimension: Eine Erfassung individueller Ausprägungen im Themenbereich linearer Funktionen. *Journal für Mathematik-Didaktik, 41*, 267-299. https://doi: 10.1007/s13138-019-00147-x

Stipek, D.C. & Hoffman, J.H. (1980). Children's achievement-related expectancies as a function of academic performance histories and sex. *Journal of Educational Psychology*, *70*, 154-166.

Sodian, B., Thoermer, C., Kircher, E., Grygier, P., & Günther, J. (2002). Vermittlung von Wissenschaftsverständnis in der Grundschule [Teaching understanding the nature of science in elementary school]. Zeitschrift für Pädagogik, *45*, 192–206.

Usher, E. L., & Pajares, F. (2009). Sources of self-efficacy in mathematics: a validation study. *Contemporary Educational Psychology, 34*, 89–101. https://doi.org/10.1016/j.cedpsych.2008.09.002

Weiß, R.H. (2006). Grundintelligenztest Skala 2 Revision (CFT 20-R). Hogrefe

Zeldin, A. L., & Pajares, F. (2000). Against the odds: self-efficacy beliefs of women in mathematical, scientific, and technological careers. *American Educational Research Journal*, *37*, 215–246. https://doi.org/10.3102%2F00028312037001215

Zimmerman, B. (1995). Self-efficacy and educational development. In A. Bandura (Ed.). *Self-efficacy in changing societies* (pp. 202-231). Cambridge University Press.

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, *27,* 172–223. https://doi.org/10.1016/j.dr2006.12.001

Appendix 1

1. Sample items

1.1 Task-specific self-efficacy item and NOS task (A03 Middle Ages; Osterhaus et al., 2020)

This is not about solving the task, but you are asked to complete the question below in terms of yourself.

Long ago, in the Middle Ages, people believed there are witches who could make people sick.	
A modern-day scientist traveled back to the Middle Ages with a time machine.	
Scientists in the Middle Ages thought that witches can make people sick. The modern-day scientist believes that bacteria can make people sick.	
The modern-day scientist shows the scientist from the Middle Ages the bacteria under the microscope and explains: "These bacteria are the reason why people get sick!"	

What will the scientist _from the Middle Ages_ say to this?

	Strongly disagree	Disagree	Agree	Strongly agree
1. I know how to deal with the task.	☐	☐	☐	☐
2. I am very familiar with such tasks and know how to solve the task.	☐	☐	☐	☐
3. I would need help to solve the task.	☐	☐	☐	☐

Now you may solve the task.

What will the scientist **from the Middle Ages** say to this?		
	He would say this	**He would <u>not</u> say this**
1. "Of course, you're right. Bacteria make people sick, not witches."	☐	☐
2. "Bacteria could be the witches' little helpers."	☐	☐
3. "It may be true that there are bacteria here, but witches are still the ones who make people sick."	☐	☐
Which is the best answer?	No._____	

1.2 Task-specific self-efficacy item and UNEX task (U1 Trees; Osterhaus et al., 2015)

This is not about solving the task, but you are asked to complete the question below in terms of yourself.

A scientist travels to a faraway planet, planet Ogi. There, he observes that all trees are very small.	
The scientist develops a tree medicine that is supposed to help the trees grow. He calls this medicine Supergrow. The scientist wants to find out whether his Supergrow works and whether it really makes the trees grow. Therefore, he conducts an experiment. He gives his Supergrow to all trees on planet Ogi.	
Six months later, the scientist travels back to planet Ogi. He observes that all trees are huge now He is convinced: "My Supergrow works!"	
Was this a good experiment?	
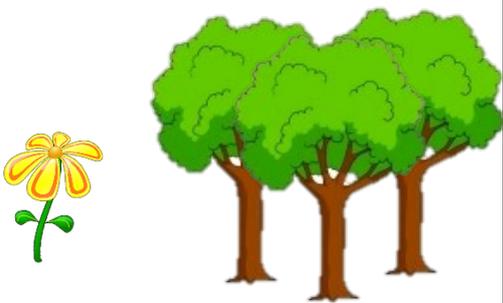
	Strongly disagree	Disagree	Agree	Strongly agree
1. I know how to deal with the task.	☐	☐	☐	☐
2. I am very familiar with such tasks and know how to solve the task.	☐	☐	☐	☐
3. I would need help to solve the task.	☐	☐	☐	☐

Now you may solve the task.

Was this a good experiment?	
☐ **Yes**	☐ **No**

Susan, Lisa, and Vera wonder whether this was a good experiment.		
Who is right and who is not?		
	Is right	**Is <u>not</u> right**
1. *Susan* says: „It was not a good experiment because he does not know how big the trees would have grown without his Supergrow."	☐	☐
2. *Lisa* says: „It was a good experiment because he found that all trees have grown huge after receiving his Supergrow."	☐	☐
3. *Vera* says: „It was a good experiment because you can only see whether things work if you test them."	☐	☐
Which of the three girls has the <u>best</u> answer?	**No._____**	