The Development of Scientific Strategy Knowledge Across Grades: A Psychometrical Approach

Ronny Scherer1 and Rüdiger Tiemann1

Abstract
In this study, we developed a new test on scientific strategy knowledge and investigated the construct validity of the resulting test scores. Moreover, measurement invariance across grade levels has been analyzed to ensure the generalizability of the assessment. Furthermore, convergent and discriminant validity were investigated. A total of $N = 1,182$ German high school students of Grade Levels 8, 10, and 12 completed tasks on strategy knowledge, fluid intelligence, content knowledge, interest in science, and scientific self-concept within a cross-sectional study. Multigroup confirmatory factor analysis was used to check for measurement invariance. Our results show that scalar invariance holds across grades and that there are significant differences in performance favoring students of higher grade levels. Furthermore, fluid intelligence and content knowledge are relevant predictors of strategy knowledge, whereas gender and motivational constructs do not show significant effects. Implications for developmental studies on strategy knowledge and assessment practice are discussed.

Keywords
measurement invariance, multigroup analysis, science education, strategy knowledge

Introduction
Metacognitive abilities are key factors in scientific problem solving. Research has provided evidence on its importance in carrying out the individual steps of problem-solving processes assumed to underlie scientific investigations (H. Kim & Pedersen, 2011; Kuhn, Iordanou, Pease, & Wirkala, 2008; Künsting, Wirth, & Paas, 2011; Thillmann, 2007; Zimmerman, 2007). Moreover, research suggests that metacognitive abilities are prerequisites for the transfer of knowledge across different steps in problem solving and can, therefore, be regarded as essential in science education (Cooper & Sandi-Urena, 2009; Gamo, Sander, & Richard, 2010; Kapa, 2007).

Efklides and Vlachopoulos (2012) differentiated three main facets of metacognition, which were also proposed by Flavell (1979) and Kuhn (2000). First, metacognitive knowledge refers to declarative knowledge about tasks (Kuhn & Pearsall, 1998; Schneider & Artelt, 2010), beliefs about knowledge and cognition (Liu, 2010), persons, and knowledge about strategies, which can be used in specific problem-solving situations (Efklides & Vlachopoulos, 2012; Neuenhaus, Artelt, Lingel, & Schneider, 2011; Taasoobshirazi & Glynn, 2009). Second, metacognitive skills refer to the application of procedural knowledge in problem-solving processes (Funke & French, 2007; Goode & Beckmann, 2010). Finally, metacognitive experiences are closely related to emotions and one’s self-efficacy (Efklides & Vlachopoulos, 2012). Based on the concept of scientific inquiry, Mayer (2007) argued that knowledge about strategies and tasks can be regarded as one crucial factor in problem solving. In this context, problem solving is defined as the ability to perform operations to bridge the gap between an initial and a goal state (e.g., Novick & Bassok, 2005). But the relationship between knowledge about strategies and the application of strategies in problem situations is not deterministic (Amsel et al., 2008; Kuhn & Pearsall, 1998; Neuenhaus et al., 2011), although correlations have been found (Funke & French, 2007; Goode & Beckmann, 2010; Schneider & Artelt, 2010).

Whereas research on metacognition has mainly focused on procedural aspects, little is known about the structure and development of the declarative components. Given the lack of assessments of knowledge about strategies (strategy knowledge), the questions of domain-specificity and how this construct could be evaluated in specific contexts have rarely been addressed (Neuenhaus et al., 2011; Rickey & Stacy, 2000; Schneider & Artelt, 2010). Researchers who are interested in the assessment of strategy knowledge have also argued that developmental research on students’ performance

1Humboldt-Universität zu Berlin, Germany

Corresponding Author:
Ronny Scherer, Humboldt-Universität zu Berlin, Humboldt Graduate School, Unter den Linden 6, 10099 Berlin, Germany.
Email: ronny.scherer@chemie.hu-berlin.de
and its relation to constructs such as intelligence and content knowledge can be regarded as desiderata in educational research (Schneider & Artelt, 2010; Van der Stel & Veenman, 2010). Taking into account these research gaps, we developed a test on strategy knowledge for the domain of chemistry and evaluated the resulting measurement model according to its factorial structure. To use the test for analyzing differences across grade levels, further statistical properties such as measurement invariance are investigated. In addition, this study provides information on the relationships between strategy knowledge and related constructs as evidence on the measure’s convergent and discriminant validity (Borsboom, Mellenbergh, & Van Heerden, 2004; Campell & Fiske, 1959; Cronbach & Meehl, 1955; Messick, 1995).

Metacognitive Knowledge About Strategies

Solaz-Portolés and SanJosé López (2008) defined declarative knowledge as “static knowledge about facts and principles that apply within a certain domain” (p. 107). It contains knowledge about variables and about how to control them to obtain information on a system or experiment (Flavell, Miller, & Miller, 2002; Schneider & Artelt, 2010). It is, therefore, part of the declarative metamemory (Flavell, 1979). According to Kuhn’s (2000) framework of metacognition, in which metacognitive awareness and controlling processes were differentiated, strategy knowledge refers to the first category of awareness. Artelt, Beinicke, Schlagmüller, and Schlagmüller (2009) as well as Schneider and Artelt (2010) followed this approach, and defined strategy knowledge by operationalizing the construct as declarative knowledge about the nature of problems, tasks, and strategic behavior. This definition also contains knowledge about the control of variables and processing information. In our study, we refer to this operationalization of the construct as a part of metacognitive knowledge.

Until now, there have been various studies on the question of whether or not certain types of knowledge or competencies are domain-general or domain-specific (De Jong & Ferguson-Hessler, 1996). From a psychological perspective, strategy knowledge might be domain-general because it refers to common problem-solving strategies that could be applied in various contexts (Van der Stel & Veenman, 2010; Veenman, Elshout, & Meijer, 1997). From an educational perspective, knowledge and competencies are always acquired in specific contexts and can, thus, be regarded as domain-specific (e.g., Hambrick, 2005; Neuenhaus et al., 2011). De Jong and Ferguson-Hessler (1996) further argued that knowledge about strategies is always bound to specific problems within a domain. In our study, we consider strategy knowledge as a domain-specific construct but argue that there might be an overlap with other domains.

There have been a few approaches to assess students’ strategy knowledge in science. For example, Thomas, Anderson, and Nashon (2008) as well as Velayutham, Aldridge, and Fraser (2011) developed paper-and-pencil tests that were based on self-reporting scales. Other studies used interview scenarios or peer tutoring procedures (see Schneider & Artelt, 2010). Recent research has focused on the assessment of strategy knowledge with questionnaires that contain specific problems and strategies (e.g., Artelt et al., 2009; Klieme et al., 2010; Shahat, Ohle, Tregast, & Fischer, 2013; Thillmann, 2007). In these approaches, students have to evaluate strategies according to their adequacy in solving a problem or task. Schneider and Artelt (2010) noted that strategy knowledge should not be assessed during problem-solving processes because it is regarded as a declarative and therefore static component of knowledge. In contrast, procedural knowledge could be assessed during problem-solving processes to capture its application in specific situations. In our study, the development of an appropriate test instrument for the domain of chemistry refers to the static assessment procedure proposed by Artelt et al. (2009).

According to Berger and Karabenick (2011) and Thillmann (2007), there are different covariates that affect students’ performance on strategy knowledge. Besides motivational and self-related constructs, domain-specific performance (e.g., school grades) and fluid intelligence show significant correlations (Hoffman & Schraw, 2009; Panaoura & Philippou, 2007; Van Kraayenoord & Schneider, 1999). In addition, De Jong and Ferguson-Hessler (1996) identified content knowledge and the accessibility of mental models as further covariates. Taking these results into account, the present study also intends to describe the relationships among strategy knowledge and covariates to check for convergent and discriminant validity (Campell & Fiske, 1959; Messick, 1995).

Developmental Aspects of Strategy Knowledge

There have been only a few attempts to investigate strategy knowledge across grade levels. In this section, the main outcomes of these studies are presented.

Kolic-Vehovec, Bajsanki, and Roncetic Zubkovic (2010) argued that the importance of metacognition in reading comprehension intensified with increasing age. In a longitudinal study, they found that patterns of age differences varied across the different components of metacognition. Furthermore, the development of metacognitive knowledge differed according to the transitions across the life span (Annevirta & Vauras, 2001; Panaoura & Philippou, 2007; Schneider, 2008, 2010). For instance, Schneider (2010) proposed a major shift in strategy knowledge at the end of kindergarten and elementary school, whereas Lemaire and Lecacheur (2011) found a development from Grades 3 to 7 in
selecting problem-solving strategies. Van der Stel and Veenman (2010) were also able to show that metacognitive skills and procedural knowledge increased over time. Further studies revealed that developmental patterns of metacognition also differed across gender groups. For instance, girls performed better in strategy knowledge tasks during high school (Artelt et al., 2009; Leutwyler, 2009). But first-grade boys significantly outperformed girls of the same age group in the domain of mathematics within a study conducted by Carr, Jessup, and Fuller (1999). In sum, research indicates that there is a growth in metacognitive knowledge and skills with increasing age or grade level.

Measurement Invariance
The growing interest in comparing students’ performance across grade levels, gender, or ethnic groups has led to the question of the psychometric properties, which have to be fulfilled to conduct these comparisons. If researchers intend to compare test performance across different groups, they have to ensure that the tests assess the same construct in each of these groups. This test feature accounts for the generalizability and structure of the construct (Messick, 1995). Psychometricians have recently focused on establishing statistical models that can be used to investigate whether or not a construct holds across groups (Brown, 2006; E. S. Kim & Yoon, 2011). In these models, the analysis of measurement invariance has become an important method to check for test validity. Invariance can also be regarded as a prerequisite for establishing vertical scales in longitudinal or cross-sectional studies and facilitates the modeling of learning progressions (Kolen & Brennan, 2004; Köller & Parchmann, 2012; Wang & Jiao, 2009). If a measure holds across grade levels, researchers could use the test for investigating developmental changes by comparing latent means, variances, and covariances. Educational researchers and psychologists have therefore applied this concept for various constructs across different domains (e.g., Bowden, Saklofske, & Weiss, 2011; Doran, Aldridge, Roesch, & Myers, 2011; Lakin, 2012; Martin, 2009).

Research Goals
In light of the research gaps on strategy knowledge, the present study focuses on the assessment of the construct across grade levels and analyzes the factorial invariance of the resulting measurement model. In this regard, we provide a methodological approach of comparing students’ performance across grades, which could be transferred to similar cross-sectional studies. More precisely, our research goals are as follows:

1. Developing a test on strategy knowledge for the domain of chemistry
2. Analyzing the resulting measurement model and its measurement invariance across grade levels
3. Comparing students’ performance of strategy knowledge across grades
4. Analyzing the effects of covariates on strategy knowledge to validate the test

In our study, we mainly focus on the evaluation of measurement invariance to legitimize comparisons in strategy knowledge across Grades 8, 10, and the upper secondary level (Grade 12). The analysis of measurement invariance provides evidence on the generalizability of test scores across grade levels and thus contributes to construct validity. Further analyses account for construct validity by analyzing the relationships among the construct and its covariates. These analyses are regarded as methods of obtaining evidence on whether the strategy knowledge scores measure the intended construct (Borsboom et al., 2004; Cronbach & Meehl, 1955). In this regard, we aim to show that strategy knowledge, fluid intelligence, and content knowledge are empirically related constructs. To sum up, this article provides a new assessment of strategy knowledge and validates the resulting test scores across three grade levels in the context of construct validity.

Method
Participants
As this study aims to compare performance on strategy knowledge across grades, we chose a cross-sectional design with three grade levels. In our study, participants were $N = 1,182$ German high school students of Grade Levels 8 ($n_8 = 453$), 10 ($n_{10} = 378$), and the upper secondary level ($n_{12} = 351$, Grade 12), who attended 1 of 61 chemistry classes. The mean age of the entire sample was 15.61 years ($SD = 1.68$) and ranged between 13 and 21 years. Of these students, 48.9% were female. Students worked on computerized versions of tests on strategy knowledge and related constructs.

Measures
Strategy knowledge. The development of the test on knowledge about problem-solving strategies (strategy knowledge) followed a common approach of assessing the construct within the domains of reading, mathematics, and physics (Artelt et al., 2009; Klieme et al., 2010; Neuenhaus et al., 2011; Thillmann, 2007). Due to the domain-specificity of strategy use (e.g., Schukajlow & Leiss, 2011), these tasks contain specific problems or scientific hypotheses as well as various strategies that might be performed to solve the problem or to check the hypotheses. Following the argumentation of Mayer (2007), these tasks refer to metacognitive knowledge about scientific problems and problem-solving procedures. The resulting measure strongly refers to metacognitive knowledge and indicates the degree of the students’ awareness of the best problem-solving methods (Schneider &
Artelt, 2010). In this framework, the construct is part of the declarative component of metacognitive knowledge and knowledge about tasks or contexts (Efklides & Vlachopoulos, 2012). The process of test development was performed in two steps: First, we developed six chemistry problems that contained a specific research question or hypothesis. According to Berry and Dienes (1991), we distinguished between implicit and explicit tasks by varying the surface structure of the stimuli. This concept was transferred to the construct of strategy knowledge and resulted in three tasks that did not contain explicit information on the variables to be controlled (Figure 1). Accordingly, students had to identify relevant variables of the system or experiment first to evaluate possible strategies of implicit tasks. In contrast, explicit information on the number and types of variables was given in three explicit tasks (Figure 2).

These tasks were administered to 20 experts in the field of science education and chemistry. Each one of them was asked to evaluate the strategies on a 6-point Likert-type scale ranging from 1 = highly adequate to 6 = insufficient according to their adequacy in solving the problem. We then analyzed whether the expert ratings significantly differed from each other by transposing the data matrix into a matrix with raters as variables and item ratings as cases. The internal consistency was computed as a measure of intraclass correlation (see Thillmann, 2007). The resulting value of .96 was significant, $F(80, 1520) = 26.94, p < .001$, and indicated a sufficient rater agreement for the entire test. To evaluate the agreement on an item-level, we subsequently checked the agreements on how raters compared pairs of two strategies. More precisely, if more than 70% of the raters evaluated Strategy A better than B, the agreement on this comparison was regarded as sufficient and was used for further analyses (Artelt et al., 2009; Scherer & Tiemann, 2012). Finally, 20 comparisons fulfilled the above mentioned criteria and were used for further analyses. The number of items obtained in this study was reasonable, as compared to Artelt et al.’s (2009) reading strategy test that contained 26 out of 77 comparisons. This procedure of item development accounted for content validity of the underlying measure (Messick, 1995).

Second, the students’ ratings were scored as follows (e.g., Artelt et al., 2009; Artelt, Schiefele, & Schneider, 2001; Thillmann, 2007): If a student rated Strategy A better than B and this relation aligned with the experts’ ratings, then the item of comparing A and B was coded to 1. In the case of equally ranked Strategies A and B, 0.5 points were yielded and 0 else. Regarding this scoring procedure, it seems apparent that items are dependent because the comparisons among strategies of one stimulus are interwoven. For example, rating Strategies A and C might be affected by rating Strategies A and B. Due to these dependencies among items and the varying numbers of items within a task, we used the $z$ transformed sum scores ($SK01$-$SK06$) as six indicators of strategy knowledge. This is a common procedure when dealing with testlets or tasks, which consist of several dependent items (Lakin, 2012; Little, Cunningham, Shahar, & Widaman, 2002; Yen, 1993).

**Covariates of strategy knowledge.** As previous research on covariates of strategy knowledge suggested (Neuenhaus et al., 2011; Phakiti, 2008; Schneider & Artelt, 2010; Van der Stel & Veenman, 2010; Veenman et al., 1997), fluid intelligence, content knowledge, self-concept, and motivational constructs can be regarded as predictors of the construct and could, therefore, be used for an external validation of the measure. Besides these covariates, we additionally assessed grades in chemistry to obtain information on the domain-specificity of strategy knowledge and used interest and enjoyment in science as indicators of motivation.

### Figure 1. Item example of the strategy knowledge test with implicit information on the variables that need to be controlled.

|   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|
| a) | “I dissolve different quantities of chemical substances in varying volumes of different solvents.” |   |   |   |   |   |
| b) | “I extract the substances’ solubility from the Internet. There is no need for an experiment.” |   |   |   |   |   |
| c) | “I dissolve different substances in different solvents and observe the solution.” |   |   |   |   |   |
| d) | “I dissolve varying quantities of a single chemical substance in a defined volume of a single solvent.” |   |   |   |   |   |
| e) | “I dissolve the same quantity of chemical substances in a defined volume of a single solvent.” |   |   |   |   |   |

Sandra hypothesizes that different factors determine the solubility of substances in water. Different procedures occur which might be performed in order to check her hypothesis in a series of experiments. Evaluate these methods with grades ranging from 1 (=very good) to 6 (=not sufficient).
Fluid intelligence. The test on fluid intelligence was based on the figural scale of a cognitive ability test, which significantly loads on the $g$ factor of intelligence (Heller & Perleth, 2000). This test consisted of 45 items, including 15 anchor items for adjacent grade levels and 10 grade-specific items. Each student had to solve 25 figural problems within a time limit of 8 min. The resulting answers were dichotomously scored.

Content knowledge and grades in chemistry. Content knowledge was assessed with 44 multiple-choice items which referred to the contents of the strategy knowledge tasks. We developed one test form for each grade level and linked these forms with 3 to 10 anchor items. Finally, each test contained 20 to 24 dichotomous items. Due to the many test contents, we expected the reliability to be low but sufficient. Furthermore, grades in chemistry were assessed to examine the influence of this performance-based variable on strategy knowledge. This variable ranged from 1 = very good/superior to 6 = insufficient and was handled as an ordinal variable. As implemented in the German curricula, students with high performance in chemistry were assigned to low grades.

Procedure
In our survey, we chose a cross-sectional design to describe and compare subpopulations of different grade levels. This design has the major advantage of assessing students’ strategy knowledge at one time point without any follow-ups and dropouts over time. In our study, we did not aim to capture individual growth patterns, but rather differences between subpopulations. The tests on strategy knowledge and covariates were administered as computerized versions in two sessions of 90 min each. In all tests, students were able to return to previous items and correct their solutions, if necessary. The resulting data were simultaneously logged and finally coded in SPSS 19 (IBM, 2010).

Statistical Analyses
Confirmatory factor analysis (CFA). To check for the structure of the strategy knowledge test, we conducted CFAs in Mplus 6.0 (Muthén & Muthén, 2010). In these models, missing values can be imputed on a model-based level. In this context, the Full Information Maximum Likelihood (FIML) procedure was used (Enders, 2010). In our study, 3.2% of the data were
missing. According to Little’s missing completely at random (MCAR) test that tests the assumption of MCAR against missing at random, our data were more likely to follow MCAR, \( \chi^2(37, N = 1,182) = 37.09, p = .51 \). Therefore, the model-based imputation of the FIML procedure was legitimate.

In our analyses, we used the robust maximum likelihood estimator and evaluated the goodness of fit by taking into account the following statistics: the Satorra–Bentler scale corrected chi-square value (SB-\( \chi^2 \)), the Comparative Fit Index (CFI), the Root Mean Square Error of Approximation (RMSEA), and the Standardized Root Mean Square Residual (SRMR; Brown, 2006). Common guidelines for an acceptable model fit require a nonsignificant SB-\( \chi^2 \) value, a CFI above .90, a RMSEA below .08, and a SRMR below .09 (Hu & Bentler, 1999; Marsh, Hau, & Grayson, 2005). However, we note that the statistical significance of the SB-\( \chi^2 \) value strongly depends on the sample size (Berger & Karabeni, 2011). For further comparisons of competing nested CFA models, we applied \( \chi^2 \)-difference tests with a Satorra–Bentler correction (Bryant & Satorra, 2012).

**Multigroup CFA.** The investigation of measurement invariance was conducted by multigroup CFA, which allows evaluating different types of invariance and yields information on model-based fit criteria (Campbell, Barry, Joe, & Finney, 2011). In this concept, four types of measurement invariance are considered (Hildebrandt, Wilhelm, & Robitzsch, 2009; Lubke, Dolan, Kelderman, & Mellenbergh, 2003): (a) configural invariance, which can be established by freely estimating factor loadings and by assuming the same number of factors and loading patterns across groups; (b) metric invariance, in which equal factor loadings are added to the configural model; (c) scalar invariance, in which equal intercepts are introduced to the metric model; and finally (d) strict invariance, in which equal residual variances are added to the metric model. Again, \( \chi^2 \)-difference tests and goodness of fit indexes can be used to compare these hierarchically ordered models (Byrne & Stewart, 2006; Cheung & Rensvold, 2002). As Brown (2006) suggested, if at least scalar invariance is met, comparisons of means, variances, and covariances can be conducted.

**Latent regression analyses.** To examine convergent and discriminant validity of the strategy knowledge measure, we conducted regression analyses with scientific self-concept, enjoyment in science, interest in science, content knowledge, fluid intelligence, grades in chemistry, and gender as manifest predictors. In this analysis, strategy knowledge was modeled as a latent construct with a single factor. These analyses were based on the best invariance model of the multigroup CFA.

**General comments on latent variable modeling and construct validity.** In the present study, we used a latent variable modeling approach. First, measurement models were specified, which comprised a latent factor and manifest indicators of strategy knowledge. Second, the model’s generalizability across grade levels was investigated by latent multigroup CFA. Third, the relations among covariates and strategy knowledge were used to check for convergent and discriminant validity. These analyses finally served as tools to investigate the test’s construct validity (Campell & Fiske, 1959; Cronbach & Meehl, 1955). In this regard, latent variables were modeled as representatives of unobservable constructs, which are measured by manifest and observable items. This procedure has the major advantage of correcting and controlling for measurement error (Borsboom, Mellenbergh, & Van Heerden, 2003). Furthermore, the latent regression analyses resulted in more precise path coefficients when using the latent measurement model of strategy knowledge. Brown (2006) also argued that using latent models reduces the dimensionality of data because latent factors represent the common variance shared by manifest indicators.

**Results**

In this section, we first present the psychometric properties and descriptive statistics of the tests on strategy knowledge and covariates. Second, we analyze the internal structure of strategy knowledge by means of CFA. Based on the outcomes of these analyses, we check for measurement invariance to justify comparisons across grade levels. Finally, latent regression models are applied with the aim of obtaining information on construct validity of the strategy knowledge measure.

**Descriptive Statistics**

**Strategy knowledge.** As shown in Table 1, the test on strategy knowledge revealed an internal consistency of \( \alpha = .77 \) for the entire sample with item-to-total correlations between .16 and .52. Furthermore, the strategy knowledge scale showed a slight ceiling effect and yielded a mean sum score of 16.29 (SD = 2.93) for the 20 items and the entire sample. The internal consistencies were sufficient and comparable with previous studies on metacognitive knowledge in science (e.g., Liu, 2010; Thillmann, 2007). Further descriptives of the scales’ sum scores are presented in Table 1. Differentiating between grade levels led to acceptable reliabilities between .76 and .81 for the test, which can be regarded as sufficient at this stage of the test development (Artelt et al., 2009; Liu, 2010).

**Fluid intelligence, content knowledge, and grades in chemistry.** Using the unidimensional Rasch model, we found a sufficient reliability of the test on fluid intelligence (WLE reliability = .90) and a lower value for the test on content knowledge (WLE reliability = .61). However, the reliability score of the knowledge test is still arguable for a test that measures knowledge in different content areas (Kalyuga, 2006). For both tests, students’ mean ability scores were transformed to zero. The students’ grades in chemistry were moderate (\( M = 2.65, SD = .89, \text{Median} = 3.00, \text{Minimum} = 1.00, \text{Maximum} = 5.00 \)).
Motivational constructs and self-concept. Based on 4-point Likert-type items, we established scales of scientific self-concept ($M = 9.61$, $SD = 3.16$, five items, Cronbach’s $\alpha = .88$), enjoyment in science ($M = 9.74$, $SD = 3.92$, six items, Cronbach’s $\alpha = .91$), and interest in science ($M = 16.31$, $SD = 5.14$, nine items, Cronbach’s $\alpha = .85$) as motivational and personality traits. Using CFA, we fitted a measurement model with three correlated factors. The resulting model revealed an acceptable fit and was, thus, accepted, SB-$\chi^2(176, N = 1,182) = 807.14$, $p < .001$, CFI = .95, RMSEA = .05, $p$(RMSEA $\leq .05$) = .45, SRMR = .04. In this model, factor loadings ranged from .36 to .69 for interest, from .65 to .83 for enjoyment in science, and from .76 to .83 for scientific self-concept. We, finally, used the sum scores of each of the scales as indicators for the underlying constructs.

Measurement Model of Strategy Knowledge: CFA

To check the structure of strategy knowledge, we established a measurement model with one latent factor for the data of the combined sample. The resulting model is shown in Figure 3. This model revealed an acceptable goodness of fit and represented the data sufficiently (see Table 2).

We then analyzed whether or not this model held across grade levels by conducting the analysis in each grade level separately. Similar to the combined sample, reasonable and acceptable model fits resulted for Grades 8 and 12 (Table 2). However, goodness of fit statistics were poor for Grade Level 10. These might have resulted from less homogeneous response patterns in the strategy knowledge items in Grade 10. But due to the fact that this model can be applied for the combined sample, Grades 8 and 12, we continued with further analyses. As Lakin (2012) simplified the statistical prerequisites of the baseline model by arguing that the measurement model must reveal an acceptable goodness of fit for each subgroup in terms of plausibility, we conclude that the proposed model of strategy knowledge could be accepted as a baseline model.

Measurement Invariance: Multigroup CFA

Having established a baseline model for each grade level and the combined sample, we analyzed different levels of measurement invariance across grades. First, we checked for configural invariance by applying a multigroup CFA model. In this analysis, all factor loadings, residuals, and intercepts were freely estimated. The resulting values of the fully standardized factor loadings across grades are shown in Table 3.

As shown in Table 3, the loadings of the $z$-standardized indicators followed the same pattern in each grade level. The highest loadings of the latent factor were found for items SK01 and SK04. Taken together, the configural invariance model revealed an acceptable fit with similar patterns of factor loadings across grades. Given that configural invariance was supported, we further constrained the factor loadings to equality across grades (metric invariance). Accordingly, the resulting model revealed acceptable fit statistics and was, thus, accepted. We then compared the two nested models of configural and metric invariance by conducting a $\Delta$SB-\$\chi^2$ difference test and found that the metric model was empirically preferred (Table 4). This indicated that the items had equal salience in Grade Levels 8, 10, and 12. Taken together, the factorial structure of the unidimensional measurement model held across grades.

As a further step, scalar invariance was tested. Again, the model revealed a remarkable goodness of fit and sufficiently represented the data (Table 2). Compared with the previous
model, the scalar model was empirically preferred (Table 4). Taken together, the measurement model of strategy knowledge, assuming one latent factor, could be used to compare group means across grades.

Finally, we applied the most restrictive model of strict invariance by constraining residuals to equality across grades. In this model, the assumption was made that the measurement of the underlying construct was not biased and revealed the same reliability and accuracy in each grade level. The resulting fit indexes were poor (Table 2) and indicated that the model was empirically not preferred (Table 4). In light of these results, the scalar model was accepted as the final model. Given the scalar invariance of the unidimensional model, we were able to compare means, variances, and covariances of students’ performance across grade levels (Byrne & Stewart, 2006; Lakin, 2012).

### Comparing Latent Means Across Grade Levels

As another step in evaluating the effects of grade level on strategy knowledge, we analyzed the differences in latent means. In these analyses, we used the scalar model and constrained latent means of one grade level as a reference to zero (Byrne & Stewart, 2006; Van de Schoot, Lugtig, & Hox, 2012). By setting Grade 8 as a reference group, we were able to compare the means of Grades 8 and 10 as well as 8 and 12. The second analysis with Grade 10 as a reference was necessary to compare Grade Levels 10 and 12. Moreover, the resulting measurement model with constrained means showed a very good model fit and was, therefore, accepted, \( SB-\chi^2(47, N = 1,177) = 65.22, p = .17, CFI = .97, RMSEA = .02, p(RMSEA \leq .05) = .99, SRMR = .04 \). To evaluate the practical importance of these differences, we computed Hedges’s \( g \) and transformed the resulting value into the standardized effect size of \( r \) (Steinmetz, Schmidt, Tina-Booh, Wieczorek, & Schwartz, 2009). The results are shown in Table 5.

There were significant differences between Grade Levels 8 and 10, 8 and 12, and 10 and 12 favoring students of higher grades. Effect sizes (\( r \)) were small to moderate and ranged between .16 and .42. The differences in latent means between Grades 10 and 12 were smaller than for Grades 8 and 10 or 8 and 12, respectively. Taken together, our data suggested an increase in strategy knowledge across grades.

### Effects of Covariates on Strategy Knowledge

The analysis of convergent and discriminant validity was conducted by introducing covariates of strategy knowledge to the scalar model for the entire sample. The resulting model represented the data with an acceptable goodness of fit and explained 39.9% of variance (Table 6). The effect size for the latent factor was medium (\( f^2 = .664 \)).

Based on this regression model, we found that content knowledge, fluid intelligence, and grades in chemistry were significant predictors. Further covariates such as gender, interest in science, scientific self-concept, and enjoyment in science did not contribute to the variance explanation in

---

**Table 2. Goodness of Fit Statistics of the Baseline Model and the Multigroup CFA Models.**

| Model                  | n     | SB-\( \chi^2 \) | df   | \( p \) | CFI   | RMSEA | 90% CI     | \( p(RMSEA \leq .05) \) | SRMR |
|-----------------------|-------|-----------------|------|--------|-------|-------|------------|--------------------------|------|
| 1. Baseline model     |       |                 |      |        |       |       |            |                          |      |
| Combined sample       | 1,177 | 31.33           | 9    | <.01   | .95   | .04   | [.02, .06] | .76                       | .03  |
| Grade 8               | 423   | 8.10            | 9    | .54    | 1.00  | .00   | .[00, .05] | .95                       | .02  |
| Grade 10              | 403   | 23.63           | 9    | .02    | .89   | .06   | [.02, .09] | .32                       | .04  |
| Grade 12              | 351   | 14.36           | 9    | .20    | .97   | .03   | [.00, .07] | .72                       | .03  |
| 2. Configural invariance | 1,177 | 46.08           | 27   | .04    | .96   | .04   | [.01, .06] | .84                       | .03  |
| 3. Metric invariance  | 1,177 | 52.75           | 37   | .17    | .98   | .02   | [.00, .04] | .99                       | .04  |
| 4. Scalar invariance  | 1,177 | 65.27           | 47   | .17    | .97   | .02   | [.00, .04] | 1.00                      | .04  |
| 5. Strict invariance  | 1,177 | 215.37          | 59   | <.001  | .67   | .07   | [.06, .08] | .01                       | .10  |

**Note.** 90% CI = 90% confidence interval of the RMSEA. The Satorra–Bentler corrected \( \chi^2 \) values (SB-\( \chi^2 \)) were based on the robust maximum likelihood estimator (Bryant & Satorra, 2012). In this analysis, 5 observations had to be excluded due to a unreasonably high number of missing values.

**Table 3. Factor Loadings of the Configural Invariance Model Across Grades.**

| Item | Grade 8 | Grade 10 | Grade 12 |
|------|---------|----------|----------|
| SK01 | .55 (.07)| .50 (.08)| .51 (.08)|
| SK02 | .28 (.08)| .21 (.08)| .18 (.08)|
| SK03 | .30 (.07)| .23 (.08)| .30 (.08)|
| SK04 | .45 (.08)| .50 (.07)| .61 (.08)|
| SK05 | .36 (.07)| .47 (.08)| .59 (.08)|
| SK06 | .38 (.06)| .45 (.07)| .41 (.08)|

**Note.** The table shows the fully standardized solution. All factor loadings are statistically significant (\( p < .001 \)).

**Table 4. Model Comparisons of Multigroup Confirmatory Factor Analyses Across Grades.**

| Model comparison | \( \Delta SB-\chi^2 \) | \( \Delta df \) | \( p \) | \( |\Delta CFI| \) | Favored model |
|------------------|------------------------|----------------|------|----------------|------|
| Baseline–configural | 13.47                  | 18             | .76  | .01            | Configural |
| Configural–metric | 4.99                   | 10             | .89  | .02            | Metric |
| Metric–scalar     | 11.26                  | 10             | .34  | .01            | Scalar |

**Note.** \( \Delta SB-\chi^2 \) = adjusted difference in Satorra–Bentler scaled \( \chi^2 \) values.
strategy knowledge. In sum, the data suggested that performance-based predictors significantly affected strategy knowledge, whereas personality-related constructs did not show effects.

**Discussion**

To compare students’ performance across subpopulations, the present study investigated whether a unidimensional measurement model of strategy knowledge held across grade levels. In these analyses, the concept of measurement invariance was applied. Furthermore, the relationships between scientific strategy knowledge and covariates were analyzed.

**Assessment of Strategy Knowledge**

We were able to develop a test on strategy knowledge for the domain of chemistry with sufficient internal consistencies across grades. The values of Cronbach’s α for the overall scale were above .70 for each grade and could, thus, be regarded as sufficient at this stage of research. For example, Thillmann (2007) found values between .58 and .85, and Artelt et al. (2009) between .56 and .81 for different domains. As Liu (2010) and Muis, Bendixen, and Haeerle (2006) discussed in their reviews, these values are common for the assessment of metacognitive knowledge. Although factor loadings and item-to-total correlations were comparably low for some items, our findings somehow align with previous test statistics which were derived from the PISA assessments of strategy knowledge in reading ($\alpha = .84$, $r_{it} = .21-.57$) and mathematics ($\alpha = .78$, $r_{it} = .03-.45$) (Klieme et al., 2010). However, at this stage of the test development, we cannot clarify whether or not this was an artifact of statistical or methodologies’ issues (e.g., item parceling). Further studies have to be conducted to improve the measurement accuracy of the strategy knowledge scale. We further discuss these results in light of the difficulty of measuring constructs of implicit metacognition (Flavell et al., 2002; Muis et al., 2006; Reber, 1993). Within implicit tasks, many processes are involved that cannot be assessed by explicit answers on specific tasks. These processes often relate to constructs such as content knowledge (De Jong & Ferguson-Hessler, 1996), epistemological views of science (Liu, 2010; Tsai, Ho, Liang, & Lin, 2011), reading ability (Artelt, Schiefele, & Schneider, 2001), and modes of knowledge representation (Alibali, Phillips, & Fischer, 2009; Dienes & Perner, 1999). The resulting effects on the strategy knowledge scale could have led to weaker statistical properties and interfered with students’ performance. In addition, students’ views of scientific knowledge and methods play an important role while evaluating strategies according to their goodness of fit (Liu, 2010).

The analysis of the structure of the strategy knowledge test was based on a single latent factor, which was measured by implicit and explicit tasks. However, there might be further dimensions of strategy knowledge. As Leutwyler (2009) and Phakiti (2008) suggested, the construct of strategy knowledge could show a higher order structure with different factors across domains and facets of the construct. However, in our study, we were did not specify a multidimensional structure as we hypothesized only one latent trait in a single domain.

At this stage of the proposed research, we are not able to empirically clarify whether or not the construct of strategy knowledge can be regarded as domain-specific or domain-general. Together with Kaberman and Dori (2009), Lingel,

### Table 5. Latent Means of Students’ Strategy Knowledge Across Grades.

| Grade | M    | SE   | p    |
|-------|------|------|------|
| 8     | .00  | .00  | —    |
| 10    | .62  | .12  | <.001|
| 12    | .93  | .15  | <.001|

8 vs. 10

| Hedges’s $g$ | -.62 |
| Pearson’s $r$ | .30  |

8 vs. 12

| Hedges’s $g$ | -.93 |
| Pearson’s $r$ | .42  |

10 vs. 12

| Hedges’s $g$ | -.32 |
| Pearson’s $r$ | .16  |

This group represents the reference group with constrained latent means. Pearson’s $r$ represents the standardized effect size which is based on Hedges’s $g$.

### Table 6. Regression Model of Strategy Knowledge for the Entire Sample.

| Predictors                | $\beta$ (SE) | p    |
|---------------------------|--------------|------|
| Enjoyment in science      | .08 (.05)    | .14  |
| Scientific self-concept   | -.06 (.05)   | .27  |
| Interest in science       | .05 (.05)    | .28  |
| Fluid intelligence        | .45 (.03)    | <.001|
| Content knowledge         | .20 (.04)    | <.001|
| Gender (0 = male)         | .01 (.04)    | .81  |
| Grade in chemistry        | -.18 (.04)   | <.001|

$R^2$ (f$^2$) .399 (.664)

Model fit

| SB-$\chi^2$ (44, N = 1,177) = 100.52, $p < .001$, CFI = .93, RMSEA = .03, SRMR = .02 |

Note. Significant path coefficients are bold (p < .001). f$^2$ represents the effect size, given by: $f^2 = R^2 / (1 - R^2)$.

1$p$(RMSEA ≤ .05) > .05. **p < .01. ***p < .001.
Neuenhaus, Artelt, and Schneider (2010), and Neuenhaus et al. (2011), we argue that there is a domain-specific component. To address this issue, further research should analyze the structure of the construct across further domains (Schneider & Artelt, 2010). In our study, we conceptualized strategy knowledge as a construct, specifically defined and assessed for chemistry. After all, our assumption has been supported by the significant correlations with students’ grades in chemistry, but still requires further investigation. Certainly, further studies on our tasks need to be conducted to validate the test measure and improve its quality. In addition, qualitative approaches could clarify further cognitive and metacognitive processes which are closely related to strategy knowledge. Moreover, it would also be interesting to investigate the relationships with further aspects of metacognition, which was proposed by Efklides and Vlachopoulos (2012).

Our study provided a test on strategy knowledge which can be regarded as objective, reliable, and valid according to the assumptions of classical test theory. Moreover, we combined research on knowledge in scientific problem solving with the theory of declarative metacognition for a specific domain. It, therefore, contributes to the field of research on metacognition in educational settings (Schneider & Artelt, 2010).

**Measurement Invariance Across Grades**

In our study, the results revealed scalar invariance of the measurement model of strategy knowledge across grades. Hence, as the construct holds across grades, we argue that the strategy knowledge test could be used to assess developmental changes of strategy knowledge with increasing grade level (Hildebrandt et al., 2009). From a methodological perspective, we also note that multigroup CFAs are appropriate models for assessing invariance and legitimatize meaningful comparisons of means, variance, and covariances (Sass, 2011; Van de Schoot et al., 2012). In light of the current discussion on how to model learning progressions, we further claim that these procedures gain importance in educational research because they provide evidence whether or not there is a construct shift over time and, thus, contribute to construct validity of test measures (Kölner, & Parchmann, 2012). Our study, therefore, provided an example of how to handle multigroup data to assess differences across grade levels.

**Latent Means Across Grades**

Our analyses also revealed a significant effect of grade level on students’ performance in strategy knowledge favoring upper grades. This finding confirms the results of a study conducted by Short, Schachtsneider, and Friebert (1993), who found that there is a development of strategy knowledge with increasing grade level in math. In addition, our finding aligns with previous longitudinal studies in other domains (Neuenhaus et al., 2011; Schneider, 2010). In light of these results, we argue that students grow in their cognitive awareness of problem-solving strategies.

Our results also show that the effects are smaller for the differences between Grade Levels 10 and 12. This might be due to the heterogeneity of the upper secondary level. In this grade level, students attended basic and advanced courses, which differed in their curricula and could be voluntarily chosen by the students. In addition, students of adjacent age groups also took part in these courses, leading to a more heterogeneous subsample of the upper secondary level. Therefore, grade level differences of one year occurred within this subgroup of participants. Furthermore, we note that our sample was derived from the German high school and was, therefore, quite selective. The effects of grade level on strategy knowledge might be different across school types or federal states. In addition, the selectiveness of the German school system might have influenced the grade level differences. For example, students of the upper secondary level were educated for the transition to the tertiary level, whereas students of the lower secondary level had to take an obligatory education in all subjects. In light of this selectivity, the effect sizes of the differences between students at Grade Level 10 and the upper secondary level could be influenced by further effects of the school system.

**Effects of Covariates on Strategy Knowledge**

All analyses of the relationships between strategy knowledge and covariates revealed correlations below .50 and, thus, indicated that the construct can be distinguished from its covariates. Within the regression models, content knowledge significantly predicted strategy knowledge across grade levels. This finding supports the results of a study conducted by Cromley and Azevedo (2011), who found that the application of domain-specific strategies is determined by students’ background knowledge. Also, the findings of H. Kim and Pedersen (2011) can be supported. We conclude that meaningful comparisons of strategies that refer to scientific problems or hypotheses require a certain level of content knowledge that can be activated. Furthermore, the construct was strongly affected by fluid intelligence. In light of theories on knowledge application and metacognition (De Jong & Ferguson-Hessler, 1996; Flavell, 1979; Neuenhaus et al., 2011), intelligence is a crucial factor that determines the application of (metacognitive) knowledge. Thillmann (2007) supported this argument for the domain of physics. It seems apparent that content knowledge and intelligence are important predictors. In addition, students’ grade in chemistry showed significant effects, whereas motivational constructs did not contribute to strategy knowledge. The latter finding indicates that our measure of strategy knowledge is more determined by cognitive factors (Scherer & Tiemann, 2012). In the context of construct validity, convergent validity was present for cognitive constructs and discriminant validity.
was found for motivational constructs. Taken together, the regression coefficients suggested that strategy knowledge is strongly determined by performance-based variables.

In light of this discussion, we also note that strategy knowledge is quite difficult to assess because it interferes with epistemologies, students’ views of strategies, and content knowledge. But to understand these processes more deeply, further analyses are necessary. Taken together, our findings align with previous research on strategy knowledge (Artelt et al., 2009; Schneider, 2008, 2010) and lead to the conclusion that our test measures a construct which does not only require intelligence and content knowledge, but far greater knowledge and skills.

**Limitations of the Present Study**

Although our data supported the theoretical assumptions and models of strategy knowledge and provided evidence on the differences of the construct across grades, this study has a number of limitations that warrant discussion. First, our results revealed that a unidimensional structure of strategy knowledge was present. However, this result should be interpreted with caution because our sample did not represent the entire German school system across all grade levels. As Rost (2009) argued, the dimensionality of a construct could be due to the selectivity of the sample. Second, further studies across all grade levels are necessary to generalize our findings. A broader range of age should be taken into account in the modeling process, as this might lead to a greater variation in performance, and to more precise measurement models.

Third, further covariates should be taken into account to obtain more precise information on construct validity of the strategy knowledge measure. It might be worth analyzing the effects of domain-specific competencies and domain-general metacognitive abilities. Results on these issues would contribute to the discussion of the domain-specificity of the construct. Fourth, due to the cross-sectional character of this study, our conclusions are limited to the comparison of grade levels. Our results are therefore only indicators for the development of strategy knowledge and refer to *interindividual* changes. Further longitudinal analyses are necessary to identify *intraindividual* growth patterns and to explain causal effects of covariates such as the degree of problem-based learning scenarios in science lessons.

**Conclusion**

Taken together, strategy knowledge remains a measure which is quite difficult to assess. In light of our modeling approach, measurement invariance is given. In general, further research on learning progressions or the development of psychological constructs should take into account invariance as a prerequisite for analyzing differences between groups. But due to the findings of Artelt et al. (2009) and Welsh and Huizinga (2005), which indicated that strategy knowledge does not necessarily imply a meaningful application of knowledge within problem situations, it would now be interesting to check for the relationship among strategy knowledge and its application in problem-solving scenarios. First, small-scale studies revealed moderate relationships (Thillmann, 2007) for physics, but no effects for chemistry (Scherer & Tiemann, 2012). These analyses could reveal significant implications for educational classroom practice and contribute to new teaching strategies or approaches on test development. For instance, studies on fostering the application of strategies in problem-solving situations could take into account that various processes are involved in strategy knowledge. The proposed assessment approach and its connection to manifest indicators could provide an instructional guide for educational assessment.

**Declaration of Conflicting Interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The author(s) received no financial support for the research, authorship, and/or publication of this article.

**References**

Alibali, M. W., Phillips, K. M. O., & Fischer, A. D. (2009). Learning new problem-solving strategies leads to changes in problem representation. *Cognitive Development, 24*, 89-101.

Amsel, E., Klaczynski, P. A., Johnston, A., Bench, S., Close, J., Sadler, E., & Walker, R. (2008). A dual-process account of the development of scientific reasoning: The nature and development of metacognitive intercession skills. *Cognitive Development, 23*, 452-471.

Annevirta, T., & Vauras, M. (2001). Metacognitive knowledge in primary grades: A longitudinal study. *European Journal of Psychology of Education, 16*, 257-282.

Artelt, C., Beinicke, A., Schlamburgler, M., & Schlamburgler, W. (2009). Diagnose von Strategiewissen beim Textverstehen [Diagnosing strategy knowledge in reading comprehension]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 41*, 96-103.

Artelt, C., Schiefele, U., & Schneider, W. (2001). Predictors of reading literacy. *European Journal of Psychology of Education, 16*, 363-383.

Berger, J.-L., & Karabenick, S. (2011). Motivation and students’ use of learning strategies: Evidence of unidirectional effects in mathematics classrooms. *Learning and Instruction, 21*, 416-428.

Berry, D., & Dienes, Z. (1991). The relationship between implicit memory and implicit learning. *British Journal of Psychology, 82*, 359-373.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*, 203-219.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*, 1061-1071.
Bowden, S., Saklofske, D., & Weiss, L. (2011). Invariance of the measurement model underlying the Wechsler Adult Intelligence Scale-IV in the United States and Canada. *Educational and Psychological Measurement, 71*, 186-199.

Brown, T. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford.

Bryant, F., & Satorra, A. (2012). Principles and practice of scaled difference chi-square testing. *Structural Equation Modeling, 19*, 372-398.

Byrne, B., & Stewart, S. (2006). The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling, 13*, 287-321.

Campbell, H., Barry, C., Joe, J., & Finney, S. (2011). Configural, metric, and scalar invariance of the Modified Achievement Goal Questionnaire across African American and White University Students. *Educational and Psychological Assessment, 68*, 988-1007. doi:10.1177/0013164408318766

Campell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validity by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Carr, M., Jessup, D., & Fuller, D. (1999). Gender differences in first-grade mathematics strategy use: Parent and teacher contributions. *Journal of Research in Mathematics Education, 30*, 20-46.

Cheung, G., & Rensvold, R. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233-255.

Cooper, M., & Sandi-Urena, S. (2009). Design and validation of an instrument to assess metacognitive skillfulness in chemistry problem solving. *Journal of Chemical Education, 86*, 240-245.

Cromley, J., & Azevedo, R. (2011). Measuring strategy use in context with multiple-choice items. *Metacognition and Learning, 6*, 155-177.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

De Jong, T., & Ferguson-Hessler, M. (1996). Types and qualities of knowledge. *Educational Psychologist, 31*, 105-113.

Dienes, Z., & Perner, J. (1999). A theory of implicit and explicit knowledge. *Behavioral and Brain Sciences, 22*, 735-755.

Doran, N., Aldridge, A., Roesch, S., & Myers, M. (2011). Factor structure and invariance of the behavioral undercontrol questionnaire. *European Journal of Psychological Assessment, 27*, 145-152.

Efklides, A., & Vlachopoulos, S. (2012). Measurement of metacognitive knowledge of self, task, and strategies in mathematics. *European Journal of Psychological Assessment, 28*, 227-239.

Enders, C. (2010). *Applied missing data analysis*. New York, NY: Guilford.

Flavell, J. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34*, 906-911.

Flavell, J., Miller, P., & Miller, S. (2002). *Cognitive development* (4th ed.). Englewood Cliffs, NJ: Prentice Hall.

Funke, J., & Frensch, P. (2007). Complex problem solving: The European Perspective—ten years after. In D. H. Jonassen (Ed.), *Learning to solve complex scientific problems* (pp. 25-47). Hillsdale, NJ: Lawrence Erlbaum.

Gamo, S., Sander, E., & Richard, J.-F. (2010). Transfer of strategy use by semantic recording in arithmetic problem solving. *Learning and Instruction, 20*, 400-410.

Goode, N., & Beckmann, J. (2010). You need to know: There is a causal relationship between structural knowledge and control performance in complex problem solving tasks. *Intelligence, 38*, 345-352.

Hammack, D. (2005). The role of domain knowledge in higher-level cognition. In O. Wilhelm & R. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 361-372). Thousand Oaks, CA: Sage.

Heller, K., & Perleth, C. (2000). *Kognitiver Fähigkeitsstest für 4-12. Klasslen, Revision (KFT 4-12+R)* [Cognitive ability test for grades 4 to 12]. Göttingen, Germany: Hogrefe.

Hildebrandt, A., Wilhelm, O., & Robitzsch, A. (2009). Complementary and competing factor analytic approaches for the investigation of measurement invariance. *Review of Psychology, 16*, 87-102.

Hoffman, B., & Schraw, G. (2009). The influence of self-efficacy and working memory capacity on problem-solving efficacy. *Learning and Individual Differences, 19*, 91-100.

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

IBM Corporation. (2010). *SPSS Statistics 19*. Armonk, NY: Author.

Kaberman, Z., & Dori, Y. (2009). Metacognition in chemical education: Question posing in the case-based computerized learning environment. *Instructional Science, 37*, 403-436.

Kapa, E. (2007). Transfer from structured to open-ended problem solving in a computerized metacognitive environment. *Learning and Instruction, 17*, 688-707.

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multigroup categorical CFA and IRT. *Structural Equation Modeling, 18*, 212-228.

Kim, H., & Pedersen, S. (2011). Advancing young adolescents’ hypothesis-development performance in a computer-supported and problem-based learning environment. *Computers & Education, 57*, 1780-1789.

Klieme, E., Artelt, C., Hartig, J., Jude, N., Köller, O., Prenzel, M., & . . . Stanat, P. (Eds.). (2010). *PISA 2009*. Münster, Germany: Waxmann.

Kolen, M., & Brennan, R. (2004). *Test equating, scaling, and linking* (2nd ed.). New York, NY: Springer Science+Business Media.

Kolic-Vehovec, S., Bajsanki, I., & Roncevic Zubkovic, B. (2010). Metacognition and reading comprehension: Age and gender differences. In A. Efklides & P. Misailidi (Eds.), *Trends and prospects in metacognition research, Part 2* (pp. 327-344). New York, NY: Springer.

Köller, O., & Parchmann, I. (2012). Competencies: The German Notion of learning outcomes. In S. Bernholt, K. Neumann, & P. Nentwig (Eds.), *Making it tangible: Learning outcomes in science education* (pp. 165-185). Münster, Germany: Waxmann.

Kuh, D. (2000). Metacognitive development. *Current Directions in Psychological Science, 9*, 178-181.

Kuhn, D., Iordanou, K., Pease, M., & Wirkala, C. (2008). Beyond control of variables: What needs to develop to achieve skilled scientific thinking? *Cognitive Development, 23*, 435-451.

Kuhn, D., & Pearsall, S. (1998). Relations between metastrategic knowledge and strategic performance. *Cognitive Development, 13*, 227-247.

Künzting, J., Wirth, J., & Paas, F. (2011). The goal specificity effect on strategy use and instructional efficiency during computer-
based scientific discovery learning. *Computers & Education*, 56, 668-679.

Lakin, J. (2012). Multidimensional ability tests and culturally and linguistically diverse students: Evidence of measurement invariance. *Learning and Individual Differences*, 22, 397-403.

Lemaire, P., & Lecacheur, M. (2011). Age-related changes in children’s executive functions and strategy selection: A study in computational estimation. *Cognitive Development*, 26, 282-294.

Leutwyler, B. (2009). Metacognitive learning strategies: Differential developmental patterns in high school. *Metacognition and Learning*, 4, 111-123.

Lingel, K., Neuenhaus, N., Artelt, C., & Schneider, W. (2010). Metakognitives Wissen in der Sekundarstufe: Konstruktion und Evaluation domänenspezifischer Messverfahren [Metacognitive knowledge in secondary education: Developing and evaluating domain-specific assessments]. In E. Klieme, D. Leutner, & M. Kenk (Eds.), *Kompetenzmodellierung* (pp. 228-238). Weinheim, Germany: Beltz.

Little, T., Cunningham, W., Shahar, G., & Widaman, K. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling*, 9, 151-173.

Liu, X. (2010). *Using and developing measurement instruments in science education*. Charlotte, NC: Information Age Publishing.

Lubke, G., Dolan, C., Kelderman, H., & Mellenbergh, G. (2003). Weak measurement invariance with respect to unmeasured variables: An implication of strict factorial invariance. *British Journal of Mathematical and Statistical Psychology*, 56, 231-248.

Marsh, H., Hau, K.-T., & Grayson, D. (2005). Goodness of fit in structural equation models. In A. Maydeu-Olivares & J. McArdle (Eds.), *Contemporary psychometrics* (pp. 275-340). Mahwah, NJ: Lawrence Erlbaum.

Martin, A. (2009). Motivation and engagement across the academic life span. A developmental construct validity study of elementary school, high school, and university/college students. *Educational and Psychological Measurement*, 69, 794-824.

Mayer, J. (2007). Erkenntnisgewinnung als wissenschaftliches Handeln: Zur Darstellung einer mathematisch-naturwissenschaftlichen Problemlosung [Inquiry as scientific problem solving]. In D. Krüger & H. Vogt (Eds.), *Theorien biologiedidaktischer Forschung* (pp. 177-186). Berlin, Germany: Springer.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, 50, 741-749.

Muis, K., Bendixen, L., & H aerle, F. (2006). Domain-generality and domain-specificity in personal epistemology research: Philosophical and empirical reflections in the development of a theoretical framework. *Educational Psychology Review*, 18, 3-54.

Muthén, L., & Muthén, B. (2010). Mplus 6.0 [Computer software]. Los Angeles, CA: Author.

Neuenhaus, N., Artelt, C., Lingel, K., & Schneider, W. (2011). Fifth graders’ metacognitive knowledge: General or domain-specific? *European Journal of Psychology of Education*, 26, 163-178.

Novick, L. R., & Bassok, M. (2005). Problem solving. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 321-349). New York, NY: Cambridge University Press.

Organisation for Economic Co-Operation and Development. (2009). *PISA 2006: Technical report*. Paris, France: Author.

Panaoura, A., & Philippou, G. (2007). The developmental change of young pupils’ metacognitive ability in mathematics in relation to their cognitive abilities. *Cognitive Development*, 22, 149-164.

Phakiti, A. (2008). Strategic competence as a fourth-order factor model: A structural equation modeling approach. *Language Assessment Quarterly*, 5, 20-42.

Reber, A. S. (1993). Implicit learning and tacit knowledge: An essay on the cognitive unconscious. Oxford, UK: Oxford University Press.

Rickey, D., & Stacy, A. (2000). The role of metacognition in learning chemistry. *Journal of Chemical Education*, 77, 915-920.

Rost, D. (2009). *Intelligenz [Intelligence]*. Weinheim, Germany: BeltzPVU.

Sass, D. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment*, 29, 347-363.

Scherer, R., & Tiemann, R. (2012). Factors of problem-solving competency in a virtual chemistry environment: The role of metacognitive knowledge about strategies. *Computers & Education*, 59, 1199-1214.

Schneider, W. (2008). The development of metacognitive knowledge in children and adolescents: Major trends and implications for education. *Mind, Brain, and Education*, 2, 114-121.

Schneider, W. (2010). The development of metacognitive competencies. In B. Glatzeder, V. Goel, & A. Müller (Eds.), *Towards a theory on thinking: Building blocks for a conceptual framework* (pp. 203-214). Berlin, Germany: Springer.

Schneider, W., & Artelt, C. (2010). Metacognition and mathematics education. *ZDM Mathematics Education*, 42, 149-161.

Schukajlow, S., & Leiss, D. (2011). Selbstberichtete Strategienutzung und mathematische Modellierungskompetenz [Self-reported strategy use and mathematical modeling competency]. *Journal für Mathematik-Didaktik*, 32, 53-77.

Shahat, M., Ohle, A., Treagust, D., & Fischer, H. (2013). Design, development and validation of a model of problem solving for Egyptian classes. *International Journal of Science and Mathematics Education*, 11, 1157-1181.

Short, E., Schatschneider, C., & Friebert, S. (1993). Relationship between memory and metamemory performance: A comparison of specific and general strategy knowledge. *Journal of Educational Psychology*, 85, 412-423.

Solaz-Portolés, J., & Sanjosé López, V. (2008). Types of knowledge and their relations to problem solving in science: Directions for practice. *Sisifo. Educational Sciences Journal*, 6, 105-112.

Steinmetz, H., Schmidt, P., Tino-Booh, A.,Wieczorek, S., & Schwartz, S. H. (2009). Testing measurement invariance using multigroup CFA: Differences between educational groups in human values measurement. *Quality & Quantity*, 43, 599-613.

Taaasoobshirazi, G., & Glynn, S. (2009). College students solving chemistry problems: A theoretical model of expertise. *Journal of Research in Science Teaching*, 46, 1070-1089.

Thillmann, H. (2007). *Selbstreguliertes Lernen durch Experimentieren* [Self-regulated learning by experimentation] (Doctoral dissertation). Universität Duisburg-Essen, Germany. Retrieved from http://duepublico.uni-duisburg-essen.de/servlets/ DeriveServlet/ Derivate-18970/Dissertation_Thillmann_online-Version.pdf

Thomas, G., Anderson, D., & Nashon, S. (2008). Development of an instrument designed to investigate elements of science students’ metacognition, self-efficacy and learning processes: The SEMLI-S. *International Journal of Science Education*, 30, 1701-1724.
Tsai, C.-C., Ho, H. N. J., Liang, J.-C., & Lin, H.-M. (2011). Scientific epistemic beliefs, conceptions of learning science and self-efficacy of learning science among high school students. *Learning and Instruction, 21*, 757-769.

Van der Stel, M., & Veenman, M. (2010). Development of metacognitive skillfulness: A longitudinal study. *Learning and Individual Differences, 20*, 220-224.

Van de Schoot, R., Lughtig, P., & Hox, J. (2012). Developmetrics: A checklist for testing measurement invariance. *European Journal of Developmental Psychology, 9*, 486-492.

Van Kraayenoord, C., & Schneider, W. (1999). Reading achievement, metacognition, reading self-concept and interest: A study of German students in grades 3 and 4. *European Journal of Psychology of Education, 14*, 305-324.

Veenman, M., Elshout, J., & Meijer, J. (1997). The generality versus domain-specificity of metacognitive skills in novice learning across domains. *Learning and Instruction, 7*, 187-209.

Velayutham, S., Aldridge, J., & Fraser, B. (2011). Development and validation of an instrument to measure students’ motivation and self-regulation in science learning. *International Journal of Science Education, 33*, 2159-2179.

Wang, S., & Jiao, H. (2009). Construct equivalence across grades in a vertical scale for a K-12 large-scale reading assessment. *Educational and Psychological Measurement, 69*, 760-777.

Welsh, M., & Huizinga, M. (2005). Tower of Hanoi disk-transfer task: Influences of strategy knowledge and learning on performance. *Learning and Individual Differences, 15*, 283-298.

Wu, M., Adams, R., Wilson, M., & Haldane, S. (2007). *ConQuest 2.0: Generalised item response modelling software*. Camberwell: Australian Council for Education Research.

Yen, W. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213.

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review, 27*, 172-223.

Author Biographies

**Ronny Scherer** is a researcher who is now working at the Centre for Educational Measurement (CEMO) at the University of Oslo, Norway. His interdisciplinary research combines STEM education, computer-based assessment, and educational measurement in the context of large-scale studies.

**Rüdiger Tiemann** is a Professor of Chemistry Education at the Humboldt-Universität zu Berlin in Germany. He is working on teacher education, problem solving, and inquiry processes in science.