Journal for Educational Research Online
Journal für Bildungsforschung Online
Volume 4 (2012), No. 2, 120–142

Eva Susanne Fritzsche, Stephan Kröner, Markus Dresel,
Bärbel Kopp & Sabine Martschinke

Confidence scores as measures of metacognitive monitoring in primary students? (Limited) validity in predicting academic achievement and the mediating role of self-concept

Abstract

Stankov and Lee (2008) have shown confidence scores to have unique predictive effects on achievement criteria when the corresponding test scores are controlled. These findings suggest that confidence scores might provide valid indicators of metacognitive monitoring. However, as confidence is related to self-concept (Kröner & Biermann, 2007), it is possible that the unique predictive effects disappear when self-concept is also controlled. This study examines whether average confidence regarding performance on the items of a spelling test showed incremental validity above and beyond the corresponding test scores in predicting academic achievement with and without control for verbal self-concept. N = 414 10-year-olds from 22 fourth grade classes in Bavarian primary schools participated in the research project. Students' confidence ratings were shown to correlate with corresponding test scores. Furthermore, when test scores were controlled, average confidence scores had unique predictive effects on academic achievement. When both test scores and self-concept were controlled, however, there was no substantial unique predictive effect of confidence. Thus, the predictive power of primary students' confidence ratings may result from their correlation with self-concept.

Dr. Eva Susanne Fritzsche (corresponding author) · Prof. Dr. Stephan Kröner, Chair of Empirical Educational Sciences, Friedrich-Alexander-University Erlangen-Nuremberg, Regensburger Str. 160, 90478 Nuremberg, Germany
e-mail: eva.fritzsche@ewf.uni-erlangen.de
stephan.kroener@ewf.uni-erlangen.de

Prof. Dr. Markus Dresel, Chair of Psychology, Augsburg University, Universitätsstraße 10, 86135 Augsburg, Germany
e-mail: markus.dresel@phil.uni-augsburg.de

Prof. Dr. Bärbel Kopp · Prof. Dr. Sabine Martschinke, Institute of Primary School Research, Friedrich-Alexander-University Erlangen-Nuremberg, Regensburger Str. 160, 90478 Nuremberg, Germany
e-mail: baerbel.kopp@ewf.uni-erlangen.de
sabine.martschinke@ewf.uni-erlangen.de

Keywords
Metacognition; Confidence ratings; Self-concept; Primary students

Antwortsicherheiten als Maß für die metakognitive Überwachung bei Grundschulkindern? (Eingeschränkte) Validität bei der Vorhersage schulischer Leistungen und die mediierende Rolle des Selbstkonzepts

Zusammenfassung
Wie Stankov und Lee (2008) zeigten, wiesen Antwortsicherheiten inkrementelle Validität in Bezug auf externe Leistungskriterien auch nach Kontrolle der Leistung auf, anhand der die Antwortsicherheiten erhoben wurden. Dies lässt sich so interpretieren, dass Antwortsicherheiten valide Indikatoren für die metakognitive Überwachung sind. Da die Antwortsicherheiten aber mit dem Selbstkonzept korreliert sind (Kröner & Biermann, 2007), könnte ihre inkrementelle Validität verschwinden, wenn zusätzlich Effekte des Selbstkonzepts kontrolliert werden. Dies wurde in der vorliegenden Studie untersucht, indem aggregierte Antwortsicherheiten und Testleistungen in einem Rechtschreibtest als Prädiktoren schulischer Leistungen verwendet wurden, und zwar mit und ohne Kontrolle des Selbstkonzepts im Fach Deutsch. Es nahmen N = 414 zehnjährige Kinder aus 22 vierten Klassen bayerischer Grundschulen an dieser Studie teil. Es zeigte sich, dass auch bei den untersuchten Grundschulkindern Antwortsicherheiten mit der jeweiligen Testleistung korrelierten. Außerdem wiesen sie inkrementelle Validität in Bezug auf schulische Leistungen auf, und zwar auch bei Kontrolle der Rechtschreibleistung. Wenn jedoch darüber hinaus Effekte des Selbstkonzepts kontrolliert wurden, verschwand dieser Effekt. Dies deutet darauf hin, dass der Erklärungswert von Antwortsicherheiten für externe Leistungskriterien auf ihre Korrelation mit dem Selbstkonzept zurückgeht.

Schlagworte
Metakognition; Antwortsicherheiten; Selbstkonzept; Grundschulkinder

1. Confidence in the context of procedural metacognition

"How confident are you that your answer is correct?" This is a prototypical question for the assessment of response confidence, or the subjective probability of having solved a task or a test item correctly. The present study investigates whether primary students' confidence scores provide valid indicators of metacognitive monitoring and can thus be used as a measure of procedural metacognition.

1.1 A model of procedural metacognition

In the domain of metacognition, we can distinguish between declarative and procedural metacognition (Flavell, Miller, & Miller, 2001). The declarative aspect of metacognition concerns the awareness of the difficulty of a given task and the necessity of expending more effort in learning difficult items than in learning easy ones, for example. The procedural aspect concerns competencies that are necessary for regulation and control of one's learning processes during self-regulated learning. In self-regulated learning, learners first need to set a goal. They then have to monitor the learning process, to evaluate their observations and, finally, to regulate the next stage of the learning process (Bandura, 1991, 2001). Adequate monitoring and evaluation of one's performance is crucial for the successful regulation of learning processes and for positive learning outcomes. However, the monitoring and evaluation components of Bandura's taxonomy have proved difficult to dissociate empirically and are combined as metacognitive monitoring in the study of Koriat, Ma'ayan and Nussinson (2006). In his model of procedural metacognition based on the model of Nelson and Narens (1990, 1994), Schneider (2008) therefore distinguished only two components: a monitoring component that combines monitoring and evaluation and a self-regulation component equivalent to Bandura's (1991, 2001) regulation component.

Several approaches have been used for the assessment of procedural metacognition in prior research: primarily global self-report measures including questionnaires, think-aloud protocols and interviews (see, e.g., Desoete, 2008), which have only occasionally been combined with more fine-grained techniques (see, e.g., Veenman, Wilhelm, & Beishuizen, 2004). Unfortunately, these widely used self-report methods have certain drawbacks concerning, for example, their validity in tapping metacognitive monitoring (Artelt, 2000a; Kröner & Fritzsche, 2012; Spörer & Brunstein, 2006; Veenman, 2005). In addition to the aforementioned global self-report measures, there is, however, an alternative approach to the assessment of the metacognitive monitoring component in applying techniques which are both fine-grained and economic: namely, obtaining confidence ratings for each item of a cognitive test (Schraw, 2009).

1.2 Approaches to confidence as an indicator of metacognitive monitoring

Confidence as the subjective probability of having solved a task or a test item correctly can be assessed in a variety of contexts and age groups. The power of confidence scores to predict performance on measures tapping both cognitive abilities and metacognition is an important topic in two research contexts described by Stankov and Lee (2008). In one of these contexts – the "ecological" or "Brunswikian" context – item-specific confidence ratings are analyzed on the response level, with a focus on experimental and environmental conditions such as

item selection or item difficulty (Gigerenzer, Hoffrage, & Kleinbölting, 1991; Juslin, 1994; Juslin & Olsson, 1997). The findings of these studies indicate that participants are principally – although not perfectly – able to monitor their achievement on cognitive tests and to express the result of the monitoring process in confidence ratings.

In the present study, however, we adopt the "person-centered" or "Thurstonian" perspective introduced by Stankov and Lee (2008). Research taking this perspective focuses on aggregated confidence scores, which are viewed as general indicators of metacognitive monitoring closely related to learners' dispositions (Diehl, Semegon, & Schwarzer, 2006; Kleitman & Stankov, 2007; Schraw, 2009; Schraw & Dennison, 1994). For example, Stankov and colleagues have described confidence "as an aspect of a metacognitive process of self-monitoring that should be thought of as residing somewhere on the borderline between personality and intelligence" (Stankov, 2000, p. 141). On the same lines, Zimmerman (1990) has emphasized that self-regulated learning depends heavily on learners' dispositions. Effects of task characteristics notwithstanding, one important group of dispositions that may be related to confidence scores are person variables such as self-efficacy beliefs or self-concept which reflect learners' subjective estimations of their abilities (Lee & Bobko, 1994). However, in contrast to self-efficacy and self-concept, which reflect general assessments of the individual's ability to regulate the process of learning and problem solving, confidence scores relate directly to the solution of a specific task. Consequently, there is a certain conceptual overlap between confidence and other dispositional estimations of one's abilities. Self-concept, in particular, has been used in the person-centered or Thurstonian approach to confidence research (Kröner & Biermann, 2007; Stankov & Crawford, 1997), with Kröner and Biermann finding substantial latent correlations between confidence and self-concept.

1.3 Empirical findings on confidence as an indicator of procedural metacognition

Following the Brunswikian approach, confidence scores have been widely used in the context of eye-witness testimony, showing that while item-specific confidence ratings do contain valid information, they are sometimes unreliable (Allwood, Ask, & Granhag, 2005; Allwood, Granhag, & Jonsson, 2006; Allwood, Knutsson, & Granhag, 2006). Roebers and colleagues have conducted several Brunswikian studies about confidence as an indicator of metacognitive monitoring, primarily in children (Howie & Roebers, 2007; Krebs & Roebers, 2010, 2012; Roderer & Roebers, 2010; Roebers, 2002; Roebers & Howie, 2003; Roebers, Schmid, & Roderer, 2009). These studies are introduced in a later section of this paper (see section 2.2).

Confidence scores have also been used as indicators of procedural metacognition following the Thurstonian approach. However, although this seems to be a promising approach (see also Schraw, 2009), there has been little empirical inves-

tigation to date of what exactly is measured by person confidence scores or of how confidence scores relate to other dispositional indicators of procedural metacognition (e.g., self-concept). One notable exception is Kleitman and Stankov's (2007) study which examined the relationship between confidence scores and two self-report measures of procedural metacognition. They found that metacognitive inventories explained unique variance in confidence scores when cognitive abilities were controlled.

In a recent study, Stankov and Lee (2008) took another interesting approach to investigating the relationship between confidence and procedural metacognition. Starting with the idea that achievement criteria can be seen as indicators not only of cognitive abilities, but also of metacognition, they investigated whether the confidence scores of college students showed incremental validity above and beyond the corresponding test scores in predicting performance in various achievement tests. They found that average confidence scores had a small but statistically significant predictive effect above and beyond the corresponding test scores. Stankov and Lee (2008) conceded that the practical importance of their findings was limited, but suggested that average confidence scores might nevertheless be relevant predictors of other criteria of educational success, such as dropout rates or times to completion in graduate schools.

In addition to its limited practical significance, the Stankov and Lee (2008) study has a further drawback: when unique effects of a variable are interpreted as evidence of "a new dimension in individual differences [...] that is related to, but distinct from both personality and ability traits" (p. 976), it is vital to carefully control for known traits. However, Stankov and Lee controlled only for test scores. They did not control for self-concept as a potentially important confounding variable. As Kröner and Biermann (2007) have shown, however, confidence scores and self-concept are substantially related, with bivariate correlations of up to $r = .40$ between self-concept of abilities and confidence scores in different cognitive tests. One goal of the present study was therefore to investigate whether the unique predictive effects of confidence scores on achievement persist when self-concept is controlled.

2. Procedural metacognition and confidence in primary students

2.1 The development of procedural metacognition in primary students

Although many studies have investigated response confidence in adult samples, knowledge of the validity of confidence scores provided by children still developing procedural metacognition is limited. It is possible that researchers have been discouraged by the results of studies showing deficits in the procedural metacognition

of children (Kron-Sperl, Schneider, & Hasselhorn, 2008) including production and utilization deficiency (Flavell et al., 2001). Nevertheless, Panaoura and Philippou's (2007) results as well as those of Veenman et al. (2004) underline that, although metacognitive abilities are still developing during primary school years, primary students already use metacognitive strategies (cf. also Desoete, 2008; Panaoura & Philippou, 2007; Roebers et al., 2009). In addition, Veenman et al. (2004) have shown that a general factor of metacognitive skills has incremental validity above and beyond intellectual abilities when predicting learning outcomes in primary students. In accordance with the findings that self-regulation skills are still under development during primary school years, Koriat, Ackerman, Lockl, and Schneider (2009) recently observed no substantial correlations between study time and judgments of learning (JOL) for students from grades one and two, but found such correlations for students from grades three to six. A further problem relates to the finding that in most cases, young children are overconfident when asked to judge their own achievement (Lipko, Dunlosky, & Merriman, 2009), a finding also displayed in overly optimistic self-concepts in primary school (Helmke, 1998). Taking together these findings, given the promising results on procedural metacognition at least in older primary students, it still seems worthwhile to look more closely at the development of both economic and fine-grained techniques to assess metacognitive abilities of primary students.

2.2 Confidence as an indicator of procedural metacognition in primary students

Is it possible to apply confidence as an indicator of procedural metacognition even in primary students? There are some studies providing empirical findings relating to this question: in these studies, confidence ratings were used as an indicator of metacognitive monitoring in a cognitive test (see, e.g., Howie & Roebers, 2007; Krebs & Roebers, 2010, 2012; Pressley, Levin, Ghatala, & Ahmad, 1987; Roderer & Roebers, 2010; Roebers, 2002; Roebers et al., 2009). In accordance with studies of production deficiency (Flavell et al., 2001), Roebers and colleagues (2009) found that children do not start to effectively translate the results of monitoring processes into self-regulation interventions until they are at least 11 years old. However, they found that even children as young as nine display considerable monitoring skills, as indicated by their ability to discriminate between correct and incorrect responses and to express this accurately in terms of response confidence. Lockl and Schneider (2003) had similar results when analyzing judgments-of-learning and study time: Even first graders were able to discriminate between easy and difficult items, but they did not use metacognitive judgments as intensively as third graders do when regulating study time. In a recent study with 8-year-old children, Krebs and Roebers (2010) replicated the finding of adequate monitoring processes and also found evidence for effective control processes in this age group. Roderer and Roebers (2010) also used a multi-method approach to metacognitive monitor-

ing, combining the assessment of confidence judgments with an eye-tracking approach. Both measures indicated improvements in monitoring competencies as the children became older, thus generating evidence regarding validity of confidence ratings as indicators of metacognitive monitoring.

Taken together, the reported findings on confidence in primary students provide initial empirical evidence for the validity of confidence ratings as indicators of metacognitive monitoring in younger children. However, most of the studies cited focused on confidence at item level, and did not investigate the psychometric properties of aggregated confidence ratings as measures of dispositional metacognitive monitoring. Such aggregated confidence ratings have potential advantages: they may replace questionnaires on metacognitive monitoring which are of questionable validity even in older children (cf. Artelt, 2000b) since confidence ratings have a close temporal link to the cognitive and metacognitive activities of the task-solution process and at the same time include instructions which are easy to understand. Overall, confidence may be a promising indicator of metacognitive monitoring in primary students. In addition, confidence ratings may provide a means of examining the effects of personality on procedural metacognition.

As has already been mentioned, there are two possible approaches to the analysis of confidence ratings. Following the Brunswikian approach, a number of researchers have compared confidence scores on correctly and incorrectly solved items (e.g., Howie & Roebers, 2007; Roebers et al., 2009). This comparison gives insight into the calibration of subjects: the higher confidence scores on correctly solved items are in comparison to confidence scores on incorrectly solved items, the better the person is calibrated and the better are the person's monitoring skills. Following the Thurstonian approach, the relationship between mean confidence and achievement scores in the related tests are investigated as indicators of calibration or "realism" (Allwood, Innes-Ker, Homgren, & Fredin, 2008). Mean confidence can also be used to investigate whether it shows incremental validity beyond external achievement criteria in predicting academic achievement (see, e.g., Stankov & Lee, 2008). As far as primary students are concerned, Roebers et al. (2009) – following the Brunswikian approach – have shown that their item-specific confidence ratings are valid indicators of metacognitive monitoring. However, it remains unclear whether – following the Thurstonian approach – Stankov and Lee's (2008) findings regarding confidence scores as indicators of metacognition as a trait can also be replicated with primary students. The present study thus set out to extend the scientific knowledge both of the construct validity of confidence scores of primary students and of their potential application as a dispositional measure of metacognitive monitoring. A measure of metacognitive monitoring would be especially valuable in evaluating training interventions in procedural metacognition. Such training interventions could be applied as early as at the end of primary school. Confidence scores are both easy to understand and efficient to collect. Nevertheless, assessments of confidence scores need to be adapted somewhat for use with primary students, as outlined below.

2.3 Assessment of confidence in primary students

Several methods have been used to assess confidence in previous studies with adults. Participants have been asked to rate their degree of confidence in various response formats – for example, in free format as a percentage (Stankov & Crawford, 1997), on a Likert scale (e.g., Gigerenzer et al., 1991), or on a visual analogous scale (e.g., Schraw & Dennison, 1994). In terms of confidence scaling, participants in some studies have been asked for a number between guessing probability (i.e., 50 % if there are two response options) and 100 % confidence (e.g., Gigerenzer et al., 1991). Other studies have used verbal anchor labels such as "very unconfident/guessed" and "absolutely confident" (e.g., Kulhavy & Stock, 1989).

In comparison to the plethora of studies on confidence in adults, studies applying confidence ratings to children are quite rare. We were able to identify some studies regarding realism of confidence scores in the context of eyewitness memory (Allwood, Granhag, et al., 2006; Allwood et al., 2008; Allwood, Jonsson, & Granhag, 2005), a study regarding the factorial structure of simultaneously analyzed confidence ratings and correctness of related achievement items (Kleitman & Moscrop, 2010), and studies focusing on metacognitive monitoring and self-regulation processes following metacognitive monitoring (Howie & Roebers, 2007; Krebs & Roebers, 2010, 2012; Pressley et al., 1987; Roderer & Roebers, 2010; Roebers, 2002; Roebers & Howie, 2003; Roebers et al., 2009; Roebers, von der Linden, & Howie, 2007). Taken together, these studies show that confidence ratings on various scales with various anchors can be successfully applied to children as young as eight years old (Allwood et al., 2008). When assessing confidence scores in children, verbal or pictorial anchors are usually applied because young children cannot be expected to understand percentage scores. Our work builds on the studies cited, focusing on the predictive validity of confidence scores with regards to external achievement criteria (cf. Roebers et al., 2009; Stankov & Lee, 2008) while controlling for related self-concept scores (cf. Kröner & Biermann, 2007).

3. Goals of the present study

The present study takes up the idea that confidence scores reflect dispositional aspects of procedural metacognition and thus may have a unique predictive effect on external achievement criteria (Stankov & Lee, 2008). In view of previous findings on the correlation between average confidence scores and self-concept (Kröner & Biermann, 2007), however, we expected this unique effect to disappear when self-concept was included as an additional control variable. In addition, we focused on primary students, because research concerning the validity of aggregated confidence scores in this age group is rare.

4. Research questions and hypotheses

(1) Are aggregated confidence scores valid predictors of metacognitive monitoring as a disposition in primary students? We expected confidence scores to show unique predictive effects on academic achievement as a criterion reflecting both cognitive ability and metacognition when related test scores were controlled.
(2) Do aggregated confidence scores still show unique predictive effects on academic achievement when self-concept is added to the model? We expected the unique effect to be absent in this context.

5. Method

5.1 Participants

Participants were N = 414 grade 4 students from 22 classes in Bavarian primary schools located in urban districts. The mean age of subjects (55 % girls) was M = 10.26 years (SD = 0.55). The proportion of students with an immigrant background was 40 %, the Highest International Socio-economic Index of Occupational Status (HISEI, Ganzeboom, De Graaf, Treiman, & De Leeuw, 1992) in the families of the children was on average 46.5 (SD = 16.9). The data analyzed in the present paper come from a larger research project on student motivation and learning outcomes.

5.2 Procedure

All students completed self-report questionnaires on two measurement occasions, in spring and summer 2008. Most of the data used to address the present research questions were collected at the second occasion. On the first measurement occasion, data on other research questions (primarily regarding classroom influences on student motivation) and background data were collected. Data collection took place during regular class time and was conducted by trained research assistants.

5.3 Variables and instruments

5.3.1 Criterion: Academic achievement

We used academic achievement, operationalized as the German grade received on the last report card, as the criterion variable. Students were asked to state this grade in the self-report questionnaire. This self-report approach has been shown to be a valid method of collecting data on school grades (Dickhäuser & Plenter,

2005). The grades awarded in German primary schools range from 1 (highest) to 6 (lowest). The grades in German were collected on the first measurement occasion; intraclass correlation (ICC) for this variable was ρ = .28.

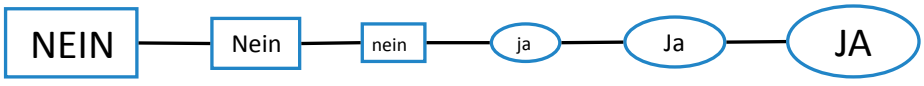
5.3.2 Predictor 1: Spelling score

Spelling was assessed using the Hamburg Spelling Test ("Hamburger Schreib-Probe 4-5", May, 2002; α = .82 [in the present sample]). In the present study, the first of two parts, covering 16 words, was administered to save testing time and to adjust demands to the capabilities and the motivation of the students involved in the study. The spelling score was collected on the second measurement occasion, ICC was ρ = .13.

5.3.3 Predictor 2: Average confidence score

Confidence scores (α = .91) were aggregated from confidence ratings on each of the HSP 4-5 items. In spite of the results of Allwood et al. (2006) suggesting the robustness of questionnaire design in relation to bias of confidence ratings, we decided to use a different design aimed at minimizing participants' extraneous cognitive load (Sweller, 1994) for the participants. Thus, we combined a six-alternative multiple-choice scale with verbal anchor labels instead of confidence ratings as percentages or on a visual analogous scale. In addition, rather than using complex anchor labels, we chose the easiest distinction we could think of: "yes" versus "no". Accordingly, we changed the prompt for the confidence ratings from "How confident are you that your answer is correct?" to "Do you think that you spelled the word correctly?". Three response options were labeled "no" and "yes", respectively, with different font sizes being used to represent a large "no", a medium "no" and a small "no", followed by a small "yes", a medium "yes" and a large "yes". Participants were instructed that the options labeled with largest font size represented the "strongest no" and "strongest yes", respectively. Responses were coded from 1 for the strongest "no" to 6 for the strongest "yes". The scale is depicted in Figure 1. The confidence scores were collected on the second measurement occasion; ICC was ρ = 01.

Figure 1: Rating scale for confidence scores



5.3.4 Predictor 3: Verbal self-concept

We used a scale developed by Martschinke, Kammermeyer, Frank, and Mahrhofer (2002) to measure verbal self-concept (α = .83 [in the present sample]). The scale included nine multiple-choice items on reading, spelling and text production, each with four response options. A pair of sentences such as "In reading, I am among the most skilled students in my class" versus "In reading, I am among the least skilled students in my class" served as scale anchors for each item. The four response options were inserted between the anchors. Children had to choose the response option that best represented their assessment of their abilities. Verbal self-concept was collected on the second measurement occasion; ICC was ρ = .04.

6. Results

6.1 Descriptive statistics and bivariate correlations

Descriptive statistics are displayed in Table 1. HSP spelling scores were fairly high, with an average of more than 12 of 16 items spelled correctly. As expected for students of this age, verbal self-concept was also quite positive. The mean German grade received on the last report card was approximately grade "3" (*M* = 2.88; *SD* = 1.00). Aggregated confidence scores reflected that, on average, the students believed they had spelled most of the HSP items correctly (*M* = 5.36; *SD* = 0.67). In adult samples, mean confidence scores are usually compared to mean achievement scores from the test which was used to assess confidence, in order to specify if the persons are biased in terms of confidence (Stankov & Crawford, 1997). In the present study, transformation of confidence scores to a numerical format is rather difficult. However, the response format enabled us to dichotomize confidence ratings at a naturally occurring threshold: between those response options labeled "yes" and those labeled "no". We were thus able to compare the proportion of "yes" confidence ratings (*M* = 0.92; *SD* = 0.15) with the proportion of correct responses (*M* = 0.80; *SD* = 0.19). As far as can be determined from the difference between "yes" confidence ratings and the proportion of correct responses (0.92 - 0.80 = 0.12), this result indicates that, in general, the participating students were overconfident. That is to say, in comparison with related achievement scores, students are too confident about their answers in the spelling test.

Table 1: Descriptive Statistics

Variables	*M*	*SD*	Min [a]	Max [a]
German grade	2.88	1.00	1	6
Spelling score	12.77	3.01	0	16
Verbal self-concept	2.97	0.54	1	4
Confidence score	5.36	0.67	1	6

Note. [a] Values for Min and Max refer to the lowest and highest possible values, not to the lowest and highest value in the present sample.

Table 2 presents the bivariate correlations of the variables used. There were statistically significant substantial relationships between all variables. Aggregated confidence correlates most highly with verbal self-concept and this correlation is descriptively even higher than the one between spelling score and confidence. Note that the correlations with school grades are negative because of the way grades are interpreted in Germany, with small values representing high achievement and high values representing low achievement.

Table 2: Bivariate Correlations

Variables	Spelling score	Verbal self-concept	Confidence score
German grade	-.52	-.50	-.21
Spelling score		.41	.35
Verbal self-concept			.44

Note. All correlations significant at the $p < .01$ level.

6.2 Hierarchical Linear Models

To address our research questions, we ran several hierarchical linear models using Mplus 5.0 to allow for the hierarchical structure of the sample (Muthén & Muthén, 2007). All predictor variables were z-standardized prior to the analyses. Academic achievement, as reflected by the German grade received on the most recent report card, was used as the criterion variable in all analyses.

First, we ran an unconditional model (model 1) without predictor variables to calculate the intraclass correlation (ICC) for school grades and to provide a baseline for comparison with the subsequent, more complex, models. We then ran four models in which the following predictors were included: spelling score (model 2); spelling score and average confidence score (model 3); spelling score and verbal self-concept (model 4); and spelling score, verbal self-concept and average confidence score (model 5). Results of the hierarchical linear models are shown in Table 3.

Model 1, the unconditional model, revealed the aforementioned ICC of $\rho = .28$ ($p < .001$) for German grade, indicating that 28 % of the variance in our criterion of academic achievement was due to differences between classes.

In Model 2, the spelling score was used to predict academic achievement. Model 2 fitted the data statistically significantly better than the unconditional model, χ^2 (1, $N = 414$) 1087.52 - 976.43 = 111.09, $p < .01$ (Hox, 2002). With only spelling score as a predictor, the model accounted for 27 % of the variance within classes and 37 % of the variance between classes (Snijders & Bosker, 1999, p. 102). Students with a spelling score one standard deviation above the mean had German grades that were, on average, about half a grade higher than those of students with average spelling score.

In Model 3, the average confidence score was added as a second predictor. Results indicate that the confidence score had unique predictive effects on the criterion variable above and beyond the corresponding spelling score. As a chi-square test showed, inclusion of the confidence score further improved the model fit, χ^2 (1, N = 414) 976.43 - 972.67 = 3.76, p = .05. Thus, in line with our first hypothesis, the results of Stankov and Lee (2008) can be generalized to apply to primary students. However, as in the Stankov and Lee (2008) study, the unique predictive effect of the confidence score was quite small, amounting to only 1 % within classes and 3 % between classes. Thus, although statistically significant, the incremental validity of confidence scores in predicting performance in a spelling test beyond the corresponding test scores can be seen as practically insignificant. Students with confidence scores one standard deviation above the mean had German grades that were, on average, only .02 higher than those of students with average confidence scores. In relation to the scaling of school grades, differences of this magnitude have no practical significance.

In Model 4, verbal self-concept replaced the confidence score as a predictor. Again, inclusion of this predictor statistically significantly improved model fit, χ^2 (1, N = 414) 976.43 - 918.13 = 58.30, p < .01. Adding self-concept also increased the amount of variance explained (by 13 % within classes and by 12 % between classes). The regression coefficients revealed that this effect was also practically significant: Students with a self-concept one standard deviation above the mean had school grades that were, on average, .31 higher than those of students with an average verbal self-concept.

We next evaluated the unique predictive effects of confidence beyond both spelling score and verbal self-concept. To this end, we included all three predictor variables in model 5 and compared the results with those of model 4, which included only self-concept and spelling score. Inclusion of the confidence score did not improve overall model fit, χ^2 (1, N = 414) 918.13 - 916.13 = 2.00, p = .16. Thus, the small amounts of additionally explained variance (1 % within classes and 2 % between classes) proved to be practically insignificant: The German grades of students with confidence scores one standard deviation above the sample mean were only .01 higher than those of students with confidence scores equivalent to the sample mean. Overall, confidence scores did not make a relevant contribution to the model as a whole beyond spelling score and self-concept. These findings are in line with our second hypothesis, which predicted that the effects of confidence scores would disappear when self-concept was controlled.

Table 3: Results of Hierarchical Linear Models with grade in German as criterion and spelling score, confidence score and verbal self-concept as predictor variables

Criterion: Grade in German (unstandardized)	Model 1 (Unconditional model)			Model 2 (Including spelling score [a])			Model 3 (Including spelling score [a] and confidence score [a])			Model 4 (Including spelling score [a] and verbal self-concept [a])			Model 5 (Including all predictor variables [a])		
Fixed effects	Coeff.	(*SE*)	*t*-ratio												
Intercept	2.94	(0.12)	24.27	2.93	(0.10)	0.44	2.93	(0.10)	30.87	2.93	(0.09)	32.35	2.93	(0.09)	32.64
Spelling score				-0.47	(0.05)	-9.22	-0.46	(0.05)	-9.20	-0.33	(0.05)	-7.38	-0.33	(0.05)	-7.38
Confidence score							-0.02	(0.00)	-10.84				-0.01	(0.00)	-2.84
Verbal self-concept										-0.31	(0.04)	-8.65	-0.31	(0.04)	-8.34
Random effects	Coeff.	(*SE*)	*t*-ratio												
Within level	0.72	(0.07)	10.66	0.56	(0.05)	12.28	0.56	(0.05)	2.31	0.49	(0.04)	11.89	0.48	(0.04)	12.05
Between level	0.29	(0.06)	4.69	0.18	(0.05)	3.66	0.17	(0.05)	3.76	0.16	(0.04)	3.90	0.15	(0.04)	4.01

Note. [a] All predictor variables were z-standardized before computing any models.

7. Discussion

7.1 Main results

The present study pursued two main goals: First, we explored whether Stankov and Lee's (2008) findings on the validity of confidence scores as a dispositional indicator of procedural metacognition following the Thurstonian approach can be generalized to apply to primary students. Our results replicated those of Stankov and Lee, indicating that confidence scores have unique predictive effects on academic achievement when related achievement data are controlled. However, as in the Stankov and Lee study, the effects proved to be minimal and of limited practical relevance. In general, these findings are in line with our first hypothesis.

Secondly, since confidence scores and self-concept are known to be related (Kröner & Biermann, 2007), we tested whether unique predictive effects of confidence scores on academic achievement were still observable when related aspects of self-concept were controlled. There are two aspects to our results concerning this research question. On the one hand, in line with Stankov and Lee (2008), our hierarchical linear analyses revealed that – when both self-concept and the corresponding test scores were controlled – confidence scores still had a statistically significant unique effect on academic achievement. On the other hand, this effect accounted for only 1 % of the variance within classes and 2 % of the variance between classes. A deviance test of model fit also showed that there was no statistically significant improvement in model fit when confidence scores were added to the model while controlling for both related test scores and self-concept. Related test scores and self-concept are, perhaps, quite a strong criterion to act as a control for confidence scores. However, as we were interested in effects of confidence scores other than self-concept, this was the most straightforward test of our second hypothesis. The results of the present study are, by and large, in line with this hypothesis.

7.2 Implications for confidence as an indicator of metacognitive monitoring

As stated above, our findings on the practical significance of confidence scores as predictors of academic achievement are in line with those of Stankov and Lee (2008), who reported a relatively small unique predictive effect of confidence scores on various achievement criteria and admitted that the "practical importance of this finding is minimal" (p. 972). However, they argued that confidence scores could be useful with respect to other criteria in the educational setting, such as in selecting students or planning interventions. Although the present findings do not rule this possibility out, they give little reason for optimism – at least for primary students as participants and for school grades as the criterion and analyzed following the Thurstonian approach. Nevertheless, it is possible that confidence scores

may prove useful for other groups of participants, for other educational criteria or for other types of analysis, especially as their incremental validity may have been difficult to detect within the present approach, as discussed below.

Apart from the Thurstonian approach, it would also be possible to analyze data following the Brunswikian approach. Thus, confidence judgments might be a valuable tool in assessing task-related metacognitive differentiation between correct and incorrect items or in analyzing confidence scores at item-specific level.

7.3 Implications for the applicability of confidence ratings to primary students

In general, since confidence judgments may be viewed as instances of metacognitive self-evaluation, our findings following the Thurstonian approach on the correlation of confidence and related achievement scores can be interpreted as renewed evidence for effective metacognition in primary students (cf. Desoete, 2008; Veenman, 2005); our results also replicate the results of Veenman et al. (2004), who showed that indicators for metacognitive abilities have unique explanatory value for achievement criteria in participants as young as primary students. These results might have been made possible due to measures we applied to ensure that the primary students in the present study fully understood the confidence assessment task. Like Roebers et al. (2009), we refrained from asking the participating students for confidence ratings as percentages (e.g., Juslin, 1994). Instead, we asked "Do you think that you spelled the word correctly?" and provided a confidence scale with the simple verbal anchors "yes" and "no".

Nevertheless, the ceiling effects found for confidence ratings in the present study might raise doubts as to whether the method of confidence scaling applied could be improved even further for future studies. With regard to confidence scaling, one could also argue that students might have avoided choosing the strongest "no" on the answering scale, because it might imply to be completely sure, that the answer is wrong. However, when instructing students how to apply the response scale for confidence scores, they were told to use the strongest "no" if they were uncertain whether the preceding word was spelled correctly. So instructions should have facilitated the subjects to be able to select the strongest "no". There are, too, other (and probably better) explanations for the ceiling effects. Specifically, the test on which the confidence ratings were based was rather easy for the participating students: on average, they solved more than 12 of 16 items correctly. High confidence ratings were therefore to be expected from children with at least basic metacognitive monitoring competencies. There would probably have been more variance between item-specific confidence scores if the test had been more difficult. To avoid ceiling effects of confidence ratings in the future, it seems as though it would be more effective to change the spelling test on which confidence ratings were based: by choosing more difficult items, for example, or by asking younger participants.

In the present study, we followed the Thurstonian approach and examined confidence scores as aggregated scores over the whole spelling test. However, it might be interesting to follow the Brunswikian approach and calculate aggregated confidence scores on correctly versus incorrectly solved items separately (Howie & Roebers, 2007; Roebers et al., 2009) and to investigate their contribution in explaining external indicators for metacognitive monitoring. Unfortunately, due to ceiling effects in the spelling test, such analyses were not viable in our study, because for many students, there were only few incorrectly solved items, leading to poor reliability of confidence scores related to incorrectly solved items. However, examining the difference between mean confidence on correctly versus incorrectly solved items as an indicator of metacognitive monitoring would be a valuable goal for further studies. Another perspective for future studies to be pursued after solving the ceiling effects in the spelling test would be to follow the Brunswikian approach and examine confidence ratings at the item level. This would provide an opportunity to examine whether the inter-personal correlations of confidence ratings and related item scores in the present study can be replicated within persons as well.

As in many studies on confidence, the students in the present sample were overconfident: the mean proportion of correct responses in the spelling test was somewhat lower than the proportion of items they believed to have answered correctly. From a pedagogical perspective, however, this is a desired outcome. Indeed, a self-concept of abilities that is "positively realistic" is thought to foster learning motivation and effort during primary schooling (Helmke, 1992), and teachers try to enhance students' self-concepts in their educational practice. Therefore, the expectation of research on self-regulated learning, that confidence scores should correspond exactly to achievement measures, should perhaps be relaxed for research in educational contexts, where slight overconfidence is often preferable to exact calibration.

Despite our somewhat disappointing findings regarding the power of primary students' confidence scores to explain their educational outcomes, we are reluctant to pass a verdict on the value of confidence ratings obtained from primary students on the basis of the present results alone: It is important to bear in mind that these students are still acquiring monitoring competencies (Schneider, 2008) and are probably not able to use the information resulting from monitoring as efficiently as adults for effective self-regulation. Therefore, the relationship between monitoring skills and positive learning outcomes in children are probably weaker than those observed in adults. Thus, perhaps even the most valid measure of metacognitive monitoring would not be able to predict learning outcomes in primary students much better. As a next step, therefore, the convergent and discriminant validity of confidence scores vis-à-vis other measures of metacognition should be assessed.

7.4 The relationship between confidence scores, self-concept, school grades and test performance

Besides the main results, the present study also reveals results about of the relationship between primary students' confidence scores, self-concept, school grades and test performance. Thus, our study extends to 10-year-olds the findings of Kröner and Biermann (2007) on the relationship of confidence scores and self-concept in adults. A further finding from studies with adults which we were able to replicate with children in the present sample is the correlation of confidence scores to related test performance. Although our 10-year-olds were generally more confident than their test performance demonstrated, confidence scores and test performance were substantially correlated with each other.

Concerning self-concept and school grades, there was evidence for incremental effects of verbal self-concept on grades in the subject German when controlling for performance in the spelling test. Thus, verbal self-concept can explain further variance above and beyond performance in the spelling test, implying that self-concept can account for variance in external achievement criteria. This is in line with published studies regarding self-concept as an indicator of achievement scores (see e.g., Guay, Marsh, & Boivin, 2003). The finding underlines the importance of self-concept as probable means for interventions in relation to improving students' achievement.

That the effects of confidence scores have been much smaller than the effects of a self-concept scale might also be due to differences between self-concept and confidence scores: For confidence scores, students probably strongly relied on the items of the spelling test, whereas for rating their self-concept, they would also have included results of social comparisons within classrooms. Furthermore, being more general might be an additional feature of self-concept as compared with confidence being responsible for a closer relationship of self-concept to grades, which in turn facilitates an explanation of the variance in grades.

7.5 Limitations of the present study

The most important limitation of the present study is the skewed distribution of confidence scores leading to restricted variance and probably leading also to less explanatory value of the external achievement criteria when compared to the corresponding test scores. The most obvious reason for this distribution would be the primary students' above average spelling achievement in the present sample. Having spelled more than 12 out of 16 words correctly should in fact naturally lead to high confidence ratings. This explanation is supported by the substantial correlation between correctness and confidence. Nevertheless, an alternative explanation for this result might be the answering scale for confidence scores: it seems possible that the anchor "no" was quite strong. Thus, it is possible that the children preferred to select one of the "yes"-answers. As outlined in the method section, to

avoid any potential problems with confidence ratings on a percent scale among primary students, we had changed the answering scale to a verbal one with the anchors "yes" and "no" as the easiest distinction we could think of. Nevertheless, it seems worthwhile to compare the confidence response scale from our study to others used in the literature. For example, Roebers et al. (2009) applied Likert-scales with sad and happy smiley emoticons as anchors.

A further limitation of the present study relates to the use of self-reported school grades as a measure of academic achievement. As Dickhäuser and Plenter (2005) could show, if self-reported grades are used, this causes no additional problem regarding reliability as compared to teacher-reported grades. However, the psychometric properties of school grades in general are somewhat controversial: they are known to be reliable and valid indicators of achievement differences within classes, but not between classes. Thus, the 28 % variance between classes for grades in German in the unconditional model can more probably be traced back to differences between teachers rather than to differences between classes in actual performance. Since we applied multilevel modeling in the present study, variance in grades was separated into variance between classes and variance within classes. As our focus was on explaining variance within classes, missing comparability of school grades between classes should not be an important concern in the present study. However, this probably does not explain the minimal effects of confidence scores found in our study, since other predictors (such as spelling score) worked reasonably well as predictors of school grades.

A further issue related to the use of school grades as a criterion is whether and to what extent students with high procedural metacognition benefit from this competence in terms of higher achievement and better grades. Procedural metacognition skills may not be equally important for the different subcomponents of achievement in German (reading, writing stories, and spelling). For example, the monitoring component seems necessary for checking spelling when writing stories. Effective procedural metacognition is probably not particularly beneficial for reading aloud, however. Thus, certain subcomponents of German grades are probably not equally influenced by metacognitive skills. An alternative explanation for confidence scores not being found to have unique predictive effects is common method bias: a closer look at the results of Stankov and Lee (2008) reveals that, in their study as well, there was no consistent evidence for the incremental validity of confidence scores in predicting achievement criteria that were unrelated to confidence and performance measures based on the TOEFL iBT exam.

A further limitation of both the present study and the Stankov and Lee (2008) study is that competencies were assessed in a single domain: that of language. It might plausibly be argued that procedural metacognition is less useful in the domain of language than elsewhere. However, as Veenman et al. (2004) stated that metacognitive skills are a general rather than a domain-specific skill, results of the present study should be transferable to further domains as well. In order to test this assumption, further research including measures of cognitive abilities and metacognition in other domains is warranted.

8. Conclusion

The present results extend the findings of Stankov and Lee (2008) to primary students, showing that confidence scores have unique, but minor, predictive effects on achievement criteria when test performance is controlled. However, in replication of Kröner and Biermann (2007), the present study also provided renewed evidence that these effects may result from the correlation of confidence scores with self-concept. Taken as a whole, results of the present study provide few grounds for optimism that confidence scores can be used as dispositional indicators of metacognitive monitoring other than for self-concept when used as aggregates scores in the Thurstonian perspective – at least for primary students. However, further studies addressing the limitations of the present study (skewed distribution of confidence scores; reliability of school grades; validity of predictors of metacognitive monitoring; focus on the domain of language; analysis following the Thurstonian approach) are needed before it is possible to draw definitive conclusions on the psychometric properties of confidence scores.

References

Allwood, C. M., Ask, K., & Granhag, P. A. (2005). The cognitive interview: Effects on the realism in witnesses' confidence in their free recall. *Psychology, Crime and Law, 11* (2), 183–198. doi: 10.1080/1068316051233132994 3

Allwood, C. M., Granhag, P. A., & Jonsson, A.-C. (2006). Child witnesses' metamemory realism. *Scandinavian Journal of Psychology, 47* (6), 461–470. doi: 10.1111/j.1467-9450.2006.00530.x

Allwood, C. M., Innes-Ker, Å. H., Homgren, J., & Fredin, G. (2008). Children's and adults' realism in their event-recall confidence in responses to free recall and focused questions. *Psychology, Crime and Law, 14*, 529–547. doi: 10.1080/10683160801961231

Allwood, C. M., Jonsson, A.-C., & Granhag, P. A. (2005). The effects of source and type of feedback on child witnesses' metamemory accuracy. *Applied Cognitive Psychology, 19* (3), 331–344. doi: 10.1002/acp.1071

Allwood, C. M., Knutsson, J., & Granhag, P. A. (2006). Eyewitnesses under influence: How feedback affects the realism in confidence judgements. *Psychology, Crime and Law, 12* (1), 25–38. doi: 10.1080/10683160512331316316

Artelt, C. (2000a). *Strategisches Lernen* [Strategic learning]. Münster: Waxmann.

Artelt, C. (2000b). Wie prädiktiv sind retrospektive Selbstberichte über den Gebrauch von Lernstrategien für strategisches Lernen? [How predictive are self-reported strategies for their actual use?] *Zeitschrift für Pädagogische Psychologie, 14* (2–3), 72–84. doi: 10.1024//1010-0652.14.23.72

Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes, 50*, 248–281.

Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual Reviews of Psychology, 52*, 1–26. doi: 10.1146/annurev.psych.52.1.1

Desoete, A. (2008). Multi-method assessment of metacognitive skills in elementary school children: how you test is what you get. *Metacognition and Learning, 3*, 189–206. doi: 10.1007/s11409-008-9026-0

Dickhäuser, O., & Plenter, I. (2005). „Letztes Halbjahr stand ich zwei". Zur Akkuratheit selbst berichteter Noten [On the accuracy of self-reported school grades]. *Zeitschrift für Pädagogische Psychologie, 19*, 219–224. doi: 10.1024/1010-0652.19.4.219

Diehl, M., Semegon, A. B., & Schwarzer, R. (2006). Assessing attention control in goal pursuit: A component of dispositional self-regulation. *Journal of Personality Assessment, 86*, 306–317. doi: 10.1207/s15327752jpa8603_06

Flavell, J. H., Miller, P. H., & Miller, S. A. (2001). *Cognitive Development*. Upper Saddle River, NJ: Prentice Hall.

Ganzeboom, H. B. G., De Graaf, P. M., Treiman, D. J., & De Leeuw, J. (1992). A standard international socio-economic index of occupational status. *Social Science Research, 21* (1), 1–56.

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

Guay, F., Marsh, H. W., & Boivin, M. (2003). Academic self-concept and academic achievement: Developmental perspectives on their causal ordering. *Journal of Educational Psychology, 95* (1), 124–136. doi: 10.1037//0022-0663.95.1.124

Helmke, A. (1992). *Selbstvertrauen und schulische Leistungen* [Self-confidence and academic achievement]. Göttingen, Germany: Hogrefe.

Helmke, A. (1998). Vom Optimisten zum Realisten? Zur Entwicklung des Fähigkeitsselbstkonzeptes vom Kindergarten bis zur 6. Klassenstufe [From optimist to realist? Development of the academic self-concept from preschoolers up to sixth grade]. In F. E. Weinert (Ed.), *Entwicklung im Kindesalter* (pp. 117–132). Weinheim: Psychologie Verlags Union.

Howie, P., & Roebers, C. M. (2007). Developmental progression in the confidence-accuracy relationship in event recall: insights provided by a calibration perspective. *Applied Cognitive Psychology, 21* (7), 871–893. doi: 10.1002/acp.1302

Hox, J. J. (2002). *Multilevel analysis. Techniques and applications*. New York, NY: Psychology Press.

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Decision Processes, 57*, 226–246. doi: 10.1006/obhd.1994.1013

Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review, 104*, 344–366. doi: 10.1037/0033-295X.104.2.344

Kleitman, S., & Moscrop, T. (2010). Self-confidence and academic achievement in primary-school children: Their relationships and links to parental bonds, intelligence, age, and gender. In A. Efklides & P. Misailidi (Eds.), *Trends and prospects in metacognition research* (pp. 293–326). New York, NY: Springer. doi: 10.1007/978-1-4419-6546-2_14

Kleitman, S., & Stankov, L. (2007). Self-confidence and metacognitive processes. *Learning and Individual Differences, 17* (2), 161–173. doi: 10.1016/j.lindif.2007.03.004

Koriat, A., Ackerman, R., Lockl, K., & Schneider, W. (2009). The memorizing effort heuristic in judgments of learning: A developmental perspective. *Journal of Experimental Child Psychology, 102*, 265–279. doi: 10.1016/j.jecp.2008.10.005

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General, 135*, 36–69. doi: 10.1037/0096-3445.135.1.36

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

Krebs, S. S., & Roebers, C. M. (2012). The impact of retrieval processes, age, general achievement level, and test scoring scheme for children's metacognitive monitoring and controlling. *Metacognition and Learning, 7* (2), 75–90. doi: 10.1007/s11409-011-9079-3

Kron-Sperl, V., Schneider, W., & Hasselhorn, M. (2008). The development and effectiveness of memory strategies in kindergarten and elementary school: Findings from the Würzburg and Göttingen longitudinal memory studies. *Cognitive Development, 23* (1), 79–104. doi: 10.1016/j.cogdev.2007.08.011

Kröner, S., & Biermann, A. (2007). The relationship between confidence and self-concept – Towards a model of response confidence. *Intelligence, 35*, 580–590. doi: 10.1016/j.intell.2006.09.009

Kröner, S., & Fritzsche, E. S. (2012). Wer riskiert eine Mahnung? Zur Validierung der NEO-PI-R-Gewissenhaftigkeitsskala und ihrer Facetten mit Hilfe von Verhaltensspuren für dispositionelle Selbstregulation [Who risks getting a written reminder? Examining the validity of the NEO-PI-R conscientiousness scale and its facets to assess dispositional self-regulation from behavioral traces]. *Diagnostica, 58* (4), 169–181. doi: 10.1026/0012-1924/a000059

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review, 1* (4), 279–308. doi: 10.1007/BF01320096

Lee, C., & Bobko, P. (1994). Self-efficacy beliefs: Comparison of five measures. *Journal of Applied Psychology, 79* (3), 364–369. doi: 10.1037/0021-9010.79.3.364

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

Lockl, K., & Schneider, W. (2003). Metakognitive Überwachungs- und Selbstkontrollprozesse bei der Lernzeiteinteilung von Kindern [Metacognitive monitoring and self-control processes for children's allocation of study time]. *Zeitschrift für Pädagogische Psychologie, 17* (3/4), 173–183. doi: 10.1024//1010-0652.17.3.173

Martschinke, S., Kammermeyer, G., Frank, A., & Mahrhofer, C. (2002). *Heterogenität im Anfangsunterricht – Welche Voraussetzungen bringen Schulanfänger mit und wie gehen Lehrerinnen damit um?* [Heterogeneity in early instruction: What characterizes students at school entry and how do teachers deal with it?] (Vol. 101). Nürnberg: Institut für Grundschulforschung der Universität Erlangen-Nürnberg.

May, P. (2002). *Hamburger Schreib-Probe 1-9 (HSP)* [Hamburg spelling test 1-9 (HSP)]. Göttingen: Hogrefe.

Muthén, L. K., & Muthén, B. O. (2007). *Mplus (Version 5.0)*. Los Angeles, CA: Muthén & Muthén.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation. Advances in research and theory* (Vol. 26, pp. 127–173). San Diego, CA: Academic Press, Inc.

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition. Knowing about knowing* (pp. 1–25). Cambridge: MIT Press.

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. doi: 10.1016/j.cogdev.2006.08.004

Pressley, M., Levin, J. R., Ghatala, E. S., & Ahmad, M. (1987). Test monitoring in young grade school children. *Journal of Experimental Child Psychology, 43* (1), 96–111.

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

Roebers, C. M. (2002). Confidence judgments in children's and adults' event recall and suggestibility. *Developmental Psychology, 38* (6), 1052–1067. doi: 10.1037//0012-1649.38.6.1052

Roebers, C. M., & Howie, P. (2003). Confidence judgments in event recall: Developmental progression in the impact of question format. *Journal of Experimental Child Psychology, 85* (4), 352–371. doi: 10.1016/S0022-0965(03)00076-6

Roebers, C. M., Schmid, C., & Roderer, T. (2009). Metacognitive monitoring and control processes involved in primary school children's test performance. *British Journal of Educational Psychology, 79*, 749–767. doi: 10.1348/978185409x429842

Roebers, C. M., von der Linden, N., & Howie, P. (2007). Favourable and unfavourable conditions for children's confidence judgments. *British Journal of Developmental Psychology, 25* (1), 109–134. doi: 10.1348/026151006x104392

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. doi: 10.1111/j.1751-228X.2008.00041.x

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning, 4* (1), 33–45. doi: 10.1007/s11409-008-9031-3

Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology, 19* (4), 460–475. doi: 10.1006/ceps.1994.1033

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modelling*. London: Sage Publications.

Spörer, N., & Brunstein, J. C. (2006). Erfassung selbstregulierten Lernens mit Selbstberichtsverfahren [Assessing self-regulated learning with self-report measures: A state-of-the-art review]. *Zeitschrift für Pädagogische Psychologie, 20* (3), 147–160. doi: 10.1024/1010-0652.20.3.147

Stankov, L. (2000). Complexity, metacognition, and fluid intelligence. *Intelligence, 28* (2), 121–143. doi: 10.1016/S0160-2896(99)00033-1

Stankov, L., & Crawford, J. D. (1997). Self-confidence and performance on tests of cognitive abilities. *Intelligence, 25*, 93–109. doi: 10.1016/S0160-2896(97)90047-7

Stankov, L., & Lee, J. (2008). Confidence and cognitive test performance. *Journal of Educational Psychology, 100*, 961–976. doi: 10.1037/a0012546

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*, 295–312. doi: 10.1016/0959-4752(94)90003-5

Veenman, M. V. J. (2005). The assessment of metacognitive skills: What can be learned from multi-method designs? In C. Artelt & B. Moschner (Eds.), *Lernstrategien und Metakognition* [Learning strategies and metacognition] (pp. 77–99). Münster: Waxmann.

Veenman, M. V. J., Wilhelm, P., & Beishuizen, J. J. (2004). The relation between intellectual and metacognitive skills from a developmental perspective. *Learning and Instruction, 14*, 89–109. doi: 10.1016/j.learninstruc.2003.10.004

Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist, 25*, 3–17. doi: 10.1207/s15326985ep2501_2