

Sociocultural Aspects of Metacognitive Monitoring

Inauguraldissertation

der Philosophisch-humanwissenschaftlichen Fakultät

der Universität Bern

zur Erlangung der Doktorwürde

vorgelegt von

Florian Jonas Bühler

Aeschi bei Spiez (BE)

Begutachtung:

Prof. Dr. Claudia. M. Roebbers und Prof. Dr. Simona Ghetti

Selbstverlag, Bern im Januar 2023

Originaldokument gespeichert auf dem Webserver der Universitätsbibliothek Bern



Dieses Werk ist unter einer Creative Commons Lizenz vom Typ Namensnennung 4.0 International zugänglich. Um eine Kopie dieser Lizenz einzusehen, konsultieren Sie <http://creativecommons.org/licenses/by/4.0/> oder wenden Sie sich brieflich an Creative Commons, Postfach 1866, Mountain View, California, 94042, USA.

Von der Philosophisch-Humanwissenschaftlichen Fakultät der Universität Bern auf Antrag von Prof. Dr. Claudia Roebers und Prof. Dr. Simona Ghetti angenommen.

Bern, den 15.03.2023

Der Dekan Prof. Dr. Stefan Troche

Acknowledgments

First and foremost, I would like to thank Claudia Roebbers for giving me the opportunity for a Ph.D. in her lab. It has been an incredibly instructive, exciting, and varied journey. Your dedication, social competencies, and humor will always be a great example for me. Next, I would like to thank Simona Ghetti, for supervising me during my six months research stay at UC Davis. From the first day, you made me part of your lab, and I very much appreciate the critical and detailed feedback you provided me. This being said, I also want to thank my wonderful team members and friends in Bern and Davis. Your support during countless coffee breaks (and a few after-work beers) made my Ph.D. experience even more fun and inspiring. A special thanks go to Sophie Wacker, Sonja Kälin, Niamh Oeri, and Oliver Bur for sharing their critical thoughts on my dissertation. Also, Oliver Bur and our fancy Italian coffee machine were great social support during lonely COVID times. Finally, I want to thank my parents, Martin and Marianne, and my sister Louisa for their unconditional support and love. You give me the courage to explore all facets of life, whether in Biel, Bern, Utrecht, Davis, or elsewhere on the globe. Thank you all for contributing to my dissertation in various ways!

The cumulative dissertation includes the following three studies:

Study 1

Buehler, F. J., van Loon, M. H., Bayard, N. S., Steiner, M., & Roebbers, C. M. (2021). Comparing metacognitive monitoring between native and non-native speaking primary school students. *Metacognition and Learning, 16*(3). <https://doi.org/10.1007/s11409-021-09261-z>

Study 2

Buehler, F. J., Orth, U., Krauss, S., Roebbers, C. M. (2022). The Longitudinal Relation between Language Abilities and Metacognitive Monitoring: Structural Differences in Native and Non-native Speakers [Manuscript under review].

Study 3

Buehler, F. J., Ghetti, S., Roebbers, C. M. (2022). Training Primary School Children's Uncertainty Monitoring [Manuscript to be submitted].

Umbrella Paper

Sociocultural Aspects of Metacognitive Monitoring

Florian Jonas Bühler

University of Bern

Abstract

Metacognitive monitoring (the ability to introspect and evaluate cognitive activities and processes) is crucial for children's self-regulated learning and academic achievement. While previous work has primarily assessed metacognitive monitoring in WEIRD (Western, Educated, Industrialized, Rich, and Democratic) populations, research with non-WEIRD populations is sparse. This is problematic as it limits the generalizability of the findings to a minority of the world population. Thus, the primary goal of the present umbrella paper is to highlight sociocultural aspects of children's metacognitive monitoring. Based on the multifaceted and multilevel model of metacognition (Efklides, 2008) and the cultural origins hypothesis (Heyes et al., 2020), I explored three research projects (Studies 1, 2, and 3) through a sociocultural lens. Study 1 revealed that native and non-native speakers do not differ in their metacognitive monitoring in memory and text comprehension tasks, which might suggest that native and non-native speakers share a highly similar sociocultural context for learning (e.g., schools). Study 2 found that native speakers' first language abilities in kindergarten predict metacognitive monitoring in grade one. Conversely, non-native speakers' overconfidence in kindergarten predicted their second language abilities in grade one. Study 3 revealed that metacognitive feedback benefits first graders' metacognitive monitoring. Taken together, our results suggest no cross-cultural differences between native and non-native speakers' metacognitive monitoring, and language and feedback as sociocultural features explaining within-cultural variance in children's metacognitive monitoring. However, more cross-cultural and within-cultural research is needed to clarify the role of sociocultural aspects for children's metacognitive monitoring development. This may benefit children's learning worldwide.

Keywords: metacognitive monitoring, sociocultural aspects, native and non-native speakers, language, feedback

Contents

1 Introduction 11

2 Sociocultural theories on metacognitive processes 13

 2.1 Empirical Evidence..... 16

3 Summary of Results 20

 3.1 Study 1 20

 3.2 Study 2..... 21

 3.3 Study 3..... 22

4 The sociocultural lens..... 23

5 Discussion 26

 5.1 Native and non-native speakers 26

 5.2 Language abilities..... 28

 5.3 Feedback..... 29

 5.4 Prospects..... 31

 5.5 Conclusion..... 33

6 References 35

7 Appendix A 45

8 Manuscripts 46

 8.1 Study 1 46

 8.2 Study 2..... 84

 8.3 Study 3..... 125

9 Erklärung zur Dissertation 164

1 Introduction

“When you know a thing, to hold that you know it; and when you do not know a thing, to allow that you do not know it: That is knowledge” (Burton, 1967, p.1060). There is nothing outdated about this quote from Confucius from many centuries ago. Research reveals that discriminating between accurate and inaccurate performance is critical for self-regulated learning (e.g., recognizing errors, allocating study time, and asking for help) and academic achievement (Coughlin et al., 2015; Destan et al., 2014; Freeman et al., 2017; Roebbers, 2017; Schraw et al., 2006). Children’s ability to introspect and evaluate cognitive activities and processes refers to metacognitive monitoring (T. O. Nelson & Narens, 1990). However, as for developmental psychology in general, research focused mainly on WEIRD (Western, Educated, Industrialized, Rich, and Democratic) samples, and non-WEIRD populations were mostly neglected (Arnett, 2008; Nielsen et al., 2017). This limits the generalizability of the findings which is problematic as non-WEIRD populations represent the majority of the world’s population (Arnett, 2008), and the cultural diversity of children in WEIRD society’s classrooms is increasing (OECD, 2018). Investigating the role of sociocultural aspects in the development of metacognitive monitoring is highly relevant to understanding and supporting children’s learning worldwide. As Rogoff et al. (2018) emphasized, examining the universality of findings is essential instead of assuming them. Therefore, the present umbrella paper aims to take a sociocultural perspective on three research projects (Study 1, Study 2, and Study 3) investigating children’s metacognitive monitoring.

Culture consists of shared values, beliefs, and practices. Importantly culture is transmitted through social interactions with individuals belonging to a cultural group (Goodnow et al., 1995; Keller & Kärtner, 2013). Sociocultural aspects of children’s cognitive development are recognized in classical developmental theories, such as the zone of proximal development and scaffolding (Vygotsky, 1978, 1987) and the ecological system theory (Bronfenbrenner & Morris, 2006). The zone of proximal development (Vygotsky, 1978)

describes the individual's region of sensitivity for learning in a specific domain and exemplifies the close relationship between culture and cognition. Learning in the zone of proximal development happens through scaffolding, which refers to social support from a more skilled individual who typically teaches culturally established thinking practices (Gauvain & Perez, 2015; Vygotsky, 1987). The ecological system theory (Bronfenbrenner & Morris, 2006) describes different social and environmental layers that affect children's development. Child development is embedded in settings and their interactions ranging from proximal (the microsystem, e.g., family, peers, school) to more distant (exo- and macrosystems, e.g., policies, laws, cultural values, and rules). From Vygotsky's (1978, 1987) and Bronfenbrenner's theories (Bronfenbrenner & Morris, 2006), it can be derived that the sociocultural context is crucial for children's cognitive development, and metacognitive monitoring should be no exception.

Indeed, empirical findings suggest that sociocultural contexts, such as families and schools, play an essential role in metacognitive development. Interactions with parents (Carr et al., 1989; Thompson & Foster, 2014) are crucial to acquire mental state language (e.g., explain, remember, learn, forget, teach), which is positively related to metacognitive development (Lockl & Schneider, 2006). Moreover, parents' strategy instructions are related to metacognitive development (Carr et al., 1989). Schooling positively affects children's metacognition (Rogoff, 1994). For instance, teachers providing mnemonic strategies and asking metacognitive questions improve first graders' metacognitive skills (Coffman et al., 2008; Grammer et al., 2013). In sum, the sociocultural context - through instructions and mental state language - drives children's metacognitive development.

To wrap up, classical developmental theories (Bronfenbrenner & Morris, 2006; Vygotsky, 1978, 1987) suggest that sociocultural context affects children's cognitive development. Empirical research reveals that metacognitive development is no exception and

is also affected by the sociocultural context, such as parents' (Carr et al., 1989; Thompson & Foster, 2014) and teachers' (Coffman et al., 2008; Grammer et al., 2013) metacognitive instructions. However, research on sociocultural aspects of metacognitive monitoring is sparse and biased towards WEIRD samples. This is problematic as practices, such as parents, and teachers' instructions, are likely to vary across cultures. Therefore, the present umbrella paper takes a sociocultural perspective on children's metacognitive monitoring development. Firstly, I will introduce two recent theories on sociocultural context and metacognition: the multifaceted and multilevel model of metacognition (Efklides, 2008) and the cultural origins hypothesis (Heyes et al., 2020). Secondly, I will summarize the results of three existing studies (Studies 1, 2, and 3) relevant to the dissertation. Thirdly, based on the multifaceted and multilevel model of metacognition (Efklides, 2008) and the cultural origins hypothesis (Heyes et al., 2020), I will elaborate on Studies 1, 2, and 3 through a sociocultural lens. Fourthly, I will discuss sociocultural aspects of metacognitive monitoring based on native and non-native speakers, language abilities, and feedback.

2 Sociocultural theories on metacognitive processes

As described previously, the importance of sociocultural context for children's cognitive development has already been recognized in theories, such as the zone of proximal development (Vygotsky, 1978, 1987) and the ecological system theory (Bronfenbrenner & Morris, 2006). These are broad theories on children's learning and development. Given the present umbrella paper's focus on metacognitive monitoring, I will introduce two recent and more specific theories on the interplay between sociocultural context and metacognition, namely the multifaceted and multilevel model of metacognition (Efklides, 2008) and the cultural origins hypothesis (Heyes et al., 2020).

Efklides (2008) multifaceted and multilevel model of metacognition describes how metacognitive knowledge, experiences, and skills interact at the nonconscious, personal-

awareness (conscious), and social levels (see Appendix A). Metacognitive knowledge includes information about tasks, strategies, and goals (Flavell, 1979) and gets updated through interactions and language use (Ebert, 2015, 2020; Lockl & Schneider, 2007; Ruffman et al., 2002). Metacognitive experiences have a cognitive (e.g., confidence judgments) and an affective character (e.g., negative affect with error detection; Efklides, 2005, 2006; Efklides & Petkaki, 2005). Metacognitive skills refer to conscious strategies to regulate cognitive processes and contribute to the co-regulation of cognition through feedback and guidance.

Most importantly, the multifaceted and multilevel mode of metacognition expands traditional models of monitoring and control (T. O. Nelson & Narens, 1990) on a social level. The social level (meta-metalevel) comprises metacognition about one's own and other cognitions. For instance, co-regulation in learning situations involves awareness of one's own and others' metacognitive experiences (Salonen et al., 2005). Monitoring and control processes are informed by social interactions and self-awareness on the personal-awareness level. The personal-awareness level (metalevel) represents conscious monitoring and control processes and is informed by the nonconscious level (object level). At the nonconscious level, monitoring and control processes are informed by unconscious cognitive and emotion regulation loops (Efklides, 2008). The model suggests that social and unconscious processes affect a person's conscious metacognitive monitoring and control. As social interactions are typically shaped by culture (Goodnow et al., 1995; Keller & Kärtner, 2013), the model is suitable for taking a sociocultural perspective on metacognitive monitoring.

The cultural origins hypothesis suggests that metacognition is acquired through cultural learning (Heyes et al., 2020). Cultural learning is based on social interaction between differently skilled individuals. The idea is that valid metacognitive decisions benefit the individual and the social group members. For instance, jurors rely on witness confidence to evaluate the credibility of their testimony (Tenney et al., 2007). In that case, accurate

confidence (metacognitive monitoring) is beneficial for the witness's credibility and also contributes to a fair lawsuit. Therefore, metacognitively more skilled individuals should be interested in teaching metacognitive skills to less skilled individuals.

The cultural origins hypothesis (Heyes et al., 2020) suggests that discrimination, interpretation, and broadcasting explain the capacity of metacognition and are acquired through cultural learning. *Discrimination* is crucial to distinguish between external (stimulus visibility vs. confidence) and internal signals (low confidence vs. fear). Discrimination is acquired through interactions with more skilled individuals who can create and label metacognitive experiences, such as certainty and uncertainty (e.g., “How certain are you?”). The *interpretation* of relevant cues for one’s metacognition, such as process fluency, response latency, and task difficulty, is crucial for metacognitive monitoring (e.g., Ackerman & Koriat, 2011; Desender et al., 2017; Koriat & Ackerman, 2010; Roebbers et al., 2019; van Loon et al., 2017). Instruction and feedback can alter the interpretation and use of metacognitive cues. For instance, the instruction “When it is easy, it is often wrong” changes the interpretation of metacognitive cues. High processing fluency - typically a cue for high confidence (e.g., Roebbers et al., 2019) - becomes a cue for low confidence, whereas low processing fluency becomes a cue for high confidence (Desender et al., 2017). *Broadcasting* involves verbally (mental state vocabulary) and non-verbally (e.g., gestures, facial expressions, posture) communicating metacognitive representations. Verbal and non-verbal communication is imitated and learned from other cultural members (Heyes et al., 2020).

Taken together, the multifaceted and multilevel model of metacognition (Efklides, 2008) and the cultural origins hypothesis (Heyes et al., 2020) suggest that metacognition is affected by social interactions and culture. Efklides's (2008) model is the first to add a social level to metacognition. It suggests that social interactions, such as co-regulation, can affect metacognitive processes at the personal awareness level. Similarly, the cultural origins

hypothesis (Heyes et al., 2020) proposes that metacognitive processes (discrimination, interpretation, and broadcasting) are acquired through cultural learning. Cultural learning involves language, imitation, teaching, and mindreading (Heyes et al., 2020). These processes vary across cultures suggesting cross-cultural differences in metacognitive processes, such as metacognitive monitoring. The multifaceted and multilevel model of metacognition (Efklides, 2008) and the cultural origins hypothesis (Heyes et al., 2020) build a theoretical framework to investigate and explain sociocultural aspects of metacognitive monitoring.

2.1 Empirical Evidence

Based on the previously outlined multilevel and multifaceted model of metacognition (Efklides, 2008) and the cultural origins hypothesis (Heyes et al., 2020), I will review empirical evidence on sociocultural aspects of metacognitive monitoring. When reviewing the literature, it is critical to differentiate between cross-cultural and within-cultural research approaches (Göncü & Gauvain, 2012). Cross-cultural research typically compares one culture with another across countries or different populations within a country (e.g., socio-economic background, native and non-native speakers). Thereby culture is seen as a stable construct. In comparison, the within-cultural approach focuses on how cognitive development is affected by culture within a culture. In that perspective, culture and cognitive development are seen as an integer interacting system that cannot be disassembled (Gauvain & Perez, 2015). Within-cultural approaches typically study how cognition is affected by cultural tools (e.g., language), institutional learning (e.g., schooling), models (who may provide feedback), or everyday cultural practices (e.g., patterns of discourse; Goodnow et al., 1995). The within-cultural focus of the present umbrella paper is on the effect of language and feedback on metacognitive monitoring. Cross-cultural and within-cultural research contributes to a better understanding of sociocultural aspects of metacognitive monitoring from slightly different angles. Therefore, I will firstly review cross-cultural and secondly within-cultural research on

metacognitive monitoring. A critical discussion highlighting the limitations and strengths of both approaches can be found in chapter 5.4.

Evidence for the cultural origins hypothesis can be found in adult studies. A recent cross-cultural study (van der Plas et al., 2022) tested the cultural origins hypothesis (Heyes et al., 2020) by comparing English and Chinese university students' metacognitive monitoring in a visual perception task. Participants saw moving dots, estimated their direction, received perceptual or social post-decision evidence (by a social agent), and indicated their confidence in the accuracy of their decision. Chinese students were less confident in inaccurate trials than English students, independently of perceptual or social post-decision evidence. This finding indicates that metacognitive differences between Chinese and English students are domain-general and not domain-specific, which supports the cultural origins hypothesis. Similarly, Muthukrishna et al. (2018) found that overprecision (being more confident in one's beliefs than justified) was lowest in Hong Kong Chinese and Japanese compared to Canadian students across two different tasks (empathy and mathematics tasks). Other studies suggest cross-cultural differences in metacognitive monitoring too. Taiwanese university students discriminated the most in confidence for accurate and inaccurate responses in a mathematics exam compared to American and Palestinian students (Lundeberg et al., 2000). Similarly, a comparison in confidence across nine world regions revealed that east Asians were the least overconfident regarding their performance in a fluid intelligence task (Stankov & Lee, 2014). Interestingly, all four studies suggest benefits in metacognitive monitoring for Asian students compared to other cultures (Lundeberg et al., 2000; Muthukrishna et al., 2018; Stankov & Lee, 2014; van der Plas et al., 2022). Cross-cultural research with adults supports the idea that sociocultural aspects affect metacognitive monitoring.

Cross-cultural research on children's metacognitive monitoring is sparse, but the few available studies provide further evidence for cross-cultural differences and sociocultural

theories on metacognition (Morony et al., 2013; Xia et al., 2022, 2023). Similarly to the previously outlined studies with adults (Lundeberg et al., 2000; Muthukrishna et al., 2018; Stankov & Lee, 2014; van der Plas et al., 2022), studies with children suggest that Asian participants monitor their uncertainty more accurately than WEIRD children. For instance, four to five-year-old Chinese children were less overconfident than Dutch children in memory and motor tasks (ball throwing; Xia et al., 2023), and East Asian adolescents (15-year-olds) were less overconfident than European students in mathematics (Morony et al., 2013). Moreover, Xia et al. (2022) found cross-cultural differences in self- and other performance estimates. Chinese four to five-year-olds estimated their memory and motor performance to be significantly worse than a peer's performance, whereas Dutch children did not differ between their own and a peer's performance estimate. These studies support the notion of cross-cultural differences in aspects of metacognitive monitoring, such as overconfidence and self and other performance estimates. Sociocultural aspects seem to affect adults' and children's metacognitive monitoring.

Despite cross-cultural differences, studies also reported cross-cultural similarities (Kim et al., 2020; Kim, Le Guen, et al., 2021; Xia et al., 2022, 2023). Dutch and Chinese children were both overconfident when predicting their performance in motor and memory tasks (Xia et al., 2022, 2023), and four-year-old Japanese, Yucatec Mayan, and German children were similarly overconfident when estimating their knowledge in a perceptual task (Kim et al., 2020; Kim, Le Guen, et al., 2021). Finally, Japanese and German three- to five-year-olds were more confident in accurate than inaccurate memories (Kim, Senju, et al., 2021). Taken together, the literature review suggests universal and culture-specific aspects of metacognitive monitoring. Universally, children and adults worldwide seem overconfident when estimating their performance and discriminate between accurate and inaccurate performances. Culture-specifically, the magnitude of overconfidence and metacognitive discrimination might vary across cultures.

However, cross-cultural evidence so far is restricted to children between three and five years old (Kim et al., 2020; Kim, Le Guen, et al., 2021; Kim, Senju, et al., 2021; Xia et al., 2022, 2023), adolescents (Morony et al., 2013), university students (Lundeberg et al., 2000; Muthukrishna et al., 2018; Stankov & Lee, 2014; van der Plas et al., 2022), and a limited number of populations (mainly Europeans and East Asians). Moreover, these results also require replication, as emphasized by different findings in similar studies. Despite coming from the same research group, investigating the same age range, and using the same task, Xia et al. (2022) did not find differences in overconfidence between Dutch and Chinese, whereas Xia et al. (2023) found that Dutch children were even more overconfident than Chinese children.

In light of within-cultural research approaches, language and feedback might be crucial sociocultural tools explaining variation in metacognitive monitoring. The multifaceted and multilevel model of metacognition suggests that language is related to developing metacognitive knowledge and experiences (Efklides, 2008). Moreover, language and feedback are crucial for cultural learning (Heyes et al., 2020). These assumptions are underlined by research showing that general language abilities (Ebert, 2015, 2020; Gonzales et al., 2021; Lecce et al., 2010; Lockl & Schneider, 2007) and mental state language (Lockl & Schneider, 2006; Schneider & Lockl, 2002) are related to metacognitive development. Also, parents' (Carr et al., 1989; Thompson & Foster, 2014) and teachers' (Coffman et al., 2008; Grammer et al., 2013) use of mental state language affects metacognitive development. Moreover, feedback benefits children's metacognition (Geurten & Meulemans, 2017; Oudman et al., 2022; van Loon et al., 2017; van Loon & Roebbers, 2017, 2020).

Within-cultural research suggests that language and feedback are crucial sociocultural features that explain within-cultural variation in metacognition but may also differ across cultures and explain cross-cultural variation in metacognition. Compared to

cross-cultural studies on metacognition, more within-cultural research is available, focusing on language and feedback in a broader age range. However, within-cultural approaches with non-WEIRD subjects are sparse. Exceptions are Gonzales et al. (2021) study looking at language and metacognition in American children from low socio-economic backgrounds and two studies investigating feedback in Chinese children (Wang & Sperling, 2021; Xia et al., 2022).

Taken together, cross-cultural and within-cultural research support the notion - of the multifaceted and multilevel model of metacognition (Efklides, 2008) and the cultural origins hypothesis (Heyes et al., 2020) - that the social and the cultural context affect metacognition and its development. However, empirical research on non-WEIRD populations is sparse. Future cross-cultural and within-cultural research should focus on non-WEIRD subjects and a comprehensive age range to clarify sociocultural aspects of metacognitive monitoring development.

3 Summary of Results

In this section, I will briefly present the results of three studies relevant to the present dissertation. In Study 1, we compared native and non-native speakers' metacognitive monitoring. In Study 2, we investigated the prospective effect of language abilities on children's metacognitive monitoring. In Study 3, we compared the benefits of performance feedback and metacognitive feedback for children's metacognitive monitoring. The full manuscripts can be found in chapter 8.

3.1 Study 1

In Study 1, we compared metacognitive monitoring between native and non-native speaking children (~10-year-olds). We assessed metacognitive monitoring in a memory task (paired-associates), including a recognition test and confidence judgments. We relied on the mean confidence difference for accurate and inaccurate trials (discriminations score) for metacognitive monitoring. A higher discrimination indicates more accurate metacognitive

monitoring. Results revealed no differences in recognition performance and metacognitive monitoring between native and non-native speakers. We replicated the findings with the same paired-associates task in a different sample. In that replication study, we additionally looked at children's metacognitive monitoring in a text comprehension task, including open questions and confidence judgments. Native speakers answered more open questions correctly than non-native speakers but did not differ in their metacognitive monitoring. In conclusion, native and non-native speaking children did not differ in their metacognitive monitoring in memory and text comprehension tasks.

3.2 Study 2

In Study 2, we investigated the prospective effect of language abilities in kindergarten on metacognitive monitoring in grade one. We were also interested in differences between native and non-native speakers' relation between language abilities and metacognitive monitoring. We relied on data from the National Educational Panel Study, a German large-scale assessment ($N = 9,159$). Metacognitive monitoring and language abilities were assessed in kindergarten and grade one. Compared to Studies 1 and 3, metacognitive monitoring was assessed on a global scale. Children were asked to estimate their overall performance after solving a math and a science task. We computed the absolute difference between estimated and actual performance as a measure of metacognitive monitoring. A score of zero represents perfectly accurate monitoring, whereas higher scores indicate less accurate metacognitive monitoring. Language abilities (standardized vocabulary and grammar tasks) were measured in the language of assessment, which is the first language for native speakers but the second language for non-native speakers. For the whole sample, cross-lagged panel models revealed that language abilities in kindergarten positively predicted metacognitive monitoring accuracy in grade one, but metacognitive monitoring in kindergarten did not predict language abilities in grade one. Multi-group analyses revealed differences in the language metacognitive monitoring relation for native and non-native speakers. Language

abilities positively predicted metacognitive monitoring for native speakers but not for non-native speakers. Conversely, language abilities negatively predicted metacognitive monitoring for non-native speakers but not for native speakers. Overconfidence in kindergarten was related to higher second language abilities of non-native speakers. Our results suggest that first language abilities positively predict metacognitive monitoring for native speakers and that overconfidence positively predicts second language abilities for non-native speakers.

3.3 Study 3

In Study 3, we aimed to improve first graders' metacognitive monitoring with feedback. For this purpose, we evaluated a six-time computer-based intervention program. We randomly assigned the participants to either a metacognitive feedback group, a performance feedback group, or an active control group. Participants in the metacognitive and performance feedback group received feedback in a memory task (paired-associates). The metacognitive feedback group received feedback on their metacognitive monitoring accuracy and recognition accuracy. The performance feedback group received feedback solely on their recognition accuracy. The active control group solved an inhibition task different from the paired-associates task. To measure and compare the training effects, we assessed children's metacognitive monitoring pre- and post-intervention with a paired-associates memory task (different from the intervention task) without feedback. As in Study 1, we relied on the mean confidence difference for accurate and inaccurate trials (discrimination score) for metacognitive monitoring accuracy, with higher discrimination indicating more accurate metacognitive monitoring. Results revealed that metacognitive monitoring accuracy increased in the metacognitive feedback group but not in the performance feedback or the active control groups. Metacognitive feedback seems to improve primary school children's metacognitive monitoring.

4 The sociocultural lens

In the following, I will elaborate on the previously summarized research projects through a sociocultural lens. It is important to note that the initial focus of Studies 1, 2, and 3 was not on sociocultural aspects of children's metacognitive monitoring. Therefore, I will take a new theoretical perspective on these studies based on the multifaceted and multilevel model of metacognition (Efklides, 2008) and the cultural origins hypothesis (Heyes et al., 2020).

In Study 1, we compared native and non-native speakers' metacognitive monitoring. Study 1 can be seen as a cross-cultural approach comparing native and non-native speakers' metacognitive monitoring. Native and non-native speakers are exposed to different languages. Language is a cultural tool incorporated into cognitive processes (K. Nelson, 1996; Tomassello, 2006; Vygotsky, 1987) and affects how children think, talk, and learn about cognitive processes (Astington & Baird, 2005; Harris et al., 2005). Language might help us to conceptualize and represent unobservable mental states, such as metacognitive monitoring. Moreover, language is a tool to communicate about mental states and learn from others, such as peers and teachers. Indeed metacognitive vocabulary and language abilities predict later metacognitive abilities (Ebert, 2015, 2020; Gonzales et al., 2021; Lecce et al., 2010; Lockl & Schneider, 2006, 2007). Native and non-native speakers might represent and conceptualize metacognitive monitoring differently based on primary language features. Furthermore, it might be easier to communicate about metacognitive experiences in one's native language than in a non-native language, facilitating native speakers to learn from peers and teachers compared to non-native speakers.

Integrating these findings in the multifaceted and multilevel model of metacognition (Efklides, 2008) suggests that differences in native and non-native speakers' language abilities might be reflected in the personal-awareness and social levels of metacognition. At the personal-awareness level, language abilities serve to conceptualize and represent

metacognitive monitoring. At the social level, language abilities are crucial to learning from others. Therefore, language differences between native and non-native speakers might cause differences in metacognitive monitoring on the personal-awareness level. Moreover, based on the cultural origins hypothesis (Heyes et al., 2020), language might affect the interpretation of relevant cues (through instruction and feedback) and broadcasting (communicating) of mental states. Therefore, native and non-native speakers might acquire metacognitive monitoring differently. Based on theoretical models and empirical findings, native and non-native speakers' might differ in their metacognitive monitoring.

In Study 2, we investigated the longitudinal relation between language abilities and native and non-native speakers' metacognitive monitoring. It is crucial to note that language abilities were assessed in the language of instruction, representing the first language for native speakers and the second language for non-native speakers. This limits the comparability between the results for native and non-native speakers. Therefore, Study 2 is more likely to represent a within-cultural study of language and metacognition in two populations (native and non-native speakers) than a cross-cultural comparison.

As outlined previously, language abilities are crucial to conceptualizing, representing, and communicating metacognitive monitoring (Ebert, 2015, 2020; Gonzales et al., 2021; Lecce et al., 2010; Lockl & Schneider, 2006, 2007). From a theoretical perspective, language abilities are represented on the personal-awareness and the social level of metacognition (Efklides, 2008) and are crucial for interpretation and broadcasting from a cultural learning perspective (Heyes et al., 2020). However, it is unclear whether abilities in the language of instruction (first vs. second language abilities) are equally relevant for native and non-native speakers' metacognition. On the one hand, it might be easier to conceptualize, represent and communicate metacognitive monitoring in the first than in the second language. On the other hand, the language of instruction is crucial for native and non-native speakers to

learn from teachers or peers at school. Overall empirical findings and theoretical models suggest language abilities as a sociocultural feature that might explain within-cultural differences in metacognitive monitoring. Therefore, we expected that language ability in kindergarten predicts metacognitive monitoring in grade one, but it is unclear whether first (native speakers) and second language abilities (non-native speakers) relate similarly to children's metacognitive monitoring.

In Study 3, we evaluated the benefits of metacognitive and performance feedback for children's metacognitive monitoring in a WEIRD sample (Swiss children). Study 3 can be seen as a within-cultural approach assessing the benefits of feedback for children's metacognitive monitoring. Feedback may benefit metacognitive monitoring as children learn to recognize valid cues (e.g., task difficulty). Previous research revealed mixed findings. Some studies suggested benefits of metacognitive feedback for children's metacognitive monitoring (Geurten & Meulemans, 2017; van Loon & Roebbers, 2020), whereas others did not (Wang & Sperling, 2021). Similarly, some studies found benefits of performance feedback for metacognitive monitoring (Oudman et al., 2022; van Loon & Roebbers, 2017), whereas others did not (Lipko et al., 2009, 2012; O'Leary & Sloutsky, 2017; Xia et al., 2022). Interestingly, the two studies with non-WEIRD participants did not find metacognitive (Wang & Sperling, 2021) or performance feedback (Xia et al., 2022) benefits for Chinese children's metacognitive monitoring. Feedback may not be equally efficient across cultures.

Feedback is typically provided at the social level intending to affect children's personal-awareness level of metacognition (Efklides, 2008; van Loon & Roebbers, 2021). Furthermore, feedback is crucial for discrimination, interpretation, and broadcasting based on the cultural origins hypothesis (Heyes et al., 2020). Labeling certainty and uncertainty experiences can help discriminate between different levels of confidence. Feedback can also hint at interpreting relevant cues of metacognitive monitoring (e.g., Desender et al., 2017).

Finally, feedback involves broadcasting mental states, such as metacognitive monitoring. Comparing performance and metacognitive feedback might help to gain a differentiated perspective on cultural learning. Some forms of feedback and instruction might be more successful than others and explain within-cultural variability in metacognitive monitoring.

5 Discussion

The present umbrella paper aims to broaden our understanding of sociocultural aspects of children's metacognitive monitoring. To address this aim, I will discuss how the results of Studies 1, 2, and 3 comply with existing within- and cross-cultural research, the multifaceted and multilevel model of metacognition (Efklides, 2008), and the cultural origins hypothesis (Heyes et al., 2020). The discussion focuses on native and non-native speakers, language abilities, and feedback. After discussing the results, I will outline prospects for cross-cultural and within-cultural research and end with a conclusion.

5.1 Native and non-native speakers

In Study 1, we found that native and non-native speakers discriminated equally well in confidence between incorrect and correct answers in memory and text comprehension tasks. This is in line with cross-cultural research, showing that children (Kim, Senju, et al., 2021) and adults (Lundeberg et al., 2000; van der Plas et al., 2022) are more confident in correct than incorrect responses across various cultures. For instance, Japanese and German children were more confident in accurate than inaccurate memories (Kim, Senju, et al., 2021). However, our results are opposed to research suggesting cross-cultural differences in aspects of children's metacognitive monitoring, such as the magnitude of overconfidence and metacognitive discrimination. Asian children were less overconfident than European children in memory, motor, and mathematics tasks (Morony et al., 2013; Xia et al., 2023). Also, studies with adults suggest that Asian students are less overconfident (Muthukrishna et al., 2018; Stankov et al., 2014) and metacognitively discriminate more than WEIRD students (Lundeberg et al., 2000; van der Plas et al., 2022). In comparison to our study, these studies

compared metacognitive monitoring across countries. Native and non-native speakers share more sociocultural features than children from different countries. For instance, they are exposed to the same teachers, and metacognition is known to be shaped by teacher instructions (Coffman et al., 2008; Grammer et al., 2013). Therefore, native and non-native speakers' metacognitive monitoring might be more similar than cross-cultural comparisons between children from different countries. Moreover, effects might be pronounced in adults as they were exposed to a particular cultural context for longer than children. Applying our results to the multifaceted and multilevel model of metacognition (Efklides, 2008) and the cultural origins hypothesis (Heyes et al., 2020) might suggest that native and non-native speakers share a similar social environment and, therefore, their metacognitive monitoring at the personal awareness level is comparable. Schools might be the main venue of cultural learning.

For future cross-cultural research, it would be fascinating to follow up on venues of cultural learning (e.g., school, home) for children's metacognitive monitoring. This would allow us to infer culture-specific (e.g., native and non-native speakers) suggestions to support children's metacognitive monitoring at school and home. Teachers' (Coffman et al., 2008; Grammer et al., 2013) and parents' (Carr et al., 1989; Thompson & Foster, 2014) instructions explain within-cultural variance in children's metacognitive development and, therefore, might also explain cross-cultural differences. To verify these assumptions, one could cross-culturally compare metacognitive instructions at school and home and their role in children's metacognitive monitoring. For instance, by observing and classifying teachers' use of mental state language in various cultures. As in Coffman et al.'s (2008) study, *the taxonomy of teacher behaviors* could be used to classify teacher conversation in four categories (instruction, cognitive structuring activities, memory requests, non-memory-relevant). The same taxonomy could also serve to cross-culturally classify and compare parents speech at home. However, an alternative classification system for parents metacognitive speech

(planning, strategy and self-monitoring) is suggested by Thompson et al. (2014). The best suiting classification could be evaluated in a pilot study. Moreover, assessing teachers and parents use of metacognitive instructions in the same study would allow to compare the role of schools and home in children's metacognitive development.

5.2 Language abilities

In Study 2, first language abilities predicted metacognitive monitoring of native speakers. Native speakers with higher language abilities in kindergarten were more accurate in their metacognitive monitoring in grade one. This is in line with previous research suggesting that language abilities are related to metacognition (Ebert, 2015, 2020; Gonzales et al., 2021; Lecce et al., 2010; Lockl & Schneider, 2006, 2007). However, second language abilities of non-native speakers did not predict their metacognitive monitoring, suggesting that first language abilities are more critical for metacognitive monitoring than second language abilities. First language abilities might explain within-cultural differences in native speakers' metacognitive monitoring. Future research should assess non-native speakers' first language abilities to clarify the role of first language abilities in non-native speakers' metacognitive monitoring.

Interestingly, metacognitive monitoring predicted second language abilities of non-native speakers, but metacognitive monitoring did not predict first language abilities of native speakers. More overconfident non-native speakers in kindergarten had higher second language abilities in grade one. This suggests that overconfidence explains within-cultural differences in non-native speakers' second language abilities. Overconfidence might lead to higher persistence and motivation when learning a second language, resulting in better second language abilities (Bjorklund & Bering, 2002; Shin et al., 2007). Future studies should include native speakers' second language abilities to clarify whether overconfidence is a general motor for second language abilities. This would be a fascinating finding since it

opposes the general assumption that more accurate monitoring benefits learning in most domains.

Integrating our findings into the multifaceted and multilevel model of metacognition (Efklides, 2008) suggests that language abilities are more likely related to the personal-awareness level than the social level of metacognition. Native and non-native speakers interact in the same language (the language of instruction) at school. Therefore, if language abilities would have operated at the social level (e.g., communicating about mental processes), second language abilities of non-native speakers should also predict their metacognitive monitoring. However, we acknowledge that our measure of language (vocabulary and grammar) is more likely to capture language abilities at the personal-awareness level than the social level. Future studies should assess language abilities relevant to the personal awareness level (e.g., vocabulary and grammar) and the social level (e.g., communication skills, use of mental state language) to clarify the role of language for metacognition. Regarding cultural learning, our results might indicate that language abilities are crucial for internal processes, such as discrimination and interpretation, and less for broadcasting. In sum, our results suggest that language is a crucial cultural feature affecting metacognition at the personal-awareness level.

5.3 Feedback

In Study 3, we were interested in how metacognitive and performance feedback affects children's metacognitive monitoring. In line with previous WEIRD research, we found that metacognitive feedback improved metacognitive monitoring (Geurten & Meulemans, 2017; van Loon & Roebbers, 2020). Interestingly, a study with Chinese 7th graders did not find benefits of metacognitive instructions for children's metacognitive monitoring in a mathematics task. Metacognitive instruction increased metacognitive bias in a three weeks lasting training (Wang & Sperling, 2021). Compared to our study, the participants monitored fairly accurately at pretest, leaving little room for improvement. Participants' higher

monitoring accuracy in Wang and Sperling's (2021) compared to our study might be explained by the older investigated age group (e.g., Roebbers, 2017) and cross-cultural differences. For instance, some research suggests that Asian children monitor more accurately than European children (Morony et al., 2013; Xia et al., 2023). The benefits of metacognitive feedback might depend on age group and sociocultural background.

Similarly to previous research, performance feedback did not improve metacognitive monitoring (Lipko et al., 2009, 2012; O'Leary & Sloutsky, 2017; Xia et al., 2022). Our results align with a cross-cultural study by Xia et al. (2022), showing that performance feedback did not decrease Chinese and Dutch children's overconfidence in motor and memory tasks. However, studies with older children (Oudman et al., 2022; van Loon & Roebbers, 2017) and studies displaying performance feedback before children's monitoring judgments (van Loon et al., 2017) revealed benefits of performance feedback for children's metacognitive monitoring. The benefits of performance feedback might depend on the age of participants and the timing of performance feedback.

Following the multifaceted and multilevel model of metacognition (Efklides, 2008), our results suggest that metacognitive feedback at the social level can alter metacognitive monitoring at the personal-awareness level. As suggested by the cultural origins hypothesis (Heyes et al., 2020), (metacognitive) feedback may have helped to discriminate between mental states, such as experiencing higher confidence for correct than incorrect answers. Moreover, metacognitive feedback may have contributed to interpreting performance accuracy as a cue for confidence. Taken together, feedback has the potential to improve children's metacognitive monitoring. The benefits of feedback might depend on feedback characteristics (quality and timing) and participants' characteristics (age group and sociocultural background), and their interactions. Feedback and participants' characteristics have the potential to explain within-cultural and cross-cultural variance in metacognitive

monitoring. Especially little is known about the role of sociocultural background in the efficiency of feedback. Therefore, future research should systematically investigate metacognitive and performance feedback in non-WEIRD participants.

5.4 Prospects

As emphasized multiple times in the present umbrella paper, future research should further assess sociocultural aspects of children's metacognitive monitoring. To gain a differentiated perspective, cross-cultural and within-cultural research approaches should be considered. I will review the limitations and strengths of both approaches and suggest future cross-cultural and within-cultural research.

Cross-cultural research typically compares cultural groups across or within countries. Thereby culture is seen as a stable construct, which has been criticized for multiple reasons (Gauvain & Perez, 2015). Firstly, this is problematic when the development of children from one culture (typically a WEIRD culture) is taken as the norm to describe and compare the development of children from another culture (typically a non-WEIRD culture). Secondly, typically cultural differences are emphasized, whereas cultural similarities are overlooked. Thirdly, a differentiated perspective on the mechanisms between culture and cognitive development is often lacking. However, cross-cultural comparisons of metacognitive monitoring could be valuable in identifying universal and culture-specific aspects of metacognitive monitoring when sampling, analyses, and interpretation of the results are made carefully with a non-judgmental and non-ethnocentric perspective on cultural differences.

Van der Plas et al.'s (2022) study on metacognitive monitoring in Chinese and English university students is an excellent example of cross-cultural research. The authors are careful in their methods and interpretations to avoid conceptualizing culture as a monolithic concept and emphasizing cultural differences. Therefore, I rely on their study to suggest future cross-cultural research on metacognitive monitoring:

(1) Van der Plas et al. (2022) rely on the cultural origins hypothesis, a fine-grained framework to cross-culturally investigate metacognitive monitoring aspects, such as discrimination, interpretation, and broadcasting. Therefore, I suggest the cultural origins hypothesis as a theoretical framework for future cross-cultural studies. (2) Van der Plas et al. (2022) matched Chinese and English students for occupation, income, demographics, and general intelligence, but most importantly, for first-order task performance. Careful matching of the samples ensures metacognitive monitoring is liable to sociocultural aspects and not due to other variables, such as first-order task performance. Hence, matching might be recommended for future cross-cultural research. (3) Van der Plas et al. (2022) assessed metacognitive monitoring with an item-by-item measure and computed metacognitive sensitivity. Metacognitive sensitivity estimates metacognitive monitoring independently from task performance. This is highly relevant as group differences in metacognitive monitoring are known to be confounded by differences in task performance (Fleming & Lau, 2014). Future cross-cultural studies should include item-by-item measures to compute metacognitive sensitivity. (4) Van der Plas et al. (2022) suggest that their findings should be replicated in non-WEIRD samples across the world. Some research has been done with East-Asian samples (e.g., Kim, Senju, et al., 2021; Morony et al., 2013; van der Plas et al., 2022; Xia et al., 2022, 2023), but other regions have mostly been neglected. To the best of my knowledge, I do not know about a single study on metacognitive monitoring with African participants. (5) Van der Plas et al. (2022) propose that their findings be replicated in different age groups. For instance, with children. Especially the transition to school is a critical phase for metacognitive development (Roebbers, 2017).

Within-cultural approaches could further deepen our understanding of the interplay between sociocultural context and metacognitive monitoring. In contrast to cross-cultural approaches, within-cultural research conceptualizes culture and cognition as a dynamic system and avoids taking one culture as the norm for other cultures (Gauvain & Perez, 2015).

Studying sociocultural aspects (e.g., language and feedback) and their effects on metacognitive monitoring may enrich our understanding of mechanisms explaining the relation between sociocultural context and metacognitive monitoring.

The previously outlined suggestions for cross-cultural research based on van der Plas et al.'s (2022) study also apply to within-cultural research. For future within-cultural research, the cultural origins hypothesis may be a relevant theoretical framework, metacognitive sensitivity is a valid measure of metacognitive monitoring, findings should be replicated with non-WEIRD participants, and studies should assess different age groups. In that regard, Gonzales et al.'s longitudinal study (2021) on vocabulary and metacognitive monitoring is a great example. They assessed American children from low socioeconomic backgrounds (non-WEIRD) and relied on metacognitive sensitivity to measure metacognitive monitoring. More such research is needed, for instance, investigating performance and metacognitive feedback in non-WEIRD samples.

5.5 Conclusion

In the present umbrella paper, I outlined sociocultural aspects of metacognitive monitoring. The multifaceted and multilevel of metacognition (Efklides, 2008) and the cultural origins hypothesis (Heyes et al., 2020) provided a theoretical framework to take a sociocultural perspective on three existing research projects. Both models suggest that metacognitive monitoring is affected by the sociocultural context. Our results are partly in line with this assumption. On the one hand, we did not find cross-cultural differences between native and non-native speakers' metacognitive monitoring (Study 1), which might suggest that sociocultural context did not affect metacognitive monitoring. However, as outlined previously, it is more likely that native and non-native speakers share a highly similar sociocultural learning environment (e.g., school) and, hence, do not differ in their metacognitive monitoring. On the other hand, we found that language (Study 2) and metacognitive feedback (Study 3) affect metacognitive monitoring. This might suggest that

language and feedback are essential sociocultural features to explain within-cultural and possibly also cross-cultural variance in metacognitive monitoring. More cross-cultural and within-cultural research on non-WEIRD children is needed to clarify universal and culture-specific aspects of metacognitive monitoring.

Coming back to Confucius's quote in the first paragraph: “When you know a thing, to hold that you know it; and when you do not know a thing, to allow that you do not know it: That is knowledge (Burton, 1967, p.1060)” may not solely apply to children’s learning but also to research on metacognition: “When we know that children's metacognitive monitoring is affected by sociocultural aspects, we should also allow that we do not know much about the universality of metacognitive monitoring. That is knowledge”.

6 References

- Ackerman, R., & Koriat, A. (2011). Response latency as a predictor of the accuracy of children's reports. *Journal of Experimental Psychology: Applied*, *17*(4), 406–417. <https://doi.org/10.1037/a0025129>
- Arnett, J. J. (2008). The Neglected 95%: Why American Psychology Needs to Become Less American. *American Psychologist*, *63*(7), 602–614. <https://doi.org/10.1037/0003-066X.63.7.602>
- Astington, J. W., & Baird, J. A. (2005). *Why Language Matters for Theory of Mind*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195159912.001.0001>
- Bjorklund, D. F., & Bering, J. M. (2002). The evolved child applying evolutionary developmental psychology to modern schooling. *Learning and Individual Differences*, *12*(4), 347–373. [https://doi.org/10.1016/S1041-6080\(02\)00047-X](https://doi.org/10.1016/S1041-6080(02)00047-X)
- Bronfenbrenner, U., & Morris, P. A. (2006). The bioecological model of human development. In W. Damon & R. M. Lerner (Eds.), *Theoretical models of human development. Volume 1 of the Handbook of child psychology* (6th ed., pp. 293–828). Wiley.
- Burton, S. (1967). *The home book of quotations: Classical and modern* (10th ed.). Dodd, Mead, & Company.
- Carr, M., Kurtz, B. E., Schneider, W., Turner, L. A., & Borkowski, J. G. (1989). Strategy Acquisition and Transfer Among American and German Children: Environmental Influences on Metacognitive Development. *Developmental Psychology*, *25*(5), 765–771. <https://doi.org/10.1037/0012-1649.25.5.765>
- Coffman, J. L., Ornstein, P. A., McCall, L. E., & Curran, P. J. (2008). Linking Teachers' Memory-Relevant Language and the Development of Children's Memory Skills. *Developmental Psychology*, *44*(6), 1640–1654. <https://doi.org/10.1037/a0013859>

- Coughlin, C., Hembacher, E., Lyons, K. E., & Ghetti, S. (2015). Introspection on uncertainty and judicious help-seeking during the preschool years. *Developmental Science, 18*(6), 957–971. <https://doi.org/10.1111/desc.12271>
- Desender, K., Van Opstal, F., & Van Den Bussche, E. (2017). Subjective experience of difficulty depends on multiple cues. *Scientific Reports, 7*, 1–14. <https://doi.org/10.1038/srep44222>
- Destan, N., Hembacher, E., Ghetti, S., & Roebbers, C. M. (2014). Early metacognitive abilities: The interplay of monitoring and control processes in 5- to 7-year-old children. *Journal of Experimental Child Psychology, 126*, 213–228. <https://doi.org/10.1016/j.jecp.2014.04.001>
- Ebert, S. (2015). Longitudinal Relations Between Theory of Mind and Metacognition and the Impact of Language. *Journal of Cognition and Development, 16*(4), 559–586. <https://doi.org/10.1080/15248372.2014.926272>
- Ebert, S. (2020). Early Language Competencies and Advanced Measures of Mental State Understanding Are Differently Related to Listening and Reading Comprehension in Early Adolescence. *Frontiers in Psychology, 11*(952), 1–18. <https://doi.org/10.3389/fpsyg.2020.00952>
- Efklides, A. (2005). Metacognitive Experiences in Problem Solving. In A. Efklides, J. Kuhl, & R. M. Sorrentino (Eds.), *Trends and Prospects in Motivation Research* (pp. 297–323). Springer. https://doi.org/10.1007/0-306-47676-2_16
- Efklides, A. (2006). Metacognition and affect: What can metacognitive experiences tell us about the learning process? *Educational Research Review, 1*(1), 3–14. <https://doi.org/10.1016/j.edurev.2005.11.001>
- Efklides, A. (2008). Metacognition - Defining Its Facets and Levels of Functioning in

Relation to Self-Regulation and Co-regulation. *European Psychologist*, 13(4), 277–287.

<https://doi.org/10.1027/1016-9040.13.4.277>

Efklides, A., & Petkaki, C. (2005). Effects of mood on students' metacognitive experiences.

Learning and Instruction, 15(5), 415–431.

<https://doi.org/10.1016/J.LEARNINSTRUC.2005.07.010>

Flavell, J. H. (1979). Metacognition and Cognitive Monitoring: A New Area of Cognitive-

Developmental Inquiry. *American Psychologist*, 34(10), 906–911.

<https://doi.org/10.1037/0003-066X.34.10.906>

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human*

Neuroscience, 8(443), 1–9. <https://doi.org/10.3389/fnhum.2014.00443>

Freeman, E. E., Karayanidis, F., & Chalmers, K. A. (2017). Metacognitive monitoring of

working memory performance and its relationship to academic achievement in Grade 4 children. *Learning and Individual Differences*, 57, 58–64.

<https://doi.org/10.1016/j.lindif.2017.06.003>

Gauvain, M., & Perez, S. (2015). Cognitive Development and Culture. In R. M. Lerner (Ed.),

Handbook of Child Psychology and Developmental Science (6th ed., pp. 1–43). Wiley.

<https://doi.org/10.1002/9781118963418.childpsy220>

Geurten, M., & Meulemans, T. (2017). The effect of feedback on children's metacognitive

judgments: a heuristic account. *Journal of Cognitive Psychology*, 29(2), 184–201.

<https://doi.org/10.1080/20445911.2016.1229669>

Göncü, A., & Gauvain, M. (2012). Sociocultural approaches to educational psychology:

Theory, research, and application. In C. B. McCormick, G. M. Sinatra, & J. Sweller

(Eds.), *APA educational psychology handbook, Vol 1: Theories, constructs, and critical issues* (pp. 125–154). American Psychological Association.

<https://doi.org/10.1037/13273-006>

Gonzales, C. R., Mercurief, A., McClelland, M. M., & Ghetti, S. (2021). The development of uncertainty monitoring during kindergarten: Change and longitudinal relations with executive function and vocabulary in children from low-income backgrounds. *Child Development, 93*(2), 1–16. <https://doi.org/10.1111/cdev.13714>

Goodnow, J. J., Miller, P. J., & Kessel, F. (1995). *Cultural practices as contexts for development*. Jossey-Bass.

Grammer, J., Coffman, J. L., & Ornstein, P. (2013). The Effect of Teachers' Memory-Relevant Language on Children's Strategy Use and Knowledge. *Child Development, 84*(6), 1989–2002. <https://doi.org/10.1111/CDEV.12100>

Harris, P. L., De Rosnay, M., & Pons, F. (2005). Language and Children's Understanding of Mental States. *Current Directions in Psychological Science, 14*(2), 69–73. <https://doi.org/10.1111/j.0963-7214.2005.00337.x>

Heyes, C., Bang, D., Shea, N., Frith, C. D., & Fleming, S. M. (2020). Knowing Ourselves Together: The Cultural Origins of Metacognition. *Trends in Cognitive Sciences, 24*(5), 349–362. <https://doi.org/10.1016/j.tics.2020.02.007>

Keller, H., & Kärtner, J. (2013). Development - The Cultural Solution of Universal Developmental Tasks. In M. J. Gelfand, C. Chiu, & Y. Hong (Eds.), *Advances in Culture and Psychology: Volume 3*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199930449.001.0001>

Kim, S., Le Guen, O., Sodian, B., & Proust, J. (2021). Are children sensitive to what they know? An insight from Yucatec Mayan children. *Journal of Cognition and Culture, 21*(3), 226–242. <https://doi.org/10.1163/15685373-12340106>

Kim, S., Senju, A., Sodian, B., Paulus, M., Itakura, S., Okuno, A., Ueno, M., & Proust, J.

- (2021). Memory Monitoring and Control in Japanese and German Preschoolers. *Memory and Cognition*. <https://doi.org/10.3758/s13421-021-01263-1>
- Kim, S., Sodian, B., Paulus, M., Senju, A., Okuno, A., Ueno, M., Itakura, S., & Proust, J. (2020). Metacognition and mindreading in young children: A cross-cultural study. *Consciousness and Cognition*, 85. <https://doi.org/10.1016/j.concog.2020.103017>
- Koriat, A., & Ackerman, R. (2010). Choice latency as a cue for children's subjective confidence in the correctness of their answers. *Developmental Science*, 13(3), 441–453. <https://doi.org/10.1111/j.1467-7687.2009.00907.x>
- Lecce, S., Zocchi, S., Pagnin, A., Palladino, P., & Taumoepeau, M. (2010). Reading Minds: The Relation Between Children's Mental State Knowledge and Their Metaknowledge About Reading. *Child Development*, 81(6), 1876–1893. <https://doi.org/10.1111/j.1467-8624.2010.01516.x>
- Lipko, A. R., Dunlosky, J., Lipowski, S. L., & Merriman, W. E. (2012). Young Children are not Underconfident With Practice: The Benefit of Ignoring a Fallible Memory Heuristic. *Journal of Cognition and Development*, 13(2), 174–188. <https://doi.org/10.1080/15248372.2011.577760>
- 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. <https://doi.org/10.1016/j.jecp.2008.10.002>
- Lockl, K., & Schneider, W. (2006). Precursors of metamemory in young children: The role of theory of mind and metacognitive vocabulary. *Metacognition and Learning*, 1, 15–31. <https://doi.org/10.1007/s11409-006-6585-9>
- Lockl, K., & Schneider, W. (2007). Knowledge About the Mind: Links Between Theory of

Mind and Later Metamemory. *Child Development*, 78(1), 148–167.

<https://doi.org/10.1111/j.1467-8624.2007.00990.x>

Lundeberg, M. A., Fox, P. W., Brown, A. C., & Elbedour, S. (2000). Cultural influences on confidence: Country and gender. *Journal of Educational Psychology*, 92(1), 152–159.

<https://doi.org/10.1037/0022-0663.92.1.152>

Morony, S., Kleitman, S., Lee, Y. P., & Stankov, L. (2013). Predicting achievement: Confidence vs self-efficacy, anxiety, and self-concept in Confucian and European countries. *International Journal of Educational Research*, 58, 79–96.

<https://doi.org/10.1016/j.ijer.2012.11.002>

Muthukrishna, M., Henrich, J., Toyokawa, W., Hamamura, T., Kameda, T., & Heine, S. J. (2018). Overconfidence is universal? Elicitation of genuine overconfidence (EGO) procedure reveals systematic differences across domain, task knowledge, and incentives in four populations. *PLOS ONE*, 13(8), 1–30.

<https://doi.org/10.1371/journal.pone.0202288>

Nelson, K. (1996). *Language in cognitive development*. Cambridge University Press.

Nelson, T. O., & Narens, L. (1990). Metamemory: A Theoretical Framework and New Findings. *Psychology of Learning and Motivation - Advances in Research and Theory*, 26, 125–173.

Nielsen, M., Haun, D., Kärtner, J., & Legare, C. H. (2017). The persistent sampling bias in developmental psychology: A call to action. *Journal of Experimental Child Psychology*, 162, 31–38. <https://doi.org/10.1016/j.jecp.2017.04.017>

O’Leary, A. P., & Sloutsky, V. M. (2017). Carving Metacognition at Its Joints: Protracted Development of Component Processes. *Child Development*, 88(3), 1015–1032.

<https://doi.org/10.1111/cdev.12644>

- OECD. (2018). *The resilience of students with an immigrant background: Factors that shape well-being*. OECD Publishing. <https://doi.org/10.1787/9789264292093-en>
- Oudman, S., van de Pol, J., & van Gog, T. (2022). Effects of self-scoring their math problem solutions on primary school students' monitoring and regulation. *Metacognition and Learning, 17*(1), 213–239. <https://doi.org/10.1007/s11409-021-09281-9>
- Roebbers, C. M. (2017). Executive function and metacognition: Towards a unifying framework of cognitive self-regulation. *Developmental Review, 45*, 31–51. <https://doi.org/10.1016/j.dr.2017.04.001>
- Roebbers, C. M., Mayer, B., Steiner, M., Bayard, N. S., & van Loon, M. H. (2019). The role of children's metacognitive experiences for cue utilization and monitoring accuracy: A longitudinal study. *Developmental Psychology, 55*(10), 2077–2089. <https://doi.org/10.1037/dev0000776>
- Rogoff, B. (1994). Developing Understanding of the Idea of Communities of Learners. *Mind, Culture, and Activity, 1*(4), 209–229. <https://www.tandfonline.com/doi/epdf/10.1080/10749039409524673?needAccess=true&role=button>
- Ruffman, T., Slade, L., & Crowe, E. (2002). The Relation between Children's and Mothers' Mental State Language and Theory-of-Mind Understanding. *Child Development, 73*(3), 734–751. <https://doi.org/10.1111/1467-8624.00435>
- Salonen, P., Vauras, M., & Efklides, A. (2005). Social interaction - What can it tell us about metacognition and coregulation in learning? *European Psychologist, 10*(3), 199–208. <https://doi.org/10.1027/1016-9040.10.3.199>
- Schneider, W., & Lockl, K. (2002). The development of metacognitive knowledge in children and adolescents. In T. J. Perfect & B. L. Schwartz (Eds.), *Applied Metacognition* (pp.

224–257). Cambridge University Press.

Schraw, G., Crippen, K. J., & Hartley, K. (2006). Promoting Self-Regulation in Science Education: Metacognition as Part of a Broader Perspective on Learning. *Research in Science Education, 36*, 111–139. <https://doi.org/10.1007/s11165-005-3917-8>

Shin, H. E., Bjorklund, D. F., & Beck, E. F. (2007). The adaptive nature of children's overestimation in a strategic memory task. *Cognitive Development, 22*(2), 197–212. <https://doi.org/10.1016/J.COGDEV.2006.10.001>

Stankov, L., & Lee, J. (2014). Overconfidence Across World Regions. *Journal of Cross-Cultural Psychology, 45*(5), 821–837. <https://doi.org/10.1177/0022022114527345>

Stankov, L., Morony, S., & Lee, Y. P. (2014). Confidence: The best non-cognitive predictor of academic achievement? *Educational Psychology, 34*(1), 9–28. <https://doi.org/10.1080/01443410.2013.814194>

Tenney, E. R., MacCoun, R. J., Spellman, B. A., & Hastie, R. (2007). Calibration trumps confidence as a basis for witness credibility. *Psychological Science, 18*(1), 46–50. <https://doi.org/10.1111/j.1467-9280.2007.01847.x>

Thompson, R. B., & Foster, B. J. (2014). Socioeconomic Status and Parent-Child Relationships Predict Metacognitive Questions to Preschoolers. *Journal of Psycholinguistic Research, 43*(4), 315–333. <https://doi.org/10.1007/s10936-013-9256-4>

Tomassello, M. (2006). *Origins of human communication*. MIT Press.

van der Plas, E., Zhang, S., Dong, K., Bang, D., Li, J., Wright, N. D., & Fleming, S. M. (2022). Identifying Cultural Differences in Metacognition. *Journal of Experimental Psychology: General, 151*(12), 3268–3280. <https://doi.org/10.1037/xge0001209>

van Loon, M. H., Destan, N., Spiess, M. A., de Bruin, A., & Roebers, C. M. (2017). Developmental progression in performance evaluations: Effects of children's cue-

utilization and self-protection. *Learning and Instruction*, 51, 47–60.

<https://doi.org/10.1016/j.learninstruc.2016.11.011>

van Loon, M. H., & Roebbers, C. M. (2017). Effects of Feedback on Self-Evaluations and Self-Regulation in Elementary School. *Applied Cognitive Psychology*, 31(5), 508–519.

<https://doi.org/10.1002/acp.3347>

van Loon, M. H., & Roebbers, C. M. (2020). Using feedback to improve monitoring judgment accuracy in kindergarten children. *Early Childhood Research Quarterly*, 53, 301–313.

<https://doi.org/10.1016/j.ecresq.2020.05.007>

van Loon, M. H., & Roebbers, C. M. (2021). Using Feedback to Support Children when Monitoring and Controlling Their Learning. In D. Moraitou & P. Metallidou (Eds.), *Trends and Prospects in Metacognition Research across the Life Span. A Tribute to Anastasia Efklides*. (pp. 161–184). Springer.

https://doi.org/10.1007/978-3-030-51673-4_8

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.

Vygotsky, L. S. (1987). *The collected works of L. S. Vygotsky: Vol. 1. Problems of general psychology*. Plenum Press.

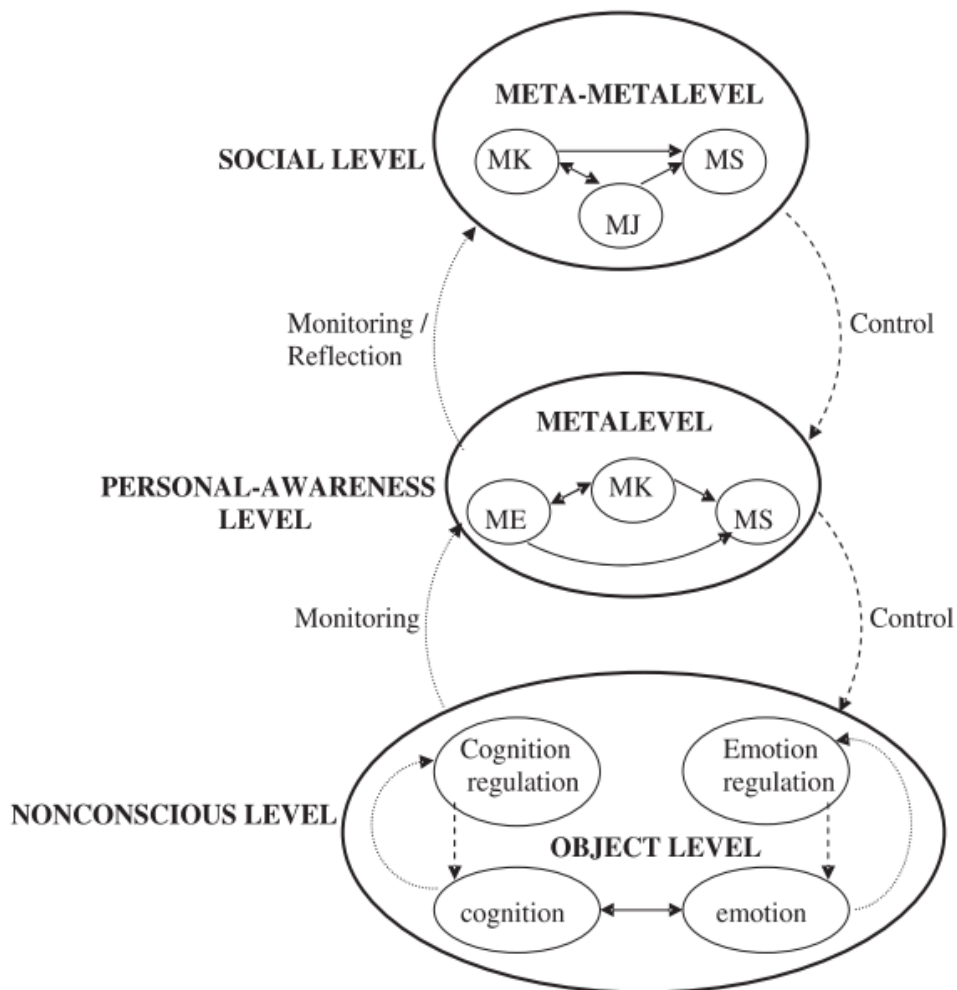
Wang, Y., & Sperling, R. A. (2021). Understanding and supporting Chinese middle Schoolers' monitoring accuracy in mathematics. *Metacognition and Learning*, 16(1), 57–88. <https://doi.org/10.1007/s11409-020-09238-4>

Xia, M., Poorthuis, A. M. G., & Thomaes, S. (2023). Why do young children overestimate their task performance? A cross-cultural experiment. *Journal of Experimental Child Psychology*, 226. <https://doi.org/10.1016/j.jecp.2022.105551>

Xia, M., Poorthuis, A. M. G., Zhou, Q., & Thomaes, S. (2022). Young children's

overestimation of performance: A cross-cultural comparison. *Child Development*, 93(2), e207–e221. <https://doi.org/10.1111/cdev.13709>

7 Appendix A



Note: ME = metacognitive experiences; MK = metacognitive knowledge;
 MS = metacognitive skills; MJ = metacognitive judgments
 ←·········· = Monitoring; ←----- = Control

Note. The multifaceted and multilevel model of metacognition. Figure copied from Efklides (2008).

8 Manuscripts

8.1 Study 1

Buehler, F. J., van Loon, M. H., Bayard, N. S., Steiner, M., & Roebbers, C. M. (2021).

Comparing metacognitive monitoring between native and non-native speaking primary school students. *Metacognition and Learning*, 16(3). <https://doi.org/10.1007/s11409-021-09261-z>

**Comparing metacognitive monitoring between native and non-native speaking
primary school students**

Florian J. Buehler¹, Mariëtte H. van Loon¹, Nathalie S. Bayard¹, Martina Steiner¹, & Claudia
M. Roebers¹

¹University of Bern, Switzerland

Abstract

Metacognitive monitoring is a significant predictor of academic achievement and is assumed to be related to language competencies. Hence, it may explain academic performance differences between native and non-native speaking students. We compared metacognitive monitoring (in terms of resolution) between native and non-native speaking fourth graders (~10 year olds) in two studies. In Study 1, we matched 30 native and 30 non-native speakers and assessed their monitoring in the context of a paired-associates task, including a recognition test and confidence judgements. Study 1 revealed that recognition and monitoring did not differ between native and non-native speaking children. In Study 2, we matched 36 native and 36 non-native speakers and assessed their monitoring with the same paired-associates task. Additionally, we included a text comprehension task with open-ended questions and confidence judgments. We replicated the findings of Study 1, suggesting that recognition and monitoring do not necessarily differ between native and non-native speakers. However, native speaking students answered more open-ended questions correctly than non-native speaking students did. Nevertheless, the two groups did not differ in monitoring their answers to open-ended questions. Our results indicate that native and non-native speaking children may monitor their metacognitive resolution equally, independent of task performance and characteristics. In conclusion, metacognitive monitoring deficits may not be the primary source of the academic performance differences between native and non-native speaking students.

Keywords: Metacognition, monitoring, language, non-native speakers, paired-associates task, text comprehension

Comparing metacognitive monitoring between native and non-native speaking primary school students

Can you imagine following a mathematics class in a foreign language when you were as young as ten years old? Across the world, many non-native speaking schoolchildren face such challenges every day. They get instructions in a language that they do not speak at home. Compared to their native speaking peers, this is likely to be an extra challenge for their learning. Not surprisingly, international studies show that non-native speaking children typically underperform in school subjects, such as reading, mathematics, and science (OECD, 2012, 2018). Although non-native speaking children build a substantial and growing population in countries of the Organisation for Economic Cooperation and Development (OECD, 2019), only very little is known about the mechanisms underlying their often observed underachievement. In this contribution, we focus on one consistent predictor of school achievement in primary school children, which is metacognitive monitoring (Freeman et al., 2017; Roebers et al., 2014), describing the ability to evaluate one's ongoing cognitive processes (Nelson & Narens, 1990; Schneider & Löffler, 2016). From a theoretical perspective, monitoring is likely related to language competencies (Ebert, 2015). Therefore, we aim to explore how language competencies are related to children's monitoring and whether monitoring differences between native and non-native speaking students may contribute to performance differences in a learning task.

Large-scale assessments such as The Programme for International Student Assessment (PISA) conducted by the OECD reveal that speaking the language of instruction at home is related to the prospect of reaching the baseline level of proficiency in the three main PISA subjects: reading, mathematics, and science (OECD, 2012, 2018). The largest differences in favour of native speaking children are typically reported in reading performance. Unlike non-native speakers, native speaking children hear and speak the language of instruction at home, from which their language skills are likely to benefit. Indeed,

speaking the language of instruction at home is strongly related to reading performance (OECD, 2012). In other words, children's linguistic environment at home seems to be vital for academic achievement and, hence, non-native speaking children may be disadvantaged in school. Towards the end of primary education, aspects of self-regulated learning, such as metacognitive monitoring skills, become increasingly important and may contribute to achievement gaps between native and non-native speaking students.

Metacognition consists of declarative (knowledge about the importance of person, task, and strategy variables for cognition; Flavell & Wellman, 1977) and procedural aspects (monitoring and regulation of memory processes; Schneider & Löffler, 2016). Declarative metacognition and self-regulated learning strategies are related to academic achievement in school aged children (Artelt et al., 2001; Schneider & Artelt, 2010; Veenman et al., 2005). Procedural metacognitive abilities were consistently found to be related to academic achievement in primary and secondary school children (Dunlosky & Metcalfe, 2009; Freeman et al., 2017; Kleitman & Gibson, 2011; Roebers et al., 2014; Stankov et al., 2012, 2014). In a very recent study, second and fourth graders' inaccurate metacognitive monitoring played a key role in understanding ineffective self-regulated learning strategies (Bayard et al., 2021). One can therefore assume that test performance in a memory task is directly related to metacognitive monitoring and control processes in primary school children (Roebers et al., 2014). Against this background, we hypothesize that differences in monitoring abilities contribute to non-native speaking children's underperformance.

Monitoring is typically assessed by asking individuals to give confidence judgments concerning their answers and relating these to actual task performance (Dunlosky et al., 2016). Within the literature, there are different approaches to quantify monitoring. The present contribution focuses on two complement measures of monitoring resolution, targeting the ability to metacognitively distinguish between correct and incorrect answers across items (Dunlosky et al., 2016; Dunlosky & Thiede, 2013). This implies that an individual gives

higher confidence judgments for answers that turn out to be correct than for those answers that turn out to be incorrect. Especially in educational and developmental contexts, monitoring resolution measures are considered to provide the most valuable insights into children's challenges when monitoring (for a review see Roebers, 2017; Schneider & Löffler, 2016).

Crucial for non-native speakers is the fact that metacognitive monitoring is typically assessed verbally (“*How sure are you that you answered this question correctly?*”). This is of high relevance as language is probably an essential variable in developing children's knowledge about mental processes (Ebert, 2015). It provides a means to think, talk, and learn about mental states and processes (Astington & Baird, 2005; Harris et al., 2005). Ebert (2015) outlined several theoretical reasons why language features might be associated with metacognition. A grammatical understanding may support children to represent mental states. Moreover, acquiring mental words may also foster a conceptual understanding, facilitating learning about unobservable cognitive processes, such as metacognitive monitoring. Finally, language skills may facilitate verbal interactions with other individuals and foster learning about the mental world, including metacognition. Against this theoretical background, it is not surprising that early language competencies were related to later metacognition (Annevirta et al., 2007; Ebert, 2015; Lecce et al., 2010; Lockl & Schneider, 2007). In conclusion, metacognitive abilities might vary across individuals with different language backgrounds and skills, such as native and non-native speaking children.

In contrast, research assessing *multilinguals* suggests an advantage for multilinguals in higher order cognitions. Meta-analyses found advantages for bilinguals in metalinguistic and metacognitive awareness, working memory, abstract and symbolic representation, attentional control, and problem solving (Adesope et al., 2010; Grundy & Timmer, 2017). Interestingly, bilinguals outperformed monolinguals in verbal and non-verbal executive function tasks, suggesting a general bilingual advantage in working memory tasks (Grundy &

Timmer, 2017). Note that bilinguals in those studies were identified as being equally (or almost equally) proficient in two languages (Adesope et al., 2010; Grundy & Timmer, 2017). In comparison, non-native speakers are second language learners in the language of instruction. It may be that speaking various languages is potentially beneficial for higher order cognition, but this is likely the case *only* when one master those languages on a proficient level. Therefore, it remains unclear whether the bilingual advantage in higher order cognitions applies to non-native speakers.

The present study

The present study focuses on metacognitive monitoring in native and non-native speaking 4th graders. This age range appears especially important as children soon face the transition into secondary education. We assessed the participants with a paired-associates task. With this task, we are avoiding effects of prior knowledge (cf. Destan et al., 2014; Roderer & Roebbers, 2010), and we can surely expect a sufficiently developed ability to metacognitive discriminate between likely correct and potentially incorrect responses. This also brings about the advantage that the participants are free to remember the content (pictures) in any language. Therefore, instruction's language should not be of high relevance for first-order performance (recognition) in the paired-associates task. Based on international assessments investigating academic achievement (OECD, 2012, 2018), we would expect first-order performance differences in favour of native speaking students, but due to the paired-associates task characteristic's such differences should not be pronounced. The scarce literature does not allow us to predict whether native and non-native speakers differ in metacognitive monitoring. Language abilities and monitoring are likely related (Ebert, 2015). Based on the findings that non-native speakers are disadvantaged in language competencies (OECD, 2012, 2018), one can expect that non-native speaking children's monitoring skills are inferior compared to native speaking children. The bilingual advantage claims that speaking multiple languages is beneficial for higher order cognitions, such as metalinguistic and

metacognitive awareness (Adesope et al., 2010; Grundy & Timmer, 2017). However, -other than bilinguals- non-native speakers are not proficient in multiple languages. Finally, the specific link between abilities in the language of instruction and monitoring was not yet investigated. We took an explorative approach to examine whether native and non-native speaking children differ in their monitoring abilities -over and above the to-be-expected performance differences. A better understanding of underlying mechanisms of monitoring might contribute to a better understanding of disadvantaged students, such as non-native speaking children (Freeman et al., 2017; Kleitman & Gibson, 2011; OECD, 2018; Roebers et al., 2014; Stankov et al., 2012, 2014).

Method - Study 1

Participants

Participants were 133 4th grade children. We recruited them from public schools in the vicinity of a mid-sized university town. Parents had signed informed consent, and children gave consent verbally before testing. Based on teachers' information, we excluded four participants with pathologies such as autism spectrum disorder or ADHD. Furthermore, we excluded three children due to technical issues and one child that broke off the task. Finally, we excluded three participants due to ceiling effects (recognition score = 100%) and two participants due to floor effects (recognition score at chance level $\leq 25\%$) in the recognition task.

To build groups of native and non-native speaking children, we asked teachers to indicate each student's mother tongue(s). Teachers retrieved such information from official documents, including demographic information about their students or by asking the students themselves. Our remaining sample consisted of 68 native speaking students, 34 non-native speaking students, and 18 multilinguals (i.e., children who speak more than one language on a native level). Multilingual children cannot be allocated to one of the two groups, as it remains

unclear whether their abilities in the language of instruction match the native or the non-native speaking group. Furthermore, the small number of multilinguals did not allow further analyses; hence, we excluded them from our analyses ($n = 18$). To ensure comparability of the two differently sized groups (68 native speakers vs. 34 non-native speakers), we matched each non-native speaking student with a native speaking peer. Non-native speaking students were individually matched to native speaking students by age (tolerance = 3 months) and gender. We could not match four non-native speaking children as their age exceeded that of any native speaking peer, therefore we excluded them. The matching led to two comparable (considering age and gender) and equally sized groups of native ($n = 30$; $M_{age} = 10.79y$; $SD_{age} = 5.73m$; 47% girls) and non-native speaking ($n = 30$; $M_{age} = 10.76y$; $SD_{age} = 5.62m$; 47% girls) participants. All native speaking children spoke German as their mother tongue. The mother tongues of the non-native speakers are indicated in Table 1.

Procedure and Materials

We conducted the study following the declaration of Helsinki. The local ethics committee (Faculty of Humanities of the University of Bern; approval number: 2016-08-00004) approved the study's procedure. We conducted a group assessment in the usual classroom setting. Two trained investigators supervised children within a class. The task was presented on a tablet computer (11.6") with a touch screen. We gave general instructions in German at the start. During the task, further instructions were given orally via headphones and visually as text on the screen (both in German). Before starting the task, children completed a practice trial to familiarize themselves with the material and the test format. The task was organized in 3 phases: The study phase, recognition, and monitoring phase (Figure 1). The task lasted approximately 30 minutes.

Study phase (Kanjis)

In the study phase, the subjects were told to remember 16 pairs of pictures and that they will be asked to recognize those pairs later in the task. The pairs were presented in random order and composed of a Kanji (a Japanese character) and its depicted meaning (a colour drawing). Each pair appeared for 5 seconds. After the study phase, subjects conducted a filler task (1 min) to prevent rehearsal and other memory strategies. The filler task consisted of an easy mouse-catching game on the tablet. The children steered a cat with one finger and tried to catch a mouse.

We piloted a large pool of item pairs beforehand to ensure sufficient variability concerning item difficulty. We included pairs with a difficulty index between .11 and .78 in the present study (Moosbrugger & Kelava, 2008). The identical task was applied in previous research (Destan et al., 2014; Destan & Roebbers, 2015).

Recognition (Kanjis)

In the recognition test, one Kanji at a time was depicted with four alternative drawings. Children were instructed to select the best alternative of the four by touching the respective drawing. All alternatives had been presented during the study time and were thus familiar to the children. A forced report selection characterized the recognition phase. When children were unsure about the correct answer, they were still asked to choose one of the four pictures. After selecting, a red frame surrounded the selected drawing. No feedback about their recognition performance was provided.

Recognition monitoring

Immediately after selecting an alternative in the recognition test, the monitoring judgment (confidence judgment) for this particular trial was collected. Children were asked: “*How sure are you that you have chosen the correct answer?*” They had to indicate their confidence on a 7-point Likert scale, presented as a thermometer, by touching the thermometer’s respective colour with their index finger. The thermometer ranged from blue

(indicating “unsure”) to red (indicating “very sure”; adapted from Koriat & Shitzer-Reichert, 2002). Children received detailed instructions, and they practiced with items before starting with the task.

Measures

For recognition, we computed the mean percentage of correctly recognized Kanjis out of the 16 to-be-remembered pairs for each participant. For metacognitive monitoring, we coded the thermometer’s confidence judgments as values ranging from 1 (very unsure) to 7 (very sure). There are many monitoring measures, and each has its strengths and weaknesses. We were specifically interested in a child-appropriate and classical monitoring measure that relates confidence judgments to item level performance, as research shows that primary school children, especially progress in their monitoring of incorrect answers (Howie & Roebbers, 2007; Roebbers et al., 2007). Thus, we focused on two resolution measures, which allow to take a differentiated perspective on confidence judgments, by contrasting judgments concerning correct and incorrect answers: (a) metacognitive discrimination primarily targeting children’s growing ability to experience and report different degrees of confidence on the continuum of confident-unconfident; (b) intra-individual Gamma correlation between recognition accuracy and the reported confidence on item level. Besides the slightly different perspective on participants monitoring resolution, the inclusion of multiple measures enables to evaluate whether they converge on the same qualitative outcomes (Dunlosky et al., 2016; Dunlosky & Thiede, 2013; Murayama et al., 2014).

For the discrimination score, we subtracted mean confidence judgments for incorrectly recognized Kanjis from mean confidence judgments for correctly recognized Kanjis (cf. Dunlosky & Thiede, 2013; Roebbers, 2002). Positive discrimination values indicate that children were reliably more confident when their answer was correct than when it was incorrect and could experience different degrees of confidence. We also computed intra-

individual Gamma correlations (Nelson, 1984) between confidence judgements and recognition (correct vs. incorrect). Gamma correlations closer to 1 indicate a more proficient monitoring resolution, whereas values closer to 0 indicate lower monitoring resolution.

Although Gamma correlations are the most frequently reported measure in metacognitive research, including students and adults (Dunlosky et al., 2016), they bear some disadvantages when used for children's data (see Roebbers & Spiess, 2017).

Analyses

We conducted a multivariate ANOVA with mother tongue (native vs. non-native speaking) as a grouping variable and recognition, the monitoring discrimination score, and Gamma correlations as dependent variables to test for group differences. We included all variables in one model to control for multiple comparisons issues.

Results – Study 1

Means for native and non-native speaking subjects are displayed in Table 2. Descriptive statistics revealed that native and non-native speaking students recognized a similar amount of Kanjis correctly. Both native, $F(1, 29) = 78.76; p < .01; \eta_p^2 = 0.73$, and non-native speakers, $F(1, 29) = 37.84; p < .01; \eta_p^2 = 0.57$, were more confident in correct recognitions compared to incorrect recognitions. Moreover, Gamma correlations between confidence judgments and recognition performance were significantly different from zero in both groups (Table 2).

A multivariate ANOVA including recognition, monitoring discrimination scores, and Gamma correlations as dependent variables with use of Pillai's trace did not show significant group differences between native and non-native speaking students, $F(3, 56) = 1.05; p = .38; \eta_p^2 = 0.05$. In sum, native and non-native speaking children did not differ significantly in recognition performance or monitoring resolution in the paired-associates task.

Discussion - Study 1

In Study 1, we investigated metacognitive monitoring of native and non-native speaking 4th graders in a paired-associates task (Kanjis). Based on the scarce literature, we took an explorative approach. Results revealed that native and non-native speaking subjects did not differ in the number of correctly recognized Kanjis. Based on confidence judgements, we computed discrimination scores and Gamma correlations as measures of monitoring resolution. Native and non-native speaking children adequately discriminated between correctly and incorrectly answered items, as indicated by both the discrimination scores and Gamma correlations. Most importantly, we did not find any differences between native and non-native speaking children in either of the two monitoring resolution measures. In other words, native and non-native speaking children monitored their recognition in the Kanji task equally well.

Unlike PISA studies, we did not find performance disadvantages for non-native speakers (OECD, 2012, 2018). We assessed subjects with a paired-associates task, which may be considered a language-reduced task, as the material was presented in the form of images. Thus, recognition performance might be independent of children's competencies in the language of instruction. As recognition performance was comparable between the two groups, the included monitoring measures are likely to have estimated children's monitoring skills about equally accurate (cf. Galvin et al., 2003; Maniscalco & Lau, 2012; Roebbers & Spiess, 2017). These aspects together might explain why we did not find any differences between language groups. Regarding metacognitive monitoring, our findings may indicate that advantages in higher order cognitions do not occur simply through exposure to multiple languages and, thus, do not necessarily emerge when comparing native with non-native speakers. The level of mastery of those languages may be crucial for benefits in higher order cognitions. Therefore, it may be that only those who speak various languages at a proficient

level –such as true bilinguals- benefit. Based on the assumption that native and non-native speakers differ in language competences, and based on previous research suggesting a theoretical and an empirical link between metacognition and language abilities (Annevirta et al., 2007; Ebert, 2015; Lecce et al., 2010; Lockl & Schneider, 2007), our results warrants replication in a more language-related task.

Therefore, we conducted a second study with an independent sample, for which we assessed children's monitoring resolution with the same paired-associates task (Kanjis) as in Study 1 and a text comprehension task. This allowed estimating the influence of a language-based task on non-native speaking children's monitoring. In contrast, to study 1, we assessed participants' abilities in the language of instruction to evaluate individual differences in language competencies between native and non-native speaking subjects. For the Kanji task, we expected to replicate the findings of Study 1, such that native and non-native speaking children would not differ in recognition performance and metacognitive monitoring resolution. We expected that native speaking students outperform non-native speaking students for the text comprehension task, as performance differences between native and non-native speakers are typically visible in language-related tasks (OECD, 2012). Regarding metacognitive monitoring, the text comprehension task's high linguistic demands may impair monitoring abilities of non-native speakers. However, it remains unclear whether monitoring competencies are affected by language abilities. Thus and again, we took an explorative approach for Study 2.

Method - Study 2

We draw the sample of Study 2 from a larger research project on children's developing metacognitive skills. Selected aspects of children's metacognitive development have been reported previously, such as recognition performance, confidence judgements (gamma correlations) and response latency (time taken for recognition and confidence

judgments in ms) for the Kanjis task and open question performance and confidence judgements (discrimination scores and gamma correlations) for the text comprehension task (Roebbers et al., 2019; Steiner et al., 2020). However, non-native speaking children were excluded in these previous reports. This manuscript's unique contribution is the focus on non-native children's monitoring, including a comparison with a subsample of native children from previous reports.

Participants

For Study 2, 151 4th graders participated. We recruited the children from public schools in the vicinity of a mid-sized university town. Parents had signed informed consent, and children gave verbal consent before testing. We excluded two participants with pathologies such as ADHD, relying on the teacher's information. As in Study 1, we excluded eight participants due to ceiling effects (recognition score = 100%) and one participant due to floor effects (recognition score at chance level $\leq 25\%$) in the recognition task, and 13 participants due to floor effects (no open question answered correctly) in the text comprehension task.

To build groups of native and non-native speaking children, we asked teachers to indicate each student's mother tongue(s). Teachers retrieved such information from official documents, including demographic information about their students or by asking the students themselves. Our sample consisted of 78 native speaking students, 38 non-native speaking students, and ten multilinguals. Multilingual children cannot be allocated to one of the two groups, as it remains unclear whether their abilities in the language of instruction match the native or the non-native speaking group. The small number of multilinguals does not allow further analyses; hence, we excluded them from our analyses ($n = 10$). To ensure comparability of the two differently sized groups (78 native speakers vs. 39 non-native speakers), we matched each non-native speaking student with a native speaking peer.

Matching was identical to Study 1. We could not match two non-native speaking children as their age exceeded that of any native speaking peer, therefore excluding them. The matching led to two comparable (considering age and gender) and equally sized groups of native ($n = 36$; $M_{age} = 10.10y$; $SD_{age} = 3.70m$; 44% girls) and non-native speaking ($n = 36$; $M_{age} = 10.14y$; $SD_{age} = 4.26m$; 44% girls) participants.

Furthermore, we asked teachers to rate participants' language competencies in school instruction language on a scale from 1 (below average) to 5 (very good). On average, teachers rated the language competences of their native speaking students ($M = 3.58$; $SD = 1.08$) higher than the language competences of their non-native speaking students ($M = 3.06$; $SD = 1.17$), $t(70) = 1.99$; $p = .05$. All native speaking children spoke German as their mother tongue. The mother tongues of the non-native speakers are indicated in Table 3.

Procedure and materials

Participants completed a paired-associates task (Kanjis identical to Study 1) and a text comprehension task. We conducted a group assessment in the usual classroom setting of the children. For testing, we split the classes into groups of 6 to 11 children, and two trained investigators supervised each group. One group started with the paired-associates task, whereas the other group started with the text comprehension task. We counterbalanced the task order. We gave General instructions orally in German at the start. Further instructions during the task were given orally via headphones and visually as text on the screen. The paired-associates task and the text comprehension task lasted approximately 30 minutes each. The materials and procedure of the paired-associates task were identical to Study 1 and are presented in Figure 1.

Text comprehension task

We gave general instructions orally in German at the start. During the task, the participants could read the instructions (in German), and they were repeated individually if

needed. The general instructions included the nature of all upcoming tests. The text comprehension task included 3 phases: a study phase (text reading), answering open-ended questions about the read texts, and a monitoring phase (Figure 2). Details about the task are reported by Steiner et al. (2020).

Study phase (Texts)

Students had to read six expository texts in German on a tablet (11.6"). Children could not move forward or backward between the texts. Study time was self-paced. However, the minimum reading duration was 10sec per text. The text font was Futura Std. Books and the size was 25 pts. Topics were animals (Bees, Bears, Dragonflies, and Camels), geographical subjects (Tropics, Desert, Egypt, Nile, Seasons, and Stars), or physiological processes (Catching a Cold, Chewing gum). Participants received the texts in random order.

We conducted a pilot study for choosing the texts and the open-ended questions for the present study. We translated and adapted the texts from previous studies (De Bruin et al., 2011; van Loon et al., 2015). We chose text-question sets that resulted in a similar amount of easy (~30%), medium (~40%), and difficult (~30%) open-ended questions. The mean length of the chosen texts was 126 words. The text's complexity was 37.81 LIX (readability index; see Björnsson, 1983), indicating that readability ranged between easy and moderate.

Open-ended questions

The text comprehension test consisted of 12 open-ended questions (2 questions per text) presented in a booklet as a paper-pencil test. Answers to open-ended questions could range from a single word to a full sentence (cf. Magliano et al., 2007). Hence, we included two kinds of open-ended questions to represent that question format fully. For each text, one of the open-ended questions required a single word ("At what time of the day are Pandas the most active?") and the other a sentence ("Why do bear loose hair during summer?").

Participants were encouraged to answer all open-ended questions. However, when they could not think of any answer, they could put a question mark instead.

Monitoring (of answers)

Immediately after answering each open-ended question, children had to rate their confidence that the answer was correct. Specifically, children were asked: “*How sure are you that your answer is correct?*” For that purpose –as in the paired-associates task- a 7-point Likert scale was presented to the right of each question. The same thermometer scale was depicted, ranging from very “unsure” to “very sure”.

Measures

We used identical measures for the paired-associates task as in Study 1. We computed recognition scores and two different monitoring resolution measures a discrimination score (difference in confidence between correctly and incorrectly recognized items) and intra-individual Gamma correlations between confidence judgments and recognition (see above).

Text comprehension performance

We coded answers to the open-ended questions as true (1) or false (0). In line with (van Loon et al., 2015), we emphasized comprehension during scoring. Thus, we scored verbatim responses as well as gist responses as correct. Two independent raters coded all answers. Interrater reliability was very high ($\kappa = 0.93$, $p < .001$).

We coded question marks as omissions (Roebbers et al., 2007). Native speaking subjects omitted 12.27% ($SD = 12.20$) and non-native speaking students 14.58% ($SD = 14$) of their answers, respectively. There was no missing data in the native speaking students and very little missing data in the non-native speaking students ($M = 0.23$; $SD = 1.39$). In the analyses, we included only completed test items because participants did not give confidence

judgements if they had not come up with an answer. For further analyses, we computed the percentage of correct answers out of all answered open-ended questions.

Text comprehension monitoring

We coded the thermometer's confidence judgments as values ranging from 1 (very unsure) to 7 (very sure). To assess text comprehension monitoring, we computed the same monitoring resolution measures as for the paired-associates task. Specifically, we subtracted mean confidence judgments for incorrectly answered open-ended questions from mean confidence judgments for correctly answered open-ended questions for a discrimination score (Dunlosky & Thiede, 2013; Roebbers, 2002). Moreover, we computed intra-individual Gamma correlations (Nelson, 1984) between confidence judgments and text comprehension (correct vs. incorrect).

Analyses

We conducted separate but identical analyses for the paired-associates task (identical analyses as in Study 1) and the text comprehension task. We conducted a multivariate ANOVA with mother tongue (native vs. non-native speaking) to test for group differences for the text comprehension task as a grouping variable and text comprehension, monitoring discrimination scores, and Gamma correlations as dependent variables. We included all variables in one model to control for multiple comparisons problems.

Results – Study 2

Means for native and non-native speaking participants are displayed in Table 2. Descriptive statistics for the paired-associates task revealed that native and non-native speaking students recognized a similar amount of Kanjis correctly. Both native, $F(1, 35) = 63.12; p < .01; \eta_p^2 = 0.64$, and non-native speakers $F(1, 35) = 29.75; p < .01; \eta_p^2 = 0.46$, were more confident in their recognition when their answers were correct compared to incorrect

recognitions. Moreover, Gamma correlations between confidence judgments and recognition performance were substantially different from zero in both groups (Table 2).

Descriptive statistics for the text comprehension task revealed that native speakers correctly answered more open-ended questions than non-native speakers. However, native, $F(1, 35) = 28.98; p < .01; \eta_p^2 = 0.45$, and non-native speakers, $F(1, 35) = 26.36; p < .01; \eta_p^2 = 0.43$, were more confident in their responses to the open-ended questions when their answers were correct compared to incorrect answers. Moreover, Gamma correlations between confidence judgments and text comprehension performance were substantial and significant in both groups (Table 2).

In a first step, we compared recognition and monitoring resolution abilities between native and non-native speaking children in the paired-associates task. We conducted a multivariate ANOVA with recognition, monitoring discrimination scores, and Gamma correlations as dependent variables. Using Pillai's trace, there was no significant group difference between native and non-native speaking students, $F(3, 68) = 0.4; p = .75; \eta_p^2 = 0.02$. Thus, as hypothesized and replicating findings from Study 1, native and non-native speaking children did not differ significantly in their metacognitive monitoring (resolution) and recognition in the paired-associates task.

We compared monitoring resolution between native and non-native speaking children for the text comprehension task in a second step. We conducted a multivariate ANOVA with text comprehension, monitoring discrimination scores, and Gamma correlations as dependent variables. Using Pillai's trace, there was a significant group difference between native and non-native speaking students, $F(3, 68) = 4.20; p < .01; \eta_p^2 = 0.16$. We followed up the multivariate ANOVA, with separate univariate tests on the dependent variables (text comprehension, monitoring discrimination scores, and Gamma correlations). Univariate tests revealed that native speakers answered significantly more open-ended questions correctly

than non-native speakers, $F(1, 70) = 11.36; p < .01; \eta_p^2 = 0.14$. Interestingly, univariate tests revealed neither group differences on the monitoring discrimination score, $F(1, 70) = 0.57; p = .45; \eta_p^2 = 0.01$, nor on Gamma correlations, $F(1, 70) = 0.45; p = .51; \eta_p^2 = 0.01$. In sum, native speaking children significantly outperformed non-native speaking children in terms of correctly answered questions. However, we did not find differences between native and non-native speakers in terms of monitoring resolution.

To gain further insights into how language competencies (teacher ratings) may explain performance and monitoring in native and non-native speakers, we computed non-parametric correlations. For native speakers, language competencies significantly correlated with performance in the text comprehension task ($r = .63; p < .01$). However, none of the other variables significantly correlated with language competencies. In other words, in both tasks (paired-associates and text comprehension), neither recognition nor monitoring resolution measures (discrimination scores and Gamma correlations) were related to language competencies in this subsample.

For non-native speakers, we found a different pattern of results. Language competencies were significantly correlated with both performance ($r = .47; p < .01$), and Gamma correlations ($r = .41; p < .05$) in the text comprehension task. Furthermore, we found marginal correlations between language competencies and monitoring discrimination scores in the paired-associates task ($r = .30; p = .07$), and in the text comprehension task ($r = .30; p = .07$). However, correlations between language abilities, recognition, and Gamma correlations in the paired-associates task were non-significant.

General discussion

In Study 2, we investigated metacognitive monitoring (resolution) of native and non-native speaking 4th graders in a paired-associates task (Kanjis) and a text comprehension task. The paired-associates task replicated the findings of Study 1. Native and non-native speaking

students did not differ in recognition and metacognitive monitoring resolution measures (discrimination scores and Gamma correlations) in the paired-associates task. In Study 2, in addition to the Kanji task used in Study 1, we included a text comprehension task. In line with our expectations, native speaking subjects answered more open-ended questions correctly than non-native speaking participants. However, native and non-native speakers did not differ in metacognitive monitoring resolution measures (discrimination scores and Gamma correlations) in the text comprehension task.

We did not find recognition differences between native and non-native speakers in the paired-associates task, but native speaking children outperformed non-native speakers in the text comprehension task. Those results align with our expectations, based on findings that performance differences are most significant in language related tasks (OECD, 2012). The included teacher ratings of language abilities confirmed that native speaking students had higher language abilities than non-native speaking students. This may be a relevant finding for future research investigating metacognition in children with various language backgrounds and abilities, as first-order task performance impacts metacognitive skills (Rinne & Mazzocco, 2014; Roebers & Spiess, 2017). Monitoring measures of native and non-native speakers may be more comparable in a paired-associates task than in a text comprehension task because first-order task performance is more similar in the paired-associates task than in the text comprehension task.

Most importantly, we did not find differences in metacognitive monitoring abilities (resolution) between native and non-native speaking students. Our results reveal that native and non-native speaking students do not differ in monitoring resolution and, hence, are both relatively well able to monitor their performance. Those findings contrast to research suggesting a multilingual advantage in higher order cognitions (Adesope et al., 2010; Grundy & Timmer, 2017). Note that the multilingual advantage occurs when the multilinguals are

assessed in their dominant language (Grundy & Timmer, 2017). We assessed all participants in the language of instruction, which was, by definition, not the dominant language of the non-native speakers. Furthermore, contrary to typical multilinguals, non-native speakers in the present study were not proficient in the instruction's language as native speakers. The bilingual advantage in higher order cognition may depend on language proficiency and the language of assessment. Therefore, it may be those non-native speakers would outperform their native speaking peers in monitoring abilities if (1) the non-native speakers would be multilingual (very proficient in more than one language) and/or (2) the non-native speakers would be assessed in their dominant language. In future research, it would be interesting to account for language proficiency and language of assessment to gain a more differentiated perspective of metacognitive monitoring in non-native speakers.

We followed up our analyses with correlations to gain more insight into the relationship between language competencies and metacognition. For the paired-associates task, our results reveal that the language abilities of non-native speakers are marginally associated with monitoring discrimination. However, we did not find any relations for the language abilities of native speakers in the paired-associates task. This finding suggests that monitoring and performance in the paired-associates task may be independent of one's language abilities. In contrast, we found significant correlations between native and non-native speakers' language abilities and performance in the text comprehension task. This is in line with research suggesting an impact of language abilities in language related tasks (OECD, 2012). Furthermore, monitoring resolution measures (discrimination scores and Gama correlations) of text comprehension were related to language abilities for non-native speakers, but not for native speakers. Our findings indicate that the relation between language abilities and metacognitive monitoring may be task- (language-reduced vs. language-based) and participant- (native vs. non-native speaking) specific.

Our results are partly in line with studies suggesting a link between language abilities and declarative metacognition (Annevirta et al., 2007; Ebert, 2015; Lecce et al., 2010; Lockl & Schneider, 2007). On the one hand, the language abilities of non-native speakers seem to be related to monitoring resolution measures in a text comprehension task. On the other hand, the present research suggests that metacognitive monitoring resolution does not necessarily differ between native and non-native speakers and, thus, monitoring abilities do not seem to be strongly affected by language competencies. This may implicate that metacognitive monitoring skills are relatively independent of the native language and the language of assessment. Once a child understood the instructions, it can ask himself how confident it is about a particular item in any language. It might be that language abilities are more closely related to declarative aspects than procedural aspects of metacognition. As Ebert (2015, p. 562) stated: «The most important variable in shaping children's knowledge about the mental world is probably language.» An interesting question for future research would be to clarify the role of language abilities for declarative and procedural aspects of metacognition.

Is it possible that the instruction language competencies are related to metacognitive abilities and thus explain performance differences? Language competencies seem to be related to metacognitive monitoring (resolution) for non-native speakers in a language related task (text comprehension). However, monitoring abilities do not seem to be generally impaired by this relationship, as indicated by similar monitoring resolution scores for native and non-native speakers. Metacognitive monitoring abilities may not be the primary source of performance differences between native and non-native speaking students. Still, metacognitive monitoring may be a valuable resource to address the performance gap among non-native speaking students. Accurate monitoring of one's task performance is an essential precondition for implementing successful control strategies, such as allocating learning time to perceived item difficulty (Destan et al., 2014; Schneider & Lockl, 2002; Schneider & Löffler, 2016). This enables an individual to learn efficiently and improve one's performance

(Dunlosky & Metcalfe, 2009). It would be interesting to assess metacognitive control processes in native and non-native speaking students in future research. This might contribute to further insights into metacognitive processes and how they are related to school performance of non-native speakers.

A strength of the present study is replicating the findings for Study 1 in a different sample in Study 2. We included a paired-associates and a text comprehension task, allowing us to take a distinguished perspective on monitoring differences between native and non-native speakers in different learning tasks. Furthermore, we made a first step connecting language abilities and procedural metacognition, a so far neglected topic in metacognition research. Despite the strengths, we need to acknowledge some limitations. We did not collect information about the socio economic status (SES) of the subjects. SES is a common confounding variable when addressing students' language skills (cf. Glick & Clark, 2012). We do not have detailed insights into how long children were used to following non-native language instructions. Therefore, it is challenging to account for individual differences in the non-native speaking group. We did not assess general cognitive abilities and could not control cognitive variables other than language when we matched the subjects. Children did not differ in performance in the paired-associates task, which may indicate similar cognitive abilities. Finally, our findings are limited to resolution measures of metacognitive monitoring in a paired-associates and a text comprehension task. To gain a more general perspective on monitoring abilities in native and non-native speakers, future research may include various measures of monitoring (e.g. resolution and calibration measures) in different cognitive tasks (e.g. recognition, text comprehension, free recall, perceptual tasks...).

Conclusion

We assessed metacognitive monitoring resolution of native and non-native speaking 4th graders in two similar yet independent samples. Our results are twofold. For one, we

showed that native speaking students outperformed their non-native speaking peers in a language related task (text comprehension) but not in a language reduced learning task (picture based paired-associates). This is in accordance with previous research, indicating that performance differences may be more pronounced in language related tasks (OECD, 2012). Most importantly, we did not find differences in metacognitive monitoring between native and non-native speaking children, independent of whether the task was language related or not. This suggests that metacognitive monitoring may not be the primary source of school performance differences between native and non-native speakers. Nevertheless, it might still be a valuable resource for non-native speaking students. Further research is needed to clarify the role of additional aspects of procedural metacognition in non-native speaking children's school performance.

References

- Adesope, O. O., Lavin, T., Thompson, T., & Ungerleider, C. (2010). A systematic review and meta-analysis of the cognitive correlates of bilingualism. *Review of Educational Research, 80*(2), 207–245. <https://doi.org/10.3102/0034654310368803>
- Annevirta, T., Laakkonen, E., Kinnunen, R., & Vauras, M. (2007). Developmental dynamics of metacognitive knowledge and text comprehension skill in the first primary school years. *Metacognition and Learning, 2*(1), 21–39. <https://doi.org/10.1007/s11409-007-9005-x>
- Artelt, C., Schiefele, U., & Schneider, W. (2001). Predictors of reading literacy. *European Journal of Psychology of Education, 16*(3), 363–383. <https://doi.org/10.1007/BF03173188>
- Astington, J. W., & Baird, J. A. (2005). *Why Language Matters for Theory of Mind*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195159912.001.0001>
- Bayard, N. S., van Loon, M. H., Steiner, M., & Roebbers, C. M. (2021). Developmental Improvements and Persisting Difficulties in Children’s Metacognitive Monitoring and Control Skills: Cross-Sectional and Longitudinal Perspectives. *Child Development, 92*(3), 1118–1136. <https://doi.org/10.1111/cdev.13486>
- Björnsson, C. H. (1983). Readability of Newspapers in 11 Languages. *Reading Research Quarterly, 18*(4), 480–497. <https://www.jstor.org/stable/747382>
- De Bruin, A. B. H., Thiede, K. W., Camp, G., & Redford, J. (2011). Generating keywords improves metacomprehension and self-regulation in elementary and middle school children. *Journal of Experimental Child Psychology, 109*(3), 294–310. <https://doi.org/10.1016/j.jecp.2011.02.005>
- Destan, N., Hembacher, E., Ghetti, S., & Roebbers, C. M. (2014). Early metacognitive abilities: The interplay of monitoring and control processes in 5- to 7-year-old children.

Journal of Experimental Child Psychology, 126, 213–228.

<https://doi.org/10.1016/j.jecp.2014.04.001>

Destan, N., & Roebbers, C. M. (2015). What are the metacognitive costs of young children's overconfidence? *Metacognition and Learning*, 10(3), 347–374.

<https://doi.org/10.1007/s11409-014-9133-z>

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition: A Textbook for Cognitive, Educational, Life Span & Applied Psychology*. Sage Publications.

Dunlosky, J., Mueller, M. L., & Thiede, K. W. (2016). Methodology for Investigating Human Metamemory: Problems and Pitfalls. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford Handbook of Metamemory*. Oxford University Press.

<https://doi.org/10.1093/oxfordhb/9780199336746.013.14>

Dunlosky, J., & Thiede, K. W. (2013). Four cornerstones of calibration research: Why understanding students' judgments can improve their achievement. *Learning and Instruction*, 24(1), 58–61. <https://doi.org/10.1016/j.learninstruc.2012.05.002>

Ebert, S. (2015). Longitudinal Relations Between Theory of Mind and Metacognition and the Impact of Language. *Journal of Cognition and Development*, 16(4), 559–586.

<https://doi.org/10.1080/15248372.2014.926272>

Flavell, J. H., & Wellman, H. M. (1977). Metamemory. In R. V. Kail & J. W. Hagen (Eds.), *Perspectives on the development of memory and cognition* (pp. 3–33). Erlbaum.

Freeman, E. E., Karayanidis, F., & Chalmers, K. A. (2017). Metacognitive monitoring of working memory performance and its relationship to academic achievement in Grade 4 children. *Learning and Individual Differences*, 57, 58–64.

<https://doi.org/10.1016/j.lindif.2017.06.003>

Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of

signal detectability: Discrimination between correct and incorrect decisions.

Psychonomic Bulletin and Review, 10(4), 843–876. <https://doi.org/10.3758/BF03196546>

Glick, J. E., & Clark, R. (2012). Cognitive Development and Family Resources Among Children of Immigrant Families. In R. King & V. Maholmes (Eds.), *The Oxford Handbook of Poverty and Child Development*. Oxford University Press.
<https://doi.org/10.1093/oxfordhb/9780199769100.013.0010>

Grundy, J. G., & Timmer, K. (2017). Bilingualism and working memory capacity: A comprehensive meta-analysis. *Second Language Research*, 33(3), 325–340.
<https://doi.org/10.1177/0267658316678286>

Harris, P. L., De Rosnay, M., & Pons, F. (2005). Language and Children's Understanding of Mental States. *Current Directions in Psychological Science*, 14(2), 69–73.
<https://doi.org/10.1111/j.0963-7214.2005.00337.x>

Howie, P., & Roebbers, C. M. (2007). Developmental Progression in the Confidence-Accuracy Relationship in Event Recall: Insights Provided by a Calibration Perspective. *Applied Cognitive Psychology*, 21, 871–893. <https://doi.org/10.1002/acp>

Kleitman, S., & Gibson, J. (2011). Metacognitive beliefs, self-confidence and primary learning environment of sixth grade students. *Learning and Individual Differences*, 21(6), 728–735. <https://doi.org/10.1016/j.lindif.2011.08.003>

Koriat, A., & Shitzer-Reichert, R. (2002). Metacognitive Judgments and their Accuracy. In P. Chambres, M. Izaute, & P.-J. Marescaux (Eds.), *Metacognition: Process, Function and Use* (pp. 1–17). Springer. https://doi.org/10.1007/978-1-4615-1099-4_1

Lecce, S., Zocchi, S., Pagnin, A., Palladino, P., & Taumoepeau, M. (2010). Reading Minds: The Relation Between Children's Mental State Knowledge and Their Metaknowledge About Reading. *Child Development*, 81(6), 1876–1893. <https://doi.org/10.1111/j.1467->

8624.2010.01516.x

Lockl, K., & Schneider, W. (2007). Knowledge About the Mind: Links Between Theory of Mind and Later Metamemory. *Child Development*, 78(1), 148–167.

<https://doi.org/10.1111/j.1467-8624.2007.00990.x>

Magliano, J. P., Millis, K., Ozurub, Y., & McNamara, D. S. (2007). A multidimensional framework to evaluate reading assessment tools. In D. S. McNamara (Ed.), *Reading Comprehension Strategies: Theory, Interventions, and Technologies* (pp. 107–136). Erlbaum.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, 21(1), 422–430. <https://doi.org/10.1016/j.concog.2011.09.021>

Moosbrugger, H., & Kelava, A. (2008). *Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien)* (2nd ed., pp. 7–26). Springer. https://doi.org/10.1007/978-3-540-71635-8_2

Murayama, K., Sakaki, M., Yan, V. X., & Smith, G. M. (2014). Type I error inflation in the traditional by-participant analysis to metamemory accuracy: A generalized mixed-effects model perspective. *Journal of Experimental Psychology: Learning Memory and Cognition*, 40(5), 1287–1306. <https://doi.org/10.1037/a0036914>

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, 95(1), 109–133.

<https://doi.org/10.1037/0033-2909.95.1.109>

Nelson, T. O., & Narens, L. (1990). Metamemory: A Theoretical Framework and New Findings. *Psychology of Learning and Motivation - Advances in Research and Theory*, 26, 125–173.

- OECD. (2012). *Untapped Skills: Realising The Potential Of Immigrant Students*. OECD Publishing. <https://doi.org/10.1787/9789264172470-en>
- OECD. (2018). *The resilience of students with an immigrant background: Factors that shape well-being*. OECD Publishing. <https://doi.org/10.1787/9789264292093-en>
- OECD. (2019). *PISA 2018 Results: Combined Executive Summaries*. OECD Publishing. https://www.oecd.org/pisa/Combined_Executive_Summaries_PISA_2018.pdf
- Rinne, L. F., & Mazzocco, M. M. M. (2014). Knowing Right from Wrong in Mental Arithmetic Judgments: Calibration of Confidence Predicts the Development of Accuracy. *PLOS ONE*, *9*(7), 1–11. <https://doi.org/10.1371/journal.pone.0098663>
- Roderer, T., & Roebbers, C. M. (2010). Explicit and implicit confidence judgments and developmental differences in metamemory: An eye-tracking approach. *Metacognition and Learning*, *5*(3), 229–250. <https://doi.org/10.1007/s11409-010-9059-z>
- Roebbers, C. M. (2002). Confidence judgments in children's and adults' event recall and suggestibility. *Developmental Psychology*, *38*(6), 1052–1067. <https://doi.org/10.1037/0012-1649.38.6.1052>
- Roebbers, C. M. (2017). Executive function and metacognition: Towards a unifying framework of cognitive self-regulation. *Developmental Review*, *45*, 31–51. <https://doi.org/10.1016/j.dr.2017.04.001>
- Roebbers, C. M., Krebs, S. S., & Roderer, T. (2014). Metacognitive monitoring and control in elementary school children: Their interrelations and their role for test performance. *Learning and Individual Differences*, *29*, 141–149. <https://doi.org/10.1016/j.lindif.2012.12.003>
- Roebbers, C. M., Mayer, B., Steiner, M., Bayard, N. S., & van Loon, M. H. (2019). The role of children's metacognitive experiences for cue utilization and monitoring accuracy: A

longitudinal study. *Developmental Psychology*, 55(10), 2077–2089.

<https://doi.org/10.1037/dev0000776>

Roebbers, C. M., & Spiess, M. (2017). The Development of Metacognitive Monitoring and Control in Second Graders: A Short-Term Longitudinal Study. *Journal of Cognition and Development*, 18(1), 110–128. <https://doi.org/10.1080/15248372.2016.1157079>

Roebbers, C. M., von der Linden, N., & Howie, P. (2007). Favourable and unfavourable conditions for children's confidence judgments. *British Journal of Developmental Psychology*, 25, 109–134. <https://doi.org/10.1348/026151006X104392>

Schneider, W., & Artelt, C. (2010). Metacognition and mathematics education. *ZDM - International Journal on Mathematics Education*, 42(2), 149–161.

<https://doi.org/10.1007/s11858-010-0240-2>

Schneider, W., & Lockl, K. (2002). The development of metacognitive knowledge in children and adolescents. In T. J. Perfect & B. L. Schwartz (Eds.), *Applied Metacognition* (pp. 224–257). Cambridge University Press.

Schneider, W., & Löffler, E. (2016). The Development of Metacognitive Knowledge in Children and Adolescents. In J. Dunlosky & S. U. K. Tauber (Eds.), *The Oxford Handbook of Metamemory*. <https://doi.org/10.1093/oxfordhb/9780199336746.013.10>

Stankov, L., Lee, J., Luo, W., & Hogan, D. J. (2012). Confidence: A better predictor of academic achievement than self-efficacy, self-concept and anxiety? *Learning and Individual Differences*, 22(6), 747–758. <https://doi.org/10.1016/j.lindif.2012.05.013>

Stankov, L., Morony, S., & Lee, Y. P. (2014). Confidence: The best non-cognitive predictor of academic achievement? *Educational Psychology*, 34(1), 9–28.

<https://doi.org/10.1080/01443410.2013.814194>

Steiner, M., van Loon, M. H., Bayard, N. S., & Roebbers, C. M. (2020). Development of

Children's monitoring and control when learning from texts: effects of age and test format. *Metacognition and Learning*, *15*, 3–27. <https://doi.org/10.1007/s11409-019-09208-5>

van Loon, M. H., Dunlosky, J., van Gog, T., van Merriënboer, J. J. G., & de Bruin, A. B. H. (2015). Refutations in science texts lead to hypercorrection of misconceptions held with high confidence. *Contemporary Educational Psychology*, *42*, 39–48. <https://doi.org/10.1016/j.cedpsych.2015.04.003>

Veenman, M. V. J., Kok, R., & Blöte, A. W. (2005). The relation between intellectual and metacognitive skills in early adolescence. *Instructional Science*, *33*(3), 193–211. <https://doi.org/10.1007/s11251-004-2274-8>

Table 1*Mother tongue of non-native speaking children, Study 1*

Language	<i>N</i>	%
Albanian	10	33.30
Kurdish	4	13.30
Serbian	2	6.70
Somali	2	6.70
Turkish	2	6.70
African Language (unknown)	1	3.30
Arabic	1	3.30
Croatian	1	3.30
Farsi	1	3.30
French	1	3.30
Hungarian	1	3.30
Polish	1	3.30
Portuguese	1	3.30
Tamil	1	3.30
Tigrinya	1	3.30
Total	30	100

Note. Teachers were asked to indicate the mother tongue of their students.

Table 2*Means of performance and monitoring measures in Study 1 and Study 2 (SD in parentheses)*

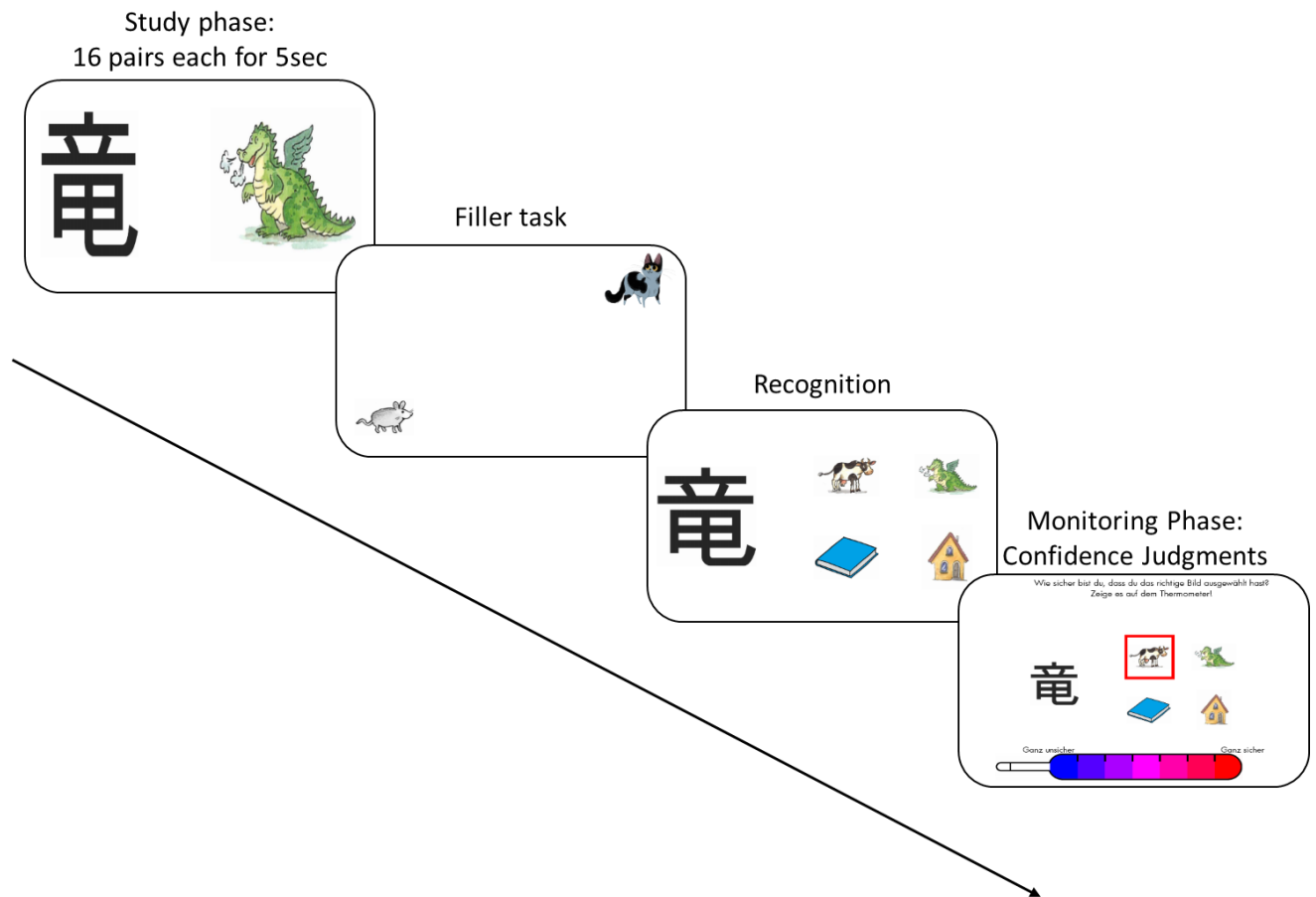
	Performance [%]	CJ correct response	CJ incorrect response	Monitoring Discrimination	Gammas
Study 1					
Paired-associates task					
Native speaking	60.21 (16.04)	5.20 (0.92)	3.72 (1.40)	1.49 (0.92)*	0.60 (0.43)**
Non-native speaking	58.33 (15.69)	5.29 (0.94)	4.23 (1.42)	1.06 (0.94)*	0.45 (0.38)**
Study 2					
Paired-associates task					
Native speaking	53.99 (16.17)	5.19 (1.03)	4.00 (1.30)	1.19 (0.90)*	0.50 (0.30)**
Non-native speaking	53.30 (14.21)	5.54 (0.81)	4.53 (1.37)	1.00 (1.10)*	0.40 (0.46)**
Study 2					
Text comprehension task					
Native speaking	54.40 (17.76) ⁺	5.27 (1.13)	4.23 (1.30)	1.05 (1.17)*	0.46 (0.50)**
Non-native speaking	38.43 (22.21)	5.06 (1.43)	3.78 (1.17)	1.29 (1.50)*	0.54 (0.47)**

Note. CJs were indicated on a 7-point Likert scale. CJ = Confidence Judgments, Monitoring Discrimination = CJ correct recognition – CJ incorrect recognition, Gammas = Intra-individual correlations between task performance and confidence judgments; * $p < .01$ native and non-native speaking participants gave significantly higher CJs when their response was correct vs. incorrect; ** $p < .001$ all Gammas were significantly different from zero; ⁺ $p < .01$ Native speakers answered significantly more open questions correctly than non-native speakers, in Study 2.

Table 3*Mother tongue of non-native speaking children, Study 2*

Language	<i>N</i>	%
Albanian	11	30.56
Italian	6	16.67
Tamil	3	8.33
French	3	8.33
Hungarian	3	8.33
Serbian	2	5.56
Spanish	2	5.56
Portuguese	2	5.56
Arabic	1	2.78
Croatian	1	2.78
Turkish	1	2.78
Urdu	1	2.78
Total	36	100

Note. Teachers were asked to indicate the mother tongue of their students.

Figure 1*Procedure of the paired-associates task*

Note. Procedure: After studying the 16 Kanji-picture pairs, children had to recognize for each Kanji the correct picture out of 4 options and provide confidence judgments.


Figure 2*Procedure of the text comprehension task*

Reading 6 Texts

Bärenfell

Alle Bären haben ein dickes Fell. Das dicke Fell schützt sie gut vor der Kälte. Es kann aber im Sommer zu warm sein. Deshalb verlieren viele Bären im Sommer einen Teil ihrer Haare.


Einige Bären haben ein Muster auf ihrem Gesicht und auf ihrer Brust. Der grosse Panda hat als einziger Bär ein Muster auf dem ganzen Körper. Sein schwarz-weisses Fell fällt im Bambuswald kaum auf. Dies ist vor allem am Abend so, wenn sich der Panda am meisten bewegt. Im Winter ist der Panda zwischen dem Schnee, den schwarzen Steinen und den Bäumen fast unsichtbar.



↓

12 Open-ended questions with Confidence Judgements


Bärenfell




Wieso verlieren viele Bären im Sommer ein Teil ihrer Haare? Wie sicher bist du dir, dass deine Antwort stimmt?

Zu welcher Tageszeit bewegen sich Pandas am meisten?

Ganz unsicher Ganz sicher



Ganz unsicher Ganz sicher



Note. Procedure: After reading each text, children had to answer open questions and provide confidence judgements. Figure adapted from Steiner et al. (2020).

8.2 Study 2

Buehler, F. J., Orth, U., Krauss, S., Roebbers, C. M. (2022). The Longitudinal Relation between Language Abilities and Metacognitive Monitoring: Structural Differences in Native and Non-native Speakers [Manuscript under review].

**The Longitudinal Relation between Language Abilities and Metacognitive
Monitoring: Structural Differences in Native and Non-native Speakers**

Florian J. Buehler¹, Ulrich Orth¹, Samantha Krauss¹, & Claudia M. Roebers¹

¹University of Bern, Switzerland

Abstract

We investigated the longitudinal relation between language abilities in kindergarten and metacognitive monitoring accuracy (the ability to evaluate ongoing cognitive processes accurately) in grade one. We analyzed data from the NEPS (National Educational Panel Study), a large-scale assessment conducted in Germany ($N = 9,159$). We examined cross-lagged panel models including receptive language abilities (vocabulary and grammar), monitoring accuracy (in a math and a science task), and first-order task performance as a control variable. All measurements were assessed in kindergarten and grade one. Cross-lagged paths revealed that earlier language abilities significantly predict later monitoring accuracy. Earlier monitoring accuracy did not predict later language abilities. Language abilities were measured in the instructional language, representing L1 language for native speakers but L2 language for the non-native speakers in the sample. Therefore, we tested whether the structural relations between language abilities and monitoring accuracy differed between native and non-native speakers. Multi-group analyses revealed that earlier language abilities predicted later monitoring accuracy in native speakers ($N = 6,399$) but not in non-native speakers ($N = 785$). Monitoring accuracy predicted the language ability of non-native but not of native speakers. Most interestingly, overconfidence in non-native speakers predicted their language abilities positively. Our results suggest that language abilities are a precursor of monitoring accuracy and that high confidence might be crucial for second language learning.

Keywords: monitoring accuracy, language abilities, native speakers, non-native speakers, task performance

The Longitudinal Relation between Language Abilities and Metacognitive Monitoring: Structural Differences in Native and Non-native Speakers

Language is a mean to think, talk and learn about mental processes (Astington & Baird, 2005; Harris et al., 2005). From a Vygotskian perspective, language is a tool for our cognition. Indeed, language is a consistent predictor for understanding mental states such as Theory of Mind and metacognition in children (Ebert, 2015, 2020; Gonzales et al., 2021; Lockl & Schneider, 2007; Milligan et al., 2007). However, little is known about the longitudinal relation of language abilities and procedural metacognition. The present research investigated the longitudinal relation between language abilities and metacognitive monitoring (the ability to evaluate ongoing cognitive processes; Nelson & Narens, 1990). Language abilities in the language of instruction differ between native and the many non-native speaking students in classrooms worldwide (OECD, 2018). Non-native speakers are instructed in their second language while they acquire their Theory of Mind and metacognitive abilities in their first language. Therefore, we investigated structural differences between language abilities and metacognitive monitoring in native and non-native speakers.

Understanding the structural relations between metacognitive monitoring and language is highly relevant because an accurate evaluation of one's cognitive processes is central for self-regulated learning, school performance, and lifelong learning (Bellon et al., 2021; Dunlosky & Metcalfe, 2009; Freeman et al., 2017; Roebbers, 2017; Roebbers et al., 2014; Schraw et al., 2006). Children have to learn to adapt to quickly changing environments, such as new teaching methods, technological tools, or a pandemic causing remote schooling. Language abilities may be vital for understanding and representing cognitive processes, such as accurate monitoring of ongoing cognitive activities (Astington & Baird, 2005; Harris et al., 2005). The relation between monitoring and language may also be bidirectional. Accurate

monitoring may be essential for language development and abilities. Indeed, research suggests that accurate monitoring is a driving force for the development of a broad range of cognitive abilities (Flavell, 1979; Flavell & Wellman, 1977; Kuhn, 2000; Rinne & Mazzocco, 2014).

Understanding the relationship between metacognitive monitoring and language is especially relevant for non-native speakers. They typically underperform in school and mostly in language-related tasks (OECD, 2018). Underperformance in language abilities may impair non-natives' participation in classroom talk, which is crucial for self-regulated and conceptual learning, such as metacognitive monitoring (e.g. Zepeda et al., 2019). Differential insights about language abilities and metacognitive monitoring for native and non-native speakers enhance our understanding of the first (L1) and second language abilities' (L2) role in metacognitive monitoring. Moreover, metacognitive monitoring may differently predict native (L1) and non-native language abilities (L2). This topic is highly relevant as the number of non-native speaking students in OECD countries' classrooms increases (OECD, 2019a). The present study offers a critical step towards understanding the longitudinal relation of language abilities and metacognitive monitoring in native and non-native speakers.

The relation between language abilities and metacognition

Not much is known about the relation between language and metacognitive monitoring. However, language abilities are a well-established predictor of Theory of Mind (ToM, Milligan et al., 2007), which is conceptually closely related to metacognition and a precursor of metacognition (Ebert, 2015; Lockl & Schneider, 2007; Schneider & Löffler, 2016). Building on this ToM research, three mechanisms may explain how language abilities contribute to the development of metacognitive monitoring (Astington & Baird, 2005; Ebert, 2015; Harris et al., 2005). First, acquiring metacognitive vocabulary (e.g., thinking, knowing, forgetting) facilitates the encoding of unobservable mental states and fosters a conceptual understanding. For instance, vocabulary for ambiguous mental states, such as higher or lower

confidence in a given answer, may enhance conceptual understanding and, thus, monitoring accuracy. Second, grammatical understanding is crucial to attribute and represent mental states. Understanding prepositions may be crucial to attribute different confidence levels to a given task performance. For instance, Kate is unsure whether the teacher announced the biology test for Friday or the following Monday. Third, language abilities enable children to communicate with various individuals and gain different perspectives on a topic. Talking and comparing their results in a given task with others may provide children with various solutions and perspectives on the same tasks, affecting their confidence in their response. Overall, language seems to be an essential tool for developing accurate metacognitive monitoring.

Few empirical studies have investigated the interplay between language abilities and metacognition. These studies indicate that language abilities in preschool contribute to later knowledge about the mind (declarative metacognition; Ebert, 2015; Lockl & Schneider, 2007; Schneider & Löffler, 2016). A recent study found that receptive grammar and vocabulary at three years predict metacognitive knowledge at nine years (Ebert, 2020). Most importantly, vocabulary in preschool predicted uncertainty monitoring in kindergarten (Gonzales et al., 2021). Other studies showed that earlier declarative metacognition in preschool and primary school was related to later language abilities (Annevirta et al., 2007; Lecce et al., 2010). Although it appears that language and metacognitions are related, the direction of the relationship is still not fully understood.

Another area of research focusing on (foreign) language learning rather than metacognition emphasizes the importance of metacognition for language abilities and language learning (Haukås et al., 2018; Wenden, 1998). Metacognitive knowledge was found to be a significant predictor of first (L1) and second language (L2) reading and writing proficiency in secondary school students (Schoonen et al., 2011; Van Gelderen et al., 2004).

In third graders, instructing metacognitive strategies enhanced vocabulary and reading comprehension (Boulware-Gooden et al., 2007). Metacognitive knowledge also predicts reading competence in many different languages (Artelt & Schneider, 2015). While the studies mentioned above targeted declarative metacognition, evidence concerning the more critical procedural metacognition for language learning is sparse (see Teng, 2019).

Native and non-native speakers

When investigating the role of language for higher-order cognitive processes such as metacognition, it is essential to differentiate between native and non-native speakers. For native speakers, the assessment language represents their first language (L1). Contrary, for non-native speakers, the assessment language represents their second language (L2). Semantics, syntax, and pragmatics are likely to differ in L1 and L2 (e.g. Ger et al., 2021). Mental state vocabulary and grammatical understanding are likely richer in L1 than in L2, and thus, native speakers may have a more sophisticated “toolbox” to represent and understand cognitive processes than non-native speakers. Moreover, it is easier to interact and communicate with the environment in L1 than in L2, and hence, native speakers may have more opportunities to enhance their understanding of metacognitive processes than non-native speakers. In sum, L1 abilities may be more relevant for monitoring development than L2 abilities, and therefore the relation between language abilities and metacognitive monitoring may differ for native and non-native speakers. A recent study supports the view of structural differences. Despite the poorer language abilities of non-native speakers, and although monitoring was assessed in their L2 language, non-native speakers’ monitoring accuracy was similar to native speakers’ in both a paired-associates and a text comprehension task. This suggests that non-natives’ L2 abilities do not play the same crucial role for monitoring accuracy as L1 abilities play for native speakers (Buehler et al., 2021). Moreover, a study assessing dual language learners’ higher-order cognitions (ToM, comprehension monitoring,

and inference making) in their L1 and L2 found a general higher-order cognitive factor with language-specific features (Kim et al., 2021).

Regarding the longitudinal effect of metacognition on language abilities, previous research showed that declarative metacognition and metacognitive regulation predict L1 and L2 abilities (Boulware-Gooden et al., 2007; Schoonen et al., 2011; Teng, 2019; Van Gelderen et al., 2004). This indicates that accurate monitoring may be relevant for later language abilities in native and non-native speakers. However, most research focused on L2 abilities. Accurate monitoring may be essential for non-native speakers as they cannot learn the language (L2) through a native-speaking environment at home. Accurate monitoring allows non-native speakers to self-regulate their learning by detecting errors and executing control behavior accordingly (e.g. Destan et al., 2014), such as learning vocabulary or grammatical rules. Contrary, monitoring accuracy may be less crucial for native speakers. They can learn the language (L1) through their native-speaking environment and are less dependent on self-regulated learning. Monitoring accuracy may thus predict language abilities for native and non-native speakers, but effects may be more pronounced for non-native than for native speakers.

Disentangling monitoring accuracy and task performance

For assessing metacognitive monitoring, subjects typically evaluate their confidence in a given answer (Confidence Judgements; Dunlosky et al., 2016). The present study focuses on absolute monitoring accuracy (calibration) to quantify monitoring accuracy. Monitoring accuracy is typically operationalized as the discrepancy between actual (first-order) task performance and estimated task performance (Confidence Judgment), with a smaller discrepancy indicating higher monitoring accuracy (Schraw, 2009). It is no surprise that monitoring accuracy and task performance are closely related (Roebbers et al., 2021). Therefore, one should be cautious when comparing monitoring accuracy across groups with

different task performances. Differences in monitoring accuracy may be based on different task performances and not on actual differences in monitoring abilities (Burson et al., 2006; Dunlosky et al., 2016).

However, evidence concerning the direction of the relation between monitoring accuracy and task performance is inconsistent. On the one hand, metacognition is related to cognitive performance (Flavell, 1979; Flavell & Wellman, 1977; Kuhn, 2000). Accurate monitoring that guides one's attention to the most relevant aspects of a task is the basis for strategy selection, study time allocation, and error detection. Consequently, accurate monitoring may predict task performance longitudinally (see Rinne & Mazzocco, 2014).

On the other hand, knowledge and skills in a specific domain are assumed to free cognitive resources for strategic and metacognitive skills (Roebbers, 2014; Schneider, 2015). A higher (first-order) task performance may deliberate cognitive resources for accurate monitoring and render monitoring into qualitatively different expert-like processes. Task performance may predict monitoring accuracy longitudinally. Indeed, earlier spelling performance predicted more accurate monitoring in second graders eight months later, but not vice versa (Roebbers & Spiess, 2017). Monitoring accuracy and task performance appear to be closely and bidirectionally related. Therefore, it is crucial to consider task performance when investigating the longitudinal relation between language abilities and monitoring accuracy.

The Present Study

We examine the longitudinal relation between language abilities and metacognitive monitoring accuracy from kindergarten to first grade. Furthermore, we compare the longitudinal relations between native and non-native speakers. We take the longitudinal relations of the first-order task performance and monitoring accuracy into account for all analyses. Insights on language abilities' role in developing metacognitive monitoring are sparse but highly relevant. Language may drive metacognitive monitoring, a critical ability

for school success and self-regulated learning (Dunlosky & Metcalfe, 2009; Roebers, 2017; Schraw et al., 2006). This may be even more crucial for non-native speakers, a growing community that typically underperforms in language-related tasks (OECD, 2018, 2019a). We relied on data from a population-based, longitudinal cohort study for analyses. Language abilities, monitoring accuracy, and task performance were assessed in kindergarten and first grade. To address our research questions, we computed cross-lagged panel models. We had the following hypotheses: Firstly, language abilities positively predict monitoring accuracy. Secondly, monitoring accuracy positively predicts language abilities. Thirdly, language abilities predict monitoring accuracy more strongly than monitoring accuracy predicts language abilities. Fourthly, the longitudinal relations between language abilities and monitoring accuracy differ between native and non-native speakers.

Method

Sample

We analyzed data from the longitudinal German National Educational Panel Study (Blossfeld & Roßbach, 2019; NEPS Network, 2021). The NEPS cohorts are representative samples assessing children and their parents throughout their lives. The present research used anonymized data from the NEPS and therefore was exempt from approval by the Ethics Committee of the authors' institution. We focused on data collected at Wave 1 (first kindergarten year), Wave 2 (second kindergarten year), and Wave 3 (grade one) in Starting Cohort 2 initiated in 2010. Children did not participate at every measurement, and new children (additional classmates) were included in the sample in grade one (Wave 3). This resulted in an analytic sample of 9,167 subjects, who participated either at W1 ($N = 2,948$), W2 ($N = 2,727$), and/or W3 ($N = 6,733$). Of the participants, 2,954 completed at least one kindergarten measurement (W1 or W2) and the grade one measurement (W3). Parents gave informed consent to participate in the study for their children and themselves.

We classified children as native or non-native speakers based on parental reports. Parents were asked about the primary language their child had learned at home during the first three years of life. This resulted in 6,403 native and 788 non-native speakers across all three waves of interest. For 1,976 children, information on their primary language was missing and we excluded these cases in models comparing native and non-native speakers. However, we relied on the entire sample in models not related to primary language ($N = 9,167$). Descriptive statistics revealed that native and non-native speakers were similar in age ($M_{W1} = 60$ months ; $M_{W2} = 71$ months; $M_{W3} = 85$ months) and gender (48% - 51% boys) across all waves. However, socio-economic status (Highest International Socio-Economic Index; HISEI) was higher for native ($HISEI = 59 - 63$) than non-native speakers ($HISEI = 43 - 51$). More information about NEPS, Starting Cohort 2 can be retrieved online at <https://www.neps-data.de>.

Measures

Our primary focus was on the relation between language abilities and monitoring accuracy. Additionally, we included task performance to disentangle task performance and monitoring accuracy. In kindergarten (W1 and W2), children were tested individually. In grade one (W3), children were tested in groups at their school.

Language Abilities

Receptive vocabulary and receptive grammar are common measures of language abilities in large-scale assessments (Berendes et al., 2013). Vocabulary is seen as one of the best measures of language abilities and is related to crystallized intelligence. Receptive grammar is important for reading comprehension and academic language abilities (Ebert & Weinert, 2013; Weinert, 2010; Weinert et al., 2019), suggesting that vocabulary and grammar are valid measures for computing general language abilities. Receptive vocabulary correlated substantially with receptive grammar, $r = .64, p < .01$ in kindergarten, and $r = .64, p < .01$ in

grade one. We operationalized language abilities as the unweighted mean of receptive vocabulary and receptive grammar.

Receptive Vocabulary. Receptive vocabulary indicates language abilities reflecting all words a person recognizes and comprehends when heard (Berendes et al., 2013). The measure for receptive vocabulary is a modified German version of the Peabody Picture Vocabulary Test (PPVT; Berendes et al., 2013; Dunn & Dunn, 2004), designed for kindergarten and grade one children (see Roßbach et al., 2005).

In kindergarten (W1), the receptive vocabulary test was administered to 2,859 children. The test consisted of 77 items requiring to choose one out of four pictures corresponding to a spoken word. The test was stopped after six consecutive incorrect answers. The test showed good item fit and reliability (EAP/PV reliability = .89; WLE reliability = .89; Fischer & Durda, 2020). We used Weighted Likelihood Estimates (WLE) in the analyses (Pohl & Carstensen, 2012).

In grade one (W3), the receptive vocabulary test was administered to 6,471 children. The test consisted of 66 items requiring to choose one out of four pictures corresponding to a spoken word. There was no stop criterion. The test showed good item fit and reliability (EAP/PV reliability = .87; WLE reliability = .87; Fischer & Durda, 2020). We used WLEs in the analyses (Pohl & Carstensen, 2012).

Receptive Grammar. Receptive grammar reflects listening comprehension on a sentence level (Lorenz et al., 2017). Children solved a shortened version of the TROG-D (see Berendes et al., 2013; Fox, 2006). Examples of sentence categories are prepositions, passive voice, personal pronouns, relative clauses, or topicalizations (Berendes et al., 2013).

In kindergarten (W1), the receptive grammar test was administered to 2,915 children. The test consisted of 48 items requiring to choose one out of four pictures corresponding to a spoken sentence. The test was stopped after five consecutive incorrect sets. A set consisted of

1 to 4 items from the same sentence category. A set was rated as incorrect when at least one of the belonging items was solved incorrectly. WLE scores are not available yet (Lorenz et al., 2017). Therefore, we computed a sum score based on 24 identical linking items assessed in kindergarten and grade one (*Cronbach's alpha* = .83). This allows comparing grammar performance between kindergarten and grade one children. We could not compute the sum score for 358 subjects because they reached the stop criterion on items presented before the 24 linking items. Finally, we computed z-scores based on the kindergarten mean and standard deviation.

In grade one (W3), the receptive grammar test was administered to 6,443 children. The test consisted of 40 items requiring to choose one out of four pictures corresponding to a spoken sentence. There was no stop criterion (Lorenz et al., 2017). As for kindergarten children, we computed a sum score based on the 24 linking items (*Cronbach's alpha* = .74). We could not compute the sum score for 927 subjects who did not reach the end of the 24 linking items due to time limits (see Lorenz et al., 2017). Finally, we computed z-scores based on the grade one mean and standard deviation.

Monitoring accuracy

Within the NEPS, monitoring accuracy is assessed as metacognitive performance judgments in various competence domains (Lockl, 2015; cf. Weinert et al., 2019). The present study focuses on the meta-level of the mathematical and scientific competence test (e.g. Nelson & Narens, 1990). These are estimations of the actual task performance. Immediately after completing the respective tests (math or science), the investigator asked the child: "What do you think: How many tasks did you solve correctly?". Each test consisted of 22 to 26 tasks (items; see below). Children gave retrospective judgments of their performance on a 5-point smiley scale. The scale ranges from no task correct (sad looking smiley coded as 1) to all correct (happy looking smiley coded as 5). In kindergarten (W1 and W2), the children pointed

to the corresponding smiley, and the investigator noted down the child's answer. In grade one (W3), children marked the corresponding smiley in their test booklets (Händel et al., 2013; Lockl, 2015, p.3).

As a measure of monitoring accuracy, deviation scores between the subjects' judgments and the actual mathematics and science task performance were computed (see Schraw, 2009). That is, the estimated proportion of correctly solved items minus the proportion of correctly solved items. Hence, the estimates of correctly solved items on the smiley scale were transformed into proportions of items solved correctly (1 = 0; 2 = 0.25; 3 = 0.5; 4 = 0.75; 5 = 1). Deviation scores range from -1 to 1. A score of 0 indicates perfect performance estimation, whereas a negative score indicates underestimation and a positive score overestimation of one's performance (Lockl, 2015 p.4). We transformed the deviation scores into absolute scores for monitoring accuracy, revealing the absolute deviation between estimated and actual performance on a range from 0 to 1. Scores closer to 0 indicate more accurate monitoring. Next, we computed the mean of absolute monitoring accuracy in mathematics and the science task for a general measure of monitoring accuracy. Monitoring accuracy in the mathematics task correlated significantly with monitoring accuracy in the science task, $r = .41$, $p < .01$ in kindergarten, and $r = .58$, $p < .01$ in grade one.

Task Performance

Task performance is based on the object-level performance (e.g. Nelson & Narens, 1990) in the mathematical and scientific competence tests. We operationalized task performance as the mean of mathematical and scientific competence. Mathematical competence correlated with scientific competence $r = .58$, $p < .01$ in kindergarten, and $r = .37$, $p < .01$ in grade one.

Mathematical Competence. The underlying framework of the tests on mathematical competence combines five mathematical content areas with six mathematical and cognitive

processes (see Neumann et al., 2013). The five content areas refer to (a) sets, numbers, and operations, (b) units and measuring, (c) space and shape, (d) change and relationships, and (e) data and chance. The six cognitive processes refer to (a) mathematical communication, (b) mathematical argumentation, (c) modeling, (d) using representational forms, (e) mathematical problem solving, and (f) technical abilities and skills.

In kindergarten (W2), the mathematics test was administered to 2,727 children. The test consisted of 26 items requiring simple multiple-choice, short constructed, matching, or sorting responses. The test showed very good item fit and good reliability (EAP/PV reliability = 0.82; WLE reliability = 0.80; Schnittjer, 2018). We used WLEs in the analyses (Pohl & Carstensen, 2012).

In grade one (W3), the mathematics test was administered to 6,510 children. The test consisted of 22 items requiring either simple (find the correct answer from several) or complex (several subtasks with two response options) multiple-choice responses. The test showed very good item fit and good reliability (EAP/PV reliability = 0.76; WLE reliability = 0.74; Schnittjer & Fischer, 2018). We used WLEs in the analyses (Pohl & Carstensen, 2012).

Scientific Competence. The underlying framework of the tests on scientific competence distinguishes between *knowledge of scientific concepts (KOS)* and *knowledge about scientific processes (KAS)*; see Hahn et al., 2013). KOS encompasses content-related components interactions, matter, development, and systems. KAS entails process-related components of scientific reasoning and scientific inquiry. KOS and KAS were applied in three chosen everyday life contexts: technology, health, and environment.

In kindergarten (W1), the scientific competence test was administered to 2,955 children. The test consisted of 26 items requiring simple multiple-choice, complex multiple-choice, or short constructed responses. Picture cards showed the response options. The test

showed very good item fit and good reliability (WLE reliability = .75 Schöps, 2013). We used WLEs in the analyses (Pohl & Carstensen, 2012).

In grade one (W3), the scientific competence test was administered to 6,734 children. The test consisted of 25 items requiring simple multiple-choice, complex multiple-choice, or short constructed responses. Picture cards showed the response options. The test showed very good item fit and good reliability (EAP/PV reliability = .73; WLE reliability = .73; Kähler, 2019). We used WLEs in the analyses (Pohl & Carstensen, 2012).

Socio-Economic Status

We used the highest International Socio-Economic Index (HISEI) of occupational status to measure socio-economic status (Ganzeboom, 2010). Parents indicated their occupation, which was then coded with the International Socioeconomic Index of Occupational Status (ISEI-08; Ganzeboom, 2010). The ISEI-08 ranks occupations based on mean income and mean level of education. Higher values indicate a higher socio-economic status. When parents differed in their ISEI-08 score, we took the highest score (HISEI). HISEI is a standard measure in large-scale assessments, such as the Programme for the International Student Assessment (PISA; OECD, 2019b).

Analytic Approach

The NEPS provides WLEs scaled with item response theory for the mathematical, scientific, and vocabulary competence measures (Pohl & Carstensen, 2012). WLEs for grammatical competence will be published in future data releases. WLEs are point estimates of individual competence scores. They are similar to sum scores of correctly answered items. However, WLEs have the advantages of facilitating the treatment of missings and the comparison of competence scores over different measurement points, cohorts, and settings (group vs. individual). WLEs are constrained to have a mean of zero, but the variance is unrestricted (Pohl & Carstensen, 2012).

We computed a cross-lagged panel model in R version 4.1.1 (R Core Team, 2021) with the lavaan package version 0.6-9 (Rosseel, 2012). We used Full Information Maximum Likelihood (FIML) to deal with missing values (Graham & Coffman, 2012). We conducted multi-group analyses to compare native and non-native speaking children, with language as the grouping variable. To test for significant group differences, we computed Chi-square difference tests. The analysis code for this study can be obtained from the first author. Access to the data can be requested here: <https://www.neps-data.de/Mainpage>. We did not preregister the study.

Results

We compared native and non-native speakers' mean values (Table 1). Independent t-tests revealed that native-speaking children in kindergarten and first grade outperformed non-native speaking children on all variables with $p < .01$. As expected, in kindergarten, native speakers outperformed non-native speakers in language abilities (mean of TROG and PPVT; $d = 1.44$), monitoring accuracy (mean of science and math monitoring accuracy; $d = -0.71$), and task performance (mean of science and math performance; $d = 0.91$). Also in first grade, native speakers outperformed non-native speakers in language abilities ($d = 1.24$), monitoring accuracy ($d = -0.57$) and task performance ($d = 0.48$).

Intercorrelations of the variables for the entire sample are presented in Table 2. All variables were moderately to highly and significantly correlated. Task performance highly correlated with monitoring accuracy and language abilities in kindergarten and grade one (correlations ranged from .52 to .80, in absolute values). This result confirms task performance as a relevant variable in the targeted interplay. We display separate correlations for native and non-native speakers in Table 3. In general, correlations among the variables were similar for native and non-native speakers, indicated by significant and moderate to high

correlations. However, monitoring accuracy in kindergarten was significantly correlated with language abilities in grade one for native speakers but not for non-native speakers.

Cross-Lagged Panel Model

Entire Sample

To examine the prospective reciprocal effects between language abilities and monitoring accuracy, we computed a cross-lagged panel model with the variables language abilities, monitoring accuracy, and task performance in kindergarten and grade one, respectively ($N = 9,159$). Figure 1 indicates the standardized estimates of the cross-lagged effects, stability effects, and correlations.

A significant cross-lagged effect emerged for the effect of earlier language abilities on later monitoring accuracy ($\beta = -0.19$). However, earlier monitoring accuracy did not predict later language abilities. Both cross-lagged effects are controlled for the prospective effect of task performance on the outcome variables. Task performance significantly predicted later language abilities ($\beta = 0.22$) and monitoring accuracy ($\beta = -0.25$). However, later task performance was not predicted by language abilities or by monitoring accuracy. Concerning stability, effects revealed significant autoregressions from kindergarten to grade one for language abilities ($\beta = 0.59$) and task performance ($\beta = 0.69$). In contrast, monitoring accuracy was not stable over time. All constructs at Wave 1 were significantly and highly correlated (ranging from .60 to .80 in absolute values). Most importantly, results showed that earlier language abilities predicted later monitoring accuracy, but earlier monitoring accuracy did not predict later language abilities.

Native vs. Non-Native Speakers

Moreover, we addressed the research question of whether the structural relation between language abilities and monitoring accuracy differs between native ($N = 6,399$) and non-native speakers ($N = 785$). We conducted a multi-group analysis to test for differences

between native and non-native speakers in the cross-lagged paths from language abilities to monitoring accuracy and vice versa. We compared two models: The coefficients for native and non-native speakers were freely estimated in the first model, whereas the coefficients were constrained to be equal in the second model. The second model fitted the data significantly worse, $\chi^2_{diff}(2, N = 7184) = 6.06, p = .048$. This result indicates that the cross-lagged paths between language abilities and monitoring accuracy differed substantially between native and non-native speakers.

For native speakers, the standardized estimates of the cross-lagged effects, stability effects, and correlations are shown in Figure 2. A significant cross-lagged effect emerged for the effect of language abilities on later monitoring accuracy ($\beta = -0.20$). However, earlier monitoring accuracy did not predict later language abilities. Both cross-lagged effects are controlled for the prospective effect of task performance on the outcome variables. Earlier task performance significantly predicted later language abilities ($\beta = 0.18$) and later monitoring accuracy ($\beta = -0.18$). Later task performance was predicted by earlier language abilities ($\beta = 0.14$), but not by earlier monitoring accuracy. Concerning stability, effects revealed significant autoregressions from kindergarten to grade one for language abilities ($\beta = 0.53$) and task performance ($\beta = 0.64$), but not for monitoring accuracy. All constructs at Wave 1 were significantly and highly correlated (ranging from .58 to .81 in absolute values). For native speakers, earlier language abilities predicted later monitoring accuracy, but earlier monitoring accuracy did not predict later language abilities.

For non-native speakers, the standardized estimates of the cross-lagged effects, stability effects, and correlations are shown in Figure 3. A significant cross-lagged effect emerged for the effect of earlier monitoring accuracy on later language abilities ($\beta = 0.62$). However, earlier language abilities did not predict later monitoring accuracy. Both cross-lagged effects are controlled for the prospective effect of task performance on the outcome

variables. Earlier task performance significantly predicted both later language abilities ($\beta = 0.60$) and later monitoring accuracy ($\beta = -0.58$). Later task performance was predicted by earlier language abilities ($\beta = 0.47$) but not by earlier monitoring accuracy. Regarding stability, estimations revealed a significant autoregression for language abilities from kindergarten to grade one ($\beta = 0.44$). Monitoring accuracy and task performance were not stable over time. All constructs at Wave 1 were significantly inter-correlated (ranging from .46 to .70 in absolute values).

Interestingly, for non-native speakers, lower monitoring accuracy in kindergarten predicted higher language abilities in grade one. Children who monitored their task performance less accurately in kindergarten showed higher L2 abilities in grade one. To better understand this surprisingly relation for non-native speakers compared to native speakers, we considered the entire range of monitoring accuracy indicating whether children were over- or underconfident. The deviation scores range from -1 to 1 , with a negative score indicating underconfidence and a positive score indicating overconfidence. Most non-native speakers were overconfident in kindergarten (96.70 %), and the mean of the monitoring accuracy deviations scores in kindergarten ($M = .47$; $SD = 0.16$) indicated that non-native speaking children were overconfident on average. Hence, overconfidence may be beneficial for the development of L2 abilities.

Discussion

The present research aimed to examine the longitudinal relations between language abilities and monitoring accuracy. We computed cross-lagged panel models with data from a German large-scale assessment. The present study is among the first to investigate the bidirectional relationship between language abilities and metacognitive monitoring accuracy in such a young age group with large-scale data. Consistent with our hypotheses, language abilities in kindergarten predicted monitoring accuracy in grade one. However, we found

structural differences between native and non-native speakers. On the one hand, earlier language abilities predicted monitoring accuracy for native, but not for non-native speakers. On the other hand, earlier monitoring accuracy predicted later language abilities for non-native, but not for native speakers.

When using the data from the entire sample (Figure 1), higher language abilities in kindergarten predicted higher monitoring accuracy in grade one, consistent with our hypotheses. This is in line with recent research showing that vocabulary in preschool predicts uncertainty monitoring in kindergarten (Gonzales et al., 2021) and with studies demonstrating that earlier language abilities contribute to later metacognitive knowledge (Ebert, 2015, 2020; Lockl & Schneider, 2007; Schneider & Löffler, 2016). Contrary to our expectations, monitoring accuracy in kindergarten did not predict language abilities in grade one. This also contrasts previous research suggesting that metacognitive knowledge predicts various language abilities and language learning (Annevirta et al., 2007; Boulware-Gooden et al., 2007; Lecce et al., 2010; Schoonen et al., 2011; Van Gelderen et al., 2004). The contrasting findings in our and previous research may be explained by conceptual differences between declarative and procedural metacognition. Declarative metacognition is highly task-specific, such as reading or writing strategies. We took a more general approach by measuring procedural metacognition in terms of the ability to evaluate ongoing cognitive processes in a math and a science task. It may thus not be surprising that reading and writing strategies are closely related to later language abilities, whereas monitoring accuracy may be not. This confirms our hypothesis that language abilities are a stronger predictor of monitoring accuracy than *vice versa*.

We included task performance in our cross-lagged model to control the well-established influence of task performance on monitoring accuracy. Our results suggest that monitoring accuracy and task performance are highly related within and across the included

measurement waves. Earlier task performance predicted later monitoring accuracy, but earlier monitoring accuracy did not predict later task performance. This supports the perspective that task performance drives monitoring accuracy (Roebbers, 2014; Roebbers & Spiess, 2017; Schneider, 2015), rather than monitoring accuracy driving task performance (Flavell, 1979; Flavell & Wellman, 1977; Kuhn, 2000; Rinne & Mazzocco, 2014). Since the studies mentioned above included different age groups, it may be that the direction of the link between task performance and monitoring accuracy is age-specific, and therefore, our findings may be considered complementary to the existing evidence.

Confirming our hypothesis regarding structural differences between native and non-native speakers, we found structural differences between language abilities and monitoring accuracy (Figures 2 and 3). The relations between language abilities and monitoring accuracy for the native speakers were equal to the relations in the entire sample. Interestingly, we found different relations for non-native speakers. In contrast to native speakers, language abilities did not predict monitoring accuracy for non-native speakers. Native and non-native speakers' language abilities were assessed in the language of instruction, being the L1 ability for native speakers but the L2 ability for non-native speakers. Previous research suggests that a common factor and language-specific features can explain dual-language learners' higher-order cognitions in L1 and L2 (Kim et al., 2021). One might argue that monitoring accuracy is a language-specific feature of higher-order cognition. Furthermore, L1 may be more crucial than L2 abilities for monitoring accuracy in L1 tasks (Buehler et al., 2021). To specify the unique contributions of L1 and L2 abilities to monitoring accuracy, future research should assess non-native speakers' L1 and L2 abilities and include measures of monitoring accuracy in L1 and L2.

However, non-natives' earlier monitoring accuracy predicted language abilities in a surprising direction. The more overconfident non-native speakers were in kindergarten, the

higher their language abilities were in grade one. This result aligns with previous research demonstrating that metacognitive knowledge predicts various language abilities and language learning (Annevirta et al., 2007; Boulware-Gooden et al., 2007; Lecce et al., 2010; Schoonen et al., 2011; Van Gelderen et al., 2004). However, this result challenges the general understanding that more accurate monitoring relates to greater cognitive outcomes (Dunlosky & Metcalfe, 2009; Roebers, 2017; Schneider & Löffler, 2016). An explanation for this finding might stem from the literature on the adaptive character of overconfidence. For example, Shin et al. (2007) found that strongly overconfident children had more recall gains across multiple recall attempts in a memory task than less overconfident children. A possible explanation might be that overconfidence increases task motivation and persistence (Bjorklund & Bering, 2002). Consequently, overconfident non-native children may be more motivated and persistent in learning, trying, speaking, practicing, and improving their L2 skills than less confident non-native children.

The relation between task performance and monitoring accuracy was similar for native and non-native speakers. Higher task performance in kindergarten predicted more accurate monitoring in grade one, but not vice versa. This is in line with research suggesting that task performance drives monitoring accuracy in the sense that individuals with better task performance are more “expert”-like and thus, their monitoring processes are assumed to be qualitatively different compared to “novice”-like poor task performers (Roebers, 2014; Roebers & Spiess, 2017; Schneider, 2015). Our finding confirms this assumption and expands the link between task performance and monitoring accuracy to non-native speakers. However, the effect was more pronounced for non-native than native speakers. This finding is of practical importance for teachers. It indicates that non-native speakers’ school achievement (i.e., task performance) is crucial for their forthcoming, higher-order cognition and self-regulated lifelong learning skills.

Several implications can deviate from the present study. Firstly, general language abilities are crucial for developing monitoring accuracy for native speakers. Our results emphasize the importance of language abilities for developing higher-order cognition, such as monitoring accuracy. It remains open to future research to investigate the relation of non-native speakers' L1 abilities to monitoring accuracy. Moreover, it would be interesting to get a more fine-grained perspective of how specific language aspects contribute to monitoring accuracy, such as vocabulary, grammar, or mental state language (e.g. Lockl & Schneider, 2006). Secondly, overconfidence may be adaptive in the context of L2 learning. Teachers and parents may be well advised to encourage and motivate non-native speakers in their L2 learning and to raise non-natives' confidence in their L2 abilities. Thirdly, task performance predicts later monitoring accuracy, especially for non-native speakers. It might be beneficial to foster actual task performance in early primary school before focusing on monitoring accuracy for non-native students. In combination with the benefits of overconfidence for L2 abilities, this might be crucial for non-native speakers.

We consider the following limitations in the present study. Not all kindergarten measures were assessed at the same time. Most measures were assessed at W1, whereas mathematical competence was assessed one year later at W2. Therefore, the variables at kindergarten age consist of measurements from W1 and W2. Moreover, to further generalize our findings, it is crucial to include measures of L1 abilities of non-native speakers in future research. The present study also offers several important strengths. The large-scale representative sample allows generalizing our findings to many children and estimating more complex structural relations between central theoretical concepts than otherwise possible. We computed a cross-lagged panel model, which allows us to control for autoregressions across time and predict the direction of effects (cross-lagged paths). Finally, we could disentangle monitoring accuracy and task performance, which is a significant challenge in monitoring accuracy measures (Dunlosky et al., 2016).

Conclusion

We assessed the longitudinal relation between language abilities and monitoring accuracy in the transition from kindergarten to formal schooling. We relied on a German representative large-scale assessment (NEPS) to compute cross-lagged panel models. Our results revealed that language abilities predict monitoring accuracy, but not vice versa. This suggests that language abilities in the instruction language (L1) are a precursor of monitoring accuracy. Moreover, multi-group analyses indicated structural differences for native and non-native speakers. Language abilities predicted monitoring accuracy for native speakers but not for non-native speakers, indicating that the difference between one's native language and the language of instruction affects this link. To further understand the contribution of L1 and L2 abilities to monitoring accuracy, future research should include measures of both L1 and L2 abilities. Surprisingly, for non-native speakers, overconfidence in kindergarten positively predicted L2 abilities in grade one. This finding highlights the importance of confidence for learning a second language. It may be crucial for teachers and parents to encourage non-native speakers in their L2 abilities.

References

- Annevirta, T., Laakkonen, E., Kinnunen, R., & Vauras, M. (2007). Developmental dynamics of metacognitive knowledge and text comprehension skill in the first primary school years. *Metacognition and Learning*, 2(1), 21–39. <https://doi.org/10.1007/s11409-007-9005-x>
- Artelt, C., & Schneider, W. (2015). Cross-Country Generalizability of the Role of Metacognitive Knowledge in Students' Strategy Use and Reading Competence. *Teachers College Record: The Voice of Scholarship in Education*, 117(1), 1–32. <https://doi.org/10.1177/016146811511700104>
- Astington, J. W., & Baird, J. A. (2005). *Why Language Matters for Theory of Mind*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195159912.001.0001>
- Bellon, E., Fias, W., & Smedt, B. De. (2021). Too Anxious to Be Confident? A Panel Longitudinal Study Into the Interplay of Mathematics Anxiety and Metacognitive Monitoring in Arithmetic Achievement. *Journal of Educational Psychology*, 113(8), 1550–1564. <https://doi.org/10.1037/edu0000704>
- Berendes, K., Weinert, S., Zimmermann, S., & Artelt, C. (2013). Assessing language indicators across the lifespan within the German National Educational Panel Study (NEPS). *Journal for Educational Research Online*, 5(2), 15–49. <https://doi.org/10.25656/01:8423>
- Bjorklund, D. F., & Bering, J. M. (2002). The evolved child applying evolutionary developmental psychology to modern schooling. *Learning and Individual Differences*, 12(4), 347–373. [https://doi.org/10.1016/S1041-6080\(02\)00047-X](https://doi.org/10.1016/S1041-6080(02)00047-X)
- Blossfeld, H.-P., & Roßbach, H.-G. (2019). *Education as a lifelong process: The German National Educational Panel Study (NEPS). Edition ZfE (2nd ed.)*. Springer VS.

- Boulware-Gooden, R., Carreker, S., Thornhill, A., & Joshi, R. M. (2007). Instruction of Metacognitive Strategies Enhances Reading Comprehension and Vocabulary Achievement of Third-Grade Students. *The Reading Teacher, 61*(1), 70–77.
<https://doi.org/10.1598/rt.61.1.7>
- Buehler, F. J., van Loon, M. H., Bayard, N. S., Steiner, M., & Roebbers, C. M. (2021). Comparing metacognitive monitoring between native and non-native speaking primary school students. *Metacognition and Learning, 16*(3). <https://doi.org/10.1007/s11409-021-09261-z>
- Burson, K. A., Larrick, R. P., & Klayman, J. (2006). Skilled or Unskilled, but Still Unaware of It: How Perceptions of Difficulty Drive Miscalibration in Relative Comparisons. *Journal of Personality and Social Psychology, 90*(1), 60–77.
<https://doi.org/10.1037/0022-3514.90.1.60>
- Destan, N., Hembacher, E., Ghetti, S., & Roebbers, C. M. (2014). Early metacognitive abilities: The interplay of monitoring and control processes in 5- to 7-year-old children. *Journal of Experimental Child Psychology, 126*, 213–228.
<https://doi.org/10.1016/j.jecp.2014.04.001>
- Dunlosky, J., & Metcalfe, J. (2009). *Metacognition: A Textbook for Cognitive, Educational, Life Span & Applied Psychology*. Sage Publications.
- Dunlosky, J., Mueller, M. L., & Thiede, K. W. (2016). Methodology for Investigating Human Metamemory: Problems and Pitfalls. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford Handbook of Metamemory*. Oxford University Press.
<https://doi.org/10.1093/oxfordhb/9780199336746.013.14>
- Dunn, L. M., & Dunn, D. M. (2004). *Peabody Picture Vocabulary Test (German Version)*. Hogrefe.

- Ebert, S. (2015). Longitudinal Relations Between Theory of Mind and Metacognition and the Impact of Language. *Journal of Cognition and Development, 16*(4), 559–586.
<https://doi.org/10.1080/15248372.2014.926272>
- Ebert, S. (2020). Early Language Competencies and Advanced Measures of Mental State Understanding Are Differently Related to Listening and Reading Comprehension in Early Adolescence. *Frontiers in Psychology, 11*(952), 1–18.
<https://doi.org/10.3389/fpsyg.2020.00952>
- Ebert, S., & Weinert, S. (2013). Predicting Reading Literacy in Primary School: The Contribution of Various Language Indicators in Preschool. In M. Pfost, C. Artelt, & S. Weinert (Eds.), *The Development of Reading Literacy from Early Childhood to Adolescence. Empirical Findings from the Bamberg BiKS Longitudinal Studies* (pp. 93–149). University of Bamberg Press.
- Fischer, L., & Durda, T. (2020). *NEPS Technical Report for Receptive Vocabulary: Scaling Results of Starting Cohort 2 for Kindergarten (Wave 1), Grade 1 (Wave 3) and Grade 3 (Wave 5) (NEPS Survey Paper No. 65)*. Leibniz Institute for Educational Trajectories, National Educational Panel Study. <https://doi.org/10.5157/NEPS:SP65:1.0>
- Flavell, J. H. (1979). Metacognition and Cognitive Monitoring: A New Area of Cognitive-Developmental Inquiry. *American Psychologist, 34*(10), 906–911.
<https://doi.org/10.1037/0003-066X.34.10.906>
- Flavell, J. H., & Wellman, H. M. (1977). Metamemory. In R. V. Kail & J. W. Hagen (Eds.), *Perspectives on the development of memory and cognition* (pp. 3–33). Erlbaum.
- Fox, A. V. (2006). *Test zur Überprüfung des Grammatikverständnisses (TROG-D)*. Schulz-Kirchner Verlag.
- Freeman, E. E., Karayanidis, F., & Chalmers, K. A. (2017). Metacognitive monitoring of

working memory performance and its relationship to academic achievement in Grade 4 children. *Learning and Individual Differences*, 57, 58–64.

<https://doi.org/10.1016/j.lindif.2017.06.003>

Ganzeboom, H. B. G. (2010). *A New International Socio-Economic Index (ISEI) Of Occupational Status For The International Standard Classification of Occupation 2008 [ISCO-2008] Constructed With Data From The ISSP 2002-2007*. Paper presented at the Annual Conference of International Social Survey Programme, Lisbon.

Ger, E., Stuber, L., Küntay, A. C., Göksun, T., Stoll, S., & Daum, M. M. (2021). Influence of causal language on causal understanding: A comparison between Swiss German and Turkish. *Journal of Experimental Child Psychology*, 210, 1–18.

<https://doi.org/10.1016/J.JECP.2021.105182>

Gonzales, C. R., Mercurief, A., McClelland, M. M., & Ghetti, S. (2021). The development of uncertainty monitoring during kindergarten: Change and longitudinal relations with executive function and vocabulary in children from low-income backgrounds. *Child Development*, 1–16. <https://doi.org/10.1111/cdev.13714>

Graham, J. W., & Coffman, D. L. (2012). Structural Equation Modeling with Missing Data. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 277–295). Guilford.

Hahn, I., Schöps, K., Rönnebeck, S., Martensen, M., Hansen, S., Saß, S., Dalehefte, I. M., & Prenzel, M. (2013). Assessing scientific literacy over the lifespan – A description of the NEPS science framework and the test development. *Journal for Educational Research Online*, 5(2), 110–138. <https://doi.org/10.25656/01:8427>

Händel, M., Artelt, C., & Weinert, S. (2013). Assessing metacognitive knowledge: Development and evaluation of a test instrument. *Journal for Educational Research Online*, 5(2), 162–188. <https://doi.org/10.25656/01:8429>

- Harris, P. L., De Rosnay, M., & Pons, F. (2005). Language and Children's Understanding of Mental States. *Current Directions in Psychological Science*, *14*(2), 69–73.
<https://doi.org/10.1111/j.0963-7214.2005.00337.x>
- Haukås, Å., Bjørke, C., & Dypedahl, M. (2018). *Metacognition in Language Learning and Teaching*. Routledge. <https://doi.org/10.4324/9781351049146-10>
- Kähler, J. (2019). *NEPS Technical Report for Science: Scaling Results of Starting Cohort 2 for Grade 1 (NEPS Survey Paper No. 58)*. Leibniz Institute for Educational Trajectories, National Educational Panel Study. <https://doi.org/10.5157/NEPS:SP58:1.0>
- Kim, Y. S. G., Wolters, A., Mercado, J., & Quinn, J. (2021). Crosslinguistic Transfer of Higher Order Cognitive Skills and Their Roles in Writing for English-Spanish Dual Language Learners. *Journal of Educational Psychology*, *114*(1), 1–15.
<https://doi.org/10.1037/edu0000516>
- Kuhn, D. (2000). Metacognitive Development. *Current Directions in Psychological Science*, *9*(5), 178–181. <https://doi.org/10.1111/1467-8721.00088>
- Lecce, S., Zocchi, S., Pagnin, A., Palladino, P., & Taumoepeau, M. (2010). Reading Minds: The Relation Between Children's Mental State Knowledge and Their Metaknowledge About Reading. *Child Development*, *81*(6), 1876–1893. <https://doi.org/10.1111/j.1467-8624.2010.01516.x>
- Lockl, K. (2015). *Assessment of procedural metacognition: Scientific Use File 2015*. Leibniz Institute for Educational Trajectories, National Educational Panel Study.
- Lockl, K., & Schneider, W. (2006). Precursors of metamemory in young children: The role of theory of mind and metacognitive vocabulary. *Metacognition and Learning*, *1*, 15–31.
<https://doi.org/10.1007/s11409-006-6585-9>
- Lockl, K., & Schneider, W. (2007). Knowledge About the Mind: Links Between Theory of

Mind and Later Metamemory. *Child Development*, 78(1), 148–167.

<https://doi.org/10.1111/j.1467-8624.2007.00990.x>

Lorenz, C., Berendes, K., & Weinert, S. (2017). *Measuring receptive grammar in kindergarten and elementary school children in the German National Educational Panel Study (NEPS Survey Paper No. 24)*. Leibniz Institute for Educational Trajectories, National Educational Panel Study. <https://doi.org/10.5157/NEPS:SP24:1.0>

Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and Theory of Mind: Meta-Analysis of the Relation Between Language Ability and False-belief Understanding. *Child Development*, 78(2), 622–646. <https://doi.org/10.1111/j.1467-8624.2007.01018.x>

Nelson, T. O., & Narens, L. (1990). Metamemory: A Theoretical Framework and New Findings. *Psychology of Learning and Motivation - Advances in Research and Theory*, 26, 125–173. [https://doi.org/10.1016/S0079-7421\(08\)60053-5](https://doi.org/10.1016/S0079-7421(08)60053-5)

NEPS Network. (2021). *National Educational Panel Study, Scientific Use File of Starting Cohort Kindergarten [Data set]*. Leibniz Institute for Educational Trajectories (LIfBi), Bamberg. <https://doi.org/10.5157/NEPS:SC2:9.0.0>

Neumann, I., Duchhardt, C., & Grüßing, M. (2013). Modeling and assessing mathematical competence over the lifespan. *Journal for Educational Research Online*, 5(2), 80–109. <https://doi.org/10.25656/01:8426>

OECD. (2018). *The resilience of students with an immigrant background: Factors that shape well-being*. OECD Publishing. <https://doi.org/10.1787/9789264292093-en>

OECD. (2019a). *PISA 2018 Results: Combined Executive Summaries*. OECD Publishing. https://www.oecd.org/pisa/Combined_Executive_Summaries_PISA_2018.pdf

OECD. (2019b). *PISA 2018 Results (Volume III): What School Life Means for Students' Lives*. OECD Publishing. <https://doi.org/10.1787/acd78851-en>.

- Pohl, S., & Carstensen, C. H. (2012). *NEPS Technical Report – Scaling the Data of the Competence Tests (NEPS Working Paper No. 14)*. Otto-Friedrich-Universität, Nationales Bildungspanel.
- R Core Team. (2021). *R: A language and environment for statistical computing [Computer software]*. <https://www.r-project.org/>
- Rinne, L. F., & Mazzocco, M. M. M. (2014). Knowing Right from Wrong in Mental Arithmetic Judgments: Calibration of Confidence Predicts the Development of Accuracy. *PLOS ONE*, *9*(7), 1–11. <https://doi.org/10.1371/journal.pone.0098663>
- Roebers, C. M. (2014). Children's deliberate memory development: The contribution of strategies and metacognitive processes. In P. Bauer & R. Fivush (Eds.), *The Wiley handbook on the development of children's memory* (pp. 865–894). Wiley Blackwell.
- Roebers, C. M. (2017). Executive function and metacognition: Towards a unifying framework of cognitive self-regulation. *Developmental Review*, *45*, 31–51. <https://doi.org/10.1016/j.dr.2017.04.001>
- Roebers, C. M., Krebs, S. S., & Roderer, T. (2014). Metacognitive monitoring and control in elementary school children: Their interrelations and their role for test performance. *Learning and Individual Differences*, *29*, 141–149. <https://doi.org/10.1016/j.lindif.2012.12.003>
- Roebers, C. M., & Spiess, M. (2017). The Development of Metacognitive Monitoring and Control in Second Graders: A Short-Term Longitudinal Study. *Journal of Cognition and Development*, *18*(1), 110–128. <https://doi.org/10.1080/15248372.2016.1157079>
- Roebers, C. M., van Loon, M. H., Buehler, F. J., Bayard, N. S., Steiner, M., & Aeschlimann, E. A. (2021). Exploring psychometric properties of children' metacognitive monitoring. *Acta Psychologica*, *220*(103399), 1–11. <https://doi.org/10.1016/j.actpsy.2021.103399>

- Roßbach, H. G., Tietze, W., & Weinert, S. (2005). *Peabody Picture Vocabulary Test – Revised. Deutsche Forschungsversion des Tests von L.M. Dunn & L.M. Dunn von 1981*. Universität Bamberg, FU Berlin.
- Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2). <https://doi.org/10.18637/jss.v048.i02>
- Schneider, W. (2015). *Memory Development from Early Childhood Through Emerging Adulthood*. Springer. <https://doi.org/10.1007/978-3-319-09611-7>
- Schneider, W., & Löffler, E. (2016). The Development of Metacognitive Knowledge in Children and Adolescents. In J. Dunlosky & S. U. K. Tauber (Eds.), *The Oxford Handbook of Metamemory*. <https://doi.org/10.1093/oxfordhb/9780199336746.013.10>
- Schnittjer, I. (2018). *NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 2 In Kindergarten (NEPS Survey Papers No. 43)*. Leibniz Institute for Educational Trajectories, National Educational Panel Study. <https://doi.org/10.5157/NEPS:SP43:1.0>
- Schnittjer, I., & Fischer, L. (2018). *NEPS Technical Report for Mathematics: Scaling Results of Starting Cohort 2 for Grade 1 (NEPS Survey Paper No. 46)*. Leibniz Institute for Educational Trajectories, National Educational Panel Study. <https://doi.org/10.5157/NEPS:SP46:1.0>
- Schoonen, R., Van Gelderen, A., Stoel, R. D., Hulstijn, J., & De Glopper, K. (2011). Modeling the Development of L1 and EFL Writing Proficiency of Secondary School Students. *Language Learning*, 61(1), 31–79. <https://doi.org/10.1111/j.1467-9922.2010.00590.x>
- Schöps, K. (2013). *NEPS Technical Report for Science – Scaling results of Starting Cohort 2 in Kindergarten (NEPS Working Paper No. 24)*. University of Bamberg, National

Educational Panel Study.

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring.

Metacognition and Learning, 4, 33–45. <https://doi.org/10.1007/s11409-008-9031-3>

Schraw, G., Crippen, K. J., & Hartley, K. (2006). Promoting Self-Regulation in Science

Education: Metacognition as Part of a Broader Perspective on Learning. *Research in Science Education*, 36, 111–139. <https://doi.org/10.1007/s11165-005-3917-8>

Shin, H. E., Bjorklund, D. F., & Beck, E. F. (2007). The adaptive nature of children's

overestimation in a strategic memory task. *Cognitive Development*, 22(2), 197–212.

<https://doi.org/10.1016/J.COGDEV.2006.10.001>

Teng, F. (2019). The role of metacognitive knowledge and regulation in mediating university

EFL learners' writing performance. *Innovation in Language Learning and Teaching*,

14(5), 436–450. <https://doi.org/10.1080/17501229.2019.1615493>

Van Gelderen, A., Schoonen, R., De Glopper, K., Hulstijn, J., Simis, A., Snellings, P., &

Stevenson, M. (2004). Linguistic Knowledge, Processing Speed, and Metacognitive

Knowledge in First- and Second-Language Reading Comprehension: A Componential

Analysis. *Journal of Educational Psychology*, 96(1), 19–30.

<https://doi.org/10.1037/0022-0663.96.1.19>

Weinert, S. (2010). Erfassung sprachlicher Fähigkeiten. In E. Walther, F. Preckel, & S.

Mecklenbräuker (Eds.), *Befragung von Kindern und Jugendlichen* (pp. 227–262).

Hogrefe. <https://fis.uni-bamberg.de/handle/uniba/4227>

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., Carstensen, C. H., & Lockl, K.

(2019). Development of Competencies Across the Life Course. In H.-P. Blossfeld & H.-

G. Roßbach (Eds.), *Education as a Lifelong Process* (3rd ed., pp. 57–81). Springer VS.

https://doi.org/10.1007/978-3-658-23162-0_4

Wenden, A. L. (1998). Metacognitive knowledge and language learning. *Applied Linguistics*, *19*(4), 515–537. <https://doi.org/10.1093/applin/19.4.515>

Zepeda, C. D., Hlutkowsky, C. O., Partika, A. C., & Nokes-Malach, T. J. (2019). Identifying Teachers' Supports of Metacognition Through Classroom Talk and Its Relation to Growth in Conceptual Learning. *Journal of Educational Psychology*, *111*(3), 522–541. <https://doi.org/10.1037/edu0000300>

Table 1*Descriptive statistics of the measures including Means and SDs in parentheses*

Variable	Kindergarten (W1 and W2)			First grade (W3)		
	All	Native	Non-native	All	Native	Non-native
<i>Language abilities</i>						
Grammar [z]	0 (1)	0.16 (0.95)	-0.69 (0.92)	0 (1)	0.2 (0.9)	-0.68 (1.03)
Vocabulary [WLE]	-0.06 (1.13)	0.25 (0.94)	-1.21 (1.01)	1.44 (0.84)	1.63 (0.73)	0.61 (0.79)
Total	-0.11 (1.04)	0.17 (0.87)	-1.11 (0.96)	0.82 (0.87)	1.02 (0.76)	0.06 (0.85)
<i>Monitoring accuracy</i>						
Science	.37 (0.17)	.34 (0.16)	.45 (0.18)	.36 (0.23)	.33 (0.21)	.45 (0.23)
Math	.48 (0.24)	.44 (0.23)	.58 (0.23)	.38 (0.24)	.35 (0.22)	.45 (0.24)
Total	.42 (0.18)	.39 (0.17)	.51 (0.17)	.37 (0.21)	.34 (0.19)	.45 (0.2)
<i>Task Performance</i>						
Science [WLE]	0 (1.04)	0.2 (1)	-0.70 (0.83)	1.38 (0.93)	1.55 (0.89)	0.9 (0.86)
Math [WLE]	-0.01 (1.17)	0.18 (1.12)	-0.59 (1.03)	1.66 (1.96)	1.83 (2.06)	1.3 (1.08)
Total	-0.02 (1)	0.19 (0.95)	-0.66 (0.81)	1.52 (1.23)	1.69 (1.25)	1.1 (0.86)

Note. W1 = wave 1; W2 = wave 2; W3 = wave 4; z = z-score; WLE = weighted likelihood estimate; Total = manifest variable based on the mean of the two variables above. Independent *t*-tests revealed significant differences between native and non-native speaking children on all variables in kindergarten and grade 1.

Table 2*Pearson Correlations of variables in the full sample*

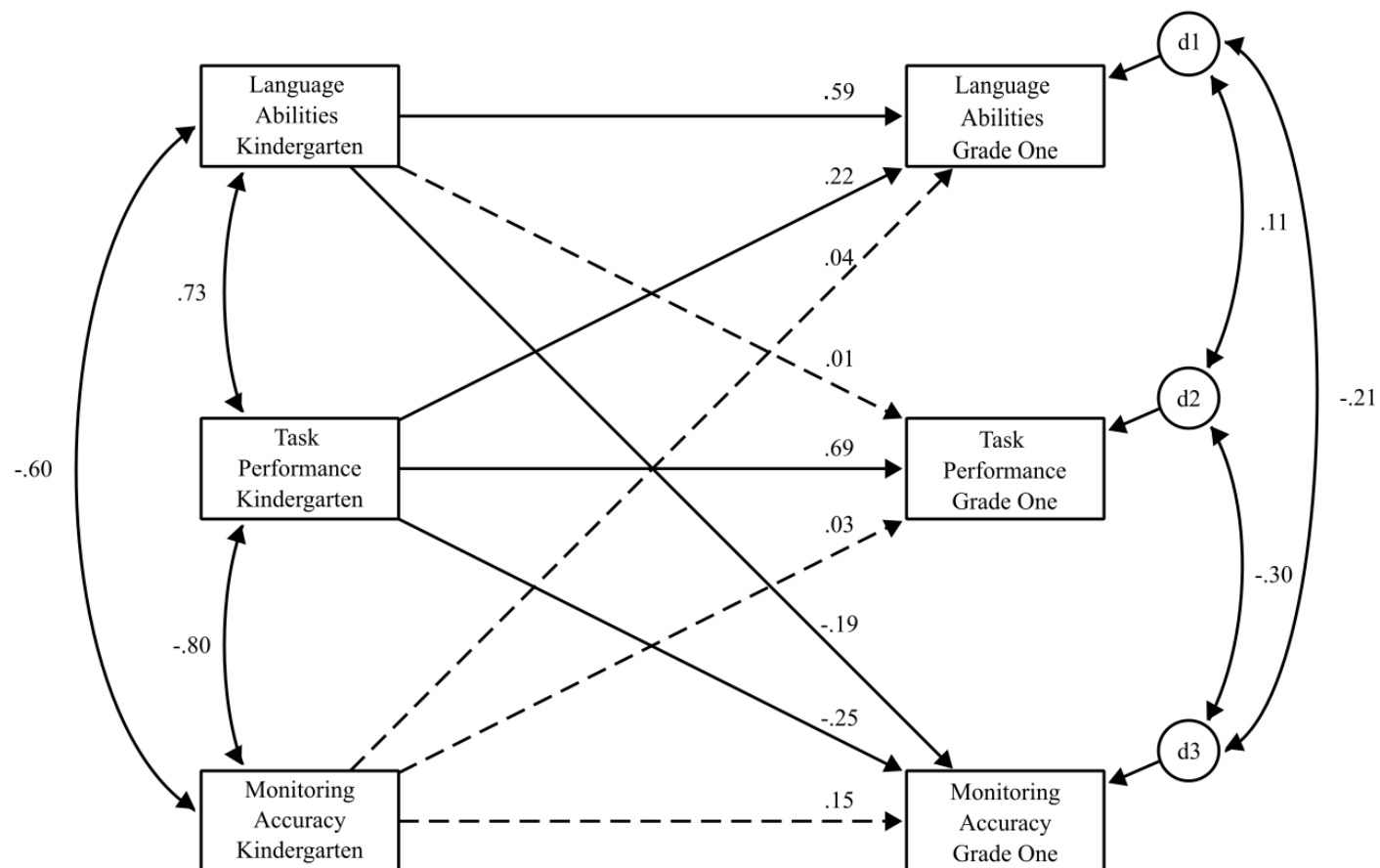
Variables	1	2	3	4	5	6
1. Language KG	-					
2. Monitoring KG	-.59**	-				
3. Performance KG	.72**	-.80**	-			
4. Language G1	.70**	-.49**	.62**	-		
5. Monitoring G1	-.52**	.51**	-.59**	-.49**	-	
6. Performance G1	.59**	-.57**	.71**	.52**	-.55**	-

Note. KG = kindergarten; G1 = grade 1. ** $p < .01$

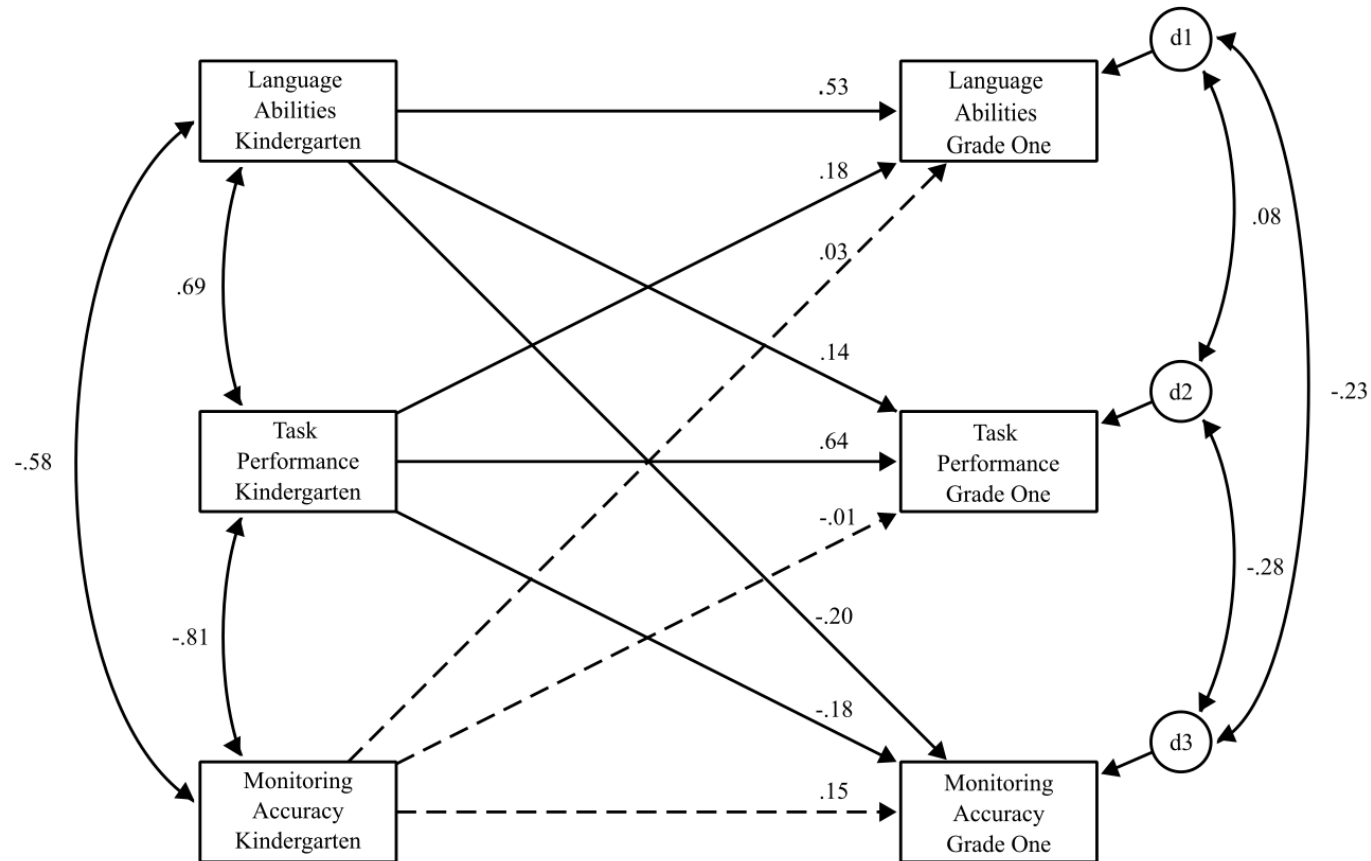
Table 3*Pearson Correlations of variables for native and non-native speaking children*

Variables	1	2	3	4	5	6
1. Language KG	-	-.45**	.64**	.61**	-.49**	.66**
2. Monitoring KG	-.57**	-	-.70**	-.22	.45**	-.36*
3. Performance KG	.68**	-.81**	-	.57**	-.61**	.61**
4. Language G1	.65**	-.47**	.57**	-	-.40**	.56**
5. Monitoring G1	-.49**	.49**	-.55**	-.44**	-	-.67**
6. Performance G1	.56**	-.56**	.69**	.46**	-.50**	-

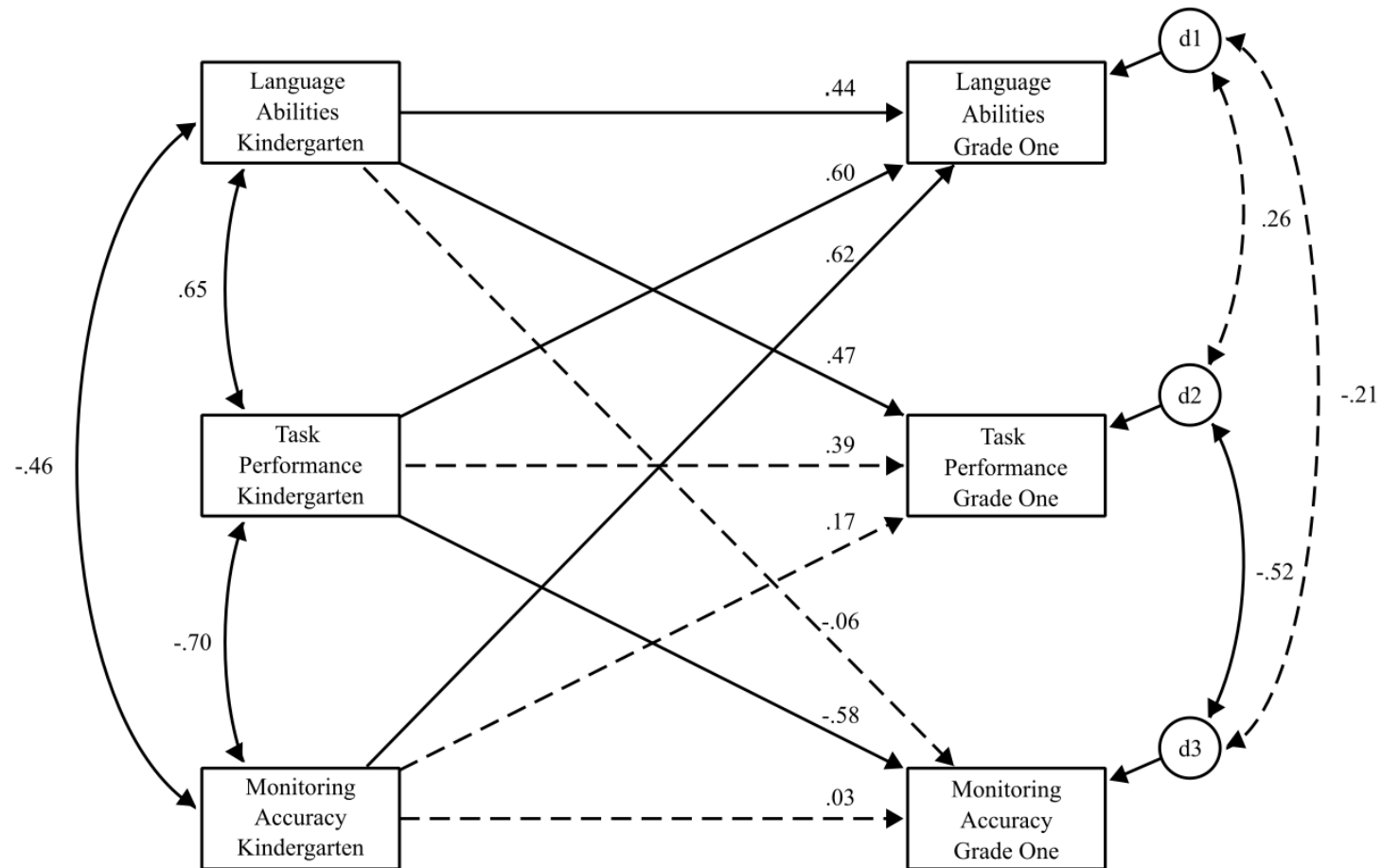
Note. Correlation for native speaking children ($N = 6,403$) are below the diagonal and correlations for non-native speakers ($N = 788$) are above the diagonal. KG = kindergarten; G1 = grade 1; * $p < .05$; ** $p < .01$

Figure 1*Cross-Lagged Panel Model for the entire sample*

Note. Cross-lagged Panel Model for Language Abilities, Monitoring Accuracy, and Task Performance. Values shown are standardized coefficients. Solid lines = significant paths ($p < .05$). Dotted lines = non-significant paths ($p > .05$).

Figure 2*Cross-Lagged Panel Model for native speaking children*

Note. Cross-lagged Panel Model for Language Abilities, Monitoring Accuracy, and Task Performance. Values shown are standardized coefficients. Solid lines = significant paths ($p < .05$). Dotted lines = non-significant paths ($p > .05$).

Figure 3*Cross-Lagged Panel Model for non-native speaking children*

Note. Cross-lagged Panel Model for Language Abilities, Monitoring Accuracy, and Task Performance. Values shown are standardized coefficients. Solid lines = significant paths ($p < .05$). Dotted lines = non-significant paths ($p > .05$).

8.3 Study 3

Buehler, F. J., Ghetti, S., Roebbers, C. M. (2022). Training Primary School Children's Uncertainty Monitoring [Manuscript to be submitted].

Training Primary School Children's Uncertainty Monitoring

Florian J. Buehler¹, Simona Ghetti², & Claudia M. Roebbers¹

¹University of Bern, Switzerland

²University of California Davis, USA

Abstract

Children's ability to accurately monitor the accuracy of their performance is crucial for self-regulated learning and academic achievement. Training children's uncertainty monitoring, or the ability to experience increased uncertainty when committing a mistake, may be beneficial, but interventions are rare. We sought to evaluate whether it was possible to train uncertainty monitoring about memory. We assigned the participants ($N = 127$; $M = 7.45$ years) to either a metacognitive feedback group, a performance feedback group, or an active control group. Participants first received a baseline recognition memory assessment and provided confidence judgments on each memory decision. Then children in the metacognitive condition received feedback on their performance on a recognition memory task and about the correspondence between their memory performance and their confidence judgments. Children in the performance condition received solely feedback on their recognition memory performance. Children in the active control group solved attention control tasks that differed from the recognition memory task. Each group completed six training sessions. Finally, children completed a new recognition memory task, including confidence judgments. Results revealed that children's uncertainty monitoring increased in the metacognitive condition but not in the performance and active control conditions. Memory accuracy in the recognition test did not increase in any of these conditions. These results underscore the importance of the correspondence between experiences of confidence and feedback to learn how to monitor uncertainty.

Keywords: uncertainty monitoring, training, metacognitive feedback, performance feedback

Training Primary School Children's Uncertainty Monitoring

Children's ability to accurately monitor their uncertainty is crucial for their self-regulated learning and academic achievement (Dunlosky & Metcalfe, 2009; Freeman et al., 2017; Schraw et al., 2006). In the context of memory decisions, uncertainty monitoring is the ability to introspect and evaluate one's memory and includes, for instance, experiencing higher confidence for correct than incorrect memories (Nelson & Narens, 1990). This is critical to recognize errors and the base for self-regulatory processes, such as allocating study time, selecting an answer for rewards, or asking for help (Destan et al., 2014; Hembacher & Ghetti, 2013). Although early in primary school, children's confidence assessments already discriminate between correct and incorrect memories (Geurten & Willems, 2016), there is still much room for developmental improvements. Younger children tend to be overconfident (Destan & Roebbers, 2015; Finn & Metcalfe, 2014) and do not attend to all of the necessary cues (e.g., retrieval fluency) to provide calibrated confidence assessments (Koriat & Ackerman, 2010). Therefore, training uncertainty monitoring may be beneficial for building early prerequisites for lifelong learning. However, interventions targeting children's uncertainty monitoring are rare. The present study compares two training protocols involving different types of feedback to an active control condition.

Improving children's uncertainty monitoring requires understanding the mechanisms underlying developmental constraints in children's untrained abilities. Previous research shows that kindergarten and primary school children do not efficiently use all relevant cues (e.g., retrieval fluency, task difficulty) to inform their confidence assessments. Even though children consider task difficulty when evaluating their memory, they remain overconfident (Destan et al., 2014; van Loon et al., 2017). Overconfidence might be explained by a positively biased memory of past test performance. For example, children aged seven to ten years have been found to overestimate the number of problems solved correctly (Finn & Metcalfe, 2014). Therefore, children likely remain overconfident even when they rely on

previous task performance and difficulty. Other studies show that kindergarten and primary school children do not rely on past task performance when confronted with the same task (Lipko et al., 2009, 2012). In summary, children have difficulty relying on valid cues for uncertainty monitoring. Thereby, feedback on past performance might help guide children's attention and cognitive resources toward relevant and unbiased cues for uncertainty monitoring.

Feedback may improve uncertainty monitoring, as suggested by Efklides' multifaceted and multilevel model of metacognition (Efklides, 2008). The model describes an individual and a social level of metacognition. At the individual level, monitoring is based on knowledge about the task, strategies, goals, confidence, and planning and regulation strategies, and task-inherent feedback. At the social level, children's metacognition is affected by interactions with others, such as peers or teachers, through feedback. The individual and the social levels are reciprocally related. Based on Efklides model, van Loon and Roebbers (2021) have suggested that feedback may be a promising approach to improve children's uncertainty monitoring by targeting the more accessible social level (e.g. van Loon & Roebbers, 2020). In Vygotsky's (1978) terms, feedback might create a zone of proximal development for children's uncertainty monitoring.

How to Encourage Uncertainty Monitoring through Feedback

Feedback may benefit performance as children learn to recognize the essential features of a cognitive task (Destan et al., 2014; Muis et al., 2015; van Loon & Roebbers, 2020). Regarding uncertainty monitoring, previous studies have primarily focused on metacognitive and performance feedback. Metacognitive feedback informs children whether their confidence corresponds to their accuracy (Geurten & Meulemans, 2017; van Loon & Roebbers, 2020). Performance feedback informs children whether their cognitive decision is accurate (O'Leary & Sloutsky, 2017; Oudman et al., 2022; van Loon et al., 2017; van Loon &

Roebbers, 2017, 2020). Each approach has advantages and disadvantages, and we review these characteristics next.

Van Loon and Roebbers (2020) compared metacognitive feedback with performance feedback about an analogical reasoning task in kindergarten children. Compared to a control group receiving no feedback, children in the monitoring and performance feedback groups exhibited better uncertainty monitoring and detected more errors after three training sessions. Moreover, the metacognitive feedback group detected more errors than the performance feedback group. However, there was still much room for improvement, as even children in the metacognitive feedback group did not recognize two-thirds of their errors. In sum, metacognitive feedback might be more beneficial than performance feedback for children's uncertainty monitoring. In line with these findings, studies with adults suggest benefits of metacognitive feedback for monitoring accuracy and even higher performance on the memory tasks associated with the feedback (Callender et al., 2016; Miller & Geraci, 2011; Nietfeld et al., 2006).

There is also evidence that under certain conditions, metacognitive feedback interferes with considering task difficulty for uncertainty monitoring. For example, Geurten and Meulemans (2017) provided 4- to 8-year-olds with metacognitive feedback in an easy or a difficult version of a memory task. After receiving feedback, they solved the opposite version of the same task (i.e., easy-difficult or difficult-easy). Participants who received metacognitive feedback on the easy task overestimated their performance on the difficult task. Participants who received metacognitive feedback on the difficult task underestimated their performance on the easy task. The authors concluded that the children relied on the previous feedback as an anchor to monitor their current memory accuracy instead of relying on task difficulty. Indeed, a control group without feedback accurately monitored their performance in the easy and difficult tasks, indicating that they relied on task difficulty for their uncertainty

monitoring judgments. Metacognitive feedback can prevent children from relying on task difficulty in uncertainty monitoring.

A second feedback approach focuses on task performance. This approach is based on the idea that children will learn to monitor their uncertainty if they attend to the most valid and informative cues, that is, the accuracy of their performance. When children are confronted with trial-by-trial feedback on their accuracy, they might learn to recognize mnemonic cues of their decisions associated with correct or incorrect outcomes, resulting in increased calibration between accuracy and confidence (Efklides & Metallidou, 2020).

In this context, some studies have reported that performance feedback decreased overconfidence in kindergarten and primary school children and increased children's error monitoring in a variety of tasks, including recognition memory, concept learning, and arithmetics (Oudman et al., 2022; van Loon et al., 2017; van Loon & Roebbers, 2017). In these three studies, children indicated monitoring judgments after responding (no feedback) and then again after being given performance feedback. This indicates that children relied on performance when they gave monitoring judgments for the second time. However, whether children benefit from performance feedback on trials that they did not receive direct feedback remains unknown. Compared to the previously outlined studies (Oudman et al., 2022; van Loon et al., 2017; van Loon & Roebbers, 2017), van Loon and Roebbers (2020) provided performance feedback after kindergarten children indicated their monitoring judgments and before the subsequent trial (analogical reasoning task). Uncertainty monitoring accuracy and error recognition were higher in the performance feedback compared to a group receiving no feedback, but children in the performance feedback group remained overconfident. Overall results revealed that kindergarten children benefit from performance feedback; however, they remain overconfident, and there is still much room for improvement. Finally, performance feedback also increased children's task performance.

Other studies have shown that performance feedback failed to reduce overconfidence in preschool and kindergarten children in memory tasks, even after receiving feedback multiple times (Lipko et al., 2009, 2012; Xia et al., 2022). In another study with five-year-olds, overconfidence did not change in a visual discrimination task when children received performance feedback (O’Leary & Sloutsky, 2017). In contrast, third graders’ overconfidence declined across task repetitions (Lipko et al., 2012). This indicates that performance feedback might benefit more advanced primary school children but not preschool and kindergarten children, perhaps because it is insufficient to direct young children’s focus to the most informative and valid cues for their confidence ratings. Moreover, it seems easier to benefit from performance feedback when feedback is provided on each trial immediately before children are asked to provide metacognitive judgments on the same trial (Oudman et al., 2022; van Loon et al., 2017; van Loon & Roebbers, 2017). Compared to when performance feedback is provided on a block of trials and children are allowed to provide overall metacognitive assessment retrospectively after the completion of a block of cognitive trials (i.e. performance postdictions; Lipko et al., 2009, 2012; O’Leary & Sloutsky, 2017; van Loon & Roebbers, 2020; Xia et al., 2022). In fact research on adults revealed that trial-by-trial performance feedback increased monitoring accuracy (Haddara & Rahnev, 2022), while, global performance feedback on the overall task performance did not increase monitoring accuracy (Miller & Geraci, 2011), underscoring the importance of the temporal contiguity between performance and metacognitive or performance feedback is received. In sum, previous research suggests that performance feedback might be most beneficial for school-aged children when performance feedback is available at the moment of monitoring judgments and when performance feedback is provided trial-by-trial.

The present research

Overall, the literature review on metacognitive feedback and performance feedback reveals that studies are sparse and findings are mixed. Critically, none of the studies investigated the effects of metacognitive or performance feedback following each task trial and over a longer time. Instead, participants received feedback only once on a few trials (O’Leary & Sloutsky, 2017; Oudman et al., 2022; van Loon et al., 2017; van Loon & Roebbers, 2017, 2020) or at the end of the task (Geurten & Meulemans, 2017; Lipko et al., 2009, 2012). Improving uncertainty monitoring likely requires multiple repetitions (van Loon & Roebbers, 2021). Finally, whether metacognitive feedback effects transfer to a task without feedback remains unknown. More research on extensive experience with metacognitive feedback is necessary to clarify the role of metacognitive feedback in children’s uncertainty monitoring.

The main goal of the present study was to compare the benefits of metacognitive feedback versus performance feedback for first graders’ uncertainty monitoring. In comparison to previous research (Geurten & Meulemans, 2017; Lipko et al., 2009, 2012; O’Leary & Sloutsky, 2017; van Loon et al., 2017; van Loon & Roebbers, 2017, 2020), we were specifically interested in contrasting the effects of two training conditions, a metacognitive feedback and performance feedback condition, compared to an active control condition. This is a critical step towards better understanding the mixed findings regarding the potential benefits of metacognitive feedback and performance feedback. Critically, we investigated how these training conditions affected children’s uncertainty monitoring and accuracy on a memory task that used different types of stimuli compared to the experimental task. In other words, we used the most conservative but robust approach to examining training effects. If training effects are observed on a different task used during a pre- and post-training session during which no feedback is received, we can be more confident that the effects will transfer across tasks. This is crucial as previous research reveals the transfer of cognitive skills across

tasks is limited (Clerc et al., 2014). The training effects within the the training task are not the focus of the present research and published elsewhere (BLINDED).

Our training conditions were delivered across six sessions on a tablet. Previous research has shown that children have positive attitudes toward tablets and that computerized tasks are suitable for training children's metacognition (Macoun et al., 2022; Muis et al., 2015). The metacognitive feedback group received feedback on their task accuracy and additional feedback about the accuracy of their confidence judgment. The performance feedback group received feedback on their task accuracy. The active control group received feedback on their accuracy in an attention control task. The comparison to an active control group is a strength of the present study. It allows us to ensure that any effect of metacognitive or performance feedback ' is not due to extraneous variables associated with repeated contact with the experimenter or participants' expectations about improvements after repeated testing (e.g., placebo effect; Shawn Green et al., 2019).

We made several predictions. We predicted that uncertainty monitoring abilities in the metacognitive and the performance feedback groups would increase from the pre- to post-training compared to the active control group. Furthermore, we expected metacognitive feedback to be more beneficial than performance feedback (van Loon & Roebbers, 2020). Although our study was primarily focused on training effects on uncertainty monitoring, we also explored whether there would be training effects on memory accuracy (Callender et al., 2016; Muis et al., 2015; Nietfeld et al., 2006; van Loon & Roebbers, 2020). To assess children's uncertainty monitoring, we computed the mean difference between confidence judgments for correct and incorrect recognition memory decisions (Dunlosky et al., 2016; Schraw, 2009), consistent with several studies on metamemory development (Bayard et al., 2021; Fandakova et al., 2017; Hembacher & Ghetti, 2014).

Method

Participants

We recruited 182 participants (age = 89.65m, $SD = 5.81m$; 52% male) from public schools in the vicinity of a mid-sized Swiss town. Participants were predominantly of white background. The mother tongue of most children was German (64%). We assessed parental education as a measure of socioeconomic status: 3% of parents had no education, 9% finished obligatory school, 28% had vocational training, 29% had a high school degree, and 31% had a university degree. This is comparable to the region's average education level (Federal Statistical Office, 2021). We assigned participants' classrooms randomly to one of the training groups: Metacognitive Feedback ($n = 67$), Performance Feedback ($n = 51$), and Active Control ($n = 64$).

Procedure

We recruited the participants through local school districts in Switzerland. Once school districts agreed to participate in our study, we contacted individual teachers within the districts, who assisted with informing families of the upcoming studies. Children whose parents or guardians agreed to participate were assessed. Children also informally agreed to participate. The local ethics committee approved the study (approval number: 2020-10-00005).

We tested children in groups in their usual classroom setting. All tasks were fully computerized and conducted on tablets (Samsung Galaxy Tab S4 and Samsung Galaxy Tab A7) with a touch screen (10.4" and 10.5"). The task instructions were given auditorily via headphones. For technical support and questions, all testings were assisted by two to three trained experimenters. We assessed children's uncertainty monitoring at pre- and posttest. Additionally, we assessed basic cognitive abilities, including measures of receptive grammar, executive functions, working memory, and fluid intelligence. Between the pre- and posttest,

participants completed six training sessions. Pre- and Posttest lasted approximately 60 minutes, and the training sessions were 15 minutes each.

Materials and Measures

Pre- and Posttest

Uncertainty monitoring. We assessed uncertainty monitoring in a paired-associates recognition memory task at pre- and posttest. Similar tasks were used in our previous studies (e.g., Buehler et al., 2021; Destan et al., 2014; Destan & Roebbers, 2015). The task consisted of 16 pairs of images, each including a Japanese Kanji symbol and an image representing its referent. We had two versions of the task so that participants learned different pairs of images in pre- and posttest. The order of the versions was counterbalanced across subjects. Task accuracy was similar at pretest (Accuracy version A = 44%; Accuracy version B = 46%). The task instructions were computerized and included a practice trial to familiarize the participants with the touch screen, the recognition test, and the confidence scale. Only participants who successfully solved the practice trial could progress to the actual task. Participants who made a mistake in the practice trial received an additional face-to-face explanation. The uncertainty monitoring task was divided into four phases: study phase, recognition test, uncertainty monitoring, and sorting task (see Figure 1). We did not analyze data from the sorting task for this report.

In the study phase, participants were told to remember each of the 16 pairs of pictures. The pairs were presented in random order. Each pair was shown for 5 sec. We piloted a large pool of item pairs beforehand to ensure sufficient variability concerning item difficulty. In the study phase, we included pairs with a difficulty index between 0.11 and 0.78. After studying the 16 picture pairs, participants executed a filler task (1 min.) to prevent rehearsal. In the filler task, the participants steered a cat with one finger and tried to catch a mouse.

In the recognition test, the participants saw one Kanji at a time and had to choose the corresponding image out of four alternative images (Figure 1). The participants were familiar with all distractors because they had all been presented during the study time. The distractors were a combination of target images for different Kanjis. The shown distractors were randomized, but the randomization was constrained, so each image was equally often shown as a distractor (two or three times per participant). Participants ought to choose one of the four images by double-clicking to continue the task. The requirement for a double click allowed participants to change their choices if they wanted. As a measure of memory accuracy, we computed the mean percentage of correctly recognized Kanjis out of the 16 to-be-remembered pairs for each participant.

For uncertainty monitoring, participants indicated how confident they felt about each recognition decision immediately after selecting an answer by using a 7-point Likert scale - presented as a thermometer- ranging from very uncertain (blue, coded as 1) to very certain (red, coded as 7), adapted from Koriat and Shitzer-Reichert (2002). Participants had to double-click to confirm their confidence judgment. To compute uncertainty monitoring, we subtracted the mean confidence judgments for incorrectly recognized Kanjis from mean confidence judgments for correctly recognized Kanjis (cf. Dunlosky & Thiede, 2013; Roebbers, 2002). Positive values indicate that participants are more confident when their memory is correct than when it is incorrect.

Training sessions

Metacognitive Feedback. We trained participants' uncertainty monitoring in a paired-associates task, which was different from that used for the pre- and posttraining assessment. Specifically, participants were told to remember 12 pairs of animals and their associated preferred food. We had six topics -one for each appointment- including different pairs of animals and food (e.g., animals from the forest, fish, birds, gnawers, African animals,

and insects). The training task was divided into four phases: study phase, recognition test, uncertainty monitoring, and metacognitive feedback (see Figure 2).

In the study phase, each animal-food pair was shown for 5 sec. After studying the picture pairs, participants executed the same mouse-catching filler task (1 min.) as in pre- and posttest to prevent rehearsal.

In the recognition test, participants chose the corresponding food out of four alternatives for each animal. The participants were familiar with the presented distractors as they were shown as target images for different animals in the learning phase. However, the presented distractors were not randomized and selected based on their perceptual similarity with the target. Participants were forced to choose one of the four images to continue the task. Participants had to double-click to select and confirm their answers.

For uncertainty monitoring, participants indicated a confidence judgment immediately after selecting an answer in the recognition test. Participants had to indicate their confidence on a 4-point Likert scale -presented as smileys- representing *very uncertain*, *uncertain*, *certain*, and *very certain*. Participants had to double-click to select their confidence judgment. We elected to use a different scale for training to prevent children from simply learning to map certain selections on the confidence scale and to ensure transfer in the posttest from the training.

After each confidence judgment in the uncertainty monitoring phase, participants received metacognitive feedback. The metacognitive feedback group received performance feedback and additional feedback on the correspondence between their performance and confidence judgment. Feedback was given visually (green tick for correct recognition and accurate monitoring judgments, red cross for incorrect recognition and inaccurate monitoring judgments) and auditive. Participants received positive feedback for correct recognition and (very) certain judgments (*Yes, that is the right food. It is good that you were (very) certain*

about your answer), and for incorrect recognition and (very) uncertain judgments (*Oh no, that is not the right food. Do not worry it is a difficult task. But it is good that you were (very) uncertain about your answer*). Participants received negative feedback for incorrect recognition and (very) certain judgements (*Oh no, that is not the right food. Don't worry it is a difficult task. But it is too bad that you were (very) certain about your answer*), and correct recognition and (very) uncertain judgments (*Yes, that is the right food. But it is too bad that you were (very) uncertain about your answer*). The feedback was meant to target the social level of uncertainty monitoring, according to Efklides's (2008) multifaceted and multilevel mode of metacognition.

Notably, for practice purposes, two characters involved in the task's cover story solved the first four trials of each training session (recognition, uncertainty monitoring, metacognitive feedback). The characters gave an example for each corresponding performance-monitoring combination (correct and very certain, correct and certain, incorrect and very uncertain, incorrect and uncertain). Consequently, participants responded to the remaining eight trials. Importantly participants did not know which of the 12 learned animal-food pairs would be used as practice trials. The two characters of the cover story were also meant to contribute to a more social atmosphere of the training sessions to target the social level of metacognition (Efklides, 2008) further.

Performance Feedback. This training was identical to the metacognitive condition except that children received exclusively feedback on the accuracy of their recognition decision. Feedback was given visually (green tick for correct recognition, red cross for incorrect recognition) and auditive via headphones (*Yes, that is the right food vs. Oh no, that is not the right food. Do not worry, it is a difficult task*).

Active control group. The active control group executed six attention control tasks on the tablets. This included three Hearts and Flowers task versions (adapted from Davidson

et al., 2006; Diamond et al., 2007) and three Simon task versions (Simon, 1990). The main difference was that the stimuli in both tasks were exchanged with two animals in each session. Moreover, children received auditive and visual performance feedback after each incorrectly solved trial. The task was shortly interrupted with a sound indicating an incorrect answer and a confused smiley appeared on the screen. In each session, different stimuli were used to maintain motivation.

Assessments of individual differences

Parental education. We asked parents in a questionnaire to indicate their highest education level: *0 = no school education; 1 = obligatory school; 2 = vocational training; 3 = High School; 4 = University*. We relied on the highest reported score by one of the parents.

Receptive Grammar. We assessed receptive grammar with a computerized version of the TROG-D (Fox-Boyer, 2011). Participants heard sentences via headphones and had to choose a corresponding picture out of four alternatives. The TROG-D includes 21 blocks with four items each. We used the first block as practice trials in which participants received feedback. The task ended after five consecutive blocks with at least one incorrectly solved item. We computed the sum score of correct blocks (all four items correct) per participant. Possible scores range from 0 to 20 (without the first practice block).

Executive functions. We assessed inhibition and shifting with the Hearts and Flowers task at pre- and posttest (Davidson et al., 2006; Diamond et al., 2007). Only the pretest is relevant for the present study, as we wanted to assess children's basic cognitive abilities. In the Hearts and Flowers task, participants reacted to a heart or a flower presented on the screen's left or right side by pressing external buttons. For hearts, participants had to press the button on the same side as the presented heart. For flowers, the participants had to press the button on the opposite side as the presented flower. The Hearts and Flowers task consisted of congruent, incongruent, and mixed blocks. In the congruent block, only Hearts

were presented (24 trials). In the Incongruent Block, only Flowers were presented (36 trials). In the mixed block, Hearts and Flowers were presented (60 trials; every fourth to sixth trial was a flower).

At the start of each trial, a fixation cross was presented for 500 ms. Next, a Heart or a Flower appeared for 600ms on the screen's right or left. The stimuli were presented until the child responded. Participants were told to respond as accurately and fast as possible in all blocks. Each block started with practice trials at the beginning (Hearts block = two times four practice trials; Flowers block = six practice trials; Mixed block = eight practice trials). The practice trial was repeated when children answered more than two of the practice trials incorrectly. The task ended when participants failed the practice trials two times.

For inhibition, we computed the mean score of correct trials and the mean reaction time of correct trials in the flower block per subject. For shifting, we computed the mean reaction time of correct trials and the mean reaction time for correct trials in the mixed block (hearts and flowers) per subject. We did not compute an inhibition or shifting score for participants who scored below chance level (< 50%) in the flowers or the mixed block ($n = 7$). Based on reaction times (RT), we excluded trials at the anticipatory level ($RT < 250$ ms), and trials with RTs higher than 2500 ms, This concerned overall 3.16% of the trials.

Working Memory. We assessed visuo-spatial working memory with a computerized position Span task (Frick & Möhring, 2016) based on the Corsi-Block-Tapping Task (Corsi, 1972). Participants saw a mole (1200 ms) that appeared and disappeared in different locations on a 4x4 grid. Then they had to indicate the locations they had seen the mole in reverse order. Participants solved three practice trials. If children solved more than one practice trial incorrectly, they received additional face-to-face instructions. The task started with a sequence of six trials with two locations. If at least three out of six trials within a sequence were solved correctly, the number of locations increased by one. The task ended when more

than three trials within a sequence were answered incorrectly (see Maurer & Roebbers, 2021). We relied on the total number of correctly remembered trials to measure working memory. Possible scores range from 0 to 36.

Fluid intelligence. We measured fluid intelligence with a computerized version of the *Odd-Item-Out* task from the RIAS (Hagmann-von Arx & Grob, 2014; Reynolds & Kamphaus, 2003). Participants had to identify an incongruous stimulus in a set of related stimuli. The task ended after three consecutively incorrectly solved matrices. The first four matrices served as practice trials on which the participants received feedback. We computed a sum score per participant as a measure of fluid intelligence. Correct answers within 30 seconds were scored with 2 points, and correct answers within 50 seconds were scored with one point. Possible scores range from 0 to 102. The practice trials are not included in the sum scores.

Statistical Analyses

We preregistered our hypotheses and analyses (<https://osf.io/f3x6k>). We conducted a 2 (Time: pre- vs. posttraining) x3 (Condition: metacognitive feedback vs. performance feedback vs. active control) mixed ANOVA with time as a within-person variable and condition as a between-person variable. The dependent variable was uncertainty monitoring computed as the mean difference in confidence judgments between correct and incorrect trials. To account for group differences at pretest, we also conducted a mixed ANCOVA testing the same model above, including individual difference variables whose average level might differ across groups. We analyzed the data with R (R Core Team, 2021; version: 4.1.1) and conducted mixed ANOVA and ANCOVA with the rstatix package (version: 0.7.0).

Results

Preliminary Analyses

We excluded children who scored below chance level ($< 25\%$; $n = 45$) and children who correctly recognized more than 75% ($n = 10$) of the items during the pretest uncertainty monitoring task. This exclusion was necessary to ensure that all children could complete the task but also exhibited incorrect answers, which is crucial for measuring uncertainty monitoring. We retained 127 participants (age = 89.38m, $SD = 5.55$ m; 53% male) (Metacognitive Feedback ($n = 51$), Performance Feedback ($n = 34$), and Active Control ($n = 42$). Compliance with the training sessions was high. In the metacognitive feedback group 92%, in the performance feedback group 94%, and in the active control group 93% attended all sessions. Children who missed sessions did not attend class because they were ill or had an appointment outside school on the assessment day.

Preliminary analyses revealed that positive uncertainty monitoring scores at pre- and posttest across all conditions, indicating that participants reported higher confidence in correct than incorrect memories at pre- ($M_{correct} = 5.14$, $SD_{correct} = 1.65$; $M_{incorrect} = 4.44$, $SD_{incorrect} = 1.78$) and posttest ($M_{correct} = 5.54$, $SD_{correct} = 1.41$; $M_{incorrect} = 4.79$, $SD_{incorrect} = 1.56$). Paired sample t-test confirmed that the confidence differences between correct and incorrect memories were significantly different from zero at pretest, $t(126) = 7.51$, $p < .001$ and posttest $t(123) = 6.57$, $p < .001$.

We compared the groups (metacognitive feedback, performance feedback, AC) with one-way ANOVAs on the dependent variables memory accuracy and uncertainty monitoring at pretest. Mean scores are displayed in Table 1. Most importantly, the groups did not differ in memory accuracy $F(2, 124) = 1.32$, $p = .20$, and uncertainty monitoring $F(2, 124) = 0.39$, $p = .68$ at pretest.

Moreover, we compared the groups (metacognitive feedback, performance feedback, active control) with one-way ANOVAs on the potential covariates of age, parental education, inhibition accuracy and reaction time, shifting accuracy and reaction time, working memory,

receptive grammar, and fluid intelligence. Mean scores are displayed in Table 1. ANOVAs revealed no significant group differences in inhibition accuracy $F(2, 107) = 0.41, p = .67$, and reaction time $F(2, 107) = 1.56, p = .22$, shifting accuracy $F(2, 91) = 0.94, p = .40$ and reaction time $F(2, 91) = 2.86, p = .06$, and working memory $F(2, 123) = 0.34, p = .72$. However, ANOVAs revealed significant group differences on highest parental education $F(2, 101) = 3.99, p = .02, \eta_p^2 = .07$, receptive grammar $F(2, 123) = 6.18, p < .01, \eta_p^2 = .09$, fluid intelligence $F(2, 123) = 3.03, p = .05, \eta_p^2 = .05$, and age $F(2, 124) = 3.59, p = .03, \eta_p^2 = .06$. Parental education was higher in the metacognitive feedback group than in the active control group (0.54, $CI [0.33, 1.05], p = .03$). Receptive grammar was higher in the metacognitive feedback group than in the performance feedback group (3.18, $CI [1.02, 5.34], p < .01$). Fluid intelligence was higher in the metacognitive feedback group than in the active control group (5.79, $CI [0.19, 11.4], p = .41$). Participants in the active control group were older than participants in the metacognitive feedback group (2.94, $CI [0.24, 5.63], p = .03$). Therefore, we included parental education, receptive grammar, fluid intelligence, and age as covariates in the analyses.

The effects of Performance and Metacognitive feedback on uncertainty monitoring

We hypothesized that participants' uncertainty monitoring benefits from feedback in the metacognitive feedback and performance feedback groups. Therefore, we expected uncertainty monitoring to increase more in the metacognitive feedback and performance feedback groups than in the active control group. We tested this hypothesis with a 2 (Time: pre- vs. posttraining) x 3 (Condition: metacognitive feedback vs. performance feedback vs. active control) mixed ANOVA with time as a within-person variable and condition as a between-person variable. The dependent variable was uncertainty monitoring. The mixed ANOVA revealed no main effects of time $F(1, 121) = 0, p = .95$ or group $F(2, 121) = 0.48, p = .62$, but a time x condition interaction approached statistical significance, $F(2, 121) = 2.97$,

$p = .06$, $\eta_p^2 = .05$. We broke down the significant interaction with simple effect analyses (repeated measure ANOVAs). Simple effect analyses revealed that participants' uncertainty monitoring in the metacognitive feedback group increased from pre- to posttest $F(1, 50) = 4.35$, $p = .04$, $\eta_p^2 = .08$. We did not find simple effects for the performance feedback $F(1, 31) = 0.35$, $p = .56$, or active control groups $F(1, 40) = 1.87$, $p = .18$. Our results confirm our hypothesis that metacognitive feedback is beneficial for children's uncertainty monitoring. However, we did not find an effect of performance feedback on uncertainty monitoring. Our results align with our hypothesis, suggesting that metacognitive feedback is more beneficial for uncertainty monitoring than performance feedback.

We conducted the same model with covariates to account for group differences at pretest. We included parental education, receptive grammar, fluid intelligence, and age as covariates. The covariate receptive grammar was significantly related to uncertainty monitoring $F(1, 96) = 6.63$, $p = .01$, $\eta_p^2 = .07$, however parental education $F(1, 96) = 3.43$, $p = .07$, fluid intelligence $F(1, 96) = 0.27$, $p = .60$, and age $F(1, 96) = 2.35$, $p = .13$, were not related to uncertainty monitoring. The mixed ANCOVA revealed no main effects of time $F(1, 96) = 0$, $p = .99$ or group $F(2, 96) = 0.11$, $p = .90$, but a significant time x condition interaction $F(2, 96) = 5.44$, $p < .01$, $\eta_p^2 = .12$. We broke down the significant interaction with simple effect analyses (repeated measure ANCOVAs). Simple effect analyses revealed that participants' uncertainty monitoring in the metacognitive feedback group increased from pre- to posttest $F(1, 36) = 4.78$, $p = .04$, $\eta_p^2 = .12$. We did not find simple effect for the performance feedback $F(1, 21) = 0.88$, $p = .36$, and active control groups $F(1, 31) = 1.58$, $p = .22$. Including covariates confirmed our main findings that metacognitive feedback, but not performance feedback, is beneficial for children's uncertainty monitoring. The results including marginal estimated means are displayed in Figure 3.

The effects of Performance and Metacognitive feedback on Memory accuracy

Our exploratory analyses involved testing for training effects on memory accuracy. We run a 2 (Time: pre- vs. posttraining) x3 (Condition: metacognitive feedback vs. performance feedback vs. active control) mixed ANCOVA with time as a within-person variable and condition as a between-person variable. The dependent variable was memory accuracy. We included parental education, receptive grammar, fluid intelligence, and age as covariates. None of the covariates was significantly related to memory accuracy: receptive grammar $F(1, 94) = 3.39, p = .069$, parental education $F(3, 94) = 0.92, p = .44$, fluid intelligence $F(1, 94) = 0.15, p = .70$, and age $F(1, 94) = 0.73, p = .40$. The mixed ANCOVA revealed no main effects of time $F(1, 94) = 0.05, p = .83$, or condition $F(2, 94) = 1.08, p = .34$, and no time x condition interaction $F(2, 94) = 1.79, p = .17$. Our results reveal that monitoring and performance feedback did not significantly affect memory accuracy. Moreover, memory accuracy was also stable in the active control condition, indicating that task experience did not affect memory accuracy. The results, including marginal estimated means, are displayed in Figure 4.

Discussion

The overarching goal of the present study was to evaluate the effects of training based on metacognitive or performance feedback on primary school children's uncertainty monitoring. We delivered our training through six computerized feedback sessions. We randomly assigned children to either metacognitive feedback, performance feedback, or an active control group. In line with our hypotheses, children's uncertainty monitoring increased from pre- to posttest in the metacognitive feedback group. Contrary to our expectations, uncertainty monitoring accuracy did not increase from the pre- to posttest in the performance feedback group. As expected, we did not find an improvement in the active control group. Our results indicate that metacognitive feedback is beneficial for children's uncertainty

monitoring; however, performance feedback is not. Finally, neither metacognitive nor performance feedback increased memory accuracy.

Metacognitive feedback improved primary school children's uncertainty monitoring. This is in line with previous research showing the benefits of metacognitive feedback for children's (Geurten & Meulemans, 2017; van Loon & Roebbers, 2020) and adult's uncertainty monitoring (Callender et al., 2016; Miller & Geraci, 2011; Nietfeld et al., 2006). At posttest, children were on average 1.01 points more confident for accurate than inaccurate memories, suggesting that there is still room for additional improvement. This is similar to van Loon and Roebbers (2020), who found that metacognitive feedback increases error detection but that still two-thirds of the errors were not recognized. Also, Oudman et al. (2022) found that metacognitive feedback reduces overconfidence but that children remain overconfident. This might indicate that improving children's monitoring accuracy requires multiple and consistent repetitions. Future interventions should investigate whether more training sessions lead to higher monitoring accuracy and whether there is more generally a dose effect in the relationship between the extent of training and improvements in uncertainty monitoring.

The Performance Feedback condition did not yield any improvements in uncertainty monitoring similar to previous research (Lipko et al., 2009, 2012; O'Leary & Sloutsky, 2017; Xia et al., 2022). In line with these previous studies, our training focused on young primary school children (younger than eight years). Studies with older children found benefits of performance feedback for uncertainty monitoring (Oudman et al., 2022; van Loon & Roebbers, 2017). Performance feedback alone might not sufficiently scaffold young children's recognition and reliance on the most informative cues to uncertainty. However, other studies with children as young as five years old did report performance feedback benefits for children's uncertainty monitoring (van Loon et al., 2017; van Loon & Roebbers, 2020). Van Loon et al. (2017) gave children performance feedback before they indicated their uncertainty

judgment, and the content of the feedback remained visible during children's monitoring judgments. This procedure arguably reduced processing demands. Thus, it may be particularly demanding to identify the cues to accuracy if the task also requires that children keep feedback in mind. This difficulty may be accentuated by the fact that we provided performance feedback after children indicated their uncertainty judgments. Our elected procedure required children to link the feedback retrospectively to memory decisions, and the intervening confidence judgment may have interfered. In sum, the benefits of performance feedback might depend on age group, the timing of performance feedback (before or after the uncertainty judgment), and the interaction of age and timing. It remains up to further research to clarify the most promising approach to improve children's uncertainty monitoring with performance feedback.

Against our expectations, metacognitive and performance feedback did not increase memory accuracy. This finding is opposed to previous research suggesting that performance feedback leads to higher accuracy on related cognitive tasks (Muis et al., 2015; van Loon & Roebbers, 2020), but in these studies, the training and the test utilized the same task. In contrast, we investigated how performance feedback during the intervention session transferred to a task on which feedback was never provided. Young children might need immediate performance feedback to increase their accuracy. Unlike adults (Callender et al., 2016; Nietfeld et al., 2006) but similar to kindergarten children (van Loon & Roebbers, 2020), metacognitive feedback did not increase task performance. This is somewhat surprising given that already three-year-old children rely on their uncertainty monitoring to control learning processes, such as allocating study time, help-seeking, and withdrawing incorrect answers, which is crucial for task performance (Destan et al., 2014; Ghetti et al., 2013; Hembacher & Ghetti, 2013). In the present study, study time was predetermined, participants could not ask for help, and answers could not be withdrawn. Our study design did not provide many

opportunities for participants to self-regulate their learning, which might explain why metacognitive feedback did not increase memory accuracy.

Van Loon and Roebbers (2021) suggested that -based on Efklides (2008) multifaceted and multilevel model of metacognition- interventions targeting the social level of metacognition might be the most promising. For our training sessions, we pre-recorded auditive feedback, which was then delivered via headphones. The recorded feedback could be seen as a highly standardized version of social interaction. In that regard, our findings align with Efklides' (2008) model, showing that metacognitive feedback at the social level can affect uncertainty monitoring at the personal-awareness level. Critically, computerized feedback might be quite different to face-to-face feedback by a social agent. For instance, a social agent can verify that a child actually understands the feedback and it might be more motivating to receive feedback from a social agent than from a recorded voice. This might explain the relatively small effect sizes for metacognitive feedback. It remains up to future research to investigate whether feedback by a social agent is more effective than recorded feedback.

The present study has various strengths. This is the first study to investigate the effects of repeated feedback during training on children's uncertainty monitoring tested on a different task. Moreover, the present study looks at how feedback-based training effects transfer to a task without feedback. The benefits of metacognitive feedback for uncertainty monitoring are especially noteworthy as the recognition task and confidence scale in the pre- and posttest differed from the training sessions. This allows for estimating the robustness of feedback effects for children's uncertainty monitoring. Including an active control group, we account for placebo effects (Shawn Green et al., 2019). Finally, our results indicate similar task difficulty (memory accuracy) at pre- and posttest. Moreover, the groups did not differ in memory accuracy in pre- and posttest. That is crucial to reliably compare uncertainty

monitoring across measurement points and groups because task performance and uncertainty monitoring are reciprocally related (e.g., Fleming & Lau, 2014; Rinne & Mazzocco, 2014; Roebbers & Spiess, 2017).

We also have to acknowledge the limitations of the present study. At the pretest, the conditions differed in various individual difference measures (parental education, receptive grammar, fluid intelligence, and age), possibly due to the random assignment of participants. However, we included these variables as covariates in our model and did not find different results depending on whether those variables were included or not. Therefore, we can assume that the training effects are not due to the different characteristics of the participants assigned to each group.

Conclusion

To the best of our knowledge, this is the first study investigating the benefits of systematic and repeated metacognitive and performance feedback for children's uncertainty monitoring. We found that metacognitive feedback -but not performance feedback- benefits children's uncertainty monitoring. It might be that young primary school children require socially embedded feedback to ameliorate uncertainty monitoring. In practice, teachers could ask children to monitor their performance in exercises and exams, followed by metacognitive feedback. Moreover, metacognitive and performance feedback did not increase memory accuracy, suggesting that perhaps more training sessions, including more items, are necessary for young children to translate their improved uncertainty monitoring into effective recognition decisions. Questions that remain unanswered involve the number of sessions and items per session required to increase uncertainty monitoring, whether feedback effects are task-specific or transfer to different tasks, and the longevity of feedback effects.

References

- Bayard, N. S., van Loon, M. H., Steiner, M., & Roebbers, C. M. (2021). Developmental Improvements and Persisting Difficulties in Children's Metacognitive Monitoring and Control Skills: Cross-Sectional and Longitudinal Perspectives. *Child Development*, 92(3), 1118–1136. <https://doi.org/10.1111/cdev.13486>
- Buehler, F. J., van Loon, M. H., Bayard, N. S., Steiner, M., & Roebbers, C. M. (2021). Comparing metacognitive monitoring between native and non-native speaking primary school students. *Metacognition and Learning*, 16(3), 749–768. <https://doi.org/10.1007/s11409-021-09261-z>
- Callender, A. A., Franco-Watkins, A. M., & Roberts, A. S. (2016). Improving metacognition in the classroom through instruction, training, and feedback. *Metacognition and Learning*, 11(2), 215–235. <https://doi.org/10.1007/s11409-015-9142-6>
- Clerc, J., Miller, P. H., & Cosnefroy, L. (2014). Young children's transfer of strategies: Utilization deficiencies, executive function, and metacognition. *Developmental Review*, 34(4), 378–393. <https://doi.org/10.1016/j.dr.2014.10.002>
- Corsi, P. M. (1972). *Memory And The Medial Temporal Region Of The Brain [unpublished doctoral thesis]*. McGill University.
- Davidson, M. C., Amso, D., Anderson, L. C., & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia*, 44, 2037–2078. <https://doi.org/10.1016/j.neuropsychologia.2006.02.006>
- Destan, N., Hembacher, E., Ghetti, S., & Roebbers, C. M. (2014). Early metacognitive abilities: The interplay of monitoring and control processes in 5- to 7-year-old children. *Journal of Experimental Child Psychology*, 126, 213–228. <https://doi.org/10.1016/j.jecp.2014.04.001>

- Destan, N., & Roebbers, C. M. (2015). What are the metacognitive costs of young children's overconfidence? *Metacognition and Learning, 10*(3), 347–374.
<https://doi.org/10.1007/s11409-014-9133-z>
- Diamond, A., Barnett, W. S., Thomas, J., & Munro, S. (2007). Preschool program improves cognitive control. *Science, 318*(5855), 1387–1388.
<https://doi.org/10.1126/science.1151148>
- Dunlosky, J., & Metcalfe, J. (2009). *Metacognition: A Textbook for Cognitive, Educational, Life Span & Applied Psychology*. Sage Publications.
- Dunlosky, J., Mueller, M. L., & Thiede, K. W. (2016). Methodology for Investigating Human Metamemory: Problems and Pitfalls. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford Handbook of Metamemory*. Oxford University Press.
<https://doi.org/10.1093/oxfordhb/9780199336746.013.14>
- Dunlosky, J., & Thiede, K. W. (2013). Four cornerstones of calibration research: Why understanding students' judgments can improve their achievement. *Learning and Instruction, 24*(1), 58–61. <https://doi.org/10.1016/j.learninstruc.2012.05.002>
- Efklides, A. (2008). Metacognition - Defining Its Facets and Levels of Functioning in Relation to Self-Regulation and Co-regulation. *European Psychologist, 13*(4), 277–287.
<https://doi.org/10.1027/1016-9040.13.4.277>
- Efklides, A., & Metallidou, P. (2020). Applying Metacognition and Self-Regulated Learning in the Classroom. In *Oxford Research Encyclopedia of Education*.
<https://doi.org/10.1093/acrefore/9780190264093.013.961>
- Fandakova, Y., Selmecky, D., Leckey, S., Grimm, K. J., Wendelken, C., Bunge, S. A., & Ghetti, S. (2017). Changes in ventromedial prefrontal and insular cortex support the development of metamemory from childhood into adolescence. *Proceedings of the*

National Academy of Sciences of the United States of America, 114(29), 7582–7587.

<https://doi.org/10.1073/pnas.1703079114>

Federal Statistical Office. (2021). *Höchste abgeschlossene Ausbildung, nach Migrationsstatus, verschiedenen soziodemografischen Merkmalen und Grossregion [Highest level of education, by migration status, various sociodemographic characteristics and region]*.

<https://www.bfs.admin.ch/bfs/en/home/statistics/population/migration-integration/integration-indicators/indicators/highest-educational-level.assetdetail.20164022.html>

Finn, B., & Metcalfe, J. (2014). Overconfidence in children's multi-trial judgments of learning. *Learning and Instruction*, 32, 1–9.

<https://doi.org/10.1016/j.learninstruc.2014.01.001>

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, 8(443), 1–9. <https://doi.org/10.3389/fnhum.2014.00443>

Fox-Boyer, A. V. (2011). *TROG-D. Test zu Überprüfung des Grammatikverständnisses [Grammar comprehension test]*. Idstein Schulz-Kirchner Verlag.

Freeman, E. E., Karayanidis, F., & Chalmers, K. A. (2017). Metacognitive monitoring of working memory performance and its relationship to academic achievement in Grade 4 children. *Learning and Individual Differences*, 57, 58–64.

<https://doi.org/10.1016/j.lindif.2017.06.003>

Frick, A., & Möhring, W. (2016). A matter of balance: Motor control is related to children's spatial and proportional reasoning skills. *Frontiers in Psychology*, 6(2049), 1–10.

<https://doi.org/10.3389/fpsyg.2015.02049>

Geurten, M., & Meulemans, T. (2017). The effect of feedback on children's metacognitive

- judgments: a heuristic account. *Journal of Cognitive Psychology*, 29(2), 184–201.
<https://doi.org/10.1080/20445911.2016.1229669>
- Geurten, M., & Willems, S. (2016). Metacognition in Early Childhood: Fertile Ground to Understand Memory Development? *Child Development Perspectives*, 10(4), 263–268.
<https://doi.org/10.1111/CDEP.12201>
- Ghetti, S., Hembacher, E., & Coughlin, C. A. (2013). Feeling uncertain and acting on it during the preschool years: A metacognitive approach. *Child Development Perspectives*, 7(3), 160–165. <https://doi.org/10.1111/cdep.12035>
- Haddara, N., & Rahnev, D. (2022). The Impact of Feedback on Perceptual Decision-Making and Metacognition: Reduction in Bias but No Change in Sensitivity. *Psychological Science*, 33(2), 259–275. <https://doi.org/10.1177/09567976211032887>
- Hagmann-von Arx, P., & Grob, A. (2014). *RIAS. Reynolds Intellectual Assessment Scales and Screening. Deutschsprachige Adaptation der Reynolds Intellectual Assessment Scales (RIAS) & des Reynolds Intellectual Screening Test (RIST) von Cecil R. Reynolds und Randy W. Kamphaus [German version]*. Huber.
- Hembacher, E., & Ghetti, S. (2013). How to bet on a memory: Developmental linkages between subjective recollection and decision making. *Journal of Experimental Child Psychology*, 115(3), 436–452. <https://doi.org/10.1016/j.jecp.2013.03.010>
- Hembacher, E., & Ghetti, S. (2014). Don't Look at My Answer: Subjective Uncertainty Underlies Preschoolers' Exclusion of Their Least Accurate Memories. *Psychological Science*, 25(9), 1768–1776. <https://doi.org/10.1177/0956797614542273>
- Koriat, A., & Ackerman, R. (2010). Choice latency as a cue for children's subjective confidence in the correctness of their answers. *Developmental Science*, 13(3), 441–453.
<https://doi.org/10.1111/j.1467-7687.2009.00907.x>

- Koriat, A., & Shitzer-Reichert, R. (2002). Metacognitive Judgments and their Accuracy. In P. Chambres, M. Izaute, & P.-J. Marescaux (Eds.), *Metacognition: Process, Function and Use* (pp. 1–17). Springer. https://doi.org/10.1007/978-1-4615-1099-4_1
- Lipko, A. R., Dunlosky, J., Lipowski, S. L., & Merriman, W. E. (2012). Young Children are not Underconfident With Practice: The Benefit of Ignoring a Fallible Memory Heuristic. *Journal of Cognition and Development, 13*(2), 174–188. <https://doi.org/10.1080/15248372.2011.577760>
- 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. <https://doi.org/10.1016/j.jecp.2008.10.002>
- Macoun, S. J., Pyne, S., MacSween, J., Lewis, J., & Sheehan, J. (2022). Feasibility and potential benefits of an attention and executive function intervention on metacognition in a mixed pediatric sample. *Applied Neuropsychology: Child, 11*(3), 240–252. <https://doi.org/10.1080/21622965.2020.1794867>
- Maurer, M. N., & Roebbers, C. M. (2021). New insights into visual-motor integration exploring process measures during copying shapes. *Psychology of Sport and Exercise, 55*(101954), 1–9. <https://doi.org/10.1016/j.psychsport.2021.101954>
- Miller, T. M., & Geraci, L. (2011). Training metacognition in the classroom: The influence of incentives and feedback on exam predictions. *Metacognition and Learning, 6*(3), 303–314. <https://doi.org/10.1007/s11409-011-9083-7>
- Muis, K. R., Ranellucci, J., Trevors, G., & Duffy, M. C. (2015). The effects of technology-mediated immediate feedback on kindergarten students' attitudes, emotions, engagement and learning outcomes during literacy skills development. *Learning and Instruction, 38*, 1–13. <https://doi.org/10.1016/j.learninstruc.2015.02.001>

- Nelson, T. O., & Narens, L. (1990). Metamemory: A Theoretical Framework and New Findings. *Psychology of Learning and Motivation - Advances in Research and Theory*, 26, 125–173.
- Nietfeld, J. L., Cao, L., & Osborne, J. W. (2006). The effect of distributed monitoring exercises and feedback on performance, monitoring accuracy, and self-efficacy. *Metacognition and Learning*, 1(2), 159–179. <https://doi.org/10.1007/s10409-006-9595-6>
- O’Leary, A. P., & Sloutsky, V. M. (2017). Carving Metacognition at Its Joints: Protracted Development of Component Processes. *Child Development*, 88(3), 1015–1032. <https://doi.org/10.1111/cdev.12644>
- Oudman, S., van de Pol, J., & van Gog, T. (2022). Effects of self-scoring their math problem solutions on primary school students’ monitoring and regulation. *Metacognition and Learning*, 17(1), 213–239. <https://doi.org/10.1007/s11409-021-09281-9>
- R Core Team. (2021). *R: A language and environment for statistical computing [Computer software]*. <https://www.r-project.org/>
- Reynolds, C. R., & Kamphaus, R. W. (2003). *RIAS. Reynolds Intellectual Assessment Scales*. PAR.
- Rinne, L. F., & Mazocco, M. M. M. (2014). Knowing Right from Wrong in Mental Arithmetic Judgments: Calibration of Confidence Predicts the Development of Accuracy. *PLOS ONE*, 9(7), 1–11. <https://doi.org/10.1371/journal.pone.0098663>
- Roebbers, C. M. (2002). Confidence judgments in children’s and adults’ event recall and suggestibility. *Developmental Psychology*, 38(6), 1052–1067. <https://doi.org/10.1037/0012-1649.38.6.1052>
- Roebbers, C. M., & Spiess, M. (2017). The Development of Metacognitive Monitoring and Control in Second Graders: A Short-Term Longitudinal Study. *Journal of Cognition and*

- Development*, 18(1), 110–128. <https://doi.org/10.1080/15248372.2016.1157079>
- Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning*, 4, 33–45. <https://doi.org/10.1007/s11409-008-9031-3>
- Schraw, G., Crippen, K. J., & Hartley, K. (2006). Promoting Self-Regulation in Science Education: Metacognition as Part of a Broader Perspective on Learning. *Research in Science Education*, 36, 111–139. <https://doi.org/10.1007/s11165-005-3917-8>
- Shawn Green, C., Bavelier, D., Kramer, A. F., Vinogradov, S., Ansorge, U., Ball, K. K., Bingel, U., Chein, J. M., Colzato, L. S., Edwards, J. D., Facoetti, A., Gazzaley, A., Gathercole, S. E., Ghisletta, P., Gori, S., Granic, I., Hillman, C. H., Hommel, B., Jaeggi, S. M., ... Witt, C. M. (2019). Improving Methodological Standards in Behavioral Interventions for Cognitive Enhancement. *Journal of Cognitive Enhancement*, 3(1), 2–29. <https://doi.org/10.1007/s41465-018-0115-y>
- Simon, J. R. (1990). The effects of an irrelevant directional cue on human information processing. *Advances in Psychology*, 65, 31–86. [https://doi.org/10.1016/S0166-4115\(08\)61218-2](https://doi.org/10.1016/S0166-4115(08)61218-2)
- van Loon, M. H., Destan, N., Spiess, M. A., de Bruin, A., & Roebbers, C. M. (2017). Developmental progression in performance evaluations: Effects of children’s cue-utilization and self-protection. *Learning and Instruction*, 51, 47–60. <https://doi.org/10.1016/j.learninstruc.2016.11.011>
- van Loon, M. H., & Roebbers, C. M. (2017). Effects of Feedback on Self-Evaluations and Self-Regulation in Elementary School. *Applied Cognitive Psychology*, 31(5), 508–519. <https://doi.org/10.1002/acp.3347>
- van Loon, M. H., & Roebbers, C. M. (2020). Using feedback to improve monitoring judgment accuracy in kindergarten children. *Early Childhood Research Quarterly*, 53, 301–313.

<https://doi.org/10.1016/j.ecresq.2020.05.007>

van Loon, M. H., & Roebbers, C. M. (2021). Using Feedback to Support Children when Monitoring and Controlling Their Learning. In D. Moraitou & P. Metallidou (Eds.), *Trends and Prospects in Metacognition Research across the Life Span. A Tribute to Anastasia Efklides*. (pp. 161–184). Springer. https://doi.org/10.1007/978-3-030-51673-4_8

Vygotsky, L. S. (1978). Interaction between Learning and Development. In M. Gauvain & M. Cole (Eds.), *Readings on the Development of Children* (pp. 34–40). Scientific American Books.

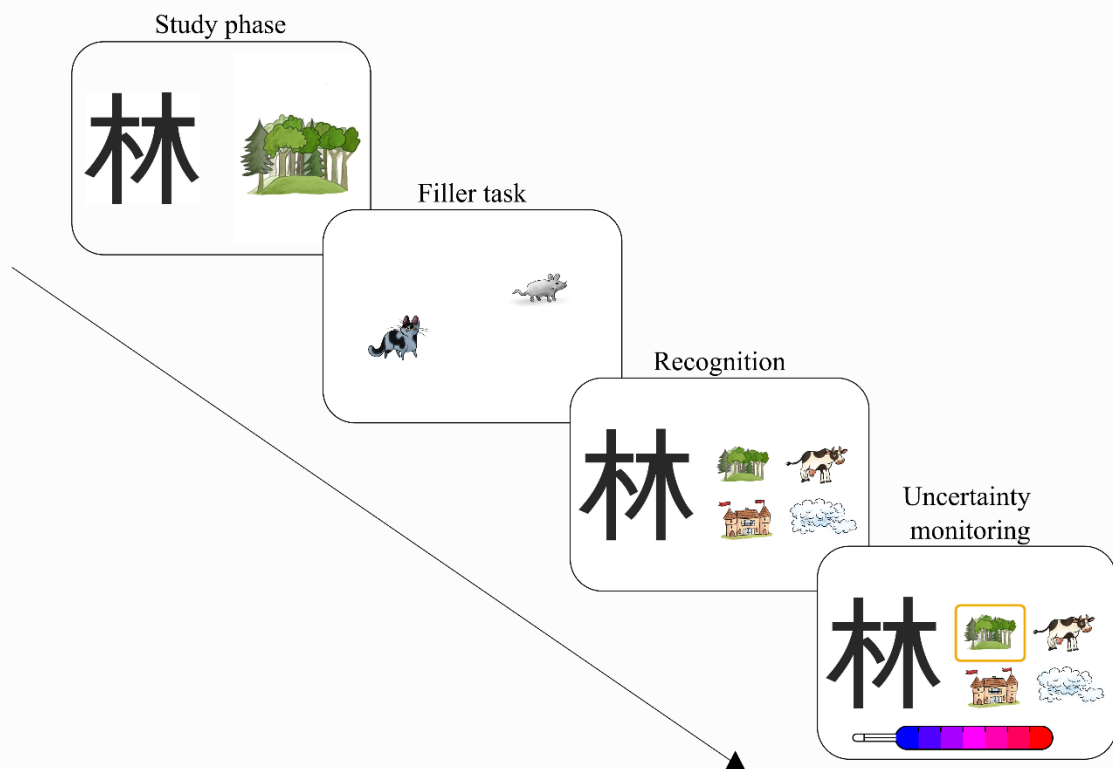
Xia, M., Poorthuis, A. M. G., Zhou, Q., & Thomaes, S. (2022). Young children's overestimation of performance: A cross-cultural comparison. *Child Development, 93*(2), e207–e221. <https://doi.org/10.1111/cdev.13709>

Table 1*Descriptive statistics including Means and SDs in parentheses*

	Metacognitive FB	Performance FB	Active Control
<i>N</i>	51	34	42
<i>Dependent variables</i>			
Memory accuracy pretest [%]	46.94 (12.08)	43.38 (10.54)	43.75 (11.38)
Memory accuracy posttest [%]	48.65 (18.85)	48.24 (16.29)	43.60 (20.13)
Uncertainty monitoring pretest	0.62 (1.01)	0.67 (1.14)	0.82 (1.04)
Uncertainty monitoring posttest	1.01 (1.19)	0.52 (1.53)	0.61 (1.12)
<i>Covariates</i>			
Age [months]	88.25 (5.4)	88.82 (5)	91.19 (5.82)
Parental education	3.22 (0.96)	2.69 (0.93)	2.68 (0.94)
Inhibition accuracy [%]	0.89 (0.16)	0.91 (0.1)	0.91 (0.16)
Inhibition RT [ms]	759.45 (151.84)	710.17 (135.98)	757.36 (141.78)
Shifting accuracy [%]	0.85 (0.07)	0.88 (0.05)	0.85 (0.1)
Shifting RT [ms]	778.69 (130.24)	708.82 (133.89)	732.16 (116.02)
Working memory	6.57 (3.55)	6.85 (3.87)	7.19 (3.57)
Receptive grammar	12.18 (3.77)	9 (4.66)	10.6 (3.93)
Fluid intelligence	42.08 (10.35)	39.09 (11.05)	36.29 (12.61)

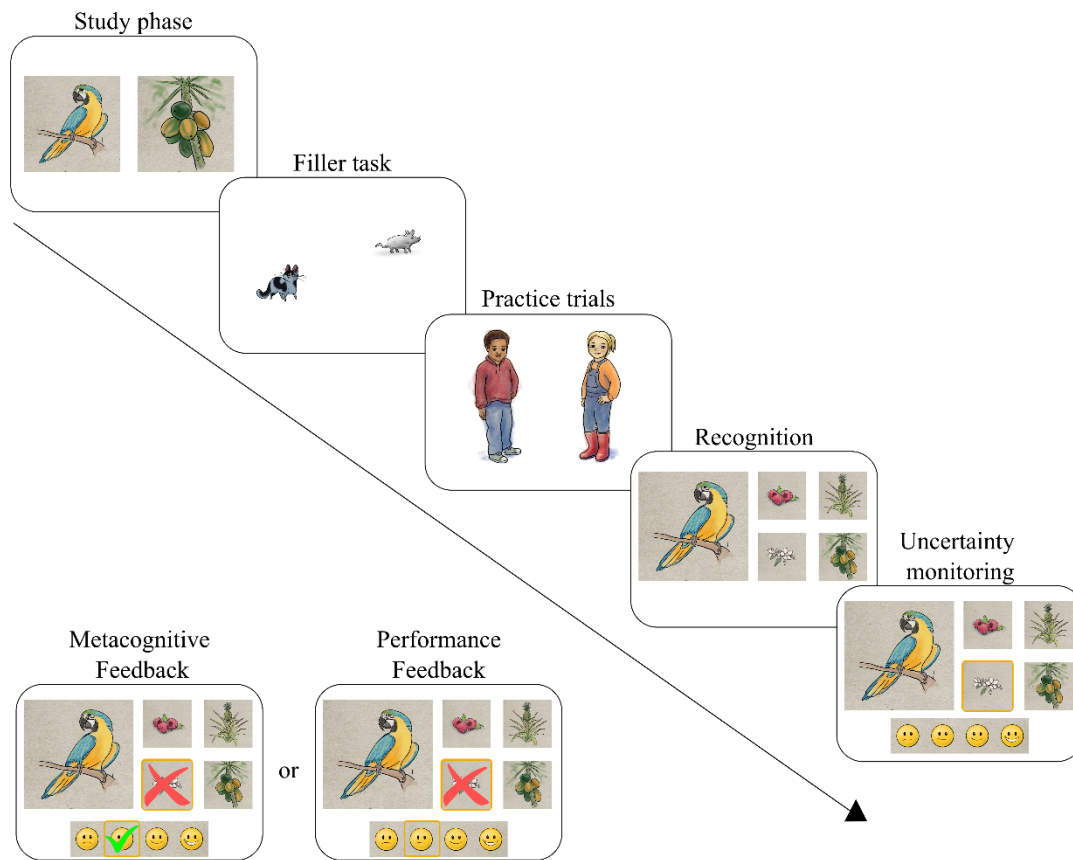
Note. FB = Feedback; RT = reaction time; Uncertainty Monitoring = Confidence Judgement correct - Confidence Judgement incorrect; significant group differences are bolt.

Figure 1
Uncertainty monitoring task at pretest and posttest

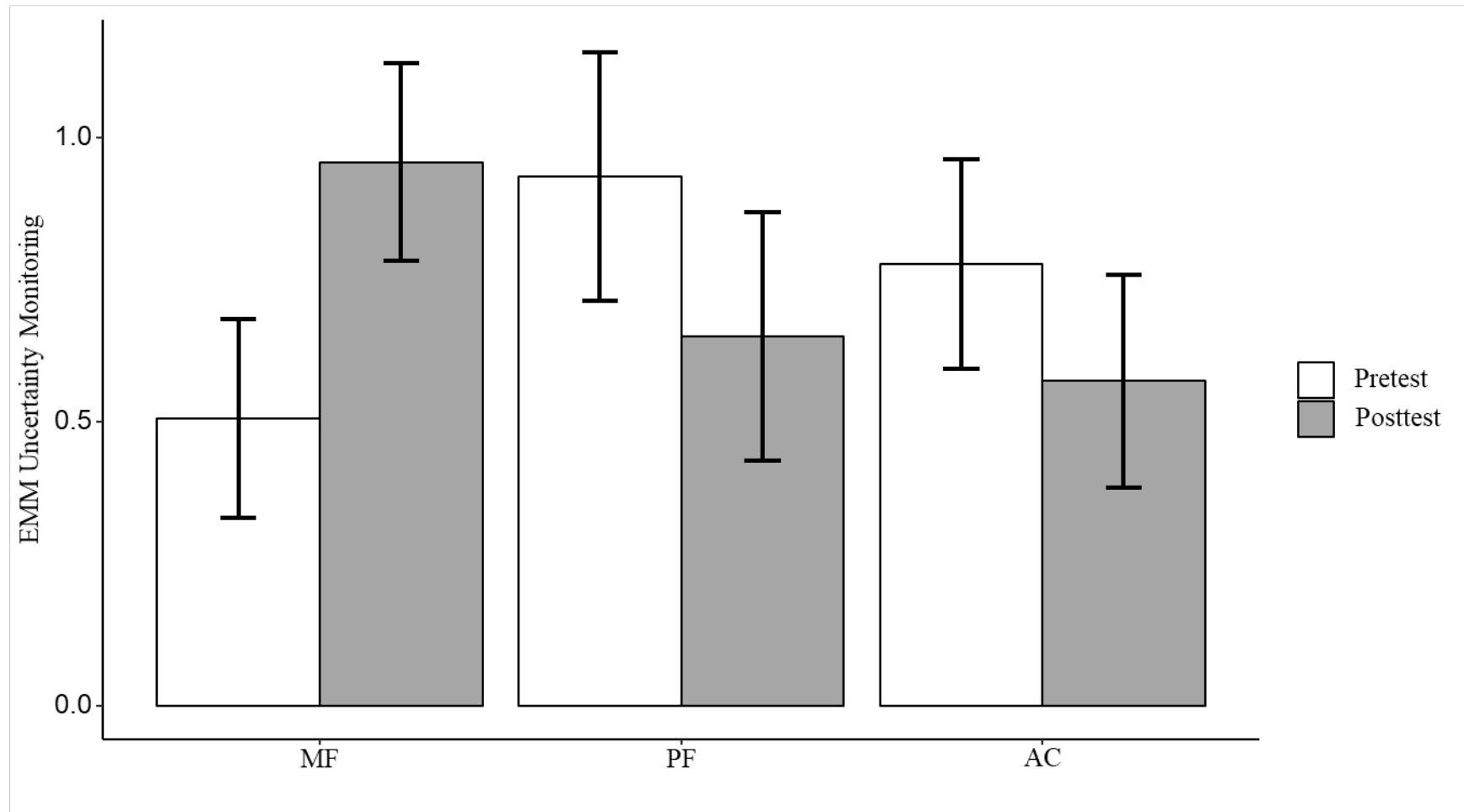


Note. Study phase: learning 16 Kanji-picture pairs; Filler task: 1 min. mouse-catching game; Recognition: recognizing the corresponding picture out of 4 options; Uncertainty Monitoring: indicating Confidence Judgments for each answer.

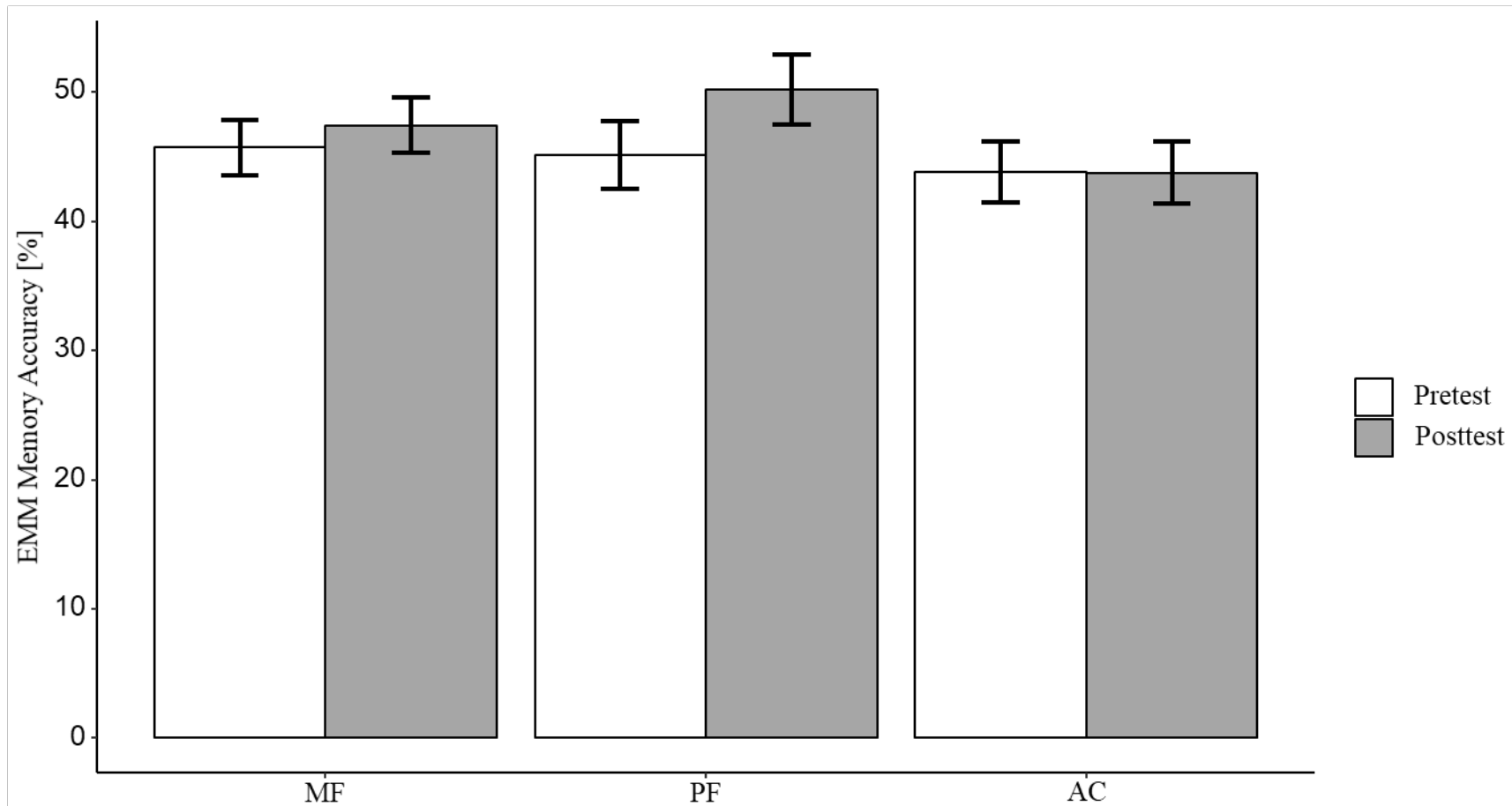
Figure 2
Training sessions



Note. Study phase: learning 12 Animal-food pairs; Filler task: 1 min. mouse-catching game; Practice trials: four trials solved by characters introduced in the cover story; Recognition: recognizing the corresponding picture out of four options; Uncertainty Monitoring: indicating Confidence Judgments for each answer; Metacognitive Feedback: Feedback on recognition and monitoring; Performance Feedback: Feedback on recognition.

Figure 3*Uncertainty Monitoring at pretest and posttest*

Note. Values shown are marginal estimated means for uncertainty monitoring with standard errors. Uncertainty monitoring = Confidence Judgments on accurate trials – Confidence Judgements on inaccurate trials. MF = Metacognitive Feedback; PF = Performance Feedback; AC = Active control group. Only the MF group increased in memory monitoring from pre- to posttest.

Figure 4*Memory accuracy at pretest and posttest*

Note. Values shown are marginal estimated means for memory accuracy with standard errors. Memory accuracy = percentage of correctly recognized trials. MF = Metacognitive Feedback; PF = Performance Feedback; AC = Active control group. None of the groups increased in memory accuracy from pre- to posttest.

9 Erklärung zur Dissertation

Philosophisch-humanwissenschaftliche Fakultät
 Dekanat
 Fabrikstrasse 8, CH-3012 Bern

u^b

b
**UNIVERSITÄT
 BERN**

Erklärung zur Dissertation

Hiermit bestätige ich, dass ich die Dissertation (Titel):

Sociocultural Aspects of Metacognitive Monitoring

im Fach Entwicklungspsychologie

unter der Leitung von Prof. Dr. Claudia Roebers

ohne unerlaubte Hilfe ausgeführt und an keiner anderen Universität zur Erlangung eines akademischen Grades eingereicht habe.

10.01.2023
 Datum


 Unterschrift