PhD-FHSE-2024-011
The Faculty of Humanities, Education and Social Sciences

DISSERTATION

Presented at 28/03/2024 in Esch-sur-Alzette

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN PSYCHOLOGIE

by

Valentin EMSLANDER
Born on 14 July 1993 in Rastatt (Germany)

An Exploration of Factors Driving School Success in Diverse Students Through Meta-Analytic and Value-Added Modeling

Dissertation defense committee

Dr Antoine Fischbach, dissertation supervisor
Associate Professor, University of Luxembourg

Dr Ronny Scherer
Professor, University of Oslo

Dr Doris Holzberger
Associate Professor, Technical University of Munich

Dr Andreas Gegenfurtner
Professor, University of Augsburg

Dr Christine Schiltz, Chairperson
Professor, University of Luxembourg

“Do whatever it takes to avoid fooling yourself

into thinking something is true that is not,

or that something is not true that is.”

Neil deGrasse Tyson

Summary

In Luxembourg, the student population is diversifying regarding socioeconomic and language background (Klein & Peltier, 2022; STATEC, 2021). This surge in diversity gave rise to educational inequality in the trilingual educational system already in primary school (see Hadjar et al., 2018; Hoffmann et al., 2018; Sonnleitner et al., 2021). While increased educational inequalities would naturally deteriorate performance in international comparisons, Luxembourg remains stable in its test results, for example, in the Programme for International Student Assessment (PISA; Weis et al., 2018). These stable test results suggest several schools must use strategies to effectively address educational inequalities and support their students against the odds. Motivated by the narrative literature review in the Theoretical Considerations section, the present thesis explores factors driving school success such as teaching quality, teacher-student relationships (TSRs), or cognitive functions, in diverse students internationally and specifically in Luxembourg. The Grand Duchy can be seen as a living laboratory for the multilingual and diverse future of educational systems in a globalized world.

Thus, the present thesis aims to identify effective educational psychological practices to address educational inequalities in a diverse, multilingual student population in the host of potential variables, schools, and stakeholders' perspectives. Two interrelated Research Strands follow this aim using open science practices. Research Strand 1 explores general social and cognitive learning processes in the international literature through meta-analytic methods in Studies 1 and 2. Study 1 delves into the link between TSRs and student outcomes. Study 2 aims to deepen our understanding of executive functions (EFs) and their relation to mathematics skills. Research Strand 2 focuses on concrete educational effectiveness measures and educational psychological practices in Luxembourg to help all students succeed. As such, our objective was to identify primary schools with stable performance against the odds (Study 3) and compare educational psychological practices in such schools to successfully address educational inequalities (Study 4).

Specifically, Study 1 (Research Strand 1) presents a preregistered systematic review of meta-analyses plus original second-order meta-analyses (SOMAs). We synthesized over 70 years of research on TSRs by aggregating 24 meta-analyses encompassing a total of 116 effect sizes based on more than 2 million prekindergarten and K-12 students. Several three-level SOMAs indicated that TSRs had similarly strong significant relations with eight clusters of outcomes: academic achievement, academic emotions, appropriate student behavior, behavior problems, EFs and self-control, motivation, school belonging and engagement, and student

well-being. Age, gender, and informant (i.e., student, peer, or teacher report) were the most frequently examined moderators in prior research, and our original moderator analyses suggested student grade level and social minority status as moderators. We further identified meaningful differences in quality between the meta-analyses, and these differences were not associated with the TSR-outcome links. Mapping the field of TSR research, Study 1 indicates how TSRs could contribute to improving student outcomes via relationship building.

Study 2 (Research Strand 1) utilized data extracted from 47 preschool studies (363 effect sizes, 30,481 participants) from 2000 to 2021. We found that, overall, EFs are significantly related to math intelligence ($\bar{r} = .34$, 95% CI [.31, .37]). This link holds for all three EF subdimensions, that is, inhibition ($\bar{r} = .30$, 95% CI [.25, .35]), shifting ($\bar{r} = .32$, 95% CI [.25, .38]), and updating ($\bar{r} = .36$, 95% CI [.31, .40]). Key measurement characteristics of EFs (e.g., task type), but neither children's age nor gender, moderated this relation. These findings indicate a positive link between EFs and math intelligence in preschool children and emphasize the importance of measurement characteristics. Further, we examined the joint relations between EFs and math intelligence with meta-analytic structural equation modeling. Evaluating different models and representations of EFs, we found no support for the expectation that the three EF subdimensions are differentially related to math intelligence.

Study 3 (Research Strand 2) examined the stability of value-added (VA) scores over time for mathematics and language learning, as VA models are widely used for accountability purposes. We drew on representative, large-scale, and longitudinal data from two cohorts of standardized achievement tests in Luxembourg ($N = 7{,}016$ students in 151 schools). We found that only 34-38% of the schools showed stable VA scores over time, with moderate rank correlations of VA scores from 2017 to 2019 of $r = .34$ for mathematics and $r = .37$ for language learning. Although they showed insufficient stability over time for high-stakes decision-making, school VA scores could be employed to identify teaching or school practices that are genuinely effective—especially in heterogeneous student populations.

Study 4 (Research Strand 2) compared what schools with high VA scores do more successfully than schools with medium or low VA scores. Based on the results of Study 3, we selected 16 schools with stable high, medium, or low VA scores as promising target schools for this comparison. The assessed variables included, for example, instructional quality (Klieme et al., 2001), school climate (Wang & Degol, 2016), TSR, boredom, and collective teacher self-efficacy (Hattie, 2008; Waack, 2018). Additionally, we measured previously identified Luxembourg specificities, such as language use and the perceived role of the school

president. In a multi-perspective, mixed-methods data collection in 49 classrooms, we conducted observations and collected questionnaire data on 511 students in Grade 2, their 410 parents, 191 classroom and subject teachers, 14 school presidents, and 13 regional directors. Our sample, roughly 10 % of all primary schools and Grade 2 students, is somewhat representative of the general student population and showed similar means in instructional quality, TSR, and school climate as in other European countries. While schools with a stable high VA score did not differ from the other two groups on most variables, we found some evidence that teachers' acknowledgment and inclusion of the student's home language could be one of the drivers of school differences and thus help reduce educational inequities.

In conclusion, the four studies identified several factors driving school success in diverse student populations using meta-analytic and VA approaches. Teachers should strengthen positive and avoid negative TSR because they correlate with crucial student variables in students of any age. EFs are already related to math skills in preschoolers but must be assessed and practiced separately, as they are distinct. Especially in multilingual classrooms, teachers should use their students' home languages to support them following the lesson, activate them cognitively, and create a welcoming learning climate. Limitations of the current thesis's meta-analytic and VA approaches indicate the need for future research to advance the field and find further the factors driving school success in diverse students.

Included and Associated Publications

Included Publications

Emslander, V., Holzberger, D., Ofstad, S. B., Fischbach, A., & Scherer, R. (2023, September 12). *Teacher-student relationships and student outcomes: A systematic review of meta-analyses and second-order meta-analysis* [Manuscript under revision at *Psychological Bulletin*]. PsyArXiv. https://doi.org/10.31234/osf.io/qxntb

Emslander, V., & Scherer, R. (2022). The relation between executive functions and math intelligence in preschool children: A systematic review and meta-analysis. *Psychological Bulletin, 148*(5-6), 337–369. https://doi.org/10.1037/bul0000369

Emslander, V., Levy, J., Scherer, R., & Fischbach, A. (2022). Value-added scores show limited stability over time in primary school. *PLoS ONE, 17*(12), e0279255. https://doi.org/10.1371/journal.pone.0279255

Emslander, V., Rosa, C., Ofstad, S. B., Levy, J., & Fischbach, A. (2023). *Instructional quality and school climate in Luxembourg's écoles fondamentales: Findings from the SIVA study*. Manuscript in principle accepted for the National Education Report 2024, Luxembourg.

Associated Preregistrations

Emslander, V., Levy, J., & Fischbach, A. (2022, March 22). *systematic identification of high "value-added" in educational contexts (SIVA)* [Preregistration]. https://doi.org/10.17605/OSF.IO/X3C48

Emslander, V., Holzberger, D., Fischbach, A., & Scherer, R. (2022, March 31). *The associations between teacher-student-relationships and student outcomes: A systematic review of meta-analyses (ReMA-TSR)* [Preregistration]. https://doi.org/10.17605/OSF.IO/J2EMF

Scherer, R., & **Emslander, V.** (2019, May 05). *The relation between neuropsychological measures of executive functioning and psychometric intelligence: a meta-analysis* [Preregistration]. PROSPERO. https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42017076291

Associated Data Sets

Emslander, V., Holzberger, D., Ofstad, S. B., Fischbach, A., & Scherer, R. (2023). *Teacher-student relationships and student outcomes: A systematic review of meta-analyses and second-order meta-analysis* [Data set]. https://doi.org/10.17605/OSF.IO/5DE6P

Emslander, V., & Scherer, R. (2022). The relation between executive functions and math intelligence in preschool children: A systematic review and meta-analysis [Data set]. https://doi.org/10.17605/OSF.IO/T67QA

Emslander, V., Levy, J., Scherer, R., & Fischbach, A. (2022). Value-added scores show limited stability over time in primary school [Data set]. https://doi.org/10.17605/OSF.IO/VWD73

Associated Presentations and Posters

Emslander, V., Holzberger, D., Ofstad, S.B., Fischbach, A., & Scherer, R. (2023, November 8-9). *What we can learn from student-teacher relationship research for Luxembourg: A systematic review of meta-analyses and second-order meta-analysis*. Presentation at the Luxembourg Educational Research Association (LuxERA) Emerging Researchers' Conference 2023 at the University of Luxembourg, Luxembourg. https://hdl.handle.net/10993/57339

Emslander, V., Levy, J., & Fischbach, A. (2023, September 18). *Was Luxemburgs Grundschulen richtig machen: Ein Value-Added-Vergleich im luxemburgischen Schulkontext* [What Luxembourg's Primary Schools Are Doing Right: A Value-Added Comparison in the Luxembourgish School Context; Poster presentation]. PAEPS 2023 19. Fachgruppentagung Pädagogische Psychologie at the Kiel University, Germany. http://hdl.handle.net/10993/56027

Emslander, V. (Ed.) (2023, August 25). *Teacher-student relationships in education—What we know and what we don't (yet) know*. Symposium at the 20th Biennial EARLI Conference, Thessaloniki, Greece. http://hdl.handle.net/10993/55881

Emslander, V., Holzberger, D., Fischbach, A., & Scherer, R. (2023, August 25). Seeing the connection: A systematic review of meta-analyses on the link between teacher-student relationships and student outcomes. In **Emslander, V.** (Ed.) (2023, August 22 – 26). *Teacher-student relationships in education—What we know and what we don't (yet) know*. Symposium at the 20th Biennial EARLI Conference, Thessaloniki, Greece. http://hdl.handle.net/10993/55883

Emslander, V., & Holzberger, D. (Eds.) (2023, February 28 – March 2). *Lehrer-Schüler-Beziehungen: Von der Generalisierbarkeit positiver Befunde [Teacher-student relationships: Of the generalizability of positive findings]*. Symposium at the 10th Conference of the Society for Empirical Educational Research (GEBF), Essen, Germany. http://hdl.handle.net/10993/54531

Emslander, V., Holzberger, D., Fischbach, A., & Scherer, R. (2023, February 28 – March 2). Lehrer-Schüler-Beziehungen und ihre Korrelate: Ein systematisches Review von Meta-Analysen [Teacher-student relationships and their correlates: A systematic review of meta-analyses]. In **Emslander, V.,** & Holzberger, D. (Eds.) (2023, February 28 – March 2). *Lehrer-Schüler-Beziehungen: Von der Generalisierbarkeit positiver Befunde [Teacher-Student Relationships: Of the Generalizability of Positive Results]*. Symposium at the 10th Conference of the Society for Empirical Educational Research (GEBF) at the University of Essen, Germany. http://hdl.handle.net/10993/54530

Emslander, V., Levy, J., & Fischbach, A. (2022, November 9-10). *What primary schools are doing right: Educational value-added in Luxembourg* [Poster presentation]. Presentation at the Luxembourg Educational Research Association (LuxERA) Conference 2022 at the University of Luxembourg, Luxembourg. http://hdl.handle.net/10993/52911

Emslander, V., Holzberger, D., Fischbach, A., & Scherer, R. (2022, November 8-9). *Teacher-student-relationships and student outcomes in heterogeneous educational settings: A systematic review of meta-analyses* [Paper presentation]. Presentation at the CIDER-LERN Conference 2022 at the University of Luxembourg, Luxembourg. http://hdl.handle.net/10993/52697

Emslander, V., Fischbach, A., & Scherer, R. (2022, June 6-9). *The associations between teacher-student-relationships and student outcomes: A systematic review of meta-analyses* [Paper presentation]. 10th SELF Biennial International Conference in Quebec City, Canada. (Withdrawn due to COVID-19 pandemic)

Emslander, V., Levy, J., Scherer, R., Brunner, M., & Fischbach, A. (2022, March 9-11). *Are value-added scores stable enough for high-stakes decisions?* [Paper presentation]. Presentation at virtual 9th Conference of the Society for Empirical Educational Research (GEBF) 2022. http://hdl.handle.net/10993/48865

Emslander, V., Levy, J., Scherer, R., Brunner, M., & Fischbach, A. (2021, November 10-11). *Stability of primary school value-added scores over time: A comparison between math and language achievement as outcome variables* [Paper presentation]. Presentation at Luxembourg Educational Research Association (LuxERA) Virtual Emerging Researchers' Conference 2021. http://hdl.handle.net/10993/48118

Emslander, V., & Scherer, R. (2021, September 14-16). *Meta-analytic structural equation models of executive functions and math intelligence in preschool children* [Paper presentation]. PAEPSY 2021 Tagung der Fachgruppe Pädagogische Psychologie, virtual conference. http://hdl.handle.net/10993/47866

Emslander, V., Levy, J., Scherer, R., Brunner, M., & Fischbach, A. (2021, September 14-16). *Stability of value-added models: Comparing classical and machine learning approaches* [Paper presentation]. PAEPSY 2021 Tagung der Fachgruppe Pädagogische Psychologie, virtual conference. http://hdl.handle.net/10993/48087

Emslander, V., & Scherer, R. (2021, July 9-12). *Measuring executive functions and their relation to math intelligence in preschool children: A meta-analysis* [Paper presentation]. 12th ITC 2021 Colloquium, virtual conference. http://hdl.handle.net/10993/47652

Emslander, V., & Scherer, R. (2021, May 18-21). *Linking executive functions and math intelligence in preschool children: A meta-analysis* [Paper presentation]. Research Synthesis & Big Data Conference 2021, virtual conference. http://dx.doi.org/10.23668/psycharchives.4805

Emslander, V., & Scherer, R. (2020, September 9-10). *Measures of executive functions and mathematical skills are distinct even at a young age: A meta-analysis with preschool children* [Paper presentation]. Frontier Research in Educational Measurement Conference in Oslo, Norway. (Cancelled due to COVID-19 pandemic)

Emslander, V., & Scherer, R. (2020, July 14-17). *Measuring executive functions and their relations to mathematical skills in preschool children: A meta-analysis* [Paper presentation]. 12th Conference of the International Test Commission in Luxembourg, Luxembourg. (Postponed due to COVID-19 pandemic)

Acknowledgements

While most books on writing academic English tell me to write the majestic “we” and address everybody who helped me in the third person to show my gratitude, I would like to take the opportunity to directly thank you, the kind people who have supported my dissertation and express my appreciation now to you.

First, I would like to express my gratitude to my doctoral patchwork family—my excellent supervisors. Thank you, Professor Antoine Fischbach, for all the trust you have consistently placed in me by giving me lots of free reigns in every venture and showing me how to be a gracious team leader. Always putting the wellbeing of everybody in the team first, you taught me how to “choose your battles” wisely.

Thank you, Professor Ronny Scherer, for being the kindest mentor any aspiring researcher could wish for and for supporting me in all things concerning academia and beyond across country and continent borders since 2016. I am glad I could learn everything about integrity in research, statistics, and open science practices from you. Thank you for always being there with an open ear and for the helpful advice on the small and big questions academia throws you.

Thank you, Professor Doris Holzberger, for always asking the right questions in our meetings, such as “Why don’t you just start writing?” or “How can we best support you now?” and stating the occasional “Just take a break when you need one.” I deeply admire your positive outlook on and optimism with academia you bestow on your students.

Further, I would like to thank the other dissertation defense committee members. Thank you, Professor Christine Schiltz, for agreeing to chair the dissertation defense committee. Thank you, Professor Andreas Gegenfurtner, for not only making my dissertation defense committee complete but also for being an inspiration in the research field of LGBTQ+ issues in education, which I am looking forward to exploring.

The SIVA project and, thus, my dissertation would not have been possible without the financial and operational support from the *Observatoire National de l’Enfance, de la Jeunesse et de la Qualité Scolaire – Section Qualité Scolaire* (OEJQS). Specifically, I would like to thank you, Jean-Marie Wirtgen, Mirko Mainini, and Martine Frising, for supporting the project through the planning and even the hands-on work of the data collection.

Thank you to the Doctoral School of Humanities and Social Sciences for making parts of my research possible through financing projects and travels to Oslo, Norway, and Munich, Germany. Thus, I would like to thank the teams of the Centre for Educational Measurement at the University of Oslo (CEMO) and the Professorship for Research on Learning and Instruction at the Technical University of Munich for welcoming me and making my research stays with them so productive and fun.

As half of the present dissertation builds on the SIVA project, the direct follow-up project of another doctoral thesis, I would like to thank Dr. Jessica Levy. Thank you, Jessica, for being the most gracious daily supervisor through the first years of my PhD journey. You were always the ideal project colleague when we made crucial decisions and the perfect work partner when things needed to get done. Thank you for creating the basis of the SIVA project through your hard work and dedication during your doctoral and postdoctoral work and for making the complex data collection possible despite the sanitary crisis in the last few years.

The entire Luxembourg Centre for Educational Testing (LUCET) team is excellent at helping under challenging times. Thus, I would also like to thank you, Cassie Rosa and Yanica Reichel, specifically for your spontaneous help during the data collection and in all things SIVA. Also, I would like to thank everybody who was directly or indirectly involved in the project and everybody helping on the way: Dr. Sonja Ugen, Dr. Rachel Wollschläger, Uli Keller, Thierry Heck, Rachid Boualam, Carol Halpern, Andrea Klein, and the entire LUCET team. Thank you, Komolakhon Israilova, Sverre Berg Ofstad, Marie Prunkl-Haberkorn, and Tony Tan, for being the best research assistants possible and your dedication to rigorous and reproducible meta-analytic research. Thank you for your kind support.

I would like to thank my colleagues, who made my PhD journey at LUCET enjoyable. Thank you, Lena Kaufmann, for all the focused moments of working together in our office and even more so for all the fun, heartfelt moments in between. Thank you, Dr. Linda Romanowska, for being a friend through the stressful and the relaxing moments of the day. Thank you, Dr. Ineke Pit-ten Cate, for showing me how to be an effective and kind teacher and mentor. Thank you, Dr. Axel Grund, for being the perfect partner-in-crime with master-thesis supervision. From outside LUCET, I would like to thank you, Janina Täschner, from Munich, and Lena Hahn, from Trier, for being bright sparring partners on academic matters and great friends.

My special thanks go to my friends from my studies in Trier and Tübingen for always keeping an eye out and being on the phone when I needed them and to my family for letting me refuel with positive energy whenever required.

Finally, and most importantly, I would like to thank you, Christoph, for all the love, support, and advice during the last ten years and for being an incredible partner in all aspects of my life.

Funding

The present dissertation was supported by the Observatoire National de l'Enfance, de la Jeunesse et de la Qualité Scolaire – Section Qualité Scolaire (OEJQS)– Section Qualité Scolaire and a several travel and research grants from the Doctoral School of Humanities and Social Sciences (DSHSS) of the University of Luxembourg.

oejqs
OBSERVATOIRE NATIONAL
DE L'ENFANCE,
DE LA JEUNESSE
ET DE LA QUALITÉ SCOLAIRE

DSHSS | DOCTORAL SCHOOL
IN HUMANITIES AND
SOCIAL SCIENCES

Table of Contents

Index of Figures

Index of Tables

Index of Tables in the Appendix

Index of Abbreviations

DORA. *San Francisco Declaration on Research Assessment*

DSHSS. Doctoral School of Humanities and Social Sciences

EFs. *Executive functions*

ÉpStan. *Luxembourg's National School Monitoring Programme (Épreuves Standardisées)*

GTS. *Global Teaching Insights Study*

K-12. *From kindergarten to 12th grade*

LUCET. *Luxembourg Centre for Educational Testing*

OECD. *Organisation for Economic Co-operation and Development*

OEJQS. *Observatoire National de l'Enfance, de la Jeunesse, et de la Qualité Scolaire – Section Qualité Scolaire*

PISA. *Programme for International Student Assessment*

SIVA. *Systematic Identification of High Value-Added in Educational Contexts*

SOMAs. *second-order meta-analyses*

TALIS. *Teaching and Learning International Survey*

TBD. *Three Basic Dimensions of Teaching Quality*

TIMSS. *Trends in International Mathematics and Science Study*

TSRs. *Teacher-Student-Relationships*

VA. *Value-added*

Authorship Transparency Statement

The present thesis is largely based on the two published manuscripts and three unpublished SIVA project reports (Emslander et al., 2020; Emslander, Levy, & Fischbach, 2021; Emslander, Levy, et al., 2023). As the SIVA project is a direct continuation of the doctoral thesis by Levy (2021), which was in part written based on the first and second SIVA project reports, there is some overlap between the two theses, which is marked by citing Levy (2021) or the respective unpublished SIVA project reports.

Valentin Emslander
Belval, Luxembourg, 2 April 2024

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this thesis, the author used GPT-3.5 from OpenAI (2023), Grammarly (https://www.grammarly.com), and Writefull (https://www.writefull.com) for proofreading purposes and to ensure linguistic precision, good style, and readability. After using these tools, the author reviewed and edited the text as needed and takes full responsibility for the content and wording of the thesis.

Valentin Emslander
Belval, Luxembourg, 2 April 2024

Declaration of Authorship

I hereby declare that the present dissertation is my own work and that, to the best of my knowledge and belief, it contains no materials previously published or written by another person except where due acknowledgment has been made in the text.

Valentin Emslander
Belval, Luxembourg, 2 April 2024

1 Educational Diversity and Inequality in Luxembourg

The Grand Duchy of Luxembourg is experiencing an increase in societal and student diversity faster than other European countries (American Psychological Association, n.d.; STATEC, 2021). This diversity encompasses nationality, socioeconomic status, language, and other factors contributing to identity (American Psychological Association, n.d.). This renders Luxembourg a superdiverse society, characterized by large parts of the (student) population belonging to ethnic minorities with diverse language backgrounds (for a definition of superdiversity, see Rogers et al., 2013). The surge in diversity can be attributed to its small size, an economic system based on cross-border work and immigration, and Luxembourg's traditional multilingualism (e.g., Hartmann-Hirsch & Amétépé, 2023). First, the small size of Luxembourg has contributed to the rapid increase in societal and student diversity. With a population of only 650,774 people, Luxembourg is one of the smallest countries in the European Union (World Bank Group, 2022). Thus, compared to its number of inhabitants, Luxembourg has one of the highest immigration rates worldwide (World Bank Group, 2022). Second, Luxembourg's economic system attracts highly trained cross-border workers and immigrants (Peroni et al., 2016), who increase migration to Luxembourg and boost its economy (Fetzer, 2011). This rise in immigration is one of the causes of the high linguistic and socioeconomic diversity in the student population (see MENJE, 2023). Third, the multilingualism of the country and the educational system could incentivize migrants with a background in any of these languages to come to Luxembourg. In the trilingual school system, the language of instruction switches twice: Initially, children are taught in Luxembourgish during early childhood education and primary school. Then, the language of instruction switches to German during primary school, and students learn French as a second language. Finally, in the highest secondary school track, French becomes the language of instruction, starting in mathematics classes in Grade 7 and all other subjects in Grade 10 (Sattler, 2022). Then, English is taught as another foreign language in Grade 8, or 9 for the students who elected Latin. This multilingualism is generally seen as an asset for all students who successfully complete their school career with highly developed language skills. This openness to languages could attract migration from French- or German-speaking and other countries. However, it poses the problem that school children must learn several languages besides their home language and creates language barriers. This high demand leaves many school children behind early in their school careers (see Hadjar et al., 2018; Hoffmann et al., 2018; Sonnleitner et al., 2021). Thus, combined with a growing number of students new to the country, the three languages of instruction (i.e., Luxembourgish, German, & French) and at least one additional foreign language (e.g., English) present both opportunities and challenges for students, teachers, schools, and the educational system as a whole.

Consequently, language barriers are a challenge for the school system, as only about 32% of primary school students speak Luxembourgish, the first language of instruction, at home (MENJE & SCRIPT, 2022). Moreover, many students whose parents migrated to Luxembourg are likely to speak none of the three languages of instruction at home (STATEC, 2021). Not speaking the language of instruction can create inequalities between students. As such, students from non-Luxembourgish backgrounds may have more difficulties following the lesson in pre-school and primary education than their Luxembourgish-background peers. In fact, not speaking the language of instruction complicates accessing high-quality education, and students with other home languages than Luxembourgish or German tend to struggle in this trilingual school system (Hadjar et al., 2018; Sonnleitner et al., 2021). Moreover, the languages of instruction are not taught as a second language but are assumed to be known by all students instead. This preconceived notion further hampers students with diverse language backgrounds from reaching their full academic potential.

Another related challenge in the Luxembourgish education system is the socioeconomic inequality among the students. Combined with the language barriers, these differences can grow to provide vastly different learning experiences for students depending on their background. In such a situation, not students' ability and motivation might determine their success, but rather their home language, socioeconomic background, or migration status—the drivers of educational inequality in Luxembourg. Educational inequality can be defined as the "…unequal distribution of academic resources… [such as high-quality teaching, technology, or funding] due to economic or social status" (Law Insider, n.d.). A comparable scenario involving language barriers and socioeconomic disparities is conceivable in other countries with a multilingual educational systems, such as Lebanon, Singapore, or South Africa (Doidge, 2014; Klöter & Saarela, 2020; Trad, 2022). The educational systems in these countries are likely to share some of the educational challenges of Luxembourg, where multi-lingual instruction and language diversity in students can create language barriers and thus prevent groups of students to access high quality education.

The confluence of increasing language diversity and educational inequality might suggest a country's decline in its performance in international comparisons. Surprisingly, the empirical data for Luxembourg's education system challenges this assumption, as highlighted in the research by Weis et al. (2020) based on PISA data. Despite the complexities resulting in educational inequalities, Luxembourg performs somewhat stably in international comparison studies. This finding implies that Luxembourg's schools have implemented effective strategies to navigate the intricate terrain of linguistic diversity. In other words, there must be pedagogical approaches that address the educational inequalities and, thus, mitigate disparities in educational student outcomes. While promising to prove invaluable in creating equal opportunities and educational quality for all students, these approaches have not yet been investigated.

If we can delve deeper into the strategies that have contributed to addressing linguistic and socioeconomic inequalities, the insights could serve as a starting point for educational institutions, school leaders, and teachers to support all their students effectively, nationally and internationally. As such, other countries with multilingual educational systems could benefit from adopting similar practices to support all their students and alleviate educational inequalities in their specific settings. In the context of Luxembourg's educational system, the stability observed in international comparisons implies that certain schools must excel at assisting students against the odds. This support leads to one of the fundamental questions in educational effectiveness research heading the following section.

1.1 What are Educational Psychological Factors of School Success That Help All Students Thrive?

In other words, what is the effective educational psychological practices to address educational inequalities in a diverse, multilingual student population in a variety of potentially interesting variables, different schools, and varying stakeholders' perspectives? The narrative literature review and the four studies in the present thesis are crucial in finding these effective practices and correlates of school success. The Theoretical Considerations aim to define the general appearance of such factors of school success through the narrative evaluation of possibly relevant variables. Studies 1 and 2 focus on two such factors and describe them in detail with the help of international scientific literature. In Study 3, we test a method to map potentially interesting primary schools and determine where a search could be most fruitful. In Study 4, we use the findings from our earlier research to search for the well-defined success factors in the systematically structured selection of the primary schools in Luxembourg.

Concretely, the present thesis will address the question above on educational psychological factors in two research strands motivated by the narrative literature review of the present thesis. Research Strand 1 will meta-analytically address students' general social and cognitive learning processes. The database for the two meta-analytic studies was drawn from the international scientific literature and does not focus on Luxembourg specifically. Concretely, the first research strand aims to:

Aim 1. Delve into the social learning processes and investigate the role of teacher-student relationships (TSRs) for diverse student outcomes in the meta-analytic literature (see Study 1).

Aim 2. Dive deeper into essential cognitive skills of executive functions (EFs) and mathematics to get a fuller picture of the cognitive learning process in the literature on preschool children (see Study 2).

Research Strand 2 addresses concrete questions of educational effectiveness in the Luxembourgish school system with value-added (VA) methods. The two included VA studies focus on the Luxembourgish context and build on the project "Systematic Identification of High Value-Added in Educational Contexts" (*SIVA*). However, their results could be tested in other multilingual educational contexts around the world, with Luxembourg representing one of multiple possible settings for such a study. The four-year SIVA project, funded by the Observatoire National de l'Enfance, de la Jeunesse et de la Qualité Scolaire – Section Qualité Scolaire (OEJQS), aimed to identify successful educational psychological strategies to help all students thrive against the odds. Such strategies were sought in target schools with stable positive VA scores by comparing them to schools with medium or low VA scores on variables such as instructional quality, school climate, and Luxembourg-specific variables in a multi-perspective, mixed-methods data collection. The author coordinated the SIVA project of the present thesis with the kind help of his colleagues and the project team and included it as Research Strand 2 in the present thesis. The project is the direct continuation of Levy's doctoral thesis (2020) and was realized with financial and operational support from the OEJQS. Within Research Strand 2, we set out to achieve the following:

Aim 3. Test the stability of Luxembourgish primary schools' VA scores, given their students' socioeconomic and linguistic backgrounds (see Study 3).

Aim 4. Compare educational psychological variables between such schools with stable high, medium, or low VA scores in a multi-perspective, mixed-methods data collection (see Study 4).

As such, the two research strands are interrelated and follow the same goal with complementary methods and perspectives. The two meta-analytic studies focus on broader questions of students' social and cognitive learning processes and complement the SIVA project, which focuses on the Luxembourgish educational context with the two VA studies. These are the steps we take to find out what some schools in Luxemburg might be doing successfully to address linguistic and socioeconomic diversity and what other schools might learn from them. Ultimately, what we find and do not find could serve as a case study for educators, policymakers, and researchers worldwide looking for ways to address these educational inequalities.

In the following, the Theoretical Considerations section presents a narrative literature review providing an overview of school effectiveness models to address educational inequalities. This narrative literature review will lay the theoretical foundation for the included Studies 1 and 2 as well as the SIVA data collection in Study 4. This thesis' Methodological Considerations toolkit will be introduced and discussed before the Present Thesis section will outline the concrete aims of this cumulative dissertation. This section will join the theoretical with the methodological considerations to motivate the core of this thesis: the four included studies. Then, the four original studies will report their individual theoretical background, methods, results, discussion, and conclusion section. A General Discussion will join the four studies' key findings and expand on their theoretical, methodological, and practical contributions, limitations, and possible future research. Finally, this thesis culminates in a short General Conclusion section to highlight its findings and future directions.

2 Theoretical Considerations

2.1 Identifying Relevant Theoretical Models of School Effectiveness

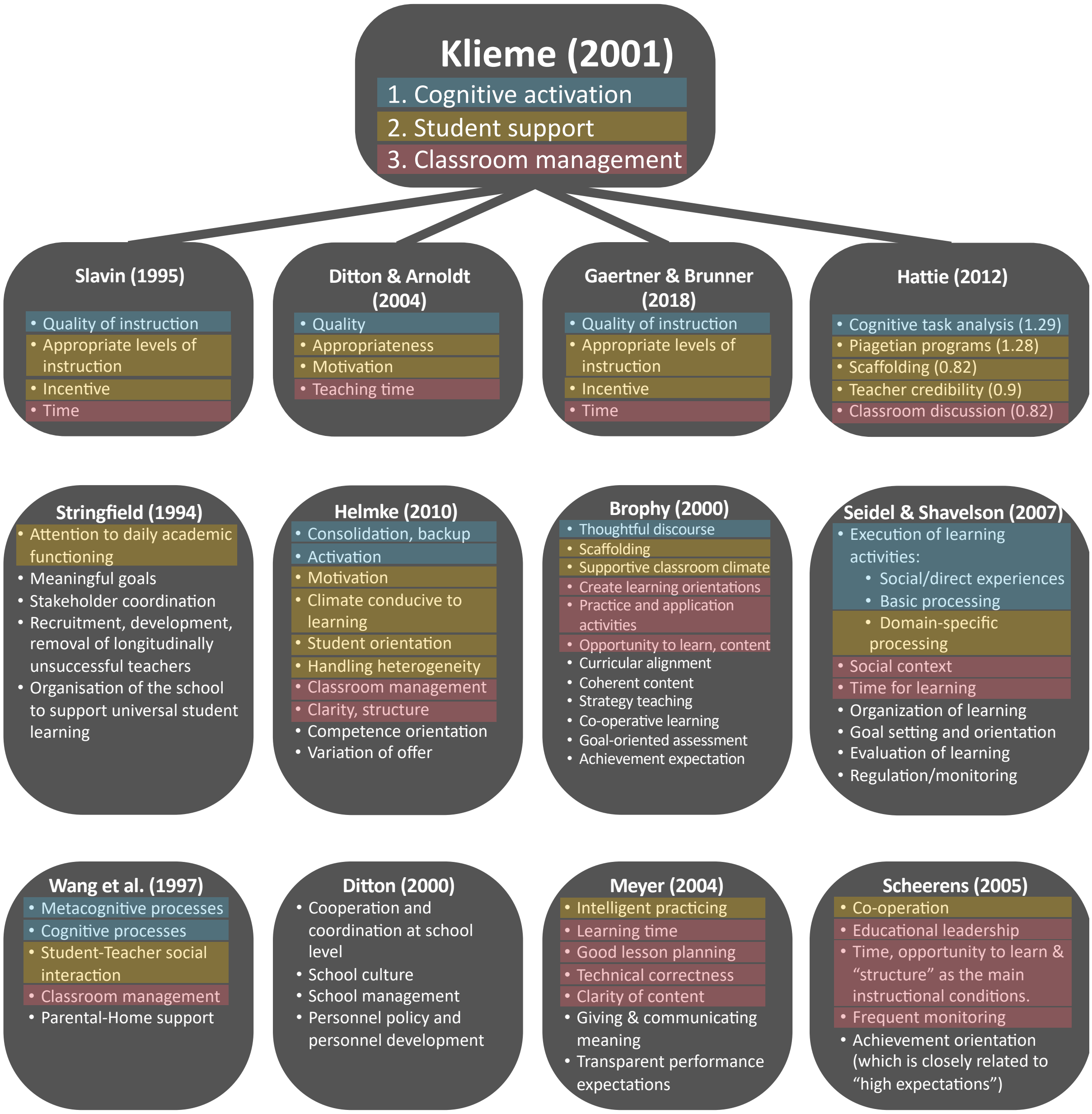
To motivate the research strands of the present thesis, we aimed to identify relevant pedagogical and psychological variables of school success. For Research Strand 1, we looked for generally relevant social and cognitive learning processes that we should investigate meta-analytically. For Research Strand 2, we aimed to identify variables of effective instruction for diverse student groups and examine their potential to help overcome educational inequalities within the SIVA project. Thus, we conducted a narrative literature review (for the first SIVA project report with the narrative literature review, see Emslander et al., 2020). Such narrative reviews contribute to educational theory building as requested by recent publications (see section Greene, 2022; Praetorius & Charalambous, 2023). The Methodological Considerations section further elaborates on this methodological approach. A summary of the narrative literature review (see Emslander et al., 2020) will be provided below. This summary will first mention the theoretical models included. Second, it will discuss the overlapping variables between important models of school effectiveness. Third, it will suggest how to combine the non-overlapping parts of the models in the data collection. Lastly, the summary will introduce

Luxembourg-specific variables that aid in capturing the particularities of the Luxembourgish educational context.

The extensive body of literature requires researchers to concentrate on prominent models of educational effectiveness. We categorized these models based on their key components, application areas, and whether they pertained to students, teachers, or schools. As a result, we identified 13 relevant models of educational effectiveness, as depicted in Figure 1.

As a staple in school success, Hattie's (2012) influential Visible Learning provides a list of meta-analytically studied correlates of student achievement (see for recent updates, Hattie, 2023; Waack, 2018). This list of learning correlates, in turn, led us to further investigate systematic research syntheses and more traditional effectiveness models in a broad sense. As further relevant meta-analytic research, we identified Seidel and Shavelson (2007) and Wang et al. (1997). As more traditional school effectiveness models, we reviewed models by Brophy (2000), Ditton (2000), Meyer (2004), Scheerens (2005), and Stringfield (1994). Additionally, we explored teaching quality models such as Klieme et al's *Three Basic Dimensions of Teaching Quality* (TBD; 2001), Ditton and Arnoldt (2004), Gaertner and Brunner (2018), Helmke (2010), Klieme et al (2001), and Slavin (1995). In a broader sense, these theoretical models of educational effectiveness constituted the basis of our initial literature review.

Figure 1. *Klieme et al's (2001) Three Basic Dimensions of Teaching Quality (TBD) model as the smallest common denominator of school effectiveness and student success models*


Klieme (2001)
1. Cognitive activation
2. Student support
3. Classroom management
Slavin (1995)
• Quality of instruction
• Appropriate levels of instruction
• Incentive
• Time
Ditton & Arnoldt (2004)
• Quality
• Appropriateness
• Motivation
• Teaching time
Gaertner & Brunner (2018)
• Quality of instruction
• Appropriate levels of instruction
• Incentive
• Time
Hattie (2012)
• Cognitive task analysis (1.29)
• Piagetian programs (1.28)
• Scaffolding (0.82)
• Teacher credibility (0.9)
• Classroom discussion (0.82)
Stringfield (1994)
• Attention to daily academic functioning
• Meaningful goals
• Stakeholder coordination
• Recruitment, development, removal of longitudinally unsuccessful teachers
• Organisation of the school to support universal student learning
Helmke (2010)
• Consolidation, backup
• Activation
• Motivation
• Climate conducive to learning
• Student orientation
• Handling heterogeneity
• Classroom management
• Clarity, structure
• Competence orientation
• Variation of offer
Brophy (2000)
• Thoughtful discourse
• Scaffolding
• Supportive classroom climate
• Create learning orientations
• Practice and application activities
• Opportunity to learn, content
• Curricular alignment
• Coherent content
• Strategy teaching
• Co-operative learning
• Goal-oriented assessment
• Achievement expectation
Seidel & Shavelson (2007)
• Execution of learning activities:
• Social/direct experiences
• Basic processing
• Domain-specific processing
• Social context
• Time for learning
• Organization of learning
• Goal setting and orientation
• Evaluation of learning
• Regulation/monitoring
Wang et al. (1997)
• Metacognitive processes
• Cognitive processes
• Student-Teacher social interaction
• Classroom management
• Parental-Home support
Ditton (2000)
• Cooperation and coordination at school level
• School culture
• School management
• Personnel policy and personnel development
Meyer (2004)
• Intelligent practicing
• Learning time
• Good lesson planning
• Technical correctness
• Clarity of content
• Giving & communicating meaning
• Transparent performance expectations
Scheerens (2005)
• Co-operation
• Educational leadership
• Time, opportunity to learn & "structure" as the main instructional conditions.
• Frequent monitoring
• Achievement orientation (which is closely related to "high expectations")

Note. Figure 1 is adapted from Emslander et al. (2020). Colors indicate parts of educational effectiveness models that overlap with Klieme et al's TBD (2001). The models of Ditton & Arnold (2004), Ditton (2000), and Helmke (2010) were translated from German.

2.2 Overlap of School Effectiveness Models as Promising Variables

From this array of theoretical models on educational effectiveness, our primary aim was to distill a robust theoretical foundation tailored to the objectives of SIVA. Thus, we searched for similarities among these theoretical models to find the smallest common denominator between the models. More specifically, we followed the rationale that constructs present in multiple of these theoretical models should be central in educational psychological research. Thus, we examined in which aspects the theoretical models overlapped. These common constructs should then be central to educational effectiveness and be the core constructs to focus on in the data collection of the SIVA project.

Upon closer examination, we discerned that most of these models shared several fundamental aspects, which could be categorized into three key constructs: (1) cognitive activation, (2) student support, and (3) classroom management, aligning with Klieme et al's TBD (2001). This observation aligns with prior research, identifying this conceptualization of instructional quality to be similar in several theoretical models (for an in-depth theoretical analysis, see Praetorius & Charalambous, 2018). Consequently, these three constructs can be considered the common denominator for most of the models we encountered and, thus, form the theoretical core of the SIVA data collection in Study 4 and the further exploration of constructs.

Klieme et al (2001) developed the model based on the German Trends in International Mathematics and Science Study (TIMSS) video study 1995 focusing on mathematics instruction in Grade 8. Since then, the model has inspired much theorizing and research on instructional quality (see Niepel, 2023). It has since advanced to be one of the most pertinent models of instructional quality (Herbert et al., 2022) with its three dimensions. The TBD have distinct features. First, the dimension of cognitive activation describes a teaching strategy to engage students cognitively, connect to their prior knowledge, and give them challenging tasks. Second, student support consists of practices for fostering an inclusive learning environment, providing guidance, and caring for the student’s competence, autonomy, and social relatedness needs. Third, classroom management refers to organizing a disturbance-free learning environment where students have as much learning time as possible (Klieme et al., 2006; Praetorius et al., 2018). The three dimensions are theoretically linked to conceptual understanding and motivation in students via processes of higher-level thinking, time spent on learning tasks, and self-determination (Alp Christ et al., 2022; Klieme et al., 2006). As such, the TBD are conceptually corroborated (see Niepel, 2023) and present the smallest common denominator of several models of educational effectiveness.

It is no coincidence, however, that several of these models have such a considerable overlap. As Gaertner and Brunner (2018) discussed, Slavin's (1995) model specifies four rather than three dimensions of teaching quality. However, these four dimensions (quality of instruction, appropriateness, incentives, and time) can be easily mapped to Klieme et al's (2001) three dimensions (see Figure 1). Ditton and Arnoldt (2004) build on Slavin's (1995) model to generate a German survey for students to give feedback on instructional quality. This survey was, in turn, the basis for Gaertner's and Brunner's (2018) research. As such, the above-described models are linked by their similar dimensions and built upon one another, making overlaps somewhat more likely. Nonetheless, the sheer amount of research, theory, and teaching guidelines along the TBD model lends this model raison d'être and warrants its inclusion in the SIVA data collection in Study 4.

Independent of the TBD model, Pianta et al. (2008) developed the *Classroom Assessment Scoring System* (not included in Figure 1 because it was not part of the original narrative review). Their system is widely used in the US for short classroom observations focused on positive teacher-student interactions in, again, three key domains: (1) instructional support, (2) emotional support, and (3) classroom organization. Therefore, the two educational effectiveness models focus on the same three constructs despite their independent origins (see Decristan & Dumont, 2023). This parallel development further corroborates the relevance of (1) cognitive activation, (2) student support, and (3) classroom management as the core variables the SIVA project should assess.

There is another set of reasons for the broad application of the TBD model in the European context (Herbert et al., 2022; Niepel, 2023). First, the model with its three dimensions is grounded in theory, with much research dedicated to its conceptual anchorage. Second, the model and its extensions specify assumptions for the relation between its three dimensions and student variables, such as the understanding and motivation of students (Alp Christ et al., 2022; Klieme et al., 2006). Third, the model is very parsimonious, making its assessment feasible and resource-conserving, which is especially helpful when assessing younger students. Fourth, the model has been tested on tens of thousands of students worldwide through its inclusion in several international large-scale studies, such as the Global Teaching Insights Study (GTS; Organisation for Economic Co-operation and Development [OECD], 2020), TIMSS, and PISA (see Niepel, 2023). Further, Scherer et al. (2016) found the TBD student questionnaires to show measurement invariance across multiple countries. This invariance suggests that the questionnaires for instructional quality are psychometrically comparable between different countries and testing languages. Therefore, the TBD build a solid practical and theoretical foundation to study educational effectiveness in Luxembourg's diverse student population.

Previous work by Praetorius and Charalambous (2018) gave a well-structured overview of assessing instructional quality in classroom observations. Their work guided us in creating the observation sheets for the SIVA study. Additionally, such studies further corroborate the importance of the TBD as essential aspects of educational effectiveness with a wide field of application and a significant amount of literature. Therefore, we constructed the data collection of Study 4 around the central TBD—cognitive activation, student support, and classroom management—adding additional constructs. These additions were either addressed in multiple of the educational effectiveness models (e.g., school climate) or relevant to capture Luxembourgish educational particularities (e.g., multilingualism in instruction).

2.3 School Success Beyond Instructional Quality

To broaden the perspective on school success beyond Klieme et al's TBD (2001), we examined Hattie's *Visible Learning* factors (2012). We considered including those constructs that exhibit the strongest positive and negative correlations with school achievement for inclusion in the SIVA data collection. Collective teacher efficacy was most strongly positively related to student achievement with Cohen's d = 1,57. Besides attention deficit hyperactivity disorder and deafness, student boredom showed the strongest negative link with student achievement, with d = -0.49. Thus, we added these two variables to the data collection of Study 4. Additionally, we explored other school effectiveness models to include more variables that instructional quality does not cover. As such, we decided to include school organization, as suggested by Ditton (2000), evaluation (e.g., Scheerens, 2005), and goals and expectations (e.g., Seidel & Shavelson, 2007). Additionally, we included parental and home support (Wang et al., 1997) as the other variables did not cover it.

Subsequently, we examined the residual parts of the models that were not overlapping. We found several of these non-overlapping variables were aspects of school climate, which were missing from the instructional quality approach we found to be the smallest common denominator. As for instructional quality, school climate has been conceptualized in a variety of well-established models (for an integrative review, see Thapa et al., 2013), and its connection with instructional quality or student achievement and motivation is well-researched (Scherer & Nilsen, 2016). The creation of school climate agencies further shows the relevance of school climate as a crucial factor for school success and student well-being. For example, the National School Climate Center at Ramapo for Children (USA) or the inclusion of school in national educational models, such as the *CARAT A School Climate Model for Luxembourg's Schools* [CARAT Ein Schulklima-Modell für Luxemburger Schulen] (SCRIPT, 2018) show it's centrality in education of diverse student populations.

Thus, we included four aspects of school climate theorized by Wang and Degol (2016) in the SIVA project. These four include safety, community, academic, and institutional aspects of school climate with three to four sub-constructs. We aimed to assess at least one subconstruct for each aspect of school climate in the data collection of Study 4. More specifically, we included measures for discipline and order for the safety aspect, measures for teacher-student and other relationships within the school setting for the community aspect, measures for leadership at the schools and professional development for the academic aspect, and measures to assess the structural organization and environmental issues such as heating or building for the institutional environment aspect. As such, we identified school climate as another aspect that spanned several influential models of educational effectiveness next to the TBD in the SIVA data collection.

2.3.1 TSRs as the Overlap of TBD and School Climate

Some definitions of school climate included variables from the school effectiveness models that Klieme et al's three basic dimensions (2001) did not cover. However, the two models also overlap in a few other crucial aspects of educational effectiveness. One of these overlaps is fostering positive TSRs.

Positive TSRs can be defined as warm, supportive, and friendly interactions between teachers and students, whereas negative TSRs are characterized by conflict and dependency (Hamre & Pianta, 2001; Pianta, 1999). Models of TBD and school climate include both positive and negative TSRs, and several meta-analyses have summarized findings on their associations with student outcomes. These meta-analyses have found correlations between TSRs and various student success indicators such as school achievement, general wellbeing, school engagement, or executive functions (Chu et al., 2010; Roorda et al., 2017; Vandenbroucke et al., 2018). Therefore, TSRs can be seen as a highly relevant variable for student success. Still, this social learning process needs further research because the international meta-analytic literature is partially contradictory and varies considerably in its definitions of TSRs, outcome variables, and findings. A comprehensive overview of TSRs and their correlates in the international meta-analytic literature is needed. Delving into questions surrounding TSRs is included in both research strands of the present thesis. We examined TSRs in-depth in Study 1 of the present thesis and added it to the SIVA data collection in Study 4. To adequately investigate the Luxembourgish context, however, the data collection should include particularities of Luxembourg’s educational system and examine their potential to address educational inequalities effectively.

2.3.2 Executive Functions as a Prerequisite for Student Success

To validate our choice of theoretical basis and included constructs, we conducted an exploratory search through the three influential (educational) psychology journals, *Psychological Bulletin*, *Psychological Science*, and *Journal of Educational Psychology.* In their five most recent volumes, we uncovered a few variables we should assess, too. One of the topics investigated most frequently in these journals was digitalization in school and beyond. Thus, we added questions concerning the teachers' approach to digital media to the current set of variables. Additionally, we found a large body of work investigating students' cognitive functions and how they contribute to student achievement and, consequently, school success.

As such, we identified executive functions as prominently researched cognitive processes that regulate human cognition and behavior (EFs; Miyake et al., 2000; Miyake & Friedman, 2012). These EFs often encompass three subdimensions: response inhibition, mental set shifting, and updating of working memory. They are prerequisites for many school-related skills, such as reading or fluid intelligence (Cassidy et al., 2016; Diamond, 2013; Follmer, 2018). Arguably most prominently, however, EFs are linked to math skills (e.g., Best et al., 2011; Cragg et al., 2017; Friso-van den Bos et al., 2013; Peng et al., 2016; Yeniad et al., 2013) and their development (van der Ven, 2011; van der Ven et al., 2012). Their link to school achievement can be explained in several different ways. First, EFs are a prerequisite for students to comprehend, manipulate, retain, and reproduce information in the school context. For instance, inhibition aids students in maintaining focus while disregarding tempting alternatives to the learning task at hand; shifting enables students to transition between tasks more effortlessly, and updating allows students to mentally manipulate information. All these cognitive functions are essential for successfully navigating a lesson and solving learning tasks, and they correlate with improved TSR (Vandenbroucke et al., 2018). Second, it is crucial for a teacher to understand the stage of their students' EF development. Only then can they provide age-appropriate learning tasks that align with the students' EFs. The cognitive capacity of students can serve as a limiting factor in the types of learning activities proposed by the teacher. By finding a difficulty level that is challenging yet attainable, the teacher can cognitively engage their students and create a stimulating learning environment (see Praetorius et al., 2018). Third, a teacher who accurately estimate a student's EFs is likely to provide more effective support. This support, in turn, can contribute to building a positive TSR. The links between TSR and other school-relevant variables will be examined in Study 1.

Thus, gaining an overview of such an important student-factor of learning as EFs is essential, as they contribute to the development of mathematical skills in preschoolers and, therefore, student success later in the school career (see Study 4). Because the SIVA project aimed to collect data on malleable variables, we only included aspects of digitalization but not students' cognitive functions in the SIVA data collection. Still, executive functions will be one of the correlates of TSR in Study 1 and the focus of our meta-analysis in Study 2.

2.4 Addressing Particularities of the Luxembourgish Educational System

In addition to the constructs internationally relevant for school success, several variables are specifically important in Luxembourg's educational system and, thus, for Study 4. In a workshop with the school quality experts from the *Observatoire National de l'Enfance, de la Jeunesse, et de la Qualité Scolaire,* we identified four central aspects in which Luxembourg differs from other European school systems. These four aspects were the changing language of instruction, the cooperation between primary schools and afternoon care, the role of the school president, and the motto of schools.

Thus, as a first block of Luxembourg-specific variables, we were interested in the languages used during the classroom observations. More specifically, we aimed to observe how much German the teacher spoke during the instruction, individual explanations, and during breaks. We focused on German because it is the official language of instruction and alphabetization in primary school. Supplementarily, we added a variable to assess what other languages were spoken. In a second block of additional variables, we wanted to examine the communication and cooperation between the school and the *maison relais*, the afternoon care, which usually cooperates with primary schools in Luxembourg. As a third block of additional variables, we aimed to inquire about the teachers' and school presidents' views on the role of the school president, the headmaster of a Luxembourgish primary school. As their role description leaves a wide range for interpretation, we added several variables to the data collection of Study 4, testing whether teachers and school presidents saw the latter's role more in administrative, pedagogical, or other aspects of school development. Lastly, we added a variable concerned with the school's vision and whether the school had a motto and a code of conduct. Besides a few more singular items, we added these four additional constructs to Study 4's data collection.

Once we finalized the variable selection for the Study 4, we compiled measurement tools from the international literature on operationalizing the respective constructs. When such assessments were not practical for use in Luxembourg, we adapted the language or other features of the measurement tool to fit Luxembourg's educational context and language requirements. Mainly for the Luxembourg-specific constructs, we create original measurement tools. As with all these measurement tools, the experts from the *Observatoire National de l'Enfance, de la Jeunesse, et de la Qualité Scolaire* supported us in considering Luxembourg's specific context, particularly regarding language. More information on the definition of TSR and EFs can be found in the preregistrations of Study 1 (accessible at https://doi.org/10.17605/OSF.IO/J2EMF) and Study 2 (accessible at https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42017076291). All constructs included in Study 4 with the respective measurement tool can be found online in our preregistration of the SIVA project (accessible at https://doi.org/10.17605/OSF.IO/X3C48).

3 Methodological Considerations

To identify potential school success factors, we made six methodological decisions. First, we have decided to focus on preschool and primary education students because the early years are crucial for later success in school (Heckman, 2008). Second, we have decided to use research synthesis methods to identify variables essential for student success and school effectiveness internationally as potential drivers of school differences in Luxembourg. Third, we have decided to apply VA methods for informative purposes, as suggested by prior research (Ferrão, 2012; Leckie & Goldstein, 2019; Levy, 2020), to identify effective target schools. Fourth, we have chosen a multi-perspective, mixed-methods design for the SIVA data collection to gain the broadest possible overview of potentially successful academic strategies in Study 4. Fifth, we have decided to follow current open science methods to preregister our studies and make our research material, collected data, and analytic codes publicly available wherever feasible. Additionally, good research practices related to the open-science idea apply to meta-analytic and VA research, respectively. These five methodological decisions are explained in greater detail below as the methodological building blocks of the present thesis.

3.1 Focus on Pre- and Primary Education for the Best Return on Investment

Whereas it was already agreed upon before the start of the SIVA project that the focus should be on younger school children, there are several reasons for this choice. First, the target group of students should be as young as possible due to economic reasons. This economic standpoint is further explained by the Heckman curve (Heckman, 2008), which shows the earlier the investment in education begins, the higher the returns on investment. Applied to a data collection in the educational field, this means that the earlier in the school career we collect data and find out what educational psychologic strategies are successful, the greater the effect might be that these strategies may have when applied successfully (see Emslander et al., 2020). Second, there is evidence that students' socioeconomic status or home language, for example, might drive educational inequalities already in the youngest students. Findings from the National School Monitoring Programme (ÉpStan ; *Épreuves Standardisées*) indicate that educational inequalities already manifest in the first years of school. More specifically, the impact of students' backgrounds on their achievement was already detected around the second learning cycle, on which the SIVA project focuses (Hoffmann et al., 2018).

Consequently, students with low socio-economic backgrounds and students who do not speak Luxembourgish at home developed less positively from Grade 1 to Grade 3 in terms of school performance when compared to their peers. These performance differences further emphasize the need to focus on the early school years to address educational inequalities in Luxembourg's diverse and multilingual student population. Third, specifically in Study 3 and Study 4, we aim to avoid overburdening the students and teachers by imposing an extensive data collection on them in the same year as the national large-scale testing. Thus, we aimed to collect data from the youngest students at an even grade level because the even grades are not taking part in the ÉpStan and would not be overburdened with standardized tests (Luxembourg Centre for Educational [Testing], LUCET, 2023). As a result, we have decided to collect data in Grade 2, referred to as learning cycle 2.2 in the Luxembourgish school system.

At the same time, there are also challenges in focusing on younger students. For example, younger students who cannot read or write sufficiently to answer questionnaires independently require more resources in an assessment. Therefore, commonly used measures based on reading and writing (e.g., paper-and-pencil questionnaires) must be rephrased in easy language and read aloud to the students. As a result, the assessment takes more time, and fewer questions can be asked. Additionally, the examiner will allot more time to test the children, posing a monetary burden. Finally, when asked about their school and teachers, younger students give rather optimistic answers (Fauth et al., 2014). At the same time, students can already quite validly provide an impression of instructional quality in Grade 3 (Fauth et al., 2014). Thus, instructional quality can be assessed with primary school students' ratings.

Although it can be more resource-intensive and bias-prone to assess younger students, targeting this population promises a higher return on investment, as they already show significant disparities. In order to not overburden our sample with educational testing, we have decided to collect data on students in Grade 2 in the learning cycle 2.2 for the SIVA project.

3.2 Research-Synthesis Methods to Identify Important Variables

One of the main questions of the present thesis is which variables are likely to drive the difference between highly effective schools and others. To select variables in a theory- and evidence-driven fashion, we use research syntheses to make an evidence-informed decision on which constructs to collect data. Generally, "Research synthesis is the practice of systematically distilling and integrating data from many studies to draw more reliable conclusions about a given research issue" (Cooper et al., 2019, the abstract). Such research syntheses are a helpful tool for practitioners or policymakers for evidence-based decision-making. For researchers, research syntheses provide an overview of a theory, a research question, or a research field (Paré & Kitsiou, 2017). More specific to the research goals, research syntheses can help identify variables closely related to school effectiveness or student success.

Other advantages of research synthesis are their generalizability and high statistical power. Research syntheses can include diverse samples, settings, and methodologies and, thus, have the potential to make claims on generalizability than primary studies (Matt & Cook, 2019). In meta-analyses, when moderator analyses confirm that results are statistically equal for different groups, these results can be generalized across the groups. This phenomenon is somewhat linked to the increased statistical power of meta-analyses. By combining multiple samples from primary studies, meta-analyses can achieve larger sample sizes than most primary studies and, thus, increase their ability to detect true effects. This increased statistical power strengthens the evidential value of research syntheses (Quintana, 2023) and quantifies their potential contribution.

The different types of research syntheses vary in their approach to searching the literature, analyzing data, or reporting methods and results (Paré et al., 2015). Using a wide array of research synthesis methods, we explore and deepen our understanding of social and cognitive learning processes in Research Strand 1. Their findings inform our variable choice of the SIVA project in Research Strand 2.

We conducted a narrative literature review to gain insight into school effectiveness theories in the Theoretical Considerations section (see above). For the two studies in Research Strand 1, however, we focused on the systematic review and statistical analysis of both primary studies and meta-analyses. As a methodological primer, (1) the benefits and shortcomings of the narrative literature reviews will be discussed below. Then, (2) the systematic review and meta-analysis methodology will be reviewed before elaborating on the methodology of (3) the systematic review of meta-analyses and the SOMA. The latter two methods combine a systematic literature search and review with a statistical analysis of effect sizes to quantify the strength of given evidence.

3.2.1 Narrative Literature Reviews Offer a Broad Theoretical Overview

To give a general overview of relevant school success theories, we conducted a narrative review of the literature as presented earlier (see the *Theoretical Considerations*). This type of research synthesis provides a qualitative overview of a theoretical question or topic. Narrative literature reviews usually do not report a systematic literature search. Consequently, we omitted a systematic search of several databases and a list of inclusion and exclusion (Rother, 2007). Even though the more systematic methodology may be missing, there are helpful guidelines on best practices in qualitative narrative reviews (e.g., Baumeister & Leary, 1997; Ferrari, 2015; for systematic-narrative hybrid research synthesis, see Turnbull et al., 2023) and the method offers several advantages.

Some of the benefits of narrative literature reviews are their potential for inclusivity and flexibility as well as their exploratory character. As such, a narrative literature review can include multiple forms of qualitative and quantitative literature to build a broad understanding of a topic (Rother, 2007). This inclusivity helps to provide context and background information to grasp a broad subject or theory. In contrast to systematic reviews, which follow a structured and predefined methodology, narrative literature reviews offer flexibility in the selection, inclusion, and interpretation of findings (Turnbull et al., 2023). As such, authors can explore common themes, patterns, or overlaps between theories and conceptual frameworks. Diverse perspectives and methodologies can be included. The lack of methodological and statistical jargon (Ferrari, 2015) often renders the narrative literature review more accessible to a broader audience, including policymakers or teachers.

While narrative reviews have these advantages, it's important to note that they may be more susceptible to bias in study selection and interpretation. Because this more qualitative approach to synthesizing the literature does not offer the necessary information to be fully reproduced, the literature selection might be biased (Rother, 2007). With no systematic literature search or screening procedure, narrative literature reviews are inherently subjective. Thus, the selection of studies and the interpretation of findings depend on the authors’ judgment (see Ferrari, 2015). This subjectivity introduces the potential for bias, as authors may unconsciously favor studies that align with their perspectives or hypotheses.

Still, a narrative review is the most common form of literature review as it constitutes the theoretical background of most academic articles. The added value of such narrative reviews is to give a broad theoretical overview of the current status of a theoretical topic and provide the reader with the relevant literature (Turnbull et al., 2023). We also conducted a narrative review to identify variables that function as a common denominator between central school effectiveness models. This narrative literature review motivated the present two research strands and corroborated the SIVA project's theoretical underpinning. As such, the narrative review of school effectiveness theories is one of the cornerstones of the present thesis.

3.2.2 Meta-Analyses Give an Overview of Narrow Empirical Research Questions

The first strand of research in the present thesis is examining social and cognitive learning processes by applying meta-analytic methods. A meta-analysis is a review work that systematically integrates the results of multiple primary studies on a defined research question by statistically aggregating their effect sizes (Borenstein et al., 2009; Cooper et al., 2019). Hence, this method empirically summarizes the research landscape on one research question. Consequently, meta-analyses can make generalizable statements across various individual studies by following a systematic sequence of steps. However, meta-analyses are not without their shortcomings. Namely, all research-synthesis methods have a high level of abstraction, risk comparing “apples with oranges,” and are challenging to conduct and report (Ioannidis, 2016; Thompson & Pocock, 1991). To combat these potential challenges of meta-analyses discussed in the literature, several streamlined checklists and guidelines help meta-analysts follow through with this intricate multi-step process (e.g., Kepes et al., 2013; Page et al., 2021; Siddaway et al., 2019).

The step-by-step process of meta-analysis typically starts with (0) familiarizing oneself with the research field before (1) formulating a research question. In this process, (2) the definition of easy-to-test research questions simplifies (3) the creation of an exact search term for the subsequent (4) systematic literature search. If the scientific articles that match the search term also align with the previously (5) defined inclusion criteria in a process called “screening,” (6) the full texts of the articles can be obtained. After that, (7) trained raters use a coding guide to fill out a coding sheet with effect sizes and any potential moderators or other characteristics from the individual studies. The (8) statistical aggregation of effect sizes into an overall effect follows. Studies with larger samples are weighted more in this aggregation process because they are usually more precise. When the included studies differ in their samples, designs, and measurement of the construct, for example, (9) moderator analyses can provide more detailed results investigating these study differences. (10) Further analyses concerning the quality of the included study, the meta-analytic robustness against outlier values, or publication bias can be conducted. For the statistical analyses above, suitable software should be used (e.g., R and the R-package metafor; R Core Team, 2019; Viechtbauer, 2010). Together with (11) the results, meta-analysts must (12) accurately report and explain the process of literature search, the exclusion of studies and effect sizes, and the statistical analyses—in short, the entire procedure (for an overview, see Cooper, 2017; Field & Gillett, 2010). When all these steps have been performed according to current quality standards and made transparent (Johnson, 2021; Page et al., 2021), meta-analyses provide a valuable high-level resource to make evidence-driven decisions on somewhat narrowly defined research questions in a specific context. Study 2 applies this method. To broaden the overview and include multiple constructs or multiple correlates of a construct in various contexts, however, a systematic review of meta-analyses and a SOMA should be undertaken.

3.2.3 Systematic Reviews of Meta-Analyses and SOMAs Provide a Broad Overview of Empirical Evidence on a Topic

Systematically reviewing systematic reviews and meta-analyzing meta-analyses are a somewhat newer phenomenon that only emerged in the 1990s (Cooper & Koenka, 2012). As in a first-order systematic review and meta-analysis, the multi-step process systematically searches the literature and statistically synthesizes the included studies. However, only meta-analyses are eligible when conducting a review of meta-analyses and a SOMA, and thus, meta-analytic effect sizes are aggregated (Schmidt & Oh, 2013). This quantitative synthesis broadens the scope of the research considerably. Going beyond aggregating findings on a specific research question, we can provide an overview of a defined research field with this combination of techniques.

Existing reviews of meta-analyses and SOMAs have already mapped considerably large parts of the educational psychology landscape (e.g., Hattie, 2008; Jansen et al., 2022; Schneider & Preckel, 2017). We can use this prior work to look for important correlates of student success we should assess in the SIVA study. Considering the review of meta-analyses conducted by Hattie (2008), for example, we can identify boredom as a decisive detrimental factor to student achievement. Conversely, collective teacher efficacy, for example, is a strong supporting factor of student success (Hattie, 2008). Using SOMAs, these findings become comparable, and we gain a robust evidence base for deciding which educational psychological constructs might make the difference between highly effective and other schools.

While the present thesis reports a review of meta-analyses and SOMAs to identify what might drive the success of some schools, there are other benefits to applying these methods in educational psychology. Such benefits include (1) summarizing the statistical evidence on outcomes correlating with a specific variable, (2) identifying important moderators within a specific research field, (3) mapping the quality of existing meta-analyses there, and (4) indicating future research that will close existing knowledge gaps in the said field (Cooper & Koenka, 2012). This second-order review and meta-analysis is the method we used in Study 1. In conclusion, systematic reviews and meta-analyses, as well as systematic reviews of meta-analyses and SOMAs, are potent sources to map findings on a research question or in a research field. They can aid in identifying potential drivers of differences between highly effective schools and others.

3.3 Value-Added Methods to Identify Highly Effective Schools

One of the most fundamental questions in educational effectiveness research is probably how we can identify which schools are effective in helping students achieve, independently of their backgrounds. Arguably, the most used quantitative methods are the so-called VA models (Chetty et al., 2014; Kane et al., 2013). Designed initially to provide fairer assessments of teacher and school effectiveness, VA models aim to evaluate educators and institutions based on their students' learning gains, representing a net impact (Driessen et al., 2016) rather than solely on their raw achievement data. Statistically speaking, VA analysis is a regression method that can account for diverse student backgrounds, prior achievement, and other preconditions when estimating school or teacher effectiveness (Emslander et al., 2020; Levy, 2020). The resulting VA scores consider several student factors the school cannot influence, such as the language the student speaks at home. Therefore, VA scores could be of great use in the Luxembourgish educational context because they promise to compare schools more fairly by taking their diverse and multilingual student bodies into account.

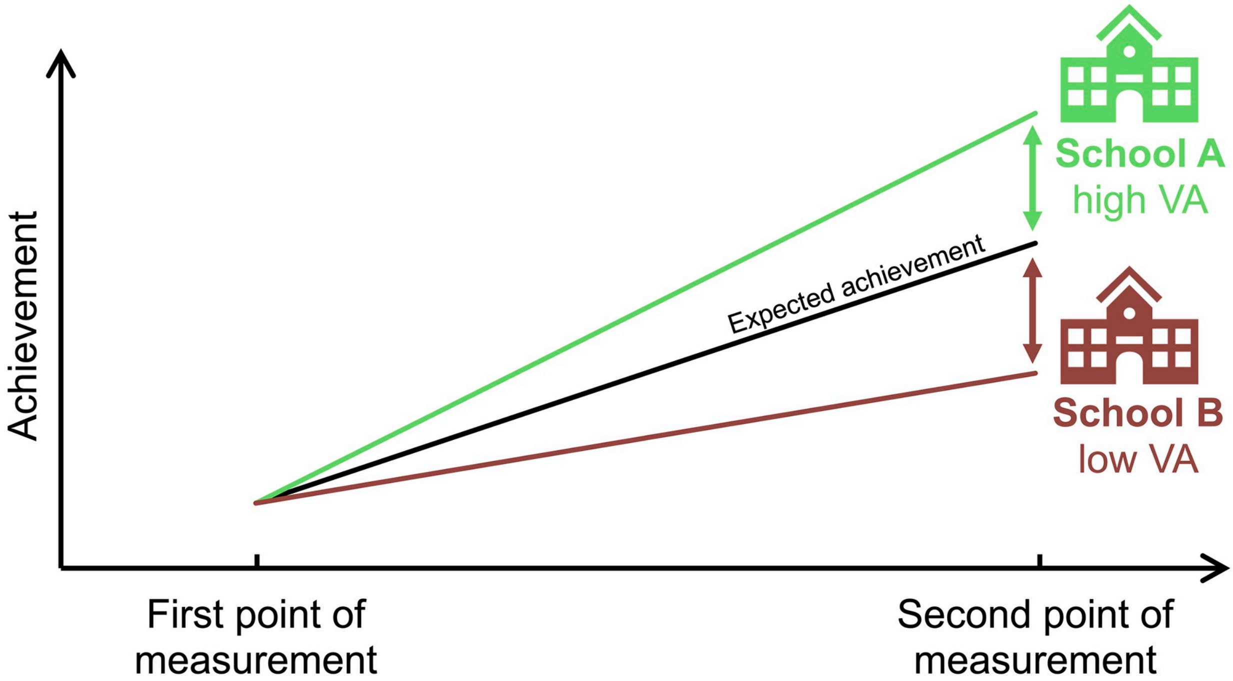
VA scores quantify the difference between the students‘ expected academic and actual academic achievement, given their backgrounds (Sanders et al., 1997). Positive VA scores indicate that students have exceeded the expected achievement based on their background characteristics, such as socioeconomic status, language spoken at home, or previous academic performance. Conversely, negative scores suggest that students have fallen short of the expected achievements of students with similar background characteristics.

To facilitate equitable school comparisons, these student VA scores can be averaged for each school (or teacher), reflecting the value a school adds to its students' progress, independently of their backgrounds (Braun, 2005; Tymms, 1999). Figure 2 compares a school with a high VA score and one with a low score. In this example, students from both schools shared similar starting backgrounds, such as previous academic achievement, and were therefore statistically expected to perform similarly. However, students from School A surpassed the statistically expected performance for a comparable group, indicating that School

A contributed *added value* to its students' academic achievements. Consequently, School A would receive a high VA score, while School B would receive a low VA score as students performed below the statistical expectation (see Emslander et al., 2020; Levy, 2020).

The SIVA project seeks to employ VA models for informative purposes rather than for evaluative purposes, aligning with recommendations from national and international researchers (Ferrão, 2012; Floden, 2012; Johnson, 2015; Leckie & Goldstein, 2019; Levy et al., 2019; Levy, 2020). Consequently, this approach can be used to systematically identify target schools that are able to *add value* to their students' achievements, from which we can then gain insights into the elements contributing to educational effectiveness. More specifically, when we aggregate all students' scores per school and then investigate what these schools with high VA scores are doing differently than the others, we might be able to learn from them how to support all schools and how to tackle educational inequalities. Therefore, employing VA models enhances our chances of uncovering the impacts of effective teaching, significant background factors, and successful school management in Luxembourg. Ferrão (2012) even noted that VA models are particularly well-suited as a tool for "progressive improvement in education" (p. 627), especially in countries with high rates of grade repetition. This detail further reinforces the rationale for using VA models for informative purposes in the Luxembourgish school context, as already in Grade 3, 17% of students are not on track (Ministry of National Education, Children and Youth, 2018), largely due to grade repetition (see Emslander et al., 2020; Levy, 2020).

Figure 2. *Illustration of high and low VA schools performing above and below what is expected of them from one measurement point to another*



Note. Figure 2 is adapted from Levy (2020). Double-headed arrows signify the VA score as the difference between a school's expected and actual achievement.

Given its promise to quantify educational effectiveness with just a straightforward score (Chetty et al., 2014), VA scores have been used for high-stakes educational decision-making in the US since the late 90s (Aslantas, 2020; Everson, 2017). The VA scores of teachers or schools serve as a basis for recognizing highly effective teachers, providing tenure, or allocating additional school funding. Conversely, teachers with low VA scores may risk losing their jobs, and schools with low VA scores lose parts of their funding (Goldhaber & Hansen, 2010; Sass, 2008). This allocation of resources hinges on the belief that changes in a student's VA score from one year to the next can be attributed exclusively to the influence of their teachers or school.

The stability of VA scores, the reason for a change in these scores, and their relation to student learning are somewhat opaque. Experts have widely criticized the use of VA scores to allocate funding to schools and to promote or demote teachers (Amrein-Beardsley, 2014; Amrein-Beardsley & Holloway, 2019). Systematic reviews in the field have investigated the instability of VA scores. These reviews have identified a lack of consensus on how to calculate VA scores as a potential source of instability (Amrein-Beardsley et al., 2023; Aslantas, 2020; Levy et al., 2019). VA scores could fluctuate due to differences in teaching and changes in school effectiveness, they could also be due to various errors in the calculation of these scores

(Loeb & Candelaria, 2012) or their instability (Emslander, Levy, Scherer, et al., 2021; Perry, 2016). Thus, researchers and policymakers have contended that VA scores are not appropriate as the sole measure for making critical decisions in educational contexts (Amrein-Beardsley, 2014; Gorard et al., 2013). Instead, they suggest limiting their use to providing insights to teachers and schools on enhancing their educational practices. Alternatively, researchers could use VA scores productively as a sampling method (Emslander, Levy, Scherer, et al., 2021; Levy et al., 2019). Taken together, whereas prior research is equivocal about whether VA scores give stable results, VA scores might still be promising for research purposes in the Luxembourgish school context. Thus, after applying state-of-the-art VA models (Emslander, Levy, Scherer, et al., 2021; Levy et al., 2020, 2022) and testing their stability (Study 3), we used VA scores to identify Luxembourg's primary schools with high, medium, and low VA scores to compare what they are doing differently in terms of the educational psychological variables identified in our literature review introduced above. Testing VA score stability in Luxembourg and examining said variables were precisely the aims of Study 3 and Study 4 in Research Strand 2, respectively.

In conclusion, VA scores are used to quantify the schools' influence on students' achievements while accounting for as many external factors as possible. Our objective in the second research strand and the SIVA project is to check for the stability of VA scores and utilize state-of-the-art VA scores (Emslander, Levy, Scherer, et al., 2021; Levy et al., 2022) as an initial point of reference for investigating effective strategies within the education system. We intend to achieve this by selecting primary schools with consistently high VA scores and comparing them with those maintaining stable medium or low VA scores, as suggested by Levy (2020). This comparative analysis aims to uncover the underlying factors that differentiate these schools. In a last step, we hope to provide recommendations on what other schools can learn from the schools with the highest VA scores, which may offer effective solutions to address educational inequalities.

3.4 Multi-Perspective and Mixed-Methods Design

Social scientists must collect data from multiple sources and perspectives to gain a broad overview of their complex research subject. In Research Strand 1, we checked meta-analytically whether different perspectives or assessment methods moderated our findings. For Research Strand 2, the data collection of the SIVA project combines data from quantitative and qualitative methods in Study 4. These sources include quantifiably questionnaire items in a closed format on well-understood educational constructs (e.g., school climate; Wang & Degol, 2016), open text responses from teachers and school presidents to learn more about their needs, and classroom observations conducted by two educational experts. Questionnaire data from the students, teachers, parents, school presidents, and regional directors combined with classroom observations allow for comparing six perspectives. This multi-perspective design provides a fine-grained insight into the classroom and school processes. It offers the possibility to compare the different stakeholders' perspectives on the same variables.

For the quantitative part of the SIVA data collection in Study 4, we ask students, teachers, parents, school presidents, and regional directors to complete standardized questionnaires. While adults can easily give valid answers in a questionnaire, collecting such data from multilingual students in Grade 2 can be more of a challenge (Levy, 2020). We address this concern by applying appealing items and an accessible four-point agreement picture scale. This scale ranges from strong disagreement, signified by an illustration of a child vigorously shaking their head, to strong agreement, signified by an illustration of a child vehemently nodding their head. This "nodding head scale" had been introduced and successfully tested in the *National School Monitoring Programme* to motivate the young test-takers (Lehnert, 2019). Additionally, trained researchers read aloud all questions and provided standardized translations if needed.

For the qualitative part of the data collection, we ask for the teachers' and school presidents' perspectives in open text fields in the questionnaire and conduct structured classroom observations in Study 4. Two trained neutral observers performed the classroom observations during a mathematics class (the choice of subject will be elaborated in Study 2). This way, we could check their interrater agreement. The observers had only a few answer options for each item on the coding sheet. This simplicity made the decisions as low-inferential as possible and, thus, increased reliability. There was additional room for open text and other observations, which were later discussed and added to the data. Hence, we collected various quantitative and qualitative data from different perspectives and sources, rendering the SIVA study a multi-perspective, mixed-methods design.

3.5 Open Science as a Means to Make Research Robust, Reproducible, and Transparent

Open science is a paradigm shift that questions the traditional research process by fostering transparency, collaboration, and unrestricted access to data and findings. Thus, it promotes a more inclusive and accelerated scientific process through principles like open access, data sharing, and collaborative methodologies. Throughout the research work within the present doctoral thesis, we have adhered to open science practices and current standards of research transparency to make our findings reproducible and easily accessible whenever possible. Open science can be defined as the “ […] scholarly research that is collaborative, transparent and reproducible and whose outputs are publicly available” (European Commission, 2018). These measures are needed, as psychological research is published more extensively than ever before; however, there is a prevailing publication bias to lament (McShane et al., 2016). This publication bias is a manifest reservation against the publication of statistically nonsignificant results. As a result, nonsignificant findings are systematically less likely to be included in scientific publications (Ferguson & Heene, 2012). Together with the phenomenon of "*p*-hacking," where results are subsequently raised above the arbitrary significance threshold of $p = .05$, for example, through statistical corrections or outlier control (Head et al., 2015), the publication bias eventually triggered a replication crisis within psychology: numerous findings on which psychological theories were based could not be replicated or were less relevant than priorly claimed (Open Science Collaboration, 2015). This crisis strengthened the open science movement in our discipline, which still calls for more transparency to make research reproducible. Thus, the studies of the present doctoral thesis aim to be fully reproducible whenever possible.

The open science movement introduces additional steps to the research workflow to achieve greater transparency (e.g., Nosek et al., 2015). After researchers have prepared all materials and before starting the data collection, they should preregister their study. Preregistration is “[t]he specification of a research design, hypotheses, and analysis plan before observing the outcomes of a study” (Nosek & Lindsay, 2018) and helps safeguard researchers from changing their Hypotheses After Results are Known. This practice is called “HARKing”. It can lead to misinterpreting research findings, *p*-hacking, and, ultimately, inflating the false positive rate and, thus, the publication bias in research. Researchers can protect themselves from these questionable research practices by making their study and analysis plans public through preregistration.

A second step involves making all collected data, analytic code, and other materials publicly available. While publicly sharing all these items is usually met with concerns over the theft of ideas or data and requires extra resources (Nguyen et al., 2023), they significantly benefit the respective research field.

1. Sharing the data facilitates collaboration and secondary use of data through more comprehensive reporting. Including primary studies, which have shared their data, in a meta-analysis is much smoother and quicker than when no data has been shared. This

openness directly aids the quality of the meta-analysis at hand. Additionally, sharing data opens many opportunities for potential collaborators to engage in the author's research. This would be especially efficient because interested colleagues can easily identify the overlapping research interests.

2. Sharing the code has three main implications for research: it makes all results easily reproducible, strengthens the trust in research, and can help detect statistical errors quickly before wrong results might have detrimental consequences (Goldacre et al., 2019).
3. Sharing other materials, such as screening and coding forms in meta-analyses or questionnaires and classroom observation sheets in field research, shares the advantages of transparency and easier collaboration but also has some unique value. As many researchers in a given fieldwork on similar questions, sharing the materials they use will reduce duplication of effort, saving time and resources for all researchers who can use the same materials. Sharing materials, in turn, can make the research process quicker and more efficient.

A final step in the open science research process would be to publish the research article open access. Open access literature refers to digital research articles accessible online free of cost and with a somewhat liberal copyright license (Suber, 2012). While open access fees can be expensive and a researcher’s attitudes and norms of a field determine the intention to publish open access (Moksness & Olsen, 2017), making one’s research publicly available has several advantages. Open-access publishing makes the full texts of articles easier to find and can, thus, increase the visibility and impact of the articles’ findings (Eysenbach, 2006; Tang et al., 2017). Further, the enhanced accessibility may help the research findings to reach a broader audience, removing some financial and geographical barriers. Influential papers, such as the San Francisco Declaration on Research Assessment (DORA), also urge researchers to publish research open access alongside devaluating traditional citation-based impact factors as a measurement of scientific quality (Cagan, 2013).

Overall, the importance of sharing research data, analytic code, and other materials as well as other open science practices have been acknowledged across fields (Chambers, 2017; Masum et al., 2013) and are increasingly seen as a chance rather than a burden (Munafò et al., 2022). Alongside several benefits of following open science processes as described above, they can facilitate reproducibility, enhance efficiency, and, thus, make research in general more reliable and collaborative. How we have implemented open science practices to improve the robustness, reproducibility, and transparency of the four included studies is shown in Figure 3.

There, open science badges indicate the open science practices applied in each included study (Kidwell et al., 2016).

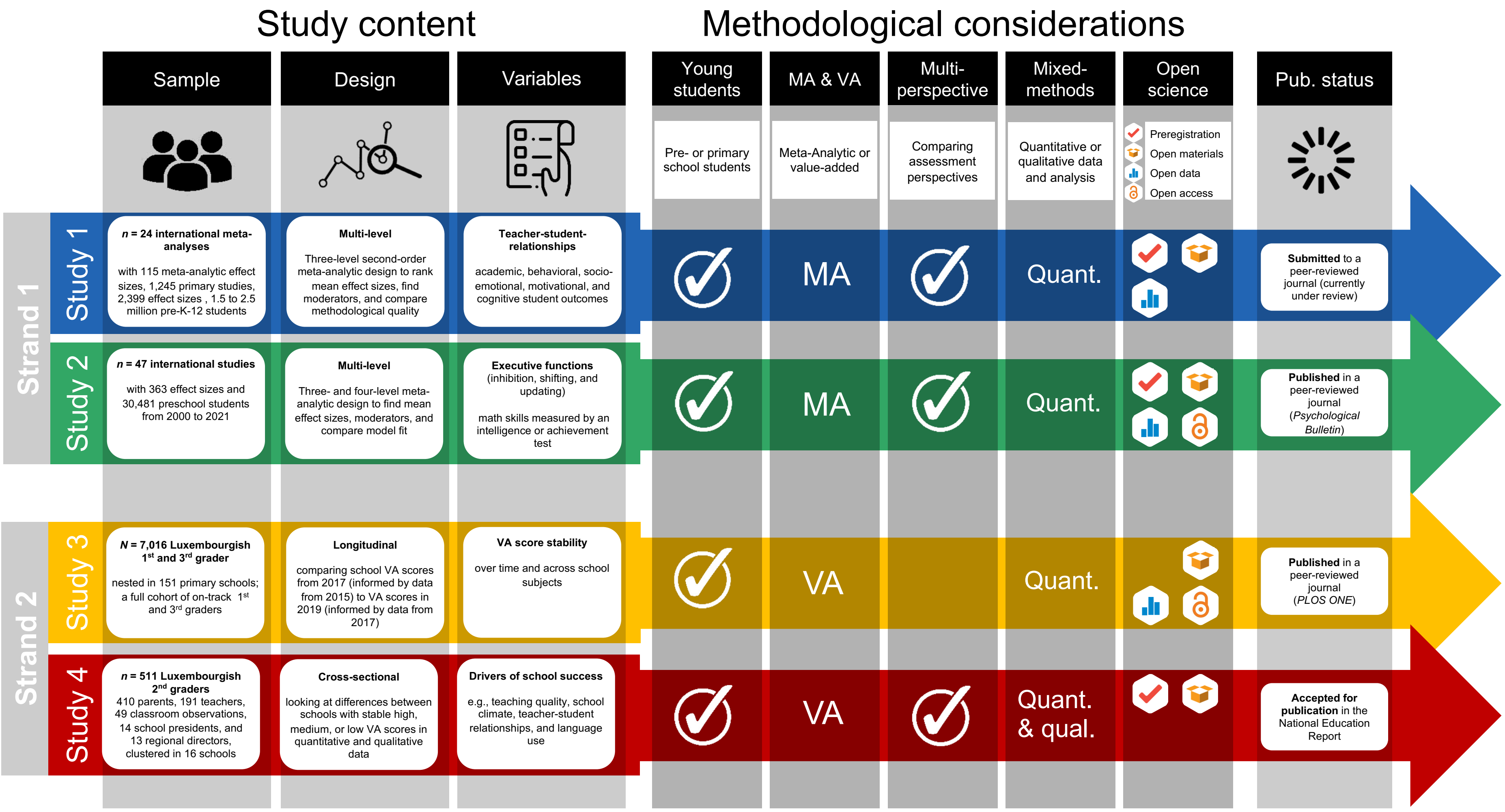
3.5.1 Open Science Practices in Meta-analytic and VA Research

Essential for the transparency in the present thesis are estimating publication bias in the meta-analytic studies and reporting null results alongside significant ones in all four studies. Estimating the publication bias within the included primary studies is imperative for meta-analytic research. Since meta-analyses build upon individual studies, they are fundamentally as good as or as poor as the individual studies they consider. Thus, a meta-analysis can also be subject to publication bias (Egger et al., 2001). The data basis of a meta-analysis can be examined for biases such as the publication bias using graphical and statistical tests to correct for its influence (Egger et al., 1997) as one of the final steps in conducting a meta-analysis. Thus, meta-analytic methods have the power to explore a research field's proneness to publication bias and assess the quality of research conducted on specific research questions to answer validity questions raised by the publication crisis.

While checking for publication bias is a crucial step in transparent meta-analytic research, and, thus for Research Strand 1, all other genuine open science practices also apply to meta-analyses. Starting with preregistration, over sharing data, analytic code, and other materials, to publishing open access—there are several helpful guides on how to conduct meta-analytic research in times of open science (Lakens et al., 2016; Moreau & Gamble, 2022; Quintana, 2015). These step-by-step instructions aid in combining the difficult task of conducting a research synthesis while adhering to open science practices to make research transparent, reproducible, and easily accessible.

The publishing of null results is imperative to overcome the strengthening of replicability in any method and discipline. Therefore, the included studies in this doctoral thesis report all results equally. In the two studies using VA methodologies in Research Strand 2, there is a particular focus on the lack of statistically significant results, posing several questions on the chosen sample and the tested variables and the validity of VA scores in educational research (see Study 3 and Study 4). In conclusion, the studies in the present doctoral thesis aim to tackle their individual research questions with meta-analytic and VA methodology following open science practices.

Figure 3. *Systematic overview of the four studies in the present thesis focusing on their content, methodological considerations, and publication status*



		Study content			Methodological considerations					
		Sample	Design	Variables	Young students	MA & VA	Multi-perspective	Mixed-methods	Open science	Pub. status
					Pre- or primary school students	Meta-Analytic or value-added	Comparing assessment perspectives	Quantitative or qualitative data and analysis	✓ Preregistration Open materials Open data Open access	
Strand 1	Study 1	***n* = 24 international meta-analyses** with 115 meta-analytic effect sizes, 1,245 primary studies, 2,399 effect sizes , 1.5 to 2.5 million pre-K-12 students	**Multi-level** Three-level second-order meta-analytic design to rank mean effect sizes, find moderators, and compare methodological quality	**Teacher-student-relationships** academic, behavioral, socio-emotional, motivational, and cognitive student outcomes	✓	MA	✓	Quant.	Preregistration Open materials Open data	**Submitted** to a peer-reviewed journal (currently under review)
	Study 2	***n* = 47 international studies** with 363 effect sizes and 30,481 preschool students from 2000 to 2021	**Multi-level** Three- and four-level meta-analytic design to find mean effect sizes, moderators, and compare model fit	**Executive functions** (inhibition, shifting, and updating) math skills measured by an intelligence or achievement test	✓	MA	✓	Quant.	Preregistration Open materials Open data Open access	**Published** in a peer-reviewed journal (*Psychological Bulletin*)
Strand 2	Study 3	***N* = 7,016 Luxembourgish 1st and 3rd grader** nested in 151 primary schools; a full cohort of on-track 1st and 3rd graders	**Longitudinal** comparing school VA scores from 2017 (informed by data from 2015) to VA scores in 2019 (informed by data from 2017)	**VA score stability** over time and across school subjects	✓	VA		Quant.	Open materials Open data Open access	**Published** in a peer-reviewed journal (*PLOS ONE*)
	Study 4	***n* = 511 Luxembourgish 2nd graders** 410 parents, 191 teachers, 49 classroom observations, 14 school presidents, and 13 regional directors, clustered in 16 schools	**Cross-sectional** looking at differences between schools with stable high, medium, or low VA scores in quantitative and qualitative data	**Drivers of school success** e.g., teaching quality, school climate, teacher-student relationships, and language use	✓	VA	✓	Quant. & qual.	Preregistration Open materials	**Accepted for publication** in the National Education Report

Note. The Figure 3 is graphically adapted from Colling (2022). For an in-depth review and an explanation of the open science badges, see Kidwell et al. (2016). MA = meta-analysis; VA = value-added; Quant. = quantitative; Qual. = qualitative.

4 The Present Thesis

Along with its general population, Luxembourg's student population is diversifying. This surge in diversity may lead to increased educational inequality caused by diverse language backgrounds or socioeconomic statuses of students in a trilingual educational system. While increased educational inequalities could naturally lead to deteriorating performance in international comparisons, such as the PISA study, Luxembourg remains stable in its test results (Weis et al., 2018). Despite a diversifying student population and more educational inequality, this lack of deterioration in test results suggests several schools in Luxembourg must be using strategies to address educational inequalities effectively and support their students against the odds. As such, the first research strand of the present thesis aims to learn about general social and cognitive learning processes. Research Strand 2 focuses on concrete educational psychological practices in Luxembourg to help all students to succeed. Together, they form the following aims of the present thesis. To (1) delve into the link between TSRs and student outcomes, (2) deepen our understanding of EFs and their relation to mathematics skills, (3) identify schools with a stable VA score, and (4) compare educational psychological strategies to address educational inequalities successfully are the objectives of the present thesis.

Addressing the four objectives, we made several methodological considerations. We defined which target population to investigate, which statistical approach to use, how to design the SIVA study effectively, and how to make our work transparent, reproducible, and openly accessible. Concerning the target population, we focus on younger students to have the highest possible return on investment (Heckman, 2008). Regarding the statistical methods, we chose research synthesis strategies of meta-analysis and SOMA in the first research strand. Additionally, we use VA scores in the second research strand to find the most effective primary schools and compare them with other primary schools to learn how effective schools successfully address educational inequalities. Regarding study design, we combine quantitative data collection statistical analyses with more qualitative classroom observations and analyses of open-text responses. Alongside this mixed methods approach, we are interested in the perspectives of different stakeholders within the school, meaning students, teachers, and third parties such as parents, expert observers, and others. Finally, aiming to make our work transparent, reproducible, and openly accessible, we follow current open science practices by preregistering our work whenever feasible, sharing the data, statistical code, and materials, and making the final article openly accessible. These methodological considerations span across the four studies within the present doctoral thesis and divide the two Research Strands of meta-analytic research on the international scientific literature (Study 1 and Study 2) and VA research in Luxembourg (Study 3 and Study 4; for an overview, see Figure 3).

4.1 Rational of the Included Studies

4.1.1 Investigating Teacher-Student-Relationships in the Meta-Analytic Literature (Study 1)

The literature review identified instructional quality (Klieme et al., 2001) and school climate (Wang & Degol, 2016) as the theoretical core for the SIVA project and Study 4. The overlap of the two models showed the importance of the relationship between teachers and students. The importance of TSRs are researched and put into legislation internationally and in Luxemburg (SCRIPT, 2018). Thus, it is crucial to clarify how and for whom TSRs work, theoretically (Study 1) and empirically (Study 4). In other words, we strive to understand the TSR research subfield better.

As introduced in the methods section, SOMAs are valid tools to give an overview of a subfield of research. Thus, Study 1 reports a systematic review of meta-analyses and several SOMAs on the link between TSRs and outcomes in preschool to K-12 students. As such, we synthesize 24 meta-analyses with 116 effect sizes based on more than 2 million prekindergarten and K-12 students, spanning 70 years of educational psychological research on TSRs and student outcomes. The student outcomes include not only academic achievement as in prior efforts (Hattie, 2008; Waack, 2018) but also academic emotions, appropriate student behavior, behavior problems, motivation, school belonging and engagement, student well-being, or EFs and self-control. Additionally, the study reviews prior moderators and calculates original moderator analyses identifying for which students and in which way TSRs are essential. As such, we find out whether younger or older students profit more from positive TSRs and whether gender differences exist. In further moderator analyses, we investigate the diverse perspectives on TSRs and compare student-, peer-, or teacher-assessments of TSRs. Further, we assess the methodological quality of the included meta-analyses and investigate their use of open science practices. The results of Study 1 offer a broad overview of the field of TSR research and address the first research aim of the present thesis.

4.1.2 Delving Into Executive Functions and Their Link to Mathematics Skills in the Preschool Literature (Study 2)

One of the student outcomes that were significantly correlated with TSRs and a prerequisite for learning, but missing in the SIVA project, are EFs (Vandenbroucke et al., 2018). EFs are part of the cognitive learning process and are an essential student factor for school success. They therefore complement the previously investigated social learning processes by exploring EFs in combination with mathematical skills. Thus, we investigate EFs in combination with mathematical skills to shed light on the cognitive prerequisites for learning and school success in this thesis. EFs are mental processes regulating human cognition and behavior (Miyake et al., 2000; Miyake & Friedman, 2012). EFs have several subprocesses, with response inhibition, mental set shifting, and updating of working memory arguably being the three most investigated. They can be considered prerequisites for many school-success-defining skills, such as reading or mathematics (Diamond, 2013; Follmer, 2018; Friso-van den Bos et al., 2013; Yeniad et al., 2013).

EFs constitute an essential prerequisite for mathematics skill development (van der Ven, 2011; van der Ven et al., 2012). Thus, they play a crucial role in school readiness and predicting academic success throughout the school career (Blair, 2002; Diamond, 2013; Duncan et al., 2007) and academic success later in the school career. Study 2 meta-analytically investigates the link between EFs and mathematics skills in preschool children. While Study 1 includes meta-analyses with students up to high school, Study 2 includes primary research articles on preschool children as the understanding of numbers already develops before school entry (Passolunghi & Lanfranchi, 2012) and is related to EFs (Geary et al., 2019), suggesting that the years before school entry are a crucial learning period with rapid development of mathematical skills and EFs (Zelazo & Carlson, 2012).

Concretely, we used data from 47 studies from 2000 to 2021 with 363 effect sizes and *N* = 30,481 participants. Using moderator analyses, we investigate whether the link between mathematics skills changes with child age or assessment perspective (e.g., child-reported vs. parent-reported EFs and mathematics skills). With MASEM we finally test whether the three subdimensions of EFs (i.e., inhibition, shifting, and updating) differ in their ability to explain variations in math intelligence, which is a controversial question among researchers. Further studies suggest that EFs are not only linked to mathematic skills, school success, and TSRs (Duncan et al., 2007; Vandenbroucke et al., 2018). They also seem associated with school climate (Piccolo et al., 2019), one of the foci of the present thesis and SIVA project and its data collection in Study 4.

4.1.3 Identify Primary Schools With High Value-Added in Luxembourg (Study 3)

After finding and further investigating the variables associated with educational effectiveness, we needed to identify the schools where we best collect the data. Study 3 lays the groundwork for using VA scores in Luxembourg for research purposes. Even though we focus only on primary schools in the SIVA project, we cannot test the full population of more than 150 primary schools in the Grand Duchy during the data collection. Thus, we want to collect data only from highly, average, or below-average effective schools and compare these three groups. One way to differentiate the three groups would be to use only schools with stable high, medium, or low VA scores.

Using VA scores to categorize schools into these groups would require VA scores to be stable over time and across the outcome domains. To increase the credibility of the VA scores, a school with a high VA score should also have a comparable score in the following years. Further, such a school would also be assumed to perform well not only in one school subject but in multiple subjects, such as both language and mathematics. However, research on school-level VA score stability over time and across subjects is still scarce, and some research even suggests that there might be a “stability problem” with VA scores (Aslantas, 2020). Suppose VA scores are unstable over time and fluctuate across outcome domains (e.g., mathematics and language learning). In that case, their use for high-stakes decision-making is in question and could have detrimental real-life implications for teachers and schools.

As a consequence, the stability of VA scores has been debated quite fiercely not only in the US (Loeb & Candelaria, 2012; Papay, 2011) but also in European countries, such as Portugal or the United Kingdom (Ferrão, 2012; Gorard et al., 2013; Perry, 2016; Thomas et al., 2007). The results of prior research are somewhat equivocal, with some researchers finding acceptable VA score stability (Ferrão, 2012; Thomas et al., 2007) and some not (Gorard et al., 2013; Perry, 2016). Thus, the stability of VA scores over time and across subjects in primary schools still needs to be tested to ensure that VA scores are informative for this educational level.

To address this research gap, we test the stability of VA scores for Luxembourgish primary schools. In Study 3, we focus on the VA scores stability of over two years and across the two subject domains of language and mathematics learning. To this end, we draw on representative, large-scale, and longitudinal data of $N = 7{,}016$ students in 151 schools from two cohorts of standardized achievement tests in Luxembourg. This sample encompasses the entire population of Luxembourg’s students who are on track from Grade 1 to Grade 3 in 2017 and 2019. While our results suggest that VA scores should not be used for high-stakes decision-making, such as teachers' tenure or school funding, we can use them to identify schools that effectively address their students' diversity and have them succeed against the odds. Further, we can compare this group of schools with stable high VA scores with those that have stable medium or low VA score over a more extended period and across subject domains.

4.1.4 Compare Schools to Identify Educational Psychological Drivers of School Success Against the Odds (Study 4)

Study 4 represents the first findings of the SIVA project (written for a broader audience interested in educational research, such as parents and teachers). Based on the results of Study 3, we identify 16 schools with stable high, medium, and low VA scores over two years and across subjects to compare them on the variables of school effectiveness that we had priorly identified: instructional quality (Klieme et al., 2001), school climate (Wang & Degol, 2016), and other important variables (e.g., boredom and collective teacher self-efficacy; Hattie, 2008; Waack, 2018) as well as Luxembourg specificities (e.g., language use and role of the school president). As explained above, we found these variables through an in-depth literature search and by combining the most pertinent teaching and school effectiveness models.

Thus, Study 4 describes the first results from the SIVA project's multi-perspective, mixed-methods data collection. Concretely, we compare the results of the 16 selected schools based on 49 classroom observations and questionnaire data on a total of 511 students in Grade 2, 410 of their parents, 191 classroom and subject teachers, 14 school presidents, and 13 regional directors, collected in early 2022 during the COVID-19 pandemic. In a five-step process, we first compare the SIVA sample to the general population to determine how representative these roughly 10% of all schools and students are for the general Luxembourgish student population. Second, we compare the mean values of the most important variables with results from other European countries to locate Luxembourg’s measures on instructional Quality, TSR, and school climate internationally. Third, we compare the three groups of primary schools with a high, medium, or low VA score and check for differences in their performance of the variables. Fourth, we use boot-strapping methods to tentatively answer whether the statistical significance of group differences changes with a higher risk of error. Fifth, we take a closer look at the qualitative open-text answers from teachers and school presidents to use their expert input to answer our research question on what might be the driver of effective schools.

In the following, the present thesis presents four studies to address its four aims. The studies are printed as published or as submitted, respectively. After the presentation of the four studies, a general discussion will summarize the results for all studies, address common limitations, discuss contributions to research and practice, and outline future research before drawing a general conclusion.

5 Study 1

Teacher-Student Relationships and Student Outcomes: A Systematic Review of Meta-Analyses and Second-Order Meta-Analysis

Valentin Emslander[a], *Doris Holzberger*[b], *Sverre Berg Ofstad*[c], *Antoine Fischbach*[a], *and Ronny Scherer*[c, d]

[a] *Luxembourg Centre for Educational Testing (LUCET) at the University of Luxembourg, Faculty of Humanities, Education and Social Sciences, University of Luxembourg, Luxembourg*

[b] *Centre for International Student Assessment (ZIB), TUM School of Social Sciences and Technology, Technical University of Munich, Germany*

[c] *Centre for Educational Measurement at the University of Oslo (CEMO), Faculty of Educational Sciences, University of Oslo, Norway*

[d] *Centre for Research on Equality in Education (CREATE), Faculty of Educational Sciences, University of Oslo, Norway*

Submitted to *Psychological Bulletin* [1]

Currently *Under Review*

This study was pre-registered at https://doi.org/10.17605/OSF.IO/J2EMF
The data, code, and material are available at https://doi.org/10.17605/OSF.IO/5DE6P
The preprint of the manuscript is available at https://doi.org/10.31234/osf.io/qxntb

[1] The numbering of headings, tables and figures has been adjusted to align with the structure of the present work. The Appendix is integrated general Appendix at the end of this thesis.

Abstract

Teacher-student relationships (TSRs) play a vital role in establishing a positive school climate and promoting positive student outcomes. Several meta-analyses have suggested significant associations between TSRs and, for example, academic achievement, a lack of disruptive behavior, school engagement, peer relationships, motivation, executive functions, and general well-being. However, these meta-analyses have differed substantially in TSR-outcome relations, moderators, and quality, thus complicating the interpretation of these findings. In this preregistered systematic review of meta-analyses plus original second-order meta-analyses (SOMAs), we aimed to (a) synthesize the meta-analytic evidence on relations between TSRs and student outcomes, (b) map influential moderators of these relations, and (c) assess the methodological quality of the meta-analyses. We synthesized over 70 years of educational research in 24 meta-analyses encompassing a total of 116 effect sizes based on more than 2 million prekindergarten and K-12 students. We conducted several three-level SOMAs and found that TSRs had similar large significant relations with eight clusters of outcomes: academic achievement, academic emotions, appropriate student behavior, behavior problems, executive functions and self-control, motivation, school belonging and engagement, and student well-being. Age, gender, and informant (student-, peer-, or teacher-assessments) were the most frequently examined moderators in prior research, and our moderator analyses suggested student grade level and social minority status as moderators. We further found large differences in quality between the meta-analyses, and these differences were not associated with the TSR-outcome relations. These results map the field of TSR research; present their relations, moderators, and meta-analytic quality; and show how TSRs can contribute to improving outcomes in students via relationship building. Future research should follow meta-analytic open science procedures to improve quality and reproducibility.

Keywords: teacher-student relationships, academic outcomes, well-being, second-order meta-analysis, school students

Public Significance Statement: The present systematic review of 24 meta-analyses gives an overview of over 70 years of research on teacher-student relationships (TSRs). We found that TSRs were associated with many crucial student characteristics, such as academic achievement and emotions, motivation, and appropriate behavior. Conversely, TSRs had negative relations with other outcomes (e.g., behavior problems at school). These relations were larger when the teacher’s viewpoint was assessed and differed between older and younger students. Ultimately, our findings provide valuable evidence for educational decision-makers and policymakers. By integrating findings from multiple meta-analyses, we offer a more comprehensive and reliable evidence base for informing policy decisions, educational practice in the classroom, and teacher education. Thus, these findings are of broad public interest and can help improve teacher education and TSR interventions.

5.1 Introduction

A good school climate is characterized by positive teacher-student relationships (TSRs), which are linked to academic, behavioral, socioemotional, motivational, and general cognitive outcomes. TSRs are defined as interactions characterized by warmth, closeness, support, and friendliness (positive TSRs), or conflict and dependency (negative TSRs; Hamre & Pianta, 2001; Pianta, 1999). Consequently, TSRs focus on affectional and emotional facets rather than administrative (e.g., giving away lunch vouchers) or lesson-based support (e.g., providing classroom organization or clarity of instruction; see Authors, 2022). Both positive and negative TSRs are well-researched, and meta-analyses have summarized findings on their associations with student outcomes. This body of meta-analytic research has brought to light how positive TSRs are connected to school engagement (Roorda et al., 2017), good peer relationships (Endedijk et al., 2022), executive functions (Vandenbroucke et al., 2018), and general well-being (Chu et al., 2010). Meta-analyses on negative TSRs have reported correlations with student anger (Nurmi, 2012) and bullying (Krause & Smith, 2022). However, these meta-analyses varied considerably in their definitions and operationalizations of TSRs, types of outcome variables, and the strengths of the relations between the TSRs and the outcomes. Moreover, the meta-analytic evidence for the TSR-outcome relation is partially contradictory. For instance, whereas some meta-analyses have suggested that positive TSRs were associated with fewer behavior problems (Moore et al., 2019; Roorda et al., 2021), others found that they were associated with more behavior problems (Nurmi, 2012).

The meta-analytic literature is also ambiguous about which moderators (e.g., age or gender) influence these relations. Cooper and Koenka (2012) suggested that differences in findings may stem from differences in approaches and the quality of the meta-analyses. To map and explain these differences, we systematically reviewed and statistically summarized the meta-analytic literature in multiple second-order meta-analyses (SOMAs; Cooper & Koenka, 2012). Over and above integrating TSR associations, we aimed to identify which associations were the largest and should therefore be tested for causal relations for use in effective TSR and life-success interventions. By comparing and contrasting findings from these diverging meta-analyses, we aimed to identify the factors contributing to variability in TSR-outcome relations (e.g., student characteristics, study designs, or publication type), thereby enabling us to evaluate the consistency and robustness of meta-analytic findings on TSRs across different studies and contexts and helping us identify areas of improvement and research gaps. Ultimately, with our SOMAs, we aimed to provide valuable evidence for educational decision-makers and policymakers. By integrating findings from multiple primary meta-analyses, we offer a more comprehensive and reliable evidence base for informing policy decisions, educational practice in the classroom, and teacher education.

Specifically, we provide a comprehensive overview of (a) the relation between TSRs and students' academic, behavioral, socioemotional, motivational, and cognitive outcomes, (b) important moderators of these relations, and (c) meta-analytic quality. In addition to mapping the moderators included in the meta-analyses, we also examined possible moderators that describe the characteristics of the meta-analyses (i.e., second-order moderators). Finally, we quantified and reviewed the methodological quality of the included meta-analyses. Table 1 presents an overview of the included meta-analyses.

5.2 Theoretical Framework

5.2.1 Defining TSRs

TSRs generally refer to the quality of the relationship(s) between teaching personnel and their student(s) in schools. As a construct, TSRs commonly focus on the affective aspects of the relationships (Roorda et al., 2021) and can thus be differentiated from administrative and instrumental support (Li et al., 2021; Vandenbroucke et al., 2018). In following the affective TSR definition, which is grounded in attachment theory, prior meta-analyses have divided TSRs into the following facets (Roorda et al., 2017; Sabol & Pianta, 2012; Vandenbroucke et al., 2018; Verschueren & Koomen, 2012):

- Closeness, representing the warmth of the interactions between teachers and students;
- Conflict, representing negative interactions; and
- Dependency, representing clingy student behavior.

Whereas there are extensive discussions about the different facets of TSRs elsewhere (see, for an overview, Wentzel, 2022), in the present paper, positive TSRs are characterized by high degrees of closeness and warmth and low degrees of conflict and dependency (Hamre & Pianta, 2001; Roorda et al., 2017). In scientific theory, TSRs have most notably been part of the concept of school climate. In their seminal review, Wang and Degol (2016) synthesized the literature on the construct, measurement, and impact of school climate and defined TSRs as "Trust, interpersonal relationships between staff and students, affiliation" (p. 4), an integral part of the community aspect of school climate. Moreover, conceptualizing TSRs as part of school climate resonates with the three basic dimensions of instructional quality—specifically, TSRs are aligned with the dimension of socioemotional teacher support (Klieme et al., 2001; Praetorius et al., 2018). Overall, TSRs can be considered key characteristics of the quality of instruction and the school climate.

Table 1. *Overview of Included Meta-Analyses*

Reference	Description & method	Age range Grade level	Examined relation (TSR & student outcome)	ks	r̄	95% CI
Ali et al., 2015	Meta-analysis of the relation between teacher acceptance and psychological adjustment & school conduct using average weighted effect sizes	9-18 MS - HS	Teacher acceptance & psychological adjustment	16	.32	[.29, .36]
		9-18 MS - HS	Teacher acceptance & school conduct	8	.22	[.17, .26]
Allen et al., 2018	Meta-analysis of the relation between teacher support and school belonging using a random-effects model	12-18 MS - HS	Teacher support & school belonging	14	.46	[.37, .54]
Cherne, 2008	Meta-analysis of single-case studies of the relation between teacher praise and student behavior. Calculated effect sizes by averaging percentage of all nonoverlapping data	4-8	Teacher praise & student behavior	52	.70	[.51, .79]
		-	Teacher praise & academic behavior	6	.75	[.58, .92]
		4-12	Teacher praise & appropriate social behavior	16	.57	[.36, .78]
		5-12	Teacher praise & inappropriate social behavior	4	.83	[.59, 1]
		-	Ability-based praise & student behavior	2	.30	[0, .89]
		5-12	Behavior-specific praise & student behavior	8	.80	[.52, 1]
		-	Effort-based praise & student behavior	2	.27	[0, .71]
		5-12	"Other" forms of praise & student behavior	3	.70	[.45, .95]
		4-12	"Undefined" forms of praise & student behavior	11	.90	[.68, 1]
Chu et al., 2010	Meta-analysis of the relation between support from teachers and school personnel and student well-being. Used weighted average effect sizes	K - HS	Teacher and school personnel support & well-being	125	.21	[.20, .21]
Cornelius-White, 2007	Meta-analysis of the relation between student-centered teaching and behavioral and cognitive outcomes. Used mean effect sizes	PRE-K - HS	Person-centered teaching & cognitive and behavioral outcomes	98	.31	[.30, .33]
		PRE-K - HS	Person-centered teaching & cognitive outcomes	71	.31	[.31, .31]
		PRE-K - HS	Person-centered teaching & behavioral outcomes	81	.35	[.35, .35]
Endedijk et al., 2022	Three-level meta-analysis of relations between TSRs and peer relationships & student behavior				.	
		3-18 PRE-S - HS	TSR & peer relationship quality	1475	28	[.26, .30]
		PRE-S - HS	TSR & student behavior	992	.255	[.23, .28]
		MS - HS	Teacher support & grades and achievement	7	0.17	[.13, .21]

Givens Rolland, 2012	Meta-analysis of the relation between teacher support and achievement & socioemotional outcomes using a mixed effects model	MS - HS	Teacher support & personal mastery	4	0.31	[.27, .36]
		MS - HS	Teacher support & intrinsic value	5	0.40	[.35, .44]
		MS - HS	Teacher support & perceived ability	5	0.38	[.34, .42]
		MS - HS	Teacher support & prosocial factors	5	0.24	[.19, .28]
Kincade et al., 2020	Meta-analyses of the effectiveness of intervention programs aimed to improve TSRs that aggregated weighted effect sizes	PRE-K - MS	TSR intervention program & TSR	9	.13	[.12, .14]
		PRE-K - MS	TSR intervention program & TSR closeness	13	.11	[.10, .12]
		PRE-K - MS	TSR intervention program & TSR conflicts	13	-.02	[-.04, -.01]
Korpershoek et al., 2016	Meta-analysis of the effect of classroom interventions that aimed to improve TSRs and student outcomes using a random-effects model	-	TSR intervention & overall outcome	1-2	.06	[-.02, .15]
		-	TSR intervention & academic outcomes	1-2	.12	[.03, .21]
		-	TSR intervention & behavioral outcomes	1-2	.03	[-.07, .13]
		-	TSR intervention & socio-emotional outcomes	1-2	.03	[-.06, .12]
		-	TSR intervention & motivational outcomes	1-2	.04	[-.05, .13]
		-	TSR intervention & other outcomes	1-2	.13	[-.05, .32]
Krause & Smith, 2022	Meta-analysis of the relation between conflicts in TSRs and bullying using a random-effects model	K - HS	TSR conflict & bullying perpetration	13	.32	[.25, .38]
		K - HS	TSR conflict & bullying victimization	12	.25	[.17, .33]
Lei et al., 2016	Meta-analysis of the relation between affective TSRs and externalizing behavior using a fixed-effects model	3-18 K - HS	Positive affective TSR & externalizing problem behavior	78	-.26	[-.27, -.25]
		3-18 K - HS	Negative affective TSR & externalizing problem behavior	71	.55	[.55, .56]
Lei et al., 2018	Meta-analysis of the relation between teacher support and students' academic emotions using a fixed-effects model	ELE - HS	Teacher support & positive academic emotions	45	.34	[.33, .35]
		ELE – HS*	Teacher support & negative academic emotions	76	-.22	[-.23, -.21]
Li et al., 2021	Three-level meta-analysis of the relation between school discipline and students' self-control	3-17 PRE-S - HS	School discipline & self-control	278	.19	[.19, .19]
		PRE-S - HS	TSR & self-control	142	.24	[.22, .24]
Moore et al., 2019	Meta-analysis of experimental and quasi-experimental single-case studies of the effect of teacher praise on student behavior. Calculated effect sizes by averaging percentage of all nonoverlapping data	K - HS	Teacher praise & appropriate behavior (overall)	7	.79$_{\text{TAU}}$	[.42, 1]
		K - HS	Teacher praise & appropriate behavior (student-level studies)	4	.68$_{\text{TAU}}$	[.48, .90]
		ELE - MS	Teacher praise & appropriate behavior (class-level studies)	3	.88$_{\text{TAU}}$	[.42, 1]
		ELE - MS	Teacher praise & disruptive behavior (overall)	7	.90$_{\text{TAU}}$	[.68, 1.17]
		ELE	Teacher praise & disruptive behavior (student-level studies)	4	.74$_{\text{TAU}}$	-
		ELE - MS	Teacher praise & disruptive behavior (class-level studies)	3	.92$_{\text{TAU}}$	[.68, 1.17]
Nurmi, 2012		-	TSR conflicts & academic achievement	10	-.20	[-.27, -.13]

	Meta-analysis of the relations between TSRs and academic, behavioral, and socioemotional outcomes using a random-effects model	-	TSR closeness & academic achievement	7	.21	[.12, .29]
		-	TSR dependency & academic achievement	2	-.19	[-.30, -.07]
		-	TSR conflicts & external problem behavior	12	.57	[.44, 0.67]
		-	TSR closeness & external problem behavior	8	-.19	[-.32, -.05]
		-	TSR dependency & external problem behavior	6	.27	[.18, .36]
		-	TSR secure attachment & external problem behavior	2	-.13	[-.34, .10]
		-	TSR anxious attachment & external problem behavior	2	-.37	[-.51, -.21]
		-	TSR conflict & anger	2	.38	[-.39, .84]
		-	TSR closeness & anger	2	-.12	[-.27, .04]
		-	TSR conflict & internal problem behavior	5	.30	[.13, .44]
		-	TSR closeness & internal problem behavior	5	-.20	[-.33, -.06]
		-	TSR dependency & internal problem behavior	3	.37	[.23, .49]
		-	TSR secure attachment & internal problem behavior	2	.10	[.07, .14]
		-	TSR anxious attachment & internal problem behavior	2	-.21	[-.54, .17]
		-	TSR conflicts & shyness	6	-.23	[-.48, .06]
		-	TSR closeness & shyness	7	-.29	[-.45, -.11]
		-	TSR dependency & shyness	2	.22	[.09, .33]
		-	TSR conflict & prosociability	2	.18	[-.03, .37]
		-	TSR closeness & prosociability	2	.19	[.03, .36]
		-	TSR dependency & prosociability	2	-.28	[-.51, -.02]
		-	TSR conflict & motivation	10	-.35	[-.44, -.26]
		-	TSR closeness & motivation	6	.17	[.00, .33]
Roorda et al., 2011	Meta-analysis of relations between TSRs and engagement & achievement using a random effects model	K - HS	Positive TSR & academic achievement	61	.16	[.13, .20]
		K - HS	Positive TSR & engagement	61	.34	[.28, .39]
		K - MS	Negative TSR & academic achievement	28	-.18	[-.22, -.15]
		K - MS	Negative TSR & engagement	18	-.31	[-.38, -.24]
Roorda et al., 2014	Meta-analysis of the relations between affective TSRs and students' school learning using mean effect sizes	K - ELE	Positive TSR & academic achievement (primary school sample)	42	.14	-
		K - ELE	Positive TSR & engagement (primary school sample)	35	.26	-
		K - ELE	Negative TSR & academic achievement (primary school sample)	24	-.19	-
		K - ELE	Negative TSR & engagement (primary school sample)	15	-.34	-
		MS - HS	Positive TSR & academic achievement (secondary school sample)	17	.16	-
		MS - HS	Positive TSR & engagement (secondary school sample)	23	.30	-
		MS - HS	Negative TSR & academic achievement (secondary school sample)	3	-.13	-
		MS - HS	Negative TSR & engagement (secondary school sample)	2	-.25	-

Roorda et al., 2017	Meta-analysis of the relations between affective TSRs and students' engagement and achievement using a random-effects model	K - HS	Positive TSR & academic performance	117	.17	[.04, .11]
		K - HS	Positive TSR & engagement	118	.35	[.25, .32]
		K - HS	Negative TSR & academic performance	41	-.16	[-.11, -.02]
		K - HS	Negative TSR & engagement	34	-.28	[-.24, -.15]
Roorda et al., 2021	Meta-analysis of the relation between student-teacher dependency and students' school adjustment using a fixed-effects model	K - ELE	Student-teacher dependency & engagement	9	-.13	[-.17, -.09]
		K - ELE	Student-teacher dependency & achievement	8	-.12	[-.15, -.09]
		K - ELE	Student-teacher dependency & externalizing behavior	19	.27	[.25, .30]
		K - ELE	Student-teacher dependency & internalizing behavior	16	.32	[.28, .35]
		K - ELE	Student-teacher dependency & prosocial behavior	7	-.17	[-.23, -.11]
		K - ELE	Student-teacher dependency & engagement	7	-.13	-
		K - ELE	Student-teacher dependency & achievement	4	-.07	-
		K - ELE	Student-teacher dependency & externalizing behavior	8	.28	-
		K - ELE	Student-teacher dependency & internalizing behavior	9	.35	-
		K - ELE	Student-teacher dependency & prosocial behavior	3	-.14	-
Strom & Bolster, 2007	Meta-analysis of the relations between supportive messages at home and in school and dropout using average weighted effect sizes	HS	Supportive communication in school & school dropout	7	.14	-
Tao et al., 2022	Meta-analysis of the relation between perceived teacher support and academic achievement using a random-effects model	ELE - HS	Teacher support & academic achievement	93	.16	[.13, .18]
Vandenbroucke et al., 2018	Meta-analysis of the relations between teacher-student interactions and executive functions using both random-effects and fixed-effect models	PRE-K - ELE	Teacher-student interaction & overall executive function	101	0.09	[.04, .13]
		PRE-K - ELE	Teacher-student interaction & general executive function measures	9	0.11	[.07, .16]
		PRE-K - ELE	Teacher-student interaction & working memory	32	0.09	[.03, .15]
		PRE-K - ELE	Teacher-student interaction & inhibition	71	0.08	[.02, .14]
		PRE-K - ELE	Teacher-student interaction & cognitive flexibility	9	0.00	[-.04, .04]
Wang et al., 2020	Three-level meta-analysis of the relations between classroom climate and students' academic & psychological well-being	K - HS	Socioemotional support & social competence	74	0.21	[.10, .32]
		K - HS	Socioemotional support & externalizing behavior	42	-0.20	[-.29, -.11]
		K - HS	Socioemotional support & socioemotional distress	42	-0.19	[-.26, -.11]
		K - HS	Socioemotional support & academic achievement	70	0.12	[.07, .17]
		K - HS	Socioemotional support & motivation and engagement	92	0.23	[.18, .29]
Wilkinson, 1980	Meta-analysis of the relations between teacher praise and student academic	ELE	Praise & student achievement	19	0.08	[.02, .13]
		ELE	Praise & reading gains	12	0.11	[.03, .19]

	achievement using average weighted	ELE	Praise & mathematics gains	8	0.18	[.08, .28]
	effect sizes	ELE	Positive classroom climate & student achievement	22	-0.04	[-.14, .06]

Note. Larger positive effect sizes indicate closer relations between TSRs and student outcomes. k_S = Number of included studies; Age range = Mean age of the youngest and the oldest sample within the meta-analysis in years; $\bar{r}$ = weighted average correlation; * = also includes a few studies with university students; $_{\text{TAU}}$ = is not a mean correlation (and will not be included in quantitative analyses) but an average Tau-U coefficient (Moore et al., 2019); PRE-K = prekindergarten, K = kindergarten, PRE-S = preschool, ELE = elementary school, MS = middle school, HS = high school.

5.2.2 Relations Between TSRs and Student Outcomes

Several meta-analyses have supported the association between positive TSRs and improved student outcomes, such as academic achievement (Roorda et al., 2017), student behavior (Endedijk et al., 2022), executive functions (Vandenbroucke et al., 2018), general well-being (Chu et al., 2010), academic emotions (Lei et al., 2018; Nurmi, 2012), and (less) bullying (Krause & Smith, 2022). Several educational and developmental psychological theories can explain these relations. In the following, we focus on the two most prominent theories that have been used to explain the impact of TSRs and discuss their influence on research on the construct of TSRs—Bronfenbrenner's (1979) bioecological model of human development and Bowlby's (1982) attachment theory.

Bronfenbrenner's (1979) bioecological model considers the driver of human development to be one's interactions with the people in one's closest environment in a complex interplay of external influences. The model describes five layers of relationships and interactions—so-called systems—around a developing child, starting from the closest layer and moving out to the most distant layer. For a student, the first layer (micro-system) is made up of the interactions they have with the people in their closest environment, such as their parents, extended family, peers, and teachers. Interactions in the school or between the school and the parents happen in the second layer of interactions (meso-system). These first two layers are further described in the third layer (exo-system) where the interaction between the micro- and meso-systems takes place. The fourth layer (macro-system) signifies the current educational policies shaping the teachers' practice and the curriculum. Lastly, the outermost fifth layer (chrono-system) puts all the interactions in a temporal perspective. Whereas these systems overlap and interact in real-life school situations, the TSRs are the some of the most important relationships for the students' development (see Bronfenbrenner & Morris, 2007). Bronfenbrenner (1979) places this relationship in the center along with the individual and the most proximal system, the micro-system. Students interact with teachers daily and for long periods of time, giving teachers great influence over their students' development (Bronfenbrenner & Ceci, 1994).

Bowlby's (1982) attachment theory focuses on the dyadic relationships between a child and their parent, as well as between a student and their teacher. The basic idea of this theory is that the parent acts as a safe base from which the child can explore the world. Thus, the parent-child interaction forms a blueprint for later interpersonal relationships—also with teachers. Whereas the core attachment process with a primary caregiver has already been formed, children can form an attachment with multiple caregivers in preschool and elementary school

(Schaffer & Emerson, 1964). These additional attachments are usually formed with actors from Bronfenbrenner's (1979) micro-system—namely, extended family, peers, and teachers. There is a large body of literature showing the positive impact of a secure attachment style in parent-child attachment (Ainsworth et al., 2014; Ainsworth & Bell, 1970; Spruit et al., 2020) but also in teacher-student attachment (Allen et al., 2018). The approach to fostering these positive attachments is the same for the primary caregiver and teachers: responding promptly, appropriately, and consistently to the needs of the child or student (Bowlby, 1982). With a secure attachment to their teacher, a child learns a sense of security, which lets them explore the environment and form stable and strong relationships with others. In this way, positive TSRs enable students to deeply engage in learning activities, whereas a negative attachment to the teacher might hamper their learning due to a lack of security and self-esteem (for a more detailed discussion, see Roorda et al., 2011). The micro- and meso-systems hold potential correlates of TSR, as academic, behavioral, socioemotional, motivational, and general cognitive student outcomes are influenced by teacher-student interactions in the classroom (Bronfenbrenner & Morris, 2007; Wang et al., 2020).

To the best of our knowledge, two reviews have mapped the extant meta-analytic evidence on TSR-achievement relations. These two reviews included different student samples (i.e., pre-school to university students in Hattie, 2008; university students in Schneider & Preckel, 2017) used achievement as the sole outcome variable, and were based on different definitions of TSRs. In his seminal work, *visible learning*, Hattie (2008) identified five meta-analyses that quantified the relation between TSRs and student achievement and Hattie (2023) additionally reports meta-analyses under teacher-student dependency and teacher-student support, six of which were also included in the present review (Cornelius-White, 2007; Kincade et al., 2020; Roorda et al., 2011, 2021; Tao et al., 2022; Vandenbroucke et al., 2018). With effects of d = 0.72 (Cornelius-White, 2007; in Hattie, 2008), d = 0.52 (Hattie, 2015; Waack, 2018), and d = 0.47 (Corwin Visible Learning Plus, 2023), TSRs were considered to have the potential to accelerate student achievement. In this latter source, TSRs were defined as the quality of the relationships between teachers and students. This review of meta-analyses also included the meta-analytic findings by Vandenbroucke et al. (2018), who reported a weighted average effect of d = 0.18, based on 23 primary studies, 23 effect sizes, and 19,906 students. This result indicated a positive but weak TSR-achievement relation. By contrast, Cornelius-White (2007) found a substantially larger relation with a weighted average effect of d = 0.72, based on 229 primary studies, 1,450 effect sizes, and 355,325 students. These two meta-analyses exemplify the diversity of meta-analytic findings on the relation between TSRs and student achievement. To explain such discrepancies in the field, it is key to examine the characteristics of the meta-analyses as possible explanations and to map and evaluate their quality and approaches.

In their systematic ranking of 105 factors associated with achievement in higher education, Schneider and Preckel (2017) found that teacher-related and instructional variables were ranked high. Among these, TSRs were ranked the 30th most important factor. In their review, TSRs were operationalized as a "teacher's concern and respect for students; friendliness" (p. 572) and contributed 11 effect sizes.

Given their inclusion criteria and focus on student achievement as the sole outcome, the two reviews provided evidence of the TSR-outcome relation based on only six meta-analyses (five included in Hattie, 2008; one included in Schneider & Preckel, 2017). Hence, this meta-analytic evidence base did not allow robust and reliable inferences to be made about the relations between student outcomes and TSRs. Moreover, they did not map the moderator analyses of the included meta-analyses or the factors explaining heterogeneity in the TSR-outcome relations at the meta-analytic level. However, such factors could potentially clarify the varying meta-analytic findings.

5.2.3 Possible Moderators of TSR-Outcome Relations

Meta-analyses of TSR-outcome relations have suggested a wide array of possible moderators, focusing on differences in students' age and gender. For instance, several meta-analyses found significantly larger correlations for girls and younger children than for boys and older children and youth (e.g., Chu et al., 2010; Krause & Smith, 2022). These findings have been questioned in other meta-analyses that found null or even negative results (e.g., Ali et al., 2015; Givens Rolland, 2012). At the same time, the quality of meta-analyses in this field has varied considerably. For instance, Roorda et al.'s (2014) and Endedijk et al.'s (2022) meta-analyses are both rather recent studies but differ greatly in methodological quality and reproducibility, potentially explaining their divergent moderator findings. In the following section, we review the evidence on dominant moderators, namely, the measurement of TSRs (the source of information about TSRs, e.g., teachers, students, or a third party), students' age (or grade level), and gender. Furthermore, we also considered less prominent moderators, including school location (urban vs. rural) and culture (Western vs. Eastern cultures).

5.2.3.1 Measurement of TSRs

Usually, TSRs are measured either between one teacher and one student (i.e., at the dyadic level) or between one teacher and their class (i.e., at the classroom level; Sabol & Pianta, 2012; Verschueren & Koomen, 2012). The most common informants are the teachers ("Do you like this specific student?” vs. “Do you like the students in your classroom?”), students (“Do you like your teacher?” vs. “Does your teacher like the students in your classroom?”), and third-party researchers or teachers (Sabol & Pianta, 2012). These third parties could be observing classroom interactions and rating aspects of the classroom interactions (e.g., warmth, conflict).

In early childhood, dyadic relationships are almost exclusively measured from the teacher’s perspective, as students are too young to provide a valid self-report (Wentzel, 2022). For children in preschool to Grade 3, the 28-item Student-Teacher Relationship Scale (STRS; Pianta, 2001) has been the most widely applied assessment tool (Wentzel, 2022). Nonetheless, the literature shows an astonishingly high degree of disagreement between teachers and students in their perspectives on their relationships (Wentzel, 2022). Thus, the informant is an important source of variation when comparing TSR measures. For a review touching on the conceptualization and measurement of TSRs, please refer to Wentzel (2022).

5.2.3.2 Students’ Age

Both Bronfenbrenner’s (1979) and Bowlby’s (1982) theories imply that the impact of teachers changes over time as the students go through adolescence. Whereas parents used to have the closest relation with the child, and thus the most proximal processes took place with parents, the focus shifts in adolescence: Peers and probably teachers become more important in the bioecological model. There is a similar shift in attachment. Whereas the primary caregiver acted as the (only) safe haven during infancy (Schaffer & Emerson, 1964), this special bond is now partially transferred to other relationships (e.g., the TSR). Hence, relations between TSRs and student outcomes may depend on students’ age. Cornelius-White (2007) suggested that “future meta-analytic research might focus on specific subsets of learner-centered behaviors to reduce heterogeneity in synthesizing results and increase the inferential potential for future syntheses” (p. 133). Relevant subsets may represent students’ grade levels (e.g., differences between primary/elementary school vs. secondary/high/middle school). At different grade levels, students may be taught by one main teacher or multiple, equally important teachers (Quin, 2017). It might be possible that a few positive TSRs become more important in middle and high school as students’ relationships with teachers become less close (Hughes & Cao, 2018). Accordingly, Roorda et al. (2017) found that TSRs had larger relations with school engagement and achievement in older students than in younger students.

5.2.3.3 Students’ and Teachers’ Gender

Both students’ and teachers’ gender are of great interest in TSR research, as evidenced by the amount of research looking into possible differences between female and male students in their relationships with their teachers. Prior meta-analyses resulted in mixed findings and indicated that male students' TSRs are linked more closely to behavior problems than female students' TSRs (Ali et al., 2015), whereas female students' TSRs have stronger relations with negative academic emotions than male students' TSRs (Lei et al., 2018). These differences could originate from gender differences in the outcomes. For instance, male students may face more behavioral problems in school (Hamre & Pianta, 2001). The differences could also be explained by gender differences in TSRs, with female students receiving more social support (Rueger et al., 2008) and profiting more from it in terms of their well-being than male students (Chu et al., 2010). Concerning teachers, Cornelius-White (2007) found that female teachers enjoyed more success than male or mixed teacher groups in building positive relationships. Taken together with gender differences in attachment styles (Levy et al., 2011), this extant literature hints that gender potentially plays a moderating role in relations between TSRs and student outcomes.

5.2.3.4 Other Moderators

We found that several other unpreregistered variables had been examined multiple times in the meta-analytic literature. The extant literature lacks integrated information about variables potentially moderating the TSR-outcome relation at the meta-analytic level. Such information is critical for mapping existing meta-analytic evidence, evaluating its credibility, and potentially explaining heterogeneity in the effect sizes between meta-analyses (see Cooper & Koenka, 2012).

A school's location (e.g., whether a school was in a rural or urban area) was examined in three of the meta-analyses with diverging results. Other frequently researched potential moderators could also be identified through the meta-analytic literature (e.g., teaching experience and ethnicity). In addition, characteristics of the meta-analytic sample (e.g., overall meta-analytic sample size), the publication of the meta-analyses (e.g., publication status), the TSR and outcome measures included in the meta-analyses (e.g., definition of the TSR, the type of outcome, and outcome informant), and the quality of the meta-analyses (e.g., transparency of the reporting, selection of a meta-analytic baseline model) may moderate TSR-outcome relations at the level of meta-analyses. Mapping variables that might influence the strength of TSR-outcome relations also on the meta-analytic level helps researchers plan their TSR-related studies and helps educators consider these factors to improve TSRs. As this type of evidence is based on several large meta-analyses, it can uncover findings that can broaden knowledge about the factors that facilitate and hamper positive TSR development. These can be highly robust findings because, if one variable moderates TSR-outcome relations in the same way in different meta-analyses, we can infer that this moderating effect will hold for different samples and across different circumstances. Thus, a list of moderators has great value for researchers and educators alike.

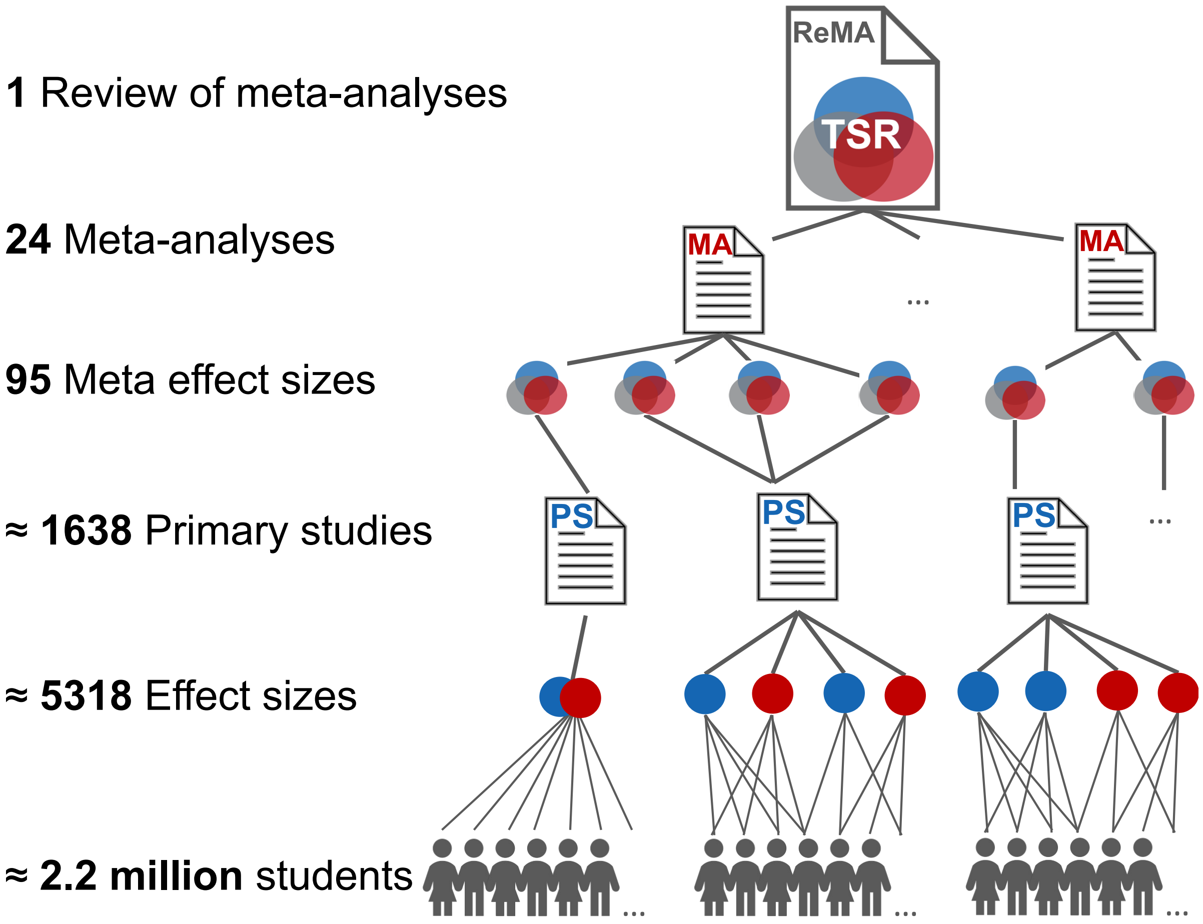
5.3 Quality of Meta-Analyses Reporting TSR-Outcome Relations

The quality and critical appraisal of primary studies and meta-analyses are crucial because educational and political stakeholders often draw on such syntheses of evidence to inform their practice (Schalken & Rietbergen, 2017). Whereas reviews of meta-analyses bring robust evidence and have high practical use, they also frequently show large differences in methodological quality in the meta-analyses in their field (as done in, e.g., Jansen et al., 2022; Schneider & Preckel, 2017). Consequently, they have also highlighted the need to consider methodological quality, reproducibility, and statistical power when evaluating the credibility and robustness of meta-analytic findings. Here, methodological quality refers to aspects of reliability, validity, and sound methodology; reproducibility can be translated into whether data and analytical code are made available; and statistical power indicates the certainty with which an effect size of interest can be detected with the sample and statistical method at hand. A critical appraisal and clear reporting of these three aspects of quality is crucial, especially in the field of TSRs, where students' school success and well-being are the outcomes.

5.4 Why Evidence From Systematic Reviews of Meta-Analyses and SOMAs Can Help Researchers Better Understand TSR-Outcome Relations

The synthesizing of research syntheses is a phenomenon that has arisen in the last 30 years (Cooper & Koenka, 2012). In parallel to the systematic review and meta-analysis, we statistically synthesized the results of the included meta-analyses narratively and statistically—thus, we conducted SOMAs. We applied the same analyses for moderators, publication bias, and study quality as in the first-order meta-analyses. Figure 4 shows the structure of the present review of meta-analyses and the SOMAs.

Figure 4. *Structure of the Data in our Review of Meta-Analyses and Second-Order Meta-Analyses*



With this combination of techniques, an overview of the TSR research field can be generated and even restructured with novel insights into relations, moderators, and the methodological quality of TSR research (see, e.g., Hattie, 2008). As we used statistical SOMAs, our results can be quantified, making them easier to compare with prior research and interpreted for use in research, teaching, and relationship building between teachers and students. All in all, there are at least four good reasons to conduct a review of meta-analyses and a SOMA in TSR research:

1. To summarize the (correlational) evidence on TSR-outcome relations that are defined by similar research questions, variables, or samples: provide mean effect sizes, compare them between different outcomes, and identify research gaps for future TSR research.
2. To identify important moderators and mediators in the TSR literature: provide a list of moderators that are of high interest in TSR research and interpret their impact on TSR-outcome relations.

3. To map the quality of existing meta-analyses in TSR research: We can explain discrepancies from differences in quality between the TSR meta-analyses and consider their contribution in light of quality indicators (Shea et al., 2007).
4. To suggest future TSR research that will close existing knowledge gaps (Cooper & Koenka, 2012): give direction with respect to which topics require more research and which TSR-outcome relations are understood well enough to direct resources elsewhere.

Prior reviews of meta-analyses and SOMAs have already done great work in mapping the educational psychology landscape (e.g., Hattie, 2008; Jansen et al., 2022; Schneider & Preckel, 2017). The present study aims to add to this work by achieving all four points listed above for the research field of TSRs.

5.5 The Present Study

In the present study, we examined the relations between TSRs and pre-K-to-12 students' academic, behavioral, socioemotional, motivational, and general cognitive outcomes in the meta-analytic literature. Specifically, we provide an overview of these relations, moderators, and methodological quality in the meta-analytic literature on TSRs, following the Meta-Analysis Reporting Standards (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008). We performed a systematic review of meta-analyses and conducted SOMAs to address the following research questions (RQs):

RQ1: According to the findings of existing meta-analyses, to what extent are TSRs related to student outcomes? (*Weighted average effect sizes*)

RQ1a: To what extent are TSRs related to academic, behavioral, socioemotional, motivational, and general cognitive outcomes?

RQ1b: To what extent are positive and negative student outcomes related to TSRs?

RQ1c: To what extent are positive and negative TSRs related to student outcomes?

RQ2: To what extent can heterogeneity in the TSR-outcome relations be explained? (*Moderator effects*)

RQ2a: Which characteristics of the primary studies (e.g., publication, sample, construct, measurement) and the meta-analyses moderate the TSR-outcome relations in the prior meta-analyses?

RQ2b: Which characteristics of the meta-analyses moderate the TSR-outcome relations in our SOMAs?

RQ3: What is the methodological quality of the meta-analyses we included? (*Quality of the meta-analyses*)

RQ3a: What are the levels of the indicators of the quality of the meta-analyses?

RQ3b: What is the statistical power of the effects reported in the meta-analyses?

RQ3c: To what extent are the meta-analytic findings on the TSR-outcome relations sensitive to the quality of the meta-analyses?

Concerning RQ1, we hypothesized that the meta-analyses would provide evidence of positive relations between TSRs and positive measures of students' academic, behavioral, socioemotional, motivational, and general cognitive outcomes (e.g., school grades, school attendance, general well-being, intelligence). Likewise, we hypothesized that the meta-analyses would provide evidence of negative relations between TSRs and negative measures of the outcomes (e.g., school dropout, aggressive behavior, bullying, sadness). Concerning RQ2, we hypothesized that the meta-analyses would provide evidence of the moderating effects of primary study characteristics, such as sample characteristics (e.g., grade level) and the measurement characteristics of TSRs. Given the limited evidence from prior research, we did not specify any hypothesis for RQ3 on the methodological quality and reproducibility of the included meta-analyses.

5.6 Method

5.6.1 Preregistration, Transparency, and Openness

We preregistered the literature search parameters, stopping rules for data collection, variables, hypotheses, and planned analyses in the Open Science Framework at https://osf.io/j2emf/?view_only=0ba8c8b41d584901a0484dcec4f64d91 [anonymized link, blinded for peer review]. Before the preregistration, we performed a scoping search to get an overview of the literature base and to check the feasibility of our systematic review of meta-analyses. In the following sections, we report the search terms, searched databases, and inclusion and exclusion criteria. We also provide the original call for meta-analyses, a preliminary codebook, a coding template with potential moderators, and the initial screening forms. The Supplemental Material contains the data (S1), the codebook (S2), and the analytical code (S3), the detailed documentation of the literature database search (S4), the back-and-forth search (S5), the gray literature search (S6), the call for meta-analyses (S7), and the updated screening forms (S8). We updated the preregistration on August 21, 2023, to report minor amendments we have made. To achieve full transparency and reproducibility, the Supplemental Material and the analytic code are accessible via the OSF at https://osf.io/6v7um/?view_only=75194e53673349f7b5b63f3bbffbe1f0 [anonymized link, blinded for peer review]. In the manuscript, we report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study.

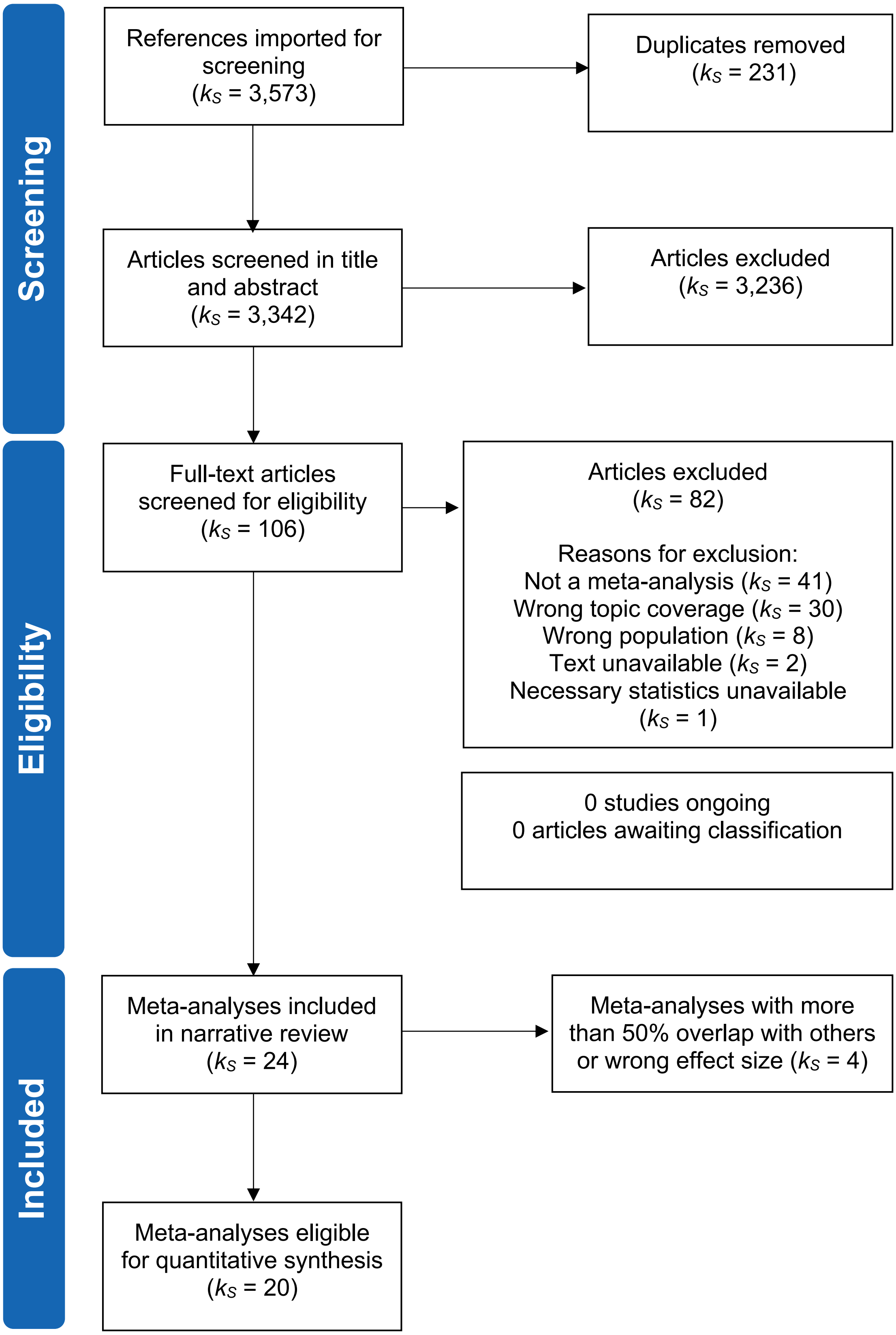
5.6.2 Literature Search

Our literature search was focused on meta-analyses that were published and unpublished (e.g., dissertations) and listed in the following databases: Education Research Complete (EBSCO), ERIC, PsycINFO, Scopus, and Web of Science. Between March 28 and April 4, 2022, a research librarian searched Titles, Abstracts, and Keywords, using the search terms in Appendix A (see Supplemental Material S4 for the exact search terms for each database). We included publications in any language, but the search terms were all in English.

To mitigate potential publication bias, we also searched for gray and unpublished literature in alternative databases with adapted search terms, and we documented the search outcomes. More specifically, we searched for gray literature in Google Scholar, EASY, PsyArXiv, and ResearchGate on March 31, 2022 (see Supplemental Material S6 for the exact search terms, number of publication entries, and results from the gray literature search. Furthermore, we contacted authors of meta-analyses in the field of TSRs and the German Psychological Society (Deutsche Gesellschaft für Psychologie, DGPs) with a 4-week deadline and a reminder after 2 weeks to elicit further (un-)published meta-analyses (see Supplemental Material S7). Finally, we hand-searched the reference lists of eight meta-analyses for additional publications (i.e., Chu et al., 2010; Cornelius-White, 2007; Kincade et al., 2020; Krause, 2020; Lei et al., 2016; Nurmi, 2012; Roorda et al., 2017; Vandenbroucke et al., 2018) in a back-and-forth search (see S5).

The main literature search yielded 4,190 publications, the gray literature search yielded an additional 9 publications, and the back-and-forth search yielded 6 more publications. We removed duplicates in Endnote (following the procedure by Bramer et al., 2016) and uploaded 3,573 publications to the software Covidence (www.covidence.org). Using Covidence, we deduplicated the publications again and submitted 3,342 publications to the Title and Abstract screening. Figure 5 summarizes the literature search, deduplication, and screening procedures for all references. Supplemental Material S9 contains a list of the excluded publications and the reasons for their exclusion.

Figure 5. *Flowchart Summarizing the Literature Search and Selection*


Screening
Eligibility
Included
References imported for screening (k_S = 3,573)
Duplicates removed (k_S = 231)
Articles screened in title and abstract (k_S = 3,342)
Articles excluded (k_S = 3,236)
Full-text articles screened for eligibility (k_S = 106)
Articles excluded (k_S = 82)
Reasons for exclusion:
Not a meta-analysis (k_S = 41)
Wrong topic coverage (k_S = 30)
Wrong population (k_S = 8)
Text unavailable (k_S = 2)
Necessary statistics unavailable (k_S = 1)
0 studies ongoing
0 articles awaiting classification
Meta-analyses included in narrative review (k_S = 24)
Meta-analyses with more than 50% overlap with others or wrong effect size (k_S = 4)
Meta-analyses eligible for quantitative synthesis (k_S = 20)

Note. k_S = Number of studies reporting meta-analyses.

5.6.3 Screening and Selecting Meta-Analyses

The screening procedure comprised two steps—an initial screening of Titles and Abstracts and a full-text screening. In the initial screening, the first author and a research assistant double-screened 689 out of 3,404 references (approximately 20% of the studies) and achieved 99% agreement with Cohen's κ = .77. Disagreements were resolved in a discussion. Next, we conducted the full-text screening. We double-screened 34 out of 109 references (approximately 31%) and achieved 91% agreement with Cohen's κ = .72. Supplemental Material S10 presents the exact agreement and Cohen's κ values for both screening steps. The numbers of references dealt with in these steps were larger than those depicted in Figure 5 because more duplicates were identified and removed during the screening. Following the two screening steps, we included all studies that fulfilled the following four criteria:

1. The meta-analysis conducted a systematic search and a quantitative/statistical analysis of the included primary studies.
2. The meta-analysis reported at least one weighted average effect size of the relation between at least one of the aspects of TSRs (definition as used in Roorda et al., 2017) and students' academic, behavioral, socioemotional, motivational, and general cognitive outcomes.
3. At least 75% of the samples in the meta-analysis consisted of young people enrolled in formal education. More specifically, we included children in prekindergarten, kindergarten, preschool, primary/elementary school, middle school, and high school (i.e., pre-K-to-12 students).
4. At least 75% of the sample in the meta-analysis were reported healthy and not diagnosed with a disorder or medical condition.

Furthermore, we excluded meta-analyses if:

1. Participants were not human.
2. The Abstract, full text, or secondary sources reporting the results of the study were not available.
3. The publication narratively synthesized research results, not providing any weighted average effect size of the intended relation (i.e., qualitative or narrative reviews).
4. TSRs were mostly operationalized as administrative or organizational (e.g., giving away lunch vouchers or providing classroom organization) rather than affective.

5. An updated and more comprehensive meta-analysis existed. In this case, we included the updated meta-analysis in the quantitative analyses, but all meta-analyses were retained in the qualitative synthesis. Meta-analyses with more primary studies were included instead of meta-analyses with fewer studies.
6. An article synthesized international large-scale assessments of different countries rather than separate studies. This criterium was included to comply with the definition of meta-analyses above. Although these studies may be eligible for individual-participant meta-analyses (see Campos et al., 2023), they often do not report the same statistics as meta-analyses and are not based on systematic literature searches or screening steps.
7. The article was a review of meta-analyses (e.g., Hattie, 2008). In these cases, only the relevant meta-analyses reported therein were included to avoid giving more weight to meta-analyses that were included both individually and in a review of meta-analyses.

A total of 24 meta-analyses were eligible for the narrative synthesis.

5.6.4 Overlap Between Meta-Analyses

After screening the meta-analyses, we reviewed their characteristics and assessed the overlap in included primary studies. First, we checked the overlap in primary studies of the included meta-analysis with the help of a cross-table in Appendix B (Pieper et al., 2014). We considered two meta-analyses with an overlap of less than 50% in primary studies to be distinct enough to quantitatively synthesize the TSR-outcome relation (see Cooper & Koenka, 2012). When two meta-analyses overlapped in 50% or more of the primary studies, we described their results narratively but did not use quantitative, inferential analyses. When a student outcome was represented by only a single meta-analytic effect, we also excluded it from any quantitative, inferential analyses. Second, in line with Pieper et al. (2014), we further considered similarities in author teams. Two meta-analyses that had a study overlap of over 50% and were authored by the same team were removed from the quantitative synthesis, and we narratively indicated the lack of independence in author teams in the narrative review sections. Third, similar to Jansen et al. (2022), we quantified the extents to which the meta-analyses were up-to-date by showing the publication range of the primary studies and meta-analyses.

Specifically, out of the 24 eligible meta-analyses, two clusters overlapped by more than 50% of the included primary studies. A total of 54% of Vandenbroucke et al.'s (2018) meta-analytic sample was contained in the meta-analysis by Li et al. (2021). Similarly, Roorda et al.'s (2017) update encompassed over 80% of the studies included in both their prior meta-analyses (i.e., Roorda et al., 2011, 2014). These earlier studies were excluded from quantitative synthesis but were retained for the narrative synthesis. Later in the process, we also excluded the meta-analysis by Moore et al. (2019), who reported tau-U, which could not be meaningfully transformed into Pearson's *r* correlations. Overall, 20 meta-analyses with $k_{ES} = 79$ effect sizes were eligible for the quantitative synthesis, and 24 meta-analyses with $k_{ES} = 116$ were submitted to the coding for the narrative synthesis.

5.6.5 Coding of Meta-Analyses

The first and the third authors independently extracted and coded the data from the full texts. Several meta-analyses did not report crucial information, such as the sample sizes, age, or other sample characteristics. In such cases, we contacted the authors and asked them to provide the missing information. After 2-6 weeks and a reminder, most authors kindly provided the missing information. In addition to the statistical relation between TSRs and outcome variables, the characteristics of the publication, sample, measurement, and meta-analysis were coded. Both coders were trained to use the coding scheme on six randomly chosen meta-analyses (25% of the 24 included references). With a new set of six meta-analyses, the two coders reached an interrater agreement of 93%. Disagreements were resolved through discussion. The coding of the methodological quality reached an agreement of 93%. In the following, we present the coding of the key variables in our systematic review. The coded data are contained in Supplemental Material S1. All coded variables, their respective categories, and examples are detailed in the Codebook in Supplemental Material S2.

5.6.5.1 Characteristics of the Meta-Analyses

For the characteristics of the meta-analyses, we coded the literary reference, publication year, publication type (journal article, conference paper, dissertation, or preprint), publication status (published study vs. gray literature), institutional affiliation of the corresponding author, effect size type (e.g., Cohen's *d*, Pearson correlation *r*), weighted average effect size(s), quality of evidence rating score (assessed with the AMSTAR as explained below; see Shea et al., 2007), the meta-analytic model used to generate the weighted average effect size (e.g., multilevel random-effects meta-analysis), and whether publication bias was analyzed (yes/no).

5.6.5.2 Characteristics of the Included Primary Studies and Samples

For the sample characteristics, we coded the gender composition (percentage of female students in the sample), the range of the publication years of the primary studies, study designs (correlational, experimental, longitudinal, interventional), the number of effect sizes from the primary studies included in the meta-analysis, the number of students/teachers/classrooms included, students' average age (in years), age range (in years), students' aptitude (special education, at-risk of low IQ, average or not specified, high IQ, intellectually gifted), whether students were reported to be part of an ethnic minority (yes/no) or of a social minority (e.g., being LGBTQ, disabled, or having a migration history; yes/no), and the grade level(s) of the sample(s) (i.e., prekindergarten, kindergarten, preschool, primary/elementary school, middle school, high school).

5.6.5.3 Measurement Characteristics

For the measurement characteristics of TSRs and the outcome variable(s), we coded the description of the TSR assessment, the level of the TSR assessment (dyadic, classroom level, both, or other), the modality of the TSR (positive vs. negative aspects), the facet of the TSR (e.g., teacher liking, teacher friendliness, and others), the informant on TSRs used in most primary studies (parents, teachers, children, peers, expert observers, a combination of these options, or other), the description of the students' outcome variable(s), the description of the measurement of the students' outcome variable(s), the level of the outcome assessment (dyadic, classroom level, mixed, or others), the type of outcome (academic, behavioral, socioemotional, motivational, and general cognitive outcome), the informant(s) of the outcome variable(s) (parents, teachers, children, peers, expert observers, a combination of these options, or other), the informant used in most primary studies (parents, teachers, children, peers, expert observers, a combination of these options, or other), and which moderators were found to be significant and nonsignificant.

To examine the relations between TSRs and the academic, behavioral, socioemotional, motivational, and general cognitive outcomes, we combined the outcome variables into clusters post hoc. With this approach, we wanted to map and categorize the types of outcomes examined in the meta-analyses while retaining the authors' definitions of the respective constructs. Inspired by Chu et al.'s (2010) and Wang et al.'s (2020) conceptualizations, we combined effect sizes by pooling similar constructs into nine outcome clusters: (a) Academic Achievement, (b) Academic Emotions, (c) Appropriate Behavior, (d) Behavior Problems (recoded), (e) Bullying (recoded), (f) Executive functions and Self-Control, (g) Motivation, (h) School Belonging and Engagement, and (i) Well-Being.

In addition to these clusters, we compared student outcomes that were originally assessed as positive or negative outcomes (recoded). Here, we distinguished between desirable or beneficial outcomes (e.g., positive school achievement and emotions) and adverse or undesirable outcomes (e.g., bullying and behavior problems). Similarly, we distinguished between positive TSRs, which are characterized by warmth and closeness, and negative TSRs (recoded), which are characterized by conflict and dependency. These clusters are described in greater detail and illustrated in Appendix C. In three cases, constructs appeared only once, could not be assigned to a cluster, and were thus omitted. The data set in Supplemental Material S1 presents the full list of choices we made to combine student outcomes and TSRs into the clusters above.

5.6.5.4 Weighted Average Effect Sizes

For the statistical indicators, we extracted the weighted average Pearson correlations $\bar{r}$ and their respective standard errors if they were available. If Pearson correlations were not available, we extracted alternative effect sizes (e.g., Cohen's *d*) and transformed them into Pearson's *r*, using the formulas described in Borenstein et al. (2009). We coded multiple weighted average effect sizes per meta-analysis if they were based on either different samples or different constructs. If two effect sizes were reported for the same sample and the same constructs, we retained the one obtained from the random-effects model rather than the fixed-effects model to account for possible heterogeneity in the meta-analytic data. The 24 meta-analyses provided 116 effect sizes for a total sample of over 2 million pre-K-to-12 students.

5.6.5.5 Overall Sample Sizes in the Meta-Analyses

Several meta-analyses reported the number of participating students only at the level of the meta-analysis and not at the level of the individual effect size (Chu et al., 2010; Givens Rolland, 2012; Li et al., 2021; Roorda et al., 2021; Wang et al., 2020). In these cases, we estimated the number of students relevant to our research question as follows: We divided the total sample by the number of all primary effect sizes and multiplied it by the number of primary effect sizes relevant to our research question. When meta-analyses also failed to provide the total number of participating students (i.e., Korpershoek et al., 2016; Wilkinson, 1980), we estimated the sample size as the average number of students per effect size across all included meta-analyses. The average number of students per meta-analytic effect sizes was over 27,000 when the estimate was computed in this way.

5.6.5.6 Quality of the Meta-Analyses

Similar to the meta-analyses of primary studies, the critical appraisal of systematic reviews and meta-analyses of meta-analyses is a key step in evaluating the quality and credibility of the evidence (Johnson, 2021). To assess the quality of evidence from the meta-analyses, we coded a total of 16 items. We adapted these items from the 11-item AMSTAR rating scale (Shea et al., 2007) and added two items about the availability of the statistical syntax and the data sets to capture reproducibility and three items on further aspects of methodological quality. The latter three items pertained to the statistical procedure the authors used to pool effect sizes, whether moderator analyses were performed, and which kind of publication bias analyses were performed in the meta-analyses (e.g., Jansen et al., 2022). All items could be answered with either yes or no and were not weighted. Moreover, we did not weight the effect sizes by primary study quality (e.g., Tod et al., 2022; Wedderhoff & Bosnjak, 2020). Given the correlational design of the meta-analyses, we used the AMSTAR rather than AMSTAR 2 checklist due to its good psychometric quality (Shea et al., 2009).

5.6.6 Analysis of the Meta-Analytic Data

As noted earlier, we synthesized the meta-analytic effect sizes quantitatively for the subset of 20 meta-analyses, whereas we synthesized all 24 meta-analyses narratively. Except for RQ2a and RQ3a, we addressed all other research questions quantitatively. Following model testing and selection, we conducted SOMAs to examine RQ1 (weighted average effect sizes) and RQ2b (moderator effects). To address RQ3a, we narratively reviewed the quality indicators of the meta-analyses and mapped the resultant levels of methodological quality. To address RQ3b, we calculated the statistical power of the reported, meta-analytic effects. Finally, to address RQ3c, we examined the possible moderating effects of methodological quality. The data set and the analytic code are stored in the Supplemental Materials S1 and S3. Parts of the analysis scripts were adapted from Emslander and Scherer (2022), Jansen et al. (2022), and Quintana (2023).

5.6.6.1 Publication Bias

Publication bias describes the phenomenon in psychological research that nonsignificant results are less likely to be published (Egger et al., 1997; Ferguson & Heene, 2012). To avoid replicating selectively published findings in the meta-analysis, we assessed publication bias at the level of the included meta-analyses instead of the primary studies. Specifically, we graphically inspected all 79 meta-analytic effect sizes using contour-enhanced funnel plots. These plots show all meta-analytic effect sizes and their standard errors centered around zero. Different shades of color indicate their level of statistical significance (Peters et al., 2008). If publication bias is present, then the distribution of effect sizes should be skewed to either side at the base of the funnel. In the absence of publication bias, the spread of the effect sizes should be roughly symmetrical, with effect sizes with larger standard errors evenly distributed around the base of the funnel (Egger et al., 1997).

5.6.6.2 SOMAs

Whereas Cooper and Koenka (2012) found three effective ways to aggregate findings from reviews and meta-analyses—namely, to focus on (a) the differences between reviews; (b) the quantitative aggregation of meta-analytic effect sizes in a SOMA; or (c) the quantitative aggregation of primary effect sizes in a new meta-analysis (see Polanin et al., 2017)—we used only their first and second approaches. Specifically, we described possible differences between meta-analyses narratively and synthesized the reported weighted average effect sizes quantitatively. To this end, we conducted a series of SOMAs and moderator analyses.

To address RQ1, we first conducted a SOMA with the $k_{\mathrm{ES}} = 79$ eligible effect sizes to estimate an overall, second-order effect. Next, we conducted multiple SOMAs with subsets of the meta-analytic data: (a) positive versus negative TSR variables; (b) positive versus negative student outcomes; and (c) outcome clusters with more than one effect size representing academic achievement, academic emotions, appropriate behavior, behavior problems, bullying, executive functions and self-control, motivation, school belonging and engagement, and well-being. To address RQ2, we narratively described important moderators in the included meta-analyses (RQ2a) and conducted a moderator analysis that was based on the SOMA baseline model (RQ2b). To address RQ3, we narratively reviewed the methodological quality and statistical power at the level of the included meta-analyses (RQ3a, RQ3b) and estimated moderator effects (RQ3c). We used the R packages “metafor” (Viechtbauer, 2010) and “dmetar” (Harrer et al., 2019) to conduct the SOMAs.

In contrast to meta-analyses of primary studies, where the number of participants can be used to calculate standard errors (Borenstein et al., 2009), we used the standard errors of the weighted average effect sizes reported in the meta-analyses instead. In a SOMA, this standard error contains both the sampling variances and the heterogeneity between or within primary studies in a meta-analysis (Schmidt & Oh, 2013). Using the weighted average effect sizes ($\bar{r}_k$) and their standard errors, we estimated a second-order meta-analytic correlation between TSRs and student outcomes ($\bar{\bar{r}}$) with inverse-variance weighting.

Given the dependencies between the multiple correlations reported in a meta-analysis, we assumed that a multilevel data structure was needed to describe the variation within and between the meta-analyses (e.g., Van den Noortgate et al., 2013). To select an appropriate meta-analytic baseline model, we compared several random-effects models and a standard fixed-effects model. For the entire data set of meta-analytic effect sizes, we assumed (a) a three-level, hierarchical data structure with variances at the levels of the primary studies, within meta-analyses, and between meta-analyses; and (b) correlated effects, because multiple effect sizes were based not only on multiple samples but also on multiple outcome measures. As a consequence, our hypothesized baseline model was a three-level random-effects model with correlated effects (CHE model; see Pustejovsky & Tipton, 2021). Given the small number of effect sizes, we assumed a constant correlation between effects and examined the sensitivity of the SOMAs to the choice of this correlation. When subsetting the data to meta-analyses that focused on specific outcomes, we re-evaluated the baseline model again and adjusted it if necessary. Hence, the SOMAs were not based on a single, unifying model.

To evaluate the models, we compared their information criteria—including Akaike's Information Criterion (AIC), the Bayesian Information Criterion (BIC), and the corrected AIC (AICc; Cavanaugh, 1997)—and performed likelihood-ratio tests. Smaller values of the information criteria indicated a preference for a model. Variances were estimated via the confidence intervals and standard errors obtained through restricted maximum-likelihood estimation with the R-Package "metafor" (Viechtbauer, 2010). To further examine the sensitivity of the SOMAs to the choice of standard error, we obtained cluster-robust standard errors using the R package "clubSandwich" (Pustejovsky, 2022), and we compare their results in Supplemental Material S11.

Given the small number of effect sizes, we assumed a constant correlation between effects and examined the sensitivity of the SOMAs to the choice of this correlation.

To evaluate the sizes of the weighted average effects extracted from the meta-analyses, we considered *r* = .10 to indicate a small effect, *r* = .20 a medium effect, and *r* = .30 a large effect (Funder & Ozer, 2019; Gignac & Szodorai, 2016; Paterson et al., 2016). We did not apply Cohen's (1988) benchmarks (*r* = .10 as small, *r* = .30 as medium, and *r* = .50 as large effects), because these effects are considered too large for the science of psychology (Richard et al., 2003). Instead, we drew on the abovementioned recommendations that are specific to the field of psychology.

We calculated Cochran's *Q* to examine whether the weighted average effect sizes were heterogeneous across meta-analyses. A statistically significant *Q* value indicates heterogeneity within the distribution of effect sizes, whereas a statistically nonsignificant *Q* value indicates homogeneity (Ellis, 2010). Furthermore, we calculated the I^2 indices for the heterogeneity estimates within ($\hat{\tau}^2_{(2)}$) and between meta-analyses ($\hat{\tau}^2_{(3)}$). The I^2 index indicates the proportion of variance introduced through heterogeneity above and beyond the standard error provided by the meta-analyses. It can be categorized as low (25%), moderate (50%), or high (75%) heterogeneity (Higgins et al., 2003). Cheung (2014) defined the level-specific I^2 indices for a three-level random-effects model as follows:

$$I^2_{(2)} = \frac{\hat{\tau}^2_{(2)}}{\hat{\tau}^2_{(2)} + \hat{\tau}^2_{(3)} + \tilde{v}} \quad \text{and} \quad I^2_{(3)} = \frac{\hat{\tau}^2_{(3)}}{\hat{\tau}^2_{(2)} + \hat{\tau}^2_{(3)} + \tilde{v}},$$

where $\tilde{v}$ is the variance component captured by the standard error of the effect size in the meta-analysis.

5.6.6.3 Moderator Analyses

To address RQ2a, we narratively reviewed all moderators that were assessed at least twice in the meta-analytic literature. To address RQ2b and RQ3c, we conducted moderator analyses for the publication, sample, measurement, and quality characteristics. We extended the three-level random-effects model (i.e., the SOMA baseline model) to a mixed-effects (meta-regression) model by adding the potential moderating variables. To facilitate the interpretation of some moderator effects, we z-transformed the students' average age, mean-centered several continuous moderators, arcsine-transformed proportions (e.g., the percentage of female students in the sample to represent the gender composition; see Schwarzer et al., 2019), and dummy-coded categorical and ordinal moderators with more than two categories (e.g., school status) into multiple binary variables. Finally, we quantified the variance explained by the moderators as the proportional reduction in the level-specific heterogeneity estimates ($R^2_{(2)}$ and $R^2_{(3)}$, respectively) when comparing the model with the moderators with the model without the moderators (see Cheung, 2014).

5.6.6.4 Evaluation of Methodological Quality and Reproducibility

To address RQ3, we evaluated the methodological quality, reproducibility, and statistical power of the included meta-analyses. More specifically, to assess the validity of the meta-analyses and to identify potential issues of quality, we used 16 items partly adapted from the AMSTAR checklist (a measurement tool for the assessment of multiple systematic reviews; Shea et al., 2007). We reviewed them narratively (RQ3a) and entered all the items separately into the moderator analyses. We also added them together to form a sum score representing overall study quality (RQ3c). A critical discussion of this approach can be found elsewhere (e.g., Wedderhoff & Bosnjak, 2020).

5.6.6.5 Calculating the Statistical Power of the Meta-Analyses

To complement our selection of methodological quality indicators, we calculated the statistical power of each included meta-analysis with the R package "metameta" (Quintana, 2023). To further supplement the visual assessment of statistical power, we constructed a fireplot to indicate the statistical power for all included meta-analyses on a range of true effect sizes and the respective mean effect sizes found by the meta-analyses (RQ3b). We constructed the fireplot at the level of the meta-analytic effect sizes with the R package "metaviz" (Kossmeier et al., 2020b, 2020a; Quintana, 2020).

5.7 Results

5.7.1 Description of the Included Meta-Analyses

Table 1 provides an overview of the included meta-analyses. The design, sample, and measurement characteristics of the 24 included meta-analyses are described in the data in Supplemental Material S1. Supplemental Material S12 shows the abstracts of the included meta-analyses. Overall, the present sample included 24 meta-analyses and 116 effect sizes. Twenty-one of the 24 meta-analyses reported multiple effect sizes, ranging from 1 to 1,475 effect sizes per meta-analysis. The effect sizes stemmed from 22 published journal articles and two gray literature doctoral dissertations (Cherne, 2008; Wilkinson, 1980). Eleven meta-analyses were first-authored by researchers from institutions in the US and Canada, whereas eight of the meta-analyses originated from Europe, four from Asia, and one from Australia. All meta-analyses drew on student samples with average IQ. None of the meta-analyses explicitly focused on ethnic minorities, and two included social minorities. More than half of the effect sizes (59%) were based on the measurements of TSRs on the dyadic level, whereas 31% used measures on the classroom level, and 10% drew on both. Most meta-analyses focused on positive rather than negative student outcomes (17 with $k_{ES} = 54$ vs. seven with $k_{ES} = 25$) and positive rather than negative TSRs (18 with $k_{ES} = 54$ vs. six with $k_{ES} = 25$).

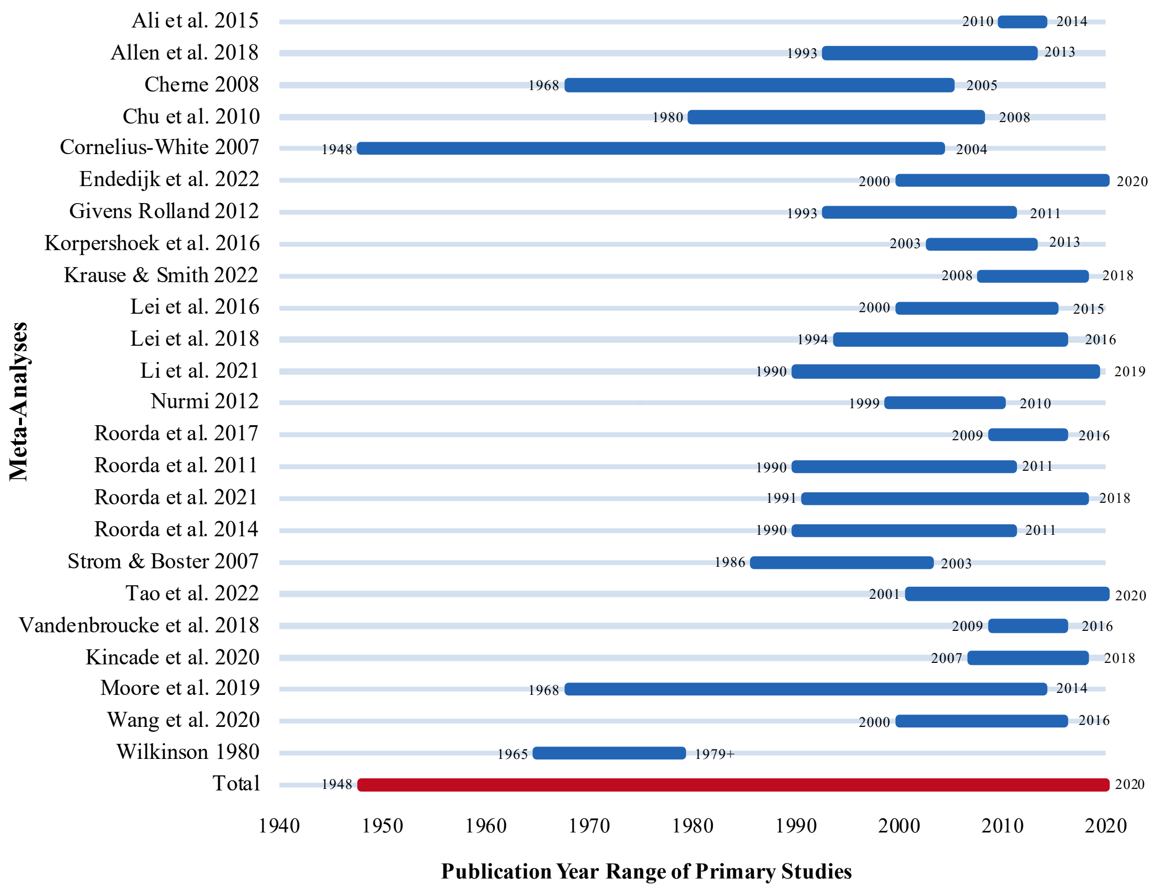
On average, the included meta-analyses drew on 231 primary studies, ranging from 16 to 1,475 (see Table 2). Our review of meta-analyses included more than 2 million students from prekindergarten to high school. The students were mostly between 3 and 18 years of age, and about half of them were girls and young women (*M* = 49.1%, *SD* = 4.5%). The included meta-analyses were published between 1980 and 2022 with an increase since 2007 (see Figure S13 in the Supplemental Material). These meta-analyses drew on primary studies published between 1948 and 2020. Thus, the publication range was 42 years for the meta-analyses and 72 years for the primary studies. Figure 6 shows the publication year range of the included meta-analyses.

Table 2. *Continuous Characteristics of the Included Meta-Analyses*

Continuous moderator	k_{MA}	k_{ES}	*M*	*SD*	*Mdn*	*Min*	*Max*
Descriptive results across all meta-analyses							
Publication year	24	95	2014.42	8,58	2016.50	1980	2022
Age (mean in years)	6	5	11.16	3.61	12.04	5.20	15.00
Age (*SD* in years)	2	2	2.79	1.74	2.79	1.56	4.02
Age range (lower bound)	9	12	4.67	3.43	3.00	2.00	12.00
Age range (upper bound)	9	12	15.94	3.34	18.00	11.00	20.00
Gender composition	8	19	49.05	4.47	50.00	42.00	54.00
Number of primary studies	24	90	11.67	6.83	11.50	1	32
Number of primary effect sizes	23	91	231.22	347.09	107.00	16	1475
Estimated sample size	23	86	97,597	148,095	57,798	382	651,014
Sample size	22	65	101,269	150,504	58,083	382	651,014
Total quality score	24	95	7.29	2.39	7.00	3.00	12.00
Descriptive results across all meta-analytic effect sizes							
Publication year	20	79	2013.25	8.00	2012	1980	2022
Age (mean in years)	3	5	11.21	1.80	11.07	8.95	13.00
Age (*SD* in years)	1	2	4.15	0.19	4.15	4.02	4.29
Age range (lower bound)	6	12	5.44	2.96	4.55	3.00	12.00
Age range (upper bound)	6	12	15.38	3.01	17.25	12.00	18.00
Gender composition	6	15	48.31	4.26	50.00	41.97	54.00
Number of primary studies	19	78	20.86	41.17	7.00	1	297
Number of primary effect sizes	20	79	57.95	199.62	8.00	2	1475
Estimated sample size	19	74	28,833	89,039	3,319	2	651,014
Sample size	15	58	34,879	99,836	3,157	2	651,014
Total quality score	20	79	7.01	1.91	6.00	3.00	12.00

Note. Publication year = the years in which the meta-analysis was published (not the primary studies therein; Age range (lower/upper bound) = the quantifies the highest and lowest end of the age range of the sample; k_{MA} = Number of included meta-analyses; k_{ES} = Number of meta-analytic effect sizes; Gender composition = Percentage of girls in the sample; Estimated sample size = average of students per effect size across all included meta-analyses.

Figure 6. *Publication Year Range of the Primary Studies in the Included Meta-Analyses*



5.7.2 Preliminary Analyses

5.7.2.1 Model Selection

As noted earlier, we first selected a baseline model for the entire data set by estimating and comparing several random-effects models and a fixed-effects model. The random-effects models included (a) a three-level random-effects model assuming only hierarchical effects, (b) a three-level random-effects model assuming a constant correlation of 0.5 between effects, and (c) a standard random-effects model ignoring the effect size multiplicity in the meta-analyses. Concerning (b), we chose a moderate average correlation between effects of 0.5 to capture both substantially and weakly correlated measures. As part of our sensitivity analyses, we further examined the impact of this choice on the pooled correlation in Supplemental Material S14. Table 3 presents the fit statistics for all statistical models. Overall, we found a preference for the three-level random-effects but chose the three-level random-effects model with correlated effects as the baseline—due to its theoretical implications—for further moderator analyses that were based on the full data set.

When subsetting the data by outcome cluster, the meta-analytic effects were likely hierarchical and no longer correlated. Hence, we evaluated (a) a three-level random-effects model with hierarchical effects, (b) a standard random-effects model, and (c) a fixed-effects model as the potential baseline model. Overall, we found that a standard random-effects model fit the data best for the outcome clusters of academic achievement, academic emotions, appropriate behavior, behavior problems (recoded to represent a positive correlation), motivation, school belonging and engagement, and well-being. A fixed-effects model fit the data best for the clusters of bullying (recoded) and executive functions and self-control, each with only two effect sizes from the same study. For the clusters of negative and positive student outcomes and for positive and negative TSRs, we found a preference for the three-level random-effects model but chose the three-level random-effects model with correlated effects as the preferred model for its theoretical implications. The model results were robust against different constant sampling correlations (Supplemental Material S14).

Table 3. *Information Criteria and Log-Likelihood Values of the Meta-Analytic Models*

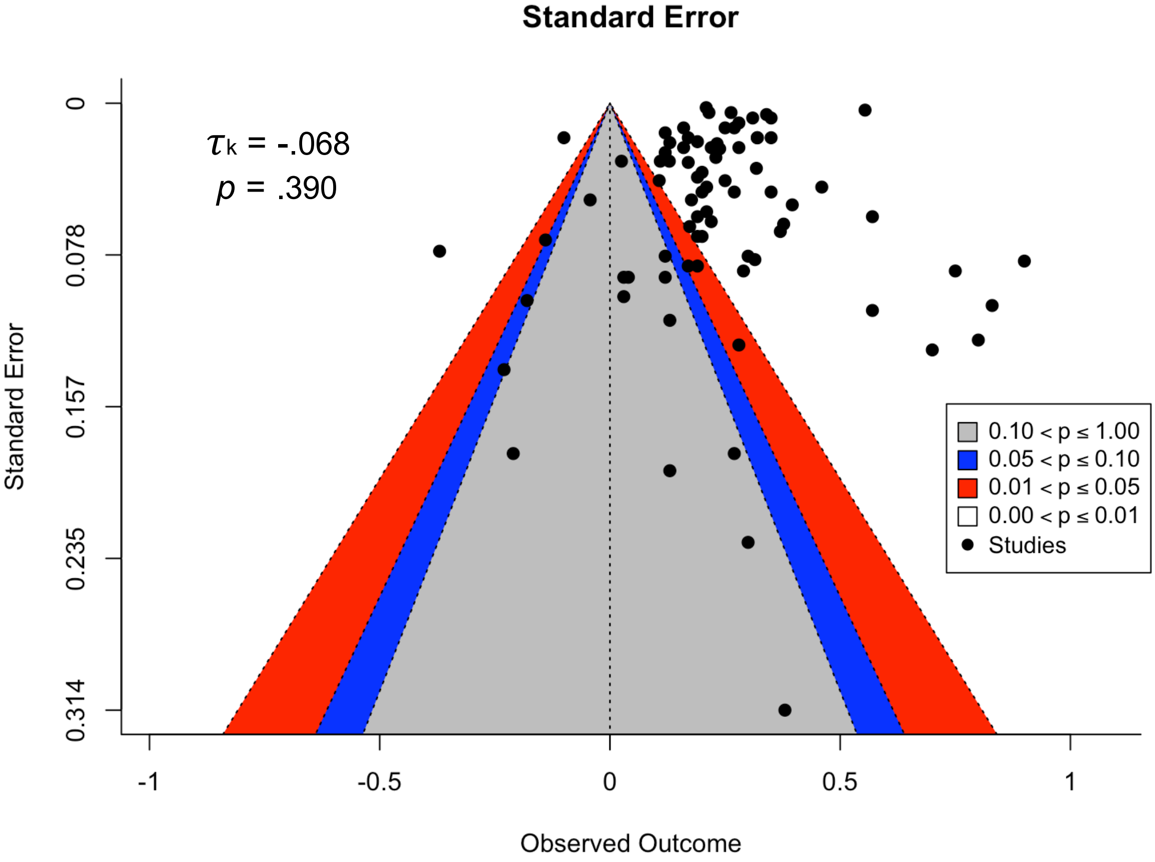
Variable	df	LogLik	AIC	BIC	AICc
Cross-classified four-level random-effects model with the hierarchical effects, $\tau^2_{(2)}$, $\tau^2_{(3)}$, and $\tau^2_{(4)}$ freely estimated	4	23.15	-38.31	-28.88	-37.76
Three-level random-effects model with hierarchical and correlated effects, $\tau^2_{(2)}$ and $\tau^2_{(3)}$ freely estimated, $r = \rho = .5$	3	21.16	-36.33	-29.26	-36.00
Three-level random-effects model with hierarchical effects, $\tau^2_{(2)}$ and $\tau^2_{(3)}$ freely estimated	3	23.15	-40.31	-33.24	-39.98
Standard random-effects model, $\tau^2_{(2)} = 0$ and $\tau^2_{(3)}$ freely estimated	2	10.17	-16.35	-11.63	-16.19
Fixed-effects model, $\tau^2_{(2)} = 0$ and $\tau^2_{(3)} = 0$	1	-4183.48	8368.95	8371.32	8369.00

Note. LogLik = Value of the Log-Likelihood; df = Degrees of freedom, AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; AICc = corrected Akaike Information Criterion.

5.7.2.2 Publication Bias Analysis

Figure 7 shows the contour-enhanced funnel plot for the full data set. Contour-enhanced funnel plots for all SOMAs are presented in Supplemental Material S15. The moderator analyses suggested no association between the sample sizes and the effect sizes. The full data set yielded a nonsignificant Kendall's τ of -.07 (p = .390; see Figure 7). Hence, the funnel plot did not exhibit substantial asymmetry. Moreover, the precision-effect test and precision-effect estimate with standard errors (PET-PEESE) indicated that there was no significant association between effect size and precision with B = -0.9, SE = 0.6, p = .144 and B = -3.5, SE = 2.5, p = .158, respectively. Overall, this evidence did not uncover substantial publication bias, and we therefore did not adjust the effect sizes for publication bias in subsequent analyses.

Figure 7. *Contour-Enhanced Funnel Plot of the Meta-Analytic Effect Sizes*

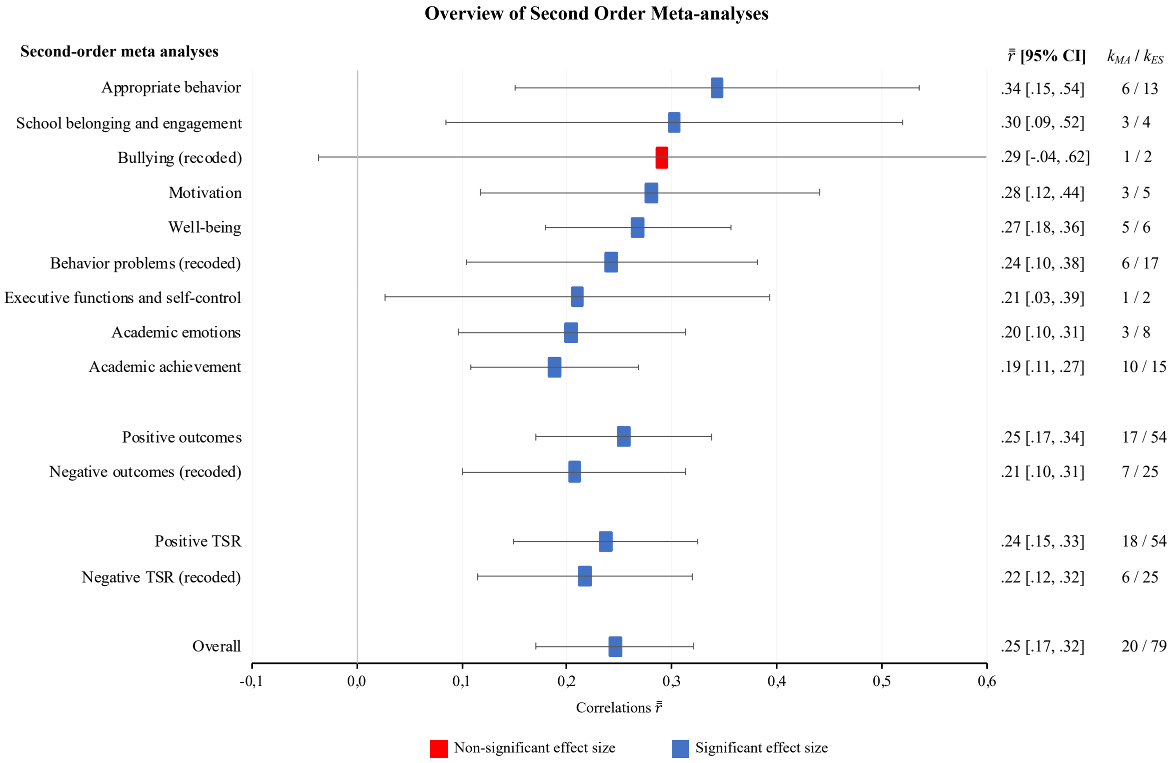


Note. Larger positive effect sizes indicate closer relations between TSRs and student outcomes. Correlation coefficients on the x-axis are plotted against the standard errors on the y-axis for every effect size. τ_K = Kendall's τ value from Begg's test.

5.7.3 Main Analysis: Pooled Correlations (RQ1)

To address RQ1, we conducted 14 SOMAs (see Figure 8)—one overall effect size for the relation between TSRs and all student outcomes, nine effect sizes for the student outcome clusters (RQ1a), two effect sizes for the positive and negative outcomes (RQ1b), and two effect sizes for positive and negative TSRs (RQ1c). Table 4 provides the respective coefficients of the 14 pooled TSR-outcome correlations, and Appendix C describes and illustrates these outcome categories. Figure S16 in the Supplemental Material presents the respective forest plots. For the SOMAs, all the expected correlations are positive to make the results more comparable, as explained above. Consequently, the correlations between positive TSRs and negative student outcomes are positive. In the narrative description, outside the SOMA context, we report the meta-analytic effect sizes in their original direction. Overall, we found an average correlation of $\bar{\bar{r}} = .25$ (95% CI [.17, .32]) for the relations between TSRs and all student outcomes (see Figure 8).

Figure 8. *Overview of the SOMAs on the TSR-Outcome Relations*



Note. The second-order meta-analytic effect sizes are distributed from largest to smallest. All effect sizes we expected to be negative were recoded to be positive (i.e., behavior problems, bullying, negative outcomes, negative TSR) to render them comparable. Thus, a positive effect size for a negative outcome (e.g., behavior problems) represents a negative correlation with TSRs. The total effect size at the bottom represents the positive weighted average of all included effect sizes. Generally, larger effect sizes signify stronger TSR-outcome relations. k_{MA} = Number of included meta-analyses; k_{ES} = Number of effect sizes; $\bar{\bar{r}}$ = Weighted average correlation.

In addressing RQ1a, the SOMAs yielded several moderate, positive, and statistically significant average correlations between TSRs and the academic, behavioral, socioemotional, motivational, and general cognitive outcomes in prekindergarten and K-12 students ranging from $\bar{\bar{r}}$ = .19 to .34. More specifically, we found positive average correlations for appropriate behavior (recoded, $\bar{\bar{r}}$ = .34, 95% CI [.15, .53]), school belonging and engagement ($\bar{\bar{r}}$ = .30, 95% CI [.09, .52]), motivation ($\bar{\bar{r}}$ = .28, 95% CI [.12, .44]), general well-being ($\bar{\bar{r}}$ = .27, 95% CI [.18, .36]), behavior problems (recoded, $\bar{\bar{r}}$ = .24, 95% CI [.10, .38]), executive functions and self-control ($\bar{\bar{r}}$ = .21, 95% CI [.18, .24]), academic emotions ($\bar{\bar{r}}$ = .20, 95% CI [.10, .31]), and academic achievement ($\bar{\bar{r}}$ = .19, 95% CI [.11, .27]). Furthermore, we found moderate average correlations for bullying (recoded, $\bar{\bar{r}}$ = .29, 95% CI [-.04, .62]), which just missed statistical significance (p = .057).

To further examine whether these pooled correlations varied across outcome clusters, we took an additional step: We extended the baseline model to a cross-classified random-effects model by adding the level of outcome clusters and removing the assumption of correlated effects. This model did not show a better fit to the data than the baseline model (see Table 3) and did not exhibit variation between outcomes ($\tau^2_{(4)}$ = .000, 95% CI [.00, .00]), which were therefore not significantly different (RQ1a: Q_M[10, 68] = 0.395, *p* = .944).

Table 4*. Results of the SOMAs of the TSR-Outcome Relations*

Clusters	k_{MA}	k_{ES}	$\bar{\bar{r}}$	95% CI	*SE*	*t*	*p*	$\tau^2_{(2)}$	$\tau^2_{(3)}$	*Q*	$I^2_{(2)}$	$I^2_{(3)}$
Academic Achievement	10	15	**.188**	[.11, .27]	.037	5.05	**<.001**	.019	-	308.75**	.97	-
Academic Emotions	3	8	**.204**	[.10, .31]	.046	4.49	**.003**	.011	-	290.63**	.98	-
Appropriate Behavior	6	13	**.343**	[.15, .53]	.088	3.88	**.002**	.089	-	135.44**	.96	
Behavior Problems (recoded)	6	17	**.242**	[.10, .38]	.065	3.70	**.002**	.068	-	3755.33**	1	-
Bullying (recoded)	1	2	.290	[-.04, .62]	.026	11.25	.057	-	-	1.69	-	-
Executive functions and Self-Control	1	2	**.210**	[.18, .24]	.014	14.58	**<.001**	-	-	2.17**	-	-
Motivation	3	5	**.280**	[.12, .44]	.058	4.80	**.009**	.013	-	17.56**	.81	-
School Belonging and Engagement	3	4	**.302**	[.09, .52]	.068	4.43	**.021**	.018	-	86.36**	.97	-
Well-Being	5	6	**.267**	[.18, .36]	.034	7.76	**.001**	.005	-	320.47*	.97	-
Negative Outcomes (recoded)	7	25	**.207**	[.10, .31]	.051	4.03	**.001**	.046	.000	8202.42**	1	.00
Positive Outcomes	17	54	**.254**	[.17, .34]	.042	6.07	**<.001**	.007	.025	1333.79**	.22	.77
Negative TSR (recoded)	6	25	**.217**	[.12, .32]	.050	4.39	**<.001**	.044	.000	1797.62**	.99	.00
Positive TSR	18	54	**.237**	[.15, .32]	.044	5.40	**<.001**	.007	.029	2038.87**	.20	.80
Overall	20	79	**.246**	[.17, .32]	.037	6.60	**<.001**	.021	.018	11770.75**	.53	.46

Note. All effect sizes we expected to be negative were recoded to be positive (i.e., behavior problems, bullying, negative outcomes, negative TSR) to render them comparable. Thus, the mean correlations of TSRs with behavior problems, bullying, and general negative student outcomes can be read as negative. Larger positive effect sizes indicate stronger TSR-outcome links. All clusters are further defined and accompanied by examples in Appendix C. $\bar{r}$ = Weighted average correlation, pooled from all effect sizes within the respective study; k_{MA} = Number of included meta-analyses; k_{ES} = Number of meta-analytic effect sizes; $\tau_{(2;3)}$ = Heterogeneity at the effect size level (2) and the study level (3), respectively; Q = Sum of squared deviations of each effect size's estimate from the overall meta-analytic estimate; $I^2{}_{(2;3)}$ = Heterogeneity indices at the effect size level (2) and the study level (3), respectively.

* $p < .05$. ** $p < .001$.

In addressing RQ1b, we found an average correlation between TSRs and positive outcomes of $\bar{\bar{r}}$ = .25 (95% CI [.17, .32]) and between TSRs and negative outcomes of $\bar{\bar{r}}$ = .21 (recoded; 95% CI [.10, .31]). In addressing RQ1c, the average correlation between student outcomes and positive TSRs was $\bar{\bar{r}}$ = .24 (95% CI [.15, .32]); for negative TSRs, it was $\bar{\bar{r}}$ = .222 (recoded; 95% CI [.12, .32]). Neither pair of pooled effects differed significantly from the other (RQ1b: Q_M[1, 77] = 0.044, p = .834; RQ1c: Q_M[1, 77] = 050, p = .824).

When looking at the single meta-analytic effect sizes, we found the four largest relations between conflicts in the TSRs and external problem behavior ($\bar{r}$ = .57; Nurmi, 2012), negative TSRs and externalizing behavior problems ($\bar{r}$ = .55; Lei et al., 2016), teacher support and school belonging ($\bar{r}$ = .46; Allen et al., 2018), and teacher support and intrinsic value in classroom activities ($\bar{r}$ = .40; Givens Rolland, 2012). We found the largest negative correlations for anxious attachment to the teacher and external problem behavior (e.g., aggression, hyperactivity; $\bar{r}$ = -.37; Nurmi, 2012), conflicts in the TSR and well-being (encompassing academic engagement, effort, effortful control, self-regulation, school avoidance, school liking, task orientation; $\bar{r}$ = -.35; Nurmi, 2012), and a negative TSR and school engagement ($\bar{r}$ = -.31 and $\bar{r}$ = -.34; Roorda et al., 2011, 2014, respectively). For a full list of all the included effect sizes, see Table 1 and Supplemental Material S1. Along with the effect sizes from Kincade et al.'s (2020) intervention study, three meta-analytic effect sizes could not be clearly categorized into any of the outcome clusters but still offered great informative value. These were the relations between students' perceptions of teacher acceptance and psychological adjustment ($\bar{r}$ = .32; Ali et al., 2015), TSRs and socioemotional outcomes ($\bar{r}$ = .03; Korpershoek et al., 2016), and supportive communication in school and school dropout ($\bar{r}$ = .14; Strom & Boster, 2007).

5.7.4 Narrative and Quantitative Moderator Analysis (RQ2)

First, we identified the most frequently analyzed moderators in the meta-analyses (RQ2a). Second, we tested which characteristics of the meta-analyses, meta-analytic samples, or constructs and their measurement moderated the TSR-outcome relation (RQ2b). To this end, we conducted SOMA moderation analyses. The corresponding data set and the analytic code are presented in Supplemental Materials S1 and S3

5.7.4.1 Narrative Review of the Moderators in the Meta-Analyses (RQ2a)

In the present review of meta-analyses, we coded all moderator analyses that were conducted in the included meta-analyses. Table 5 shows these moderators for the TSR-outcome relations ordered by the moderators, whereas Appendix D shows them ordered by the meta-analyses, and Supplemental Material S17 shows the moderator sections from the included meta-analyses color-coded for each moderator. For more information about each moderator, please refer to the moderators column in our data set in Supplemental Material S1 or to the original article for each meta-analysis. The results of the moderator analysis are presented in Table 5, and we discuss selected example moderators below.

Table 5. *Moderators Analyzed in Included Meta-Analyses (Ordered by Moderator)*

Moderating variable	Significant/nonsignificant moderation in k_{MA}	Description of the direction of the moderation (meta-analytic level)
Student characteristics		
Students' age	10/5	3 favored primary school students, 1 favored lower primary school students, 4 favored older students, 2 were inconclusive
Students' gender	6/2	3 favored female students, and 3 favored male students
Students' SES	3/2	2 favored lower SES students, 1 favored higher SES students
Students' ethnic minority status	3/1	1 favored minorities, 1 favored nonminorities, 1 was inconsistent
Teacher characteristics		
Teachers' gender	3/1	1 favored male teachers, 1 favored female teachers, 1 was inconsistent
Teaching experience	2/2	2 slightly favored teachers with more experience
Teachers' ethnicity minority status	4/0	1 favored minorities, 3 favored nonminorities
Social-ecological factors		
School location	1/2	1 favored students in rural areas
Culture and Country	2/4	2 found conflicting results
Measurement characteristics		
Informant	5/0	3 favored teacher reports, 1 favored multiple informants, 1 found conflicting results as to whether the same informant was used to measure the TSR and the outcome
Outcome type	2/0	1 favored using test scores, 1 favored grades to measure academic achievement
Publication characteristics		
Publication year	0/2	No significant effects found
Publication status	0/2	No significant effects found
Sample size	0/4	No significant effects found

Note. The table reports the number of meta-analyses (column 2) and outcomes (column 3) for which the moderator (column 1) had a significant/nonsignificant influence. The direction of the moderating effect is further described in columns 4 and 5. The list above shows moderators that were tested in at least 2 of the included meta-analysis. The list contains only moderators of the TSR-outcome link. k_{MA} = Number of included meta-analyses.

Students' Age and Grade Level. Because younger children typically attend preschool, kindergarten, or lower grades, whereas older children usually attend middle or high school, we looked at the moderating effects of age and grade level on the TSR-outcome relations jointly. Age/grade level were the most frequently examined moderators and occurred in 15 (out of the 24 included) meta-analyses. Four meta-analyses found a larger relation for younger students, four for older students, two were inconclusive, and five were statistically nonsignificant. Meta-analyses reporting stronger relations in younger students focused on the following outcomes: academic achievement (Cherne, 2008; Wilkinson, 1980), appropriate behavior (Cherne, 2008), executive functions (Vandenbroucke et al., 2018), and bullying (Krause & Smith, 2022). The meta-analyses that found larger relations for older students did so for academic achievement (Givens Rolland, 2012; Roorda et al., 2011, 2014; Tao et al., 2022), well-being (Chu et al., 2010), and behavior problems (Roorda et al., 2021). Somewhat mixed moderator effects were reported for academic emotions (Lei et al., 2016, 2018), with a tendency toward larger relations with increasing age. Strom and Boster (2007) and Roorda et al. (2017) did not find any moderating effects of age in their analysis on the relation between supportive communication and school dropout or on the relations between TSRs and either academic achievement or engagement.

Students' Gender. The proportions of male/female students in the primary study samples was the second most frequently examined moderator. Nine meta-analyses used student gender as a moderator (Appendix D), eight of which specifically looked at the TSR-outcome relation in their analyses (Table 5). Three meta-analyses found slightly stronger effects for all-female versus all-male samples (Lei et al., 2016, 2018) and majority-female samples (Roorda et al., 2014), implying that female students profit more from better TSRs than male students do. Lei et al. (2016) found that positive TSRs had a larger relation with externalizing behavior problems in an all-female sample but negative TSRs did not. Lei et al. (2018) reported that TSRs were more strongly associated with *negative* academic emotions in an all-female sample than in an all-male sample, but TSRs had the same associations with *positive* academic emotions in male and female samples. Roorda et al. (2014) found stronger relations between TSRs and achievement and engagement in samples with a majority of female students in four of six TSR-outcome combinations.

Contrary to these findings, three meta-analyses found that positive TSRs had larger relations with executive functions (Vandenbroucke et al., 2018) and school conduct (Ali et al., 2015) and for three combinations of positive and negative TSRs with academic achievement and school engagement (Roorda et al., 2011) in studies that included more male students.

Roorda et al. (2011) also found larger effects in samples with more male students for three combinations (positive/negative TSRs with academic achievement and school engagement), but larger effects in samples with more female students for the relation between positive TSRs and achievement. Cornelius-White (2007), Korpershoek et al. (2016), and Roorda et al. (2021) did not find gender differences in relations between TSRs and positive student outcomes.

Students' Socioeconomic Status (SES). SES was the third most frequently assessed moderator and was included in five of the included meta-analyses. Two meta-analyses (Roorda et al., 2011, 2014) with seven effect sizes uncovered stronger TSR-outcome relations in students with lower SES. Vandenbroucke (2018) obtained mixed results with weaker TSR relations with executive functions in low-SES samples in a between-study comparison but stronger TSR-outcome relations for low-SES students within studies. Wilkinson (1980) identified slightly higher correlations with reading gains for low-SES students compared with high-SES students. For mathematics gains, she found a slightly larger effect on middle-SES students compared with low- and high-SES students. Korpershoek et al. (2016) did not find any moderating effect of SES on the relations between TSRs and behavioral, socioemotional, motivational, and other outcomes.

Students' Ethnic Minority Status. Moderation analyses of students' ethnic groups yielded mixed results with one meta-analysis finding larger TRS-outcome relations for students with an ethnic minorities status, one for nonminorities, and one yielding inconsistent findings. Specifically, Roorda et al. (2021) found that, in studies with more Caucasian student samples, student-teacher dependency had a stronger positive association with academic achievement and a stronger negative association with behavior problems. They did not find significant moderating effects on school engagement, internalizing behavior, or prosocial behavior. Similarly, Roorda et al. (2011) and Roorda et al. (2014) found that students' ethnic minority status made no difference in the TSR-outcome relation in three out of four and three out of five analyses, respectively. Cornelius-White (2007) did not find any moderating effect of student ethnicity on the TSR-outcome links. Hence, for most student outcomes, student ethnicity did not moderate the TSR-outcome relation.

Teachers' Gender. Eight meta-analyses investigated the moderating effect of teachers' gender and found mixed results. Roorda et al. (2021) found stronger associations in samples with fewer female teachers for student-teacher dependency's relations with school engagement and achievement. Cornelius-White (2007) identified female teachers as showing stronger associations with learner-centered features than male teachers or unspecified samples. Roorda et al. (2011) found the somewhat inconclusive result that samples with more male teachers showed larger TSR-school engagement associations, whereas this gender effect did not occur when studies with secondary-school samples were excluded.

Teaching Experience. Four meta-analyses examined teaching experience as a potential moderator. Roorda et al. (2011) and Roorda et al. (2014) found that the positive TSR-achievement relation became stronger with every year of teaching experience. However, many moderating effects in these meta-analyses—namely, the ones describing the relations between positive TSRs and school engagement or between negative TSRs and school engagement and achievement—yielded nonsignificant effects. At the same time, Roorda et al. (2021) and Cornelius-White (2007) did not find any moderating effect of teaching experience.

Teachers' Ethnic Minority Status. Out of the four meta-analyses investigating teachers' ethnic minority status as a moderator, three found larger TSR-outcome effects in samples with larger proportions of ethnic majority teachers (i.e., Roorda et al., 2011, 2014, 2021). Cornelius-White (2007) found the opposite, namely, larger effects in samples with more teachers of color than when the ethnic minority was not specified. There were no significant effects when comparing teachers of color with Caucasian teachers.

School Location. Three meta-analyses examined differences between schools located in urban, suburban, and rural areas as well as the potential effect of schools located in different regions of North America. These meta-analyses generated dispersed results. Allen et al. (2018) found significantly stronger TSR-school-belonging relations in rural areas than in urban areas but mixed results for suburban areas. At the same time, Givens Rolland (2012) reported significant correlations between TSRs and achievement for schools with a mixed or unreported location but nonsignificant correlations for suburban and urban schools. She further found larger relations between TSRs and prosocial factors in urban schools than in schools in suburban and mixed or unreported areas. Cornelius-White (2007) did not find any moderating effects of school location.

Culture and Country. Three meta-analyses focused on culture (Lei et al., 2016, 2018; Tao et al., 2022) and another three on the country in which the student samples were located (Allen et al., 2018; Korpershoek et al., 2016; Strom & Boster, 2007). Culture was defined as whether the sample from the primary study had been recruited from an Eastern culture (broadly, East Asia), a Western culture (e.g., Europe or North America), or another culture (e.g., Turkey). Lei et al. (2016) found that Western students showed larger relations between positive TSRs and externalizing behavior, whereas Eastern student samples showed larger relations between negative TSRs and externalizing behavior problems. In their 2018 meta-analysis, Lei et al. reported that Western student samples showed larger relations between TSRs and positive academic emotions, whereas Eastern student samples showed larger relations between TSRs and negative academic emotions. Tao et al. (2022) did not find moderating effects of culture, and Korpershoek et al. (2016), Allen et al. (2018), and Strom and Boster (2007) did not detect any differences between countries in the TSR-outcome relations.

Informant. Informant effects (i.e., whether teachers, student peers, or students rated the TSRs) were found in four out of the five meta-analyses that investigated this variable. All four meta-analyses presented statistically significant moderating effects but with mixed directions. Cornelius-White (2007) found larger effects when multiple perspectives were combined or research focused on observers' perspectives than teachers' or students' perspectives for behavioral or affective outcomes. Roorda et al. (2011) found that the TSR informant was a significant moderator in all their moderator analyses. When looking at the relation between TSRs and engagement, they found larger correlations when both TSRs and achievement were rated by the same informant. When looking at the relation with school achievement, by contrast, they found larger correlations for different informants. Endedijk et al. (2022) found the largest correlations between TSRs and peer relations when teachers rated the TSR, lower correlations for student ratings, and the lowest correlations for classmate/peer ratings. Focusing on the relation between conflictual TSRs and peer aggression, Krause and Smith (2022) found larger effect sizes for teacher- or peer-reports than for student self-reports of the TSR. Lei et al. (2016) also found the larger effect for teacher ratings compared with student-, peer-, or parent-ratings on the association between TSRs and academic emotions.

Overall, teacher ratings yielded stronger TSR-outcome relations than any other informant in most meta-analyses. Results were inconclusive whether the same or different informants for TSR and the outcome led to stronger associations.

Outcome Type. Tao et al. (2022) found significantly larger associations between TSRs and teacher-assigned grades over standardized test scores. By contrast, Givens Rolland (2012) found that studies using standardized test scores showed significant positive associations with TSRs, whereas studies using teacher-assigned grades showed no such association.

Publication Year and Status. Neither Allen et al. (2018) nor Cornelius-White (2007) found that publication year moderated the TSR-outcome relation. Similarly, neither Cherne et al. (2008) nor Cornelius-White (2007) found publication status to be a moderator.

Sample Size. Four meta-analyses investigated the moderating effect of sample size as part of their publication bias assessment (see Table 5). Cornelius-White (2007), Roorda et al. (2011, 2021), and Vandenbroucke et al. (2018) concluded that sample size did not have a significant effect.

5.7.4.2 Second-Order Moderation Analyses (RQ2b)

The meta-analytic effect sizes in our data set were highly heterogeneous (p_Q < .001; see Table 4 for the exact *Q* values). For the three-level random-effects meta-analysis, the heterogeneity at the level of the meta-analytic effect sizes and the level of the meta-analyses was 49% and 51% of the total variance, respectively. Given the large heterogeneity in the data, we conducted second-order moderator analyses for RQ2b and found that some characteristics significantly moderated the TSR-outcome relation.

Continuous Moderators. To explain this heterogeneity, we examined a total of six continuous moderators described in Table 6. Our analyses showed that publication year, average age (in years), gender composition, estimated sample size, sample size, and the methodological quality total score did not moderate the TSR-outcome link. Table 6 presents the respective results of the second-order continuous moderator analyses.

Table 6. *Results of the Continuous Moderators on the Effect-Sizes Level*

Continuous moderator	k_{MA}	k_{ES}	*B*	*SE*	95% CI	*p*
Publication year	20	79	0.01	0.03	[-0.05, 0.08]	.713
Age (mean in years)	3	5	0.03	0.03	[-0.05, 0.11]	.321
Gender composition	6	15	0.34	0.65	[-1.07, 1.76]	.608
Estimated sample size	19	74	0.00	0.00	[-0.00, 0.00]	.861
Sample size	15	58	0.00	0.00	[-0.00, 0.00]	.737
Total quality score	20	79	0.02	0.02	[-0.01, 0.05]	.137

Note. A significant *p*-value indicates that there was a statistically significant difference between the levels of the moderator. Publication year stands for the years in which the meta-analysis was published; k_{MA} = Number of included meta-analyses; k_{ES} = Number of effect sizes; Gender composition = Percentage of female students in the sample; Estimated sample size = average of students per effect size across all included meta-analyses.

Categorical Moderators. We further investigated a total of 20 categorical moderators. Table 7 reports the results of the respective analyses, and Appendix E shows the levels of all the categorial moderators. We excluded moderators with fewer than four effect sizes per category (e.g., social minority status of the sample; see Bakermans-Kranenburg et al., 2003).

The following categorical characteristics did not show statistically significant moderating effects on the TSR-outcome relation (see Table 7): publication type, publication status, affiliation, country, grade level mode, prekindergarten, kindergarten, preschool, elementary school, TSR level, TSR modality, TSR informant mode, outcome cluster, outcome modality, outcome informant mode, outcome type mode, and moderator analysis [yes/no]. Although statistically nonsignificant, the mean effect sizes tended to be larger in the unpublished doctoral dissertations in our sample (i.e., Cherne, 2008; Wilkinson, 1980) than in the published journal articles ($\bar{\bar{r}}$ = .42 vs. $\bar{\bar{r}}$ = .23, p = .13; see Appendix E). Regarding school status, meta-analytic effect sizes based on samples including middle school ($\bar{\bar{r}}$ = .25, 95% CI [.20, .30]) were significantly larger than those based only on other grade levels ($\bar{\bar{r}}$ = .15, 95% CI [.08, .22]). This moderator explained heterogeneity mostly between meta-analyses ($R^2_{(2)}$ = 4%; $R^2_{(3)}$ = 94%). Similarly, meta-analyses of samples of high school students also reported significantly larger mean effect sizes ($\bar{\bar{r}}$ = .26, 95% CI [.21, .31]) than meta-analyses of samples of students from other school levels ($\bar{\bar{r}}$ = .15, 95% CI [.09, .22]), with large amounts of the explained variance between meta-analyses ($R^2_{(2)}$ = 6%, $R^2_{(3)}$ = 100%).

5.7.5 Methodological Quality of Included Meta-Analyses (RQ3)

5.7.5.1 Description of the Meta-Analytic Quality (RQ3a)

Figure 9 summarizes the methodological quality criteria ordered by the included meta-analyses, and Figure 10 presents the number of criteria that were fulfilled. For the detailed coding of the quality indicators, we refer readers to Appendix F and Supplemental Material S18.

Overall, the methodological quality of the included meta-analyses varied. On average, the meta-analyses fulfilled seven of 16 quality indicators, and 10 of the 24 included meta-analyses fulfilled at least half of the 16 quality criteria (Figure 9). Moreover, seven of the 16 quality and reproducibility indicators were reached by at least half of the meta-analyses (Figure 10). Most meta-analyses reported moderator analyses (k_{MA} = 19), provided a list of included studies (k_{MA} = 21) and clear descriptions of the inclusion and exclusion criteria (k_{MA} = 22), and reported confidence levels or standards errors (k_{MA} = 22). Only a few meta-analyses had made their data freely available online (k_{MA} = 3), included a list of excluded studies (k_{MA} = 2), and considered primary study quality (k_{MA} = 2). Moreover, only a few meta-analyses reported that they had used two raters to select studies and extract data (k_{MA} = 6). No meta-analysis fulfilled all 16 quality criteria. Looking at the extremes of the distribution, two meta-analyses achieved a total of 12 of the 16 quality criteria (i.e., Endedijk et al., 2021; Krause & Smith, 2022), whereas two meta-analyses met only three out of the 16 indicators (Roorda et al., 2014; Wilkinson, 1980).

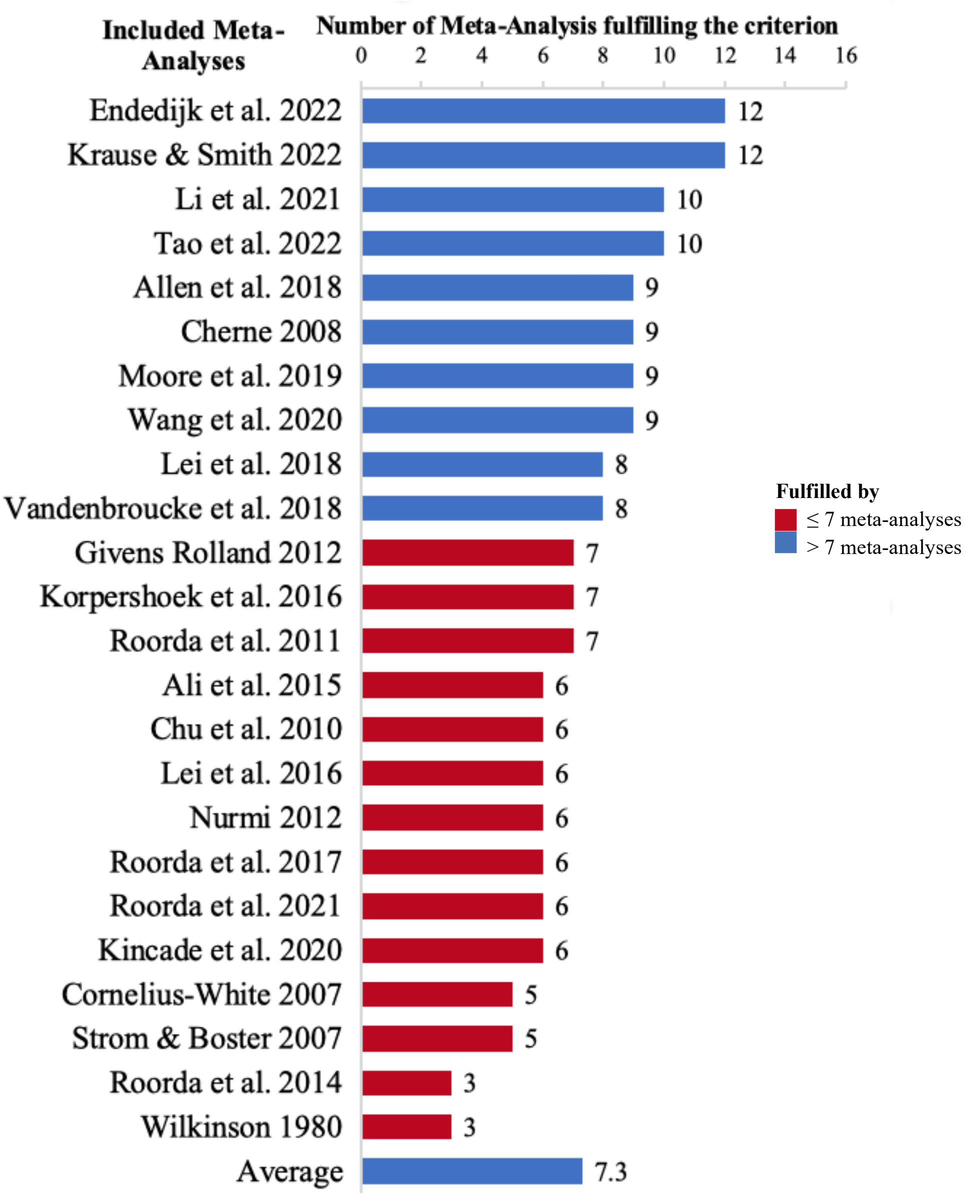
Table 7. *Results of the Categorical Moderators*

Moderator	k_{MA}	k_{ES}	$\bar{\bar{r}}$ [95% CI]	SE	$R^2_{(2)}$	$R^2_{(3)}$	p
Baseline ($I^2_{(2;3)}$: 53%; 46%)	20	79	.25 [.17, .32]	0.04	–	–	–
Publication type	20	79	–	–	.00	.13	**.10**
Publication status	20	79	–	–	.00	.13	**.10**
Affiliation	20	79	–	–	.03	.00	**.99**
Country	20	79	–	–	.00	.00	**.95**
Social minority	20	79	-	-	.02	.13	**.06**
Grade level mode	11	27	-	-	.57	.55	**.28**
Prekindergarden	19	71	-	-	.00	1	**.48**
Kindergarten	18	66	-	-	.00	1	**.89**
Preschool	18	66	-	-	.00	1	**.81**
Elementary school	18	66	-	-	.00	1	**.42**
Middle school	19	71	-	-	.04	1	**.02***
No	5	37	.15 [.08, .22]	0.03	-	-	–
Yes	14	34	.25 [.20, .30]	0.03	-	-	–
High school	19	71	-	-	.06	1	**.01***
No	5	39	.15 [.09, .22]	0.03	-	-	–
Yes	14	32	.26 [.21, .31]	0.03	-	-	–
TSR level	18	54	-	-	.62	.00	**0.21**
TSR modality	24	79	-	-	.00	.00	**.82**
TSR informant mode	15	65	-	-	.00	.00	**.21**
Outcome cluster	41	75	-	-	.00	.00	**.94**
Outcome modality	24	79	-	-	.00	.00	**.83**
Outcome informant mode	17	59	-	-	.00	.84	**.90**
Outcome type mode	37	73	-	-	.00	.00	**.95**
Moderator analysis (yes/no)	23	61	-	-	.00	.00	**1**

Note. The levels of the nonsignificant categorical moderators are shown in Appendix E. Larger positive effect sizes indicate stronger TSR-outcome links. A significant p-value indicates that there was a statistically significant difference between the levels of the moderator. $I^2_{(2;3)}$ = Heterogeneity indices at levels two and three, respectively; k_{MA} = Number of included meta-analyses; k_{ES} = Number of effect sizes; $\bar{\bar{r}}$ = Weighted average correlation; $R^2_{(2;3)}$ = Variance explained within (Level 2) and between (Level 3) meta-analyses.

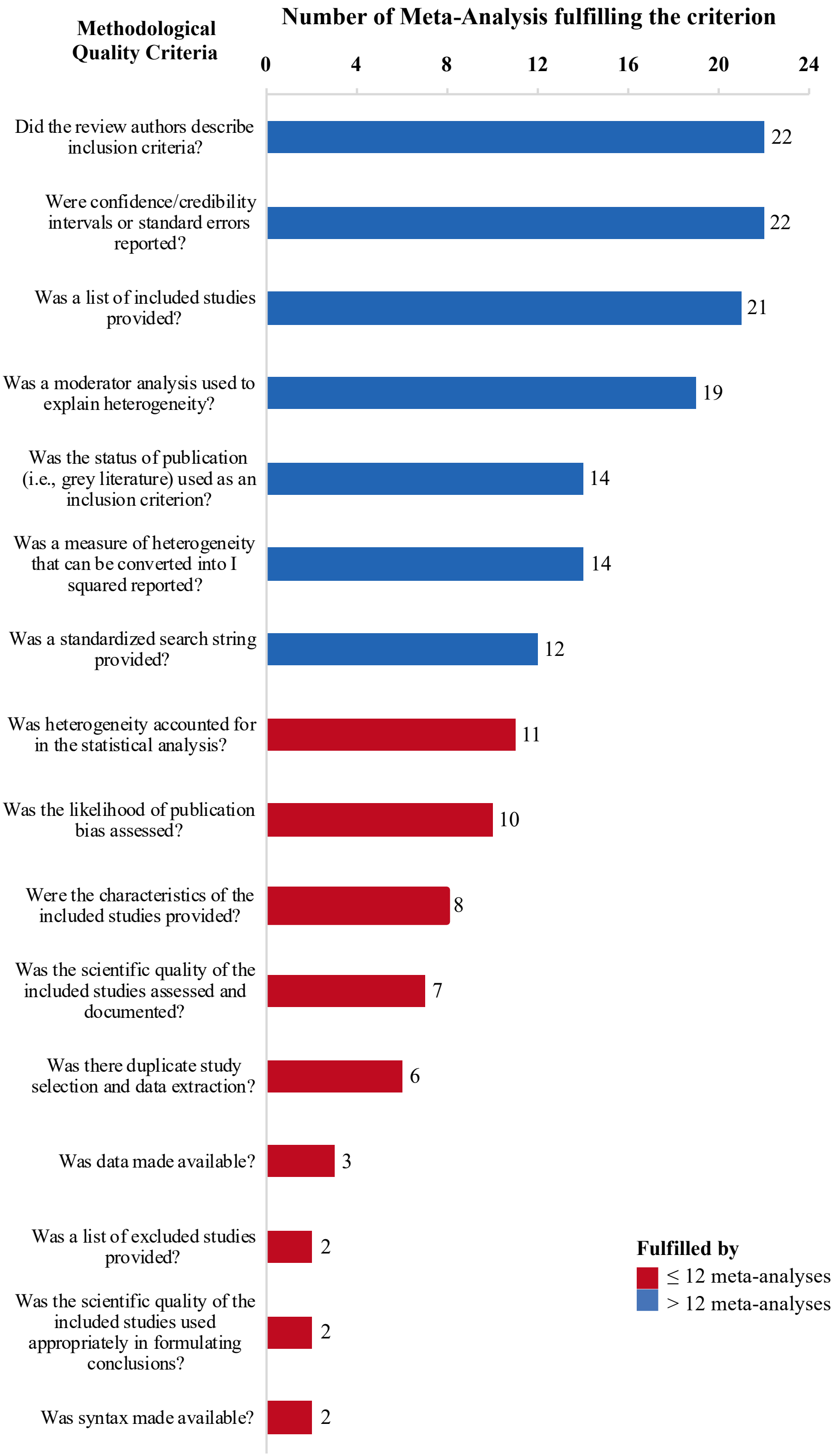
* $p < .05$.

Figure 9. *Bar Charts Summarizing the Methodological Quality and Reproducibility Indicators Each Included Meta-Analysis Fulfilled*



Note. Meta-analyses that fulfilled only half or fewer than half of the methodological quality and reproducibility indicators are colored red.

Figure 10. *Bar Charts Summarizing the Number of Meta-Analyses Fulfilling Each of Our 16 Methodological Quality and Reproducibility Indicators*

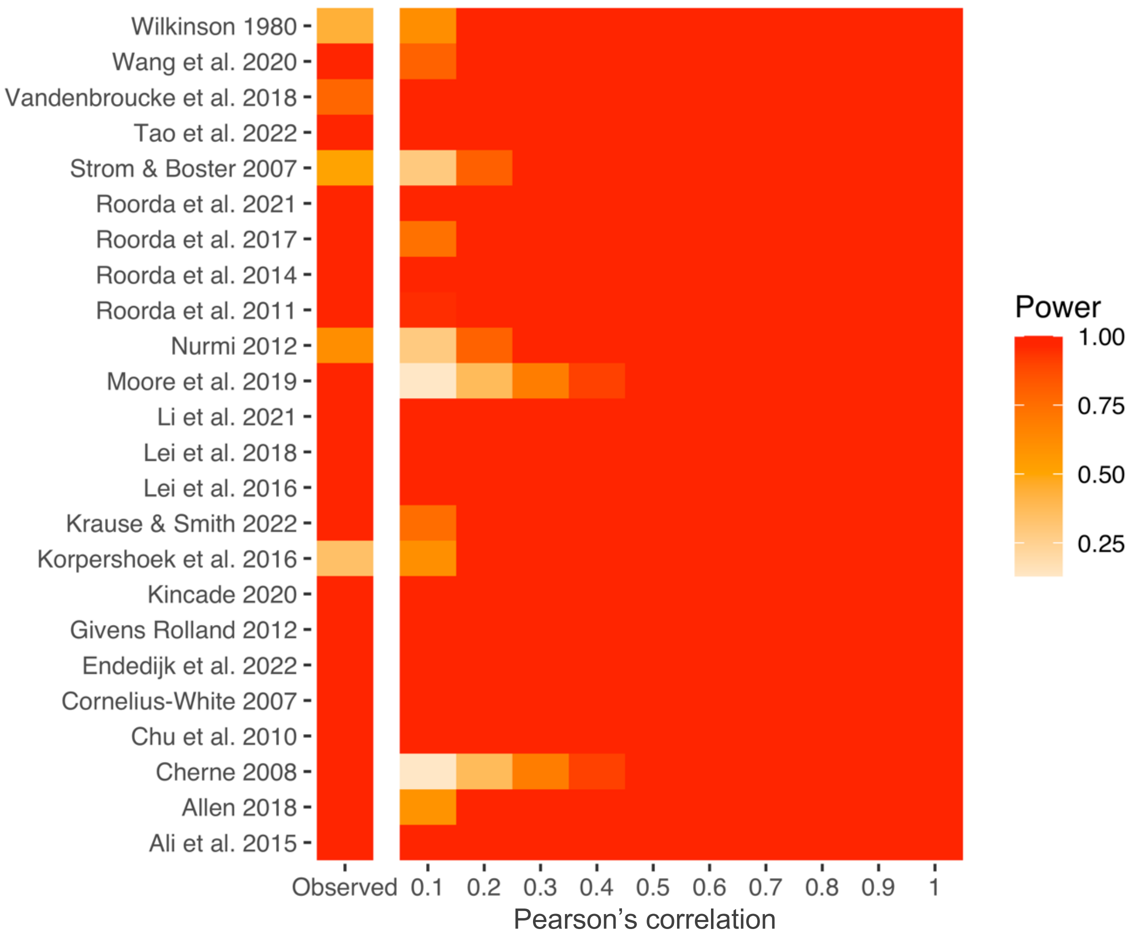


Note. Methodological quality and reproducibility indicators fulfilled by only half or fewer than half of the meta-analysis are colored red.

5.7.5.2 Statistical Power of the Included Meta-Analyses (RQ3b)

Given the lack of reports of statistical power (Valentine et al., 2010), we conducted power analyses for the included meta-analyses. Figure 11 shows a resultant fireplot of the median statistical power. By and large, this plot indicated high statistical power in our sample of meta-analyses. Specifically, 20 meta-analyses had a power of at least 80% to detect the mean effect they actually found (see Figure 11), whereas four meta-analyses were statistically underpowered.

Figure 11. *Fireplot of the Median Statistical Power of all Included Meta-Analyses*



Note. The fireplot displays statistical power for the observed mean effect size of the included meta-analyses on the left and for a range of hypothetical effect sizes (Quintana, 2023). Darker shaded cells identify larger statistical power. Lighter cells indicate smaller statistical power.

5.7.5.3 Sensitivity of the Meta-Analytic Results to Methodological Quality (RQ3c)

Even though the total methodological quality score did not moderate the TSR-outcome relations (see RQ2b), we also examined the possible moderating effects of the single quality indicators reported in Table 8. Thirteen of the 16 quality indicators did not show significant moderating effects. However, we found significantly smaller correlations in meta-analyses that did not report primary study quality ($\bar{\bar{r}}$ = .19, 95% CI [.11, .27]) than in those that did ($\bar{\bar{r}}$ = .35, 95% CI [.12, .57]). Similarly, the correlations were larger for meta-analyses that considered study quality in their conclusion ($\bar{\bar{r}}$ = .70, 95% CI [.48, .91]) than for those that did not ($\bar{\bar{r}}$ = .21, 95% CI [.17, .26]).

We also found a substantial correlation between the number of quality criteria that had been fulfilled and publication year, r = .54 (95% CI [.36, .68]). Hence, older meta-analyses were more likely to have lower quality scores than more recent ones. Figure S13 in the Supplemental Material presents this association and additionally shows a peak in recent years with a steady increase from 1980 (i.e., Wilkinson, 1980) to 2022, with Endedijk (2022) and Krause and Smith (2022) each fulfilling 12 quality indicators.

Table 8. *Results of the Quality Indicators as Moderators*

Quality indicators	Q_M	No			Yes			
		$\bar{\bar{r}}$	(95% CI)	k_{ES}	$\bar{\bar{r}}$	(95% CI)	k_{ES}	p
Inclusion criteria described	0.41	.31	(.08, .54)	7	.24	(-.24, .71)	72	.52
Double screening and coding	1.03	.22	(.13, .31)	57	.30	(.05, .56)	22	.31
Standardized search string	0.06	.23	(.12, .34)	27	.25	(-.01, .51)	52	.81
Unpublished articles included	1.56	.20	(.09, .30)	50	.29	(.04, .53)	29	.22
List of included studies	0.15	.20	(-.05, .44)	6	.25	(-.25, .75)	73	.70
List of excluded studies	0.29	.25	(.17, .33)	56	.17	(-.22, .55)	23	.59
Study description included	0.08	.24	(.15, .33)	48	.26	(.00, .52)	31	.78
Study quality assessed	**4.74**	**.19**	**(.11, .27)**	55	**.35**	**(.12, .57)**	24	**.03***
Study quality considered	**32.10**	**.21**	**(.17, .26)**	71	**.70**	**(.48, .91)**	8	**.00****
Heterogeneity index reported	0.12	.26	(.13, .39)	44	.23	(-.06, .52)	35	.73
Heterogeneity accounted for	0.03	.25	(.15, .35)	52	.24	(-.02, .49)	27	.86
Moderator analysis conducted	1.93	.15	(.00, .30)	8	.27	(-.05, .59)	48	.17
Publication bias assessed	0.05	.24	(.13, .34)	56	.25	(.00, .51)	23	.83
CI or SE reported	3.90	-.14	(-.53, .25)	1	.26	(-.54, 1.05)	78	.05
Data available	0.07	.25	(.17, .33)	74	.22	(-.09, .52)	5	.79
Syntax available	0.00	.24	(.16, .32)	75	.24	(-.09, .58)	4	.99

Note. The *p-value* refers to the test of differences in weighted average for each methodological quality indicator. The indicators were adapted from the AMSTAR rating scale by Shea et al (2007). The significant difference between meta-analyses reporting confidence intervals and those that did not, should not be interpreted, as the former only occurred once (see Bakermans-Kranenburg et al., 2003). The exact wording can be found in Figure 10 and Table S18 in the Supplemental Material. Q_M = Test statistic of the omnibus test of moderators.; $\bar{\bar{r}}$ = Weighted average correlation.
* $p < .05$. ** $p < .001$.

5.8 Discussion

5.8.1 Summary of Findings

Our preregistered review of meta-analyses provides a first quantitative umbrella study of the research on the relations between TSRs and student outcomes. Synthesizing over 70 years of educational research in 24 meta-analyses and with more than 2 million prekindergarten and K-12 students, we answered three pending research questions. First, we synthesized correlations for nine clusters of student outcomes and identified the descriptively largest significant relations (in descending order) between TSRs and appropriate behavior, school belonging and engagement, motivation, general well-being, behavior problems, executive functions and self-control, academic emotions, and academic achievement (RQ1). There were no statistically significant differences between these relations and, thus, all TSR-outcome relations are similarly important. Second, we narratively reviewed important moderators of the TSR-outcome relations in the meta-analyses and conducted second-order meta-analytic moderator analyses to quantitatively examine moderators at the level of meta-analyses. Prior meta-analyses had frequently investigated student age and gender with mixed findings, which our analyses showed as well. Further, they found larger TSR-outcome correlations for teacher-reported TSRs and for teachers without ethnic minority status. We also identified samples that yielded larger correlations, including middle school and high school students. At the same time, many variables did not moderate the TSR-outcome relations (RQ2). Third, we examined the methodological quality of the included meta-analyses and found large quality differences and areas for improvement, such as thorough reporting to facilitate reproducibility (RQ3).

5.8.2 Seven Discussion Points

In line with previous reviews of meta-analyses (Jansen et al., 2022; Schneider & Preckel, 2017), we structured our discussion around seven discussion points. The first point concerns the weighted average effect sizes in RQ1. The second, third, and fourth discussion points pertain to the variables that we found to (not) moderate the TSR-outcome relations in RQ2. The fifth point concerns the quality of the included meta-analyses from RQ3. Discussion Point 6 begins with a discussion of the impact of our findings on future TSR interventions, and Point 7 addresses potential implications for educational policy.

5.8.2.1 1. Which student outcomes have the closest associations with TSRs?

We found the largest relations between TSRs and academic achievement, academic emotions, appropriate student behavior, behavior problems, executive functions and self-control, motivation, and student well-being (Figure 8). These findings show that TSRs play a vital role in many aspects of student characteristics and are linked not only to achievement but also to the students’ emotions, behavior, and general cognitive outcomes in school, all of which are important aspects of students’ well-being (Chu et al., 2010; Wang et al., 2020).

Surprisingly, bullying did not show a significant relation with TSRs. This lack of relation could be due to two issues: First, the number of effect sizes in the outcome cluster was small (see Figure 8). Whereas some researchers might question the meta-analysis of only a few effect sizes, Valentine et al. (2010) argued that two effect sizes are sufficient for such a synthesis and that alternative synthesis strategies are usually subpar. At the same time, to ensure comparability and a meaningful interpretation, the authors cautioned against synthesizing effect sizes from studies that are vastly different. Second, the confidence intervals of the included effect sizes varied substantially and were generally large, which introduced more variation into the SOMA.

5.8.2.2 2. Do TSR-outcome relations change with students’ age?

The most frequently tested moderator in the included meta-analyses was the age of the students. Our review of prior research showed that the moderating effect of age is unclear. This lack of clarity was surprising because the extant literature largely agrees that students become more independent from their teachers over time and, thus, also less influenced by positive or negative TSRs (e.g., Cherne, 2008; Krause & Smith, 2022; Vandenbroucke et al., 2018; Wilkinson, 1980) and TSRs tend to decrease (Bosman et al., 2018). In our second-order meta-analytic moderator analysis, we found that when children from kindergarten, preschool, or elementary school were included in the meta-analytic samples, the correlations were smaller than when students from middle or high school were included. This finding too was unexpected, as both attachment theory and bioecological theory—alongside the moderator analyses of some of the included meta-analyses—suggest that younger students are more sensitive to negative TSRs and may actually be more in need of positive TSRs than older students (Bowlby, 1982; Bronfenbrenner, 1979). At the same time, there are also good reasons for TSRs to become more important over time. Chu et al. (2010) discussed that teachers and students have more of an equal footing in their relationship as the students move through their school careers, thus pointing to the great potential of building better relationships to enhance well-being. Roorda et al. (2011) suggested that positive TSRs could support the students in meeting the higher expectations of secondary school and benefit their development (Roorda et al., 2011).

Another hypothesis is that younger students typically hold positive perceptions of their teachers and the relationships they share with them (Fauth et al., 2014). If this is the case for a sample of students, then there is not much variation in TSRs, and small deviations from the average can become influential. In secondary school, students may generally have neutral perceptions of TSRs and be less engaged, but they might appreciate a specific, positive TSR that supported them during major life events (Hamre & Pianta, 2001; Roorda et al., 2011). In this sense, teachers may serve as protective factors during challenging times, such as periods of high-stakes testing, school stress, bullying, or a coming out (Krause & Smith, 2022; Murdock & Bolch, 2005; Tao et al., 2022).

5.8.2.3 3. Why are there no gender differences in TSR-outcome associations?

The second most frequently tested moderating variable was student gender, and there was no clear direction or difference between male and female students in TSR-outcome correlations. The question is why one might expect gender differences in TSRs at all. Lei et al. (2016), for example, suggested that female students cared more about positive TSRs than male students (Hu et al., 2015), and that female students might be more influenced by TSRs in an externalizing problem behavior context (Deater-Deckard & Dodge, 1997). Vandenbroucke et al. (2018) discussed that children with difficulties (e.g., male students with more problems in executive functions; Diamond & Lee, 2011) or male students with lower SES could profit more from positive TSRs (Roorda et al., 2011). Vandenbroucke et al. (2018) also questioned whether gender was a significant moderator only because samples with a majority of male students also tended to be older, and age was a significant moderator.

The meta-analyses included in our systematic review are good examples that gender differences in TSR research are somewhat unclear and, when they exist, tend to rely a great deal on the specific student outcome. Craig Aulisi et al. (2023) showed that finding significant moderating effects of gender in meta-analyses involving psychological constructs is actually rare. In their review of 286 tests of the moderating effect of the gender composition in 50 meta-analyses, they only rarely identified a moderating effect of gender even though large differences in effect sizes existed between genders. Craig Aulisi et al. (2023) argued that the lack of between-study variation in the female-to-male ratio might cause this lack of moderating effect. In the meta-analyses we reviewed, the included primary studies seemed to have very similar female-to-male ratios, making potential gender differences difficult to detect without additional single-gender primary research.

The question is whether future meta-analysts should keep looking for gender differences, even though we did not find overarching evidence for their moderating effects. In our SOMA, we found only nongeneralizable or even null results, which could be considered support for the gender similarities hypothesis—that is, the hypothesis that women and men are more similar rather than substantially different in most psychological constructs (Hyde, 2014).

These findings are just as important as finding that gender acts as a moderating variable. Future research should carefully assess the need to test for gender differences and should exercise caution with regard to implementing single-gender TSR-building programs (Hyde, 2014). Future research will also need to appraise gender diversity (not perceiving gender as binary; Hyde et al., 2019) in psychological research and use this new knowledge to advance TSRs (Cameron & Stinson, 2019; Garvey et al., 2019; McGuire et al., 2019).

5.8.2.4 4. Why future research should focus on TSRs and their link to teacher outcomes?

As noted earlier, when looking at the teacher-related variables in prior meta-analyses, we found two clear results: Three out of four meta-analyses favored nonminority teachers over teachers with a minority status, and three out of five meta-analyses favored teacher reports of TSRs over those from students or peers (see Table 5). What was not coded in most meta-analyses was the level at which the TSRs were reported (i.e., whether the reports were based on classroom-level or dyadic measures; Wentzel, 2022) and whether there was a match between the minority status of the teachers and their students. The level of reporting needs further exploration within and across meta-analyses because these two levels may capture different aspects of TSRs (Sabol & Pianta, 2012). Moreover, the two levels may interact because students tend to be more similar in their TSR reports within a classroom than between classrooms. Through the general classroom climate, students in one classroom may also share a more similar interaction style with their teacher, resulting in more similar bonds on the dyadic level. Randomized controlled trials could account for this dyadic dependency within a classroom (Sabol & Pianta, 2012). Ultimately, teachers need to be put more into the focus of TSR research. This starts by clearly reporting and differentiating between TSRs on the dyadic and the classroom-level.

In most primary studies and meta-analyses, the link between TSRs and teacher variables is largely missing. As Cornelius-White (2007) already stated, positive TSRs are a two-way street in that they influence but are also influenced by both teachers and students alike. Teachers are key players in TSRs, and it is likely that happy, intelligent, healthy, and motivated teachers can form more positive TSRs (see, e.g., Bardach & Klassen, 2020; Chamizo-Nieto et al., 2021; Peláez-Fernández et al., 2021). This effect could also work in the opposite direction—positive TSRs may help prevent teachers' burnout, enhance their motivation, facilitate more effective teaching, and offer opportunities for them to develop their skills and well-being (Li et al., 2022; Milatz et al., 2015; Pianta, 1999; Roorda et al., 2011; Spilt et al., 2011). In recent years, Li et al. (2022) found that positive TSRs are linked to complex teaching practices, and students who engage emotionally in the lesson promote better TSRs. Furthermore, positive and caring TSRs activate the positive effects of teachers' expectations on student outcomes (Johnston et al., 2022). Future research is thus needed to further identify the associations of teacher variables with TSRs, clarify the direction of these associations, and develop classroom interventions, teacher training, and policy advice.

5.8.2.5 5. How can the methodological quality of meta-analyses on TSRs be improved further?

Assessing and considering study quality is important to ensure the validity of research syntheses (Johnson, 2021; Scherer & Emslander, 2023b). In general, methodological quality and reproducibility differed between the included meta-analyses but did not moderate TSR-outcome relations in our second-order meta-analytic moderator analysis. Hence, our results did not systematically differ between studies with high and low quality. Furthermore, most included meta-analyses were sufficiently powered. Whereas these findings strengthen the validity of the present findings, we still detected several shortcomings in quality, reproducibility, and power. Adhering to common reporting guidelines and open science practices in both primary and meta-analytic research will make future research both easier and more meaningful. In line with Schalke and Rietbergen (2017), we suggest that authors should be aware of high-standard guidelines for meta-analyses, such as AMSTAR (Shea et al., 2007).

As the use of a single quality sum score has several problems (Scherer & Emslander, 2023b; Wedderhoff & Bosnjak, 2020), meta-analysts could examine the correlation between the individual quality indicators to check whether multiple scores would be appropriate. In the correlogram in Supplemental Material S19, for example, one could argue for three separate sum scores and the exclusion of a few counterintuitive quality indicators. These practices could help corroborate the face validity of the quality scores and might yield results that are easier to interpret substantively. However, more research is needed to provide a best practice example of how to derive quality scores from multiple indicators (e.g. Scherer & Emslander, 2023).

To some extent, differences in quality and reproducibility were correlated with publication year, which might be a confounder (see Figure S20 in the Supplemental Material for the correlation plot). Meta-analyses published in very short articles (i.e., Roorda et al., 2014) before the widespread use of the Internet (i.e., Wilkinson, 1980), let alone the emergence of the open science movement advocating for thorough reporting for reproducibility (Nosek et al., 2015; Open Science Collaboration, 2015), could not have met all the quality criteria. These advances promote the dissemination of knowledge and the assessment of newer meta-analyses. For example, original data, analytic code, measurement instruments, and other materials can be directly stored on open-access platforms, such as the Open Science Framework, which ensures the availability of key information for meta-analysts. Aggregated and individual participant data can both be utilized (Campos et al., 2023), making community-augmented meta-analyses possible (see Waack, 2018). In our view, to drive improvements in research quality, it is crucial to embrace preregistration, ensure online availability, uphold the principles of open science, and adhere to reporting guidelines, such as MARS (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008) and PRISMA (Moher et al., 2009; Page et al., 2021). By adopting these measures, we strive to approach an optimal framework for distinguishing high-quality meta-analyses in psychology. To take these next steps toward reproducible research, primary study authors, meta-analysts, reviewers, and editors need to work together and streamline the scientific processes that should be followed when conducting meta-analyses.

Whereas the statistical power of most included meta-analyses was high (80%), some were underpowered (see Figure 11). However, statistical power is still not commonly checked in meta-analyses. Only in recent years have guidelines on how to assess the statistical power of one or multiple meta-analyses been made prominent (Valentine et al., 2010), and R packages have been developed to visualize it (e.g., the metameta package by Quintana, 2023). As a consequence, future meta-analyses and SOMAs could include statistical power as a quality criterion, map it across meta-analyses, and utilize it for sensitivity analyses to further put into perspective the value of the evidence of the included research (Johnson, 2021; Scherer & Emslander, 2023b).

5.8.2.6 6. How do these findings inform interventions to improve TSRs and student outcomes?

The present findings on the TSR-outcome relations and important moderators can be used to refine TSR interventions, help TSR building, and enrich teacher training. At least two types of interventions are imaginable: those focusing on improving TSRs and those aiming to further enhance a specific student outcome through positive TSRs. Building TSRs through interventions or specific teacher training would be most promising if aimed at improving students' behavior and well-being. Robinson (2022) already explored how to best support teachers in their relationship-building. The question that remains is which student outcomes to focus on (RQ1) and for which students, teachers, and contexts (RQ2) such interventions or training might be effective.

In our review, students' appropriate behavior, school belonging and engagement, motivation, general well-being, and behavior problems showed the closest (bidirectional) associations with TSRs. Thus, these variables could be auspicious targets to improve through TSRs in a school context, alongside executive functions and self-control, academic achievement, and academic emotions, all of which were also statistically significantly related to TSRs and did not largely differ. Specifically, future interventions may explore how to effectively focus on the positive aspects of TSRs and positive student outcomes, as these variables showed a considerably larger effect than their negative counterparts.

Our review of moderators and our SOMAs provide valuable information about the potential efficiency of TSR interventions for specific individuals and contexts. Whereas there seems to be no need to differentiate between female and male students, there seems to be an effect of grade level. Students in middle and high school tended to have larger TSR-outcome associations, potentially indicating a larger return on investment in these older student groups. Thus, the larger connections between TSRs and outcomes of middle- and high-school students should be causally tested in future intervention studies and could inform learner-centered teacher education (Cornelius-White, 2007).

As discussed above, however, teachers' well-being should also be considered an important outcome of such interventions. Further, teachers need to engage in direct and proactive interventions to increase the TSRs (Kincade et al., 2020). Recent developments in TSR interventions are well underway (Keane & Evans, 2022; Robinson, 2022), calling for future research to develop easy-to-administer programs designed to target specific student outcomes and aspects of TSRs (see, e.g., Keane & Evans, 2022). Along with these recent findings, the present review offers a basis from which to launch future research on effective interventions.

5.8.2.7 7. What is the impact of our findings on educational policy?

In educational policy, the TSR is a crucial factor for school success and student well-being. Its importance is signified in the creation of agencies, such as the National School Climate Center at Ramapo for Children (USA), and school climate reports, such as the *CARAT A School Climate Model for Luxembourg's Schools* [CARAT Ein Schulklima-Modell für Luxemburger Schulen] (SCRIPT, 2018). The latter argues that schooling should be structured to support positive interpersonal relationships that can promote performance and health. Our findings support this claim, as we found that TSRs had medium and large associations with academic achievement and students' well-being (see Figure 8). Based on our findings, reports (e.g., the *CARAT*) should include more behavioral and socioemotional aspects of students' thriving, besides the traditional dimensions of performance and health.

Our review of meta-analyses provides some empirical basis from which to further refine educational policies on TSRs and school climate at large. One possible direction could be to emphasize the importance of TSRs in teacher training. Here, the medium and large associations (cf. Cohen, 1988; see Gignac & Szodorai, 2016) of TSRs with appropriate behavior (recoded, $\bar{\bar{r}}$ = .34), school belonging and engagement ($\bar{\bar{r}}$ = .30), motivation ($\bar{\bar{r}}$ = .28), and general well-being ($\bar{\bar{r}}$ = .27) could be helpful for preservice teachers. Furthermore, these findings should be an incentive for practicing teachers to invest more time in relationship building between teachers and students, even though the direction of this association is yet unclear.. During teacher evaluations, additional support and resources should be provided for teachers identified as needing improvement in this area through professional development programs.

In looking beyond the K-12 school years, future research may further extend the idea of emphasizing TSRs to other age groups and learning situations. Through the Bologna Process (Bonjean, 2018), for example, student-centered learning and, thus, the importance of TSRs have made their way into higher education in the European Higher Education Area. New studies could generate valuable insights into (a) which of the beneficial associations of positive TSRs are applicable in higher education, (b) which variables moderate them, and possibly (c) the impact of the Bologna process on TSRs in higher education.

By and large, we are still lacking a way to effectively translate the findings from psychological meta-analyses on TSRs into educational practice. Stellar examples of how to do this important work in a broader context are the *What Works Clearinghouse* in the US (What Works Clearinghouse, 2023) and the *Clearing House Instruction* in Germany (Clearing House Unterricht, 2023). They review, synthesize, and make research findings accessible to teachers and policymakers to see cutting-edge research applied in schools, psychological practices, and political debates. Especially for TSRs with their high relevance for everyday teaching, such clearinghouses can help disseminate research evidence to teachers and teacher education.

5.8.3 Limitations and Future Directions

5.8.3.1 Difficult Interpretation Due to a High Level of Abstraction

Several limitations are worth mentioning, as reviews of meta-analyses and SOMAs condense all the limitations of the included primary studies and meta-analyses (Polanin et al., 2017). This high level of abstraction applies to all second-order reviews and meta-analyses, leading to several imprecisions in the sample or the definition of constructs and their measurement. For example, even though we systematically excluded samples with special needs or medical conditions, they might still have been included in some of the primary study samples and might therefore be included in our sample of meta-analyses. Thus, this potential inclusion of students with special needs may affect the generalizability of our findings to this group. Furthermore, we applied bootstrapping methods to approximate the sample size when the exact number of students had not been reported. Moreover, one weighted average effect size in the SOMA might be driven largely by a specifically precise meta-analysis with a larger weight. Alternatively, single meta-analyses might lack adequate statistical power, or the vote-counting method could be overinterpreted as an inferential technique. Due to the high level of abstraction, we used a vote-counting method for the narrative review of important moderators (RQ2a). However, this vote-counting method has been identified as potentially controversial. For example, this method has properties that seriously limit the validity of an analysis by potentially ignoring the possibility of Type I and Type II errors and having less statistical power in conditions of low-to-moderate statistical power in the included meta-analyses (Valentine et al., 2010). We therefore backed up our review of moderators by conducting SOMAs. As a consequence, our review of meta-analyses and SOMAs offer a rich and multifaceted evidence base for future research.

5.8.3.2 SOMAs Implicitly Reproduce Language and Other Biases

Like most of the included meta-analyses, our review of meta-analyses used only English search terms. While explicitly welcoming meta-analyses written in another language (i.e., Roorda et al., 2014), we would have missed any meta-analysis that did not use English Keywords, an alternative Title, or Abstract. In addition, some included meta-analyses explicitly excluded primary studies that were not reported in English (e.g., Kincade et al., 2020; Moore et al., 2019). This limitation to English may have implicitly skewed our meta-analytic sample and may have overproportionately included more students, researchers, and research from English-speaking and Western countries. Future meta-analyses should include different language cultures by at least accepting publications in languages other than English and thus improving the generalizability of our research results (see Ryan et al., 2023).

5.8.3.3 Causality and the Lack Thereof

In psychological research, the question of causality is pertinent. Especially in educational contexts, research designs permitting causal inferences (e.g., experiments with randomized control trials) could be difficult to realize due to ethical concerns. For instance, assigning a teacher who may intentionally not care about students in order to create negative TSRs would be unethical. And even so-called “natural” experiments, in which students would be naturally selected into experimental and control groups are rare (see Grosz et al., 2023). Recently, more primary studies and, consequently, most meta-analyses have been based on correlational designs. This use of correlational designs is a growing trend, and recommendations for practice on the basis of correlations should be made with caution (Brady et al., 2023). However, both experimental and correlational studies—and the meta-analyses thereof—can have practical implications (Grosz, 2023). The limitations of experiments (i.e., their internal, external, and construct validity, replicability, or oversimplification of causal effects) need to be considered when evaluating their use compared with a correlational study (Diener et al., 2022). Thus, future TSR research may integrate evidence obtained from experimental *and* correlational studies and use a wide array of established methods, including natural experiments, regression discontinuity designs, or structural equation modeling, to investigate complex causal relations (Diener et al., 2022; Dumas & Edelsbrunner, 2023; Grosz, 2023; Grosz et al., 2023).

Nevertheless, some reviews have already provided insights into the causal relations between TSRs and student outcomes: Cherne (2008) reported evidence on the effect of teacher praise for increasing social and academic behaviors across different age groups and types of disabilities. However, Moore et al. (2019) reviewed 11 experimental studies and concluded that teacher praise cannot yet be considered an evidence-based practice for this student population. Korpershoek et al. (2016) found that classroom management strategies focusing on students' social-emotional development contributed most to the effectiveness of interventions. Finally, Kincade et al. (2020) evaluated approaches for improving TSRs and found that proactive and direct teacher practices had the largest impact. Two additional meta-analyses by Roorda et al. (Roorda et al., 2017, 2021) focused partially on longitudinal primary studies and explicitly tested different design effects. They found slightly different results for longitudinal studies when comparing them with cross-sectional designs. However, these meta-analyses differed in their methodological approaches, construct definitions, and conclusions, so that generalizing about the causes of good TSRs is hardly possible. We therefore argue that there is a need for more studies that allow causal inferences to be drawn on whether (a) positive TSRs lead to positive student outcomes, (b) the TSR-outcome relation is reciprocal, or (c) positive student characteristics facilitate TSR building.

5.9 Conclusions

With this preregistered systematic review of meta-analyses and SOMAs, we summarized 70 years of meta-analytic research on TSR-outcome relations, moderators, and methodological quality. On the basis of 116 effect sizes with more than 2 million prekindergarten and K-12 students, we found substantial relations between TSRs and student outcomes, which were homogeneous across outcomes. Due to their overlap, the psychological constructs underlying reports of TSRs could have properties that are similar to those of the student outcomes. Consequently, TSRs are not largely differentially related to or effective for different student outcomes but share a more generalizable association. In light of bioecological and attachment theories (Bowlby, 1982; Bronfenbrenner, 1979), our findings should encourage teacher educators to convey the importance of TSRs and how TSRs are linked to students' success and well-being.

Whereas we could explain heterogeneity through teacher and student sample characteristics, much of it remained unexplained. We drew two possible conclusions from this result: First, individual characteristics are key for describing the TSR-outcome relations and may call for adaptive interventions that are aimed at improving TSRs (e.g., age-adapted). Second, causes of heterogeneity within and across meta-analyses are not yet fully explored, are complex to uncover (e.g., gender differences; see Craig Aulisi et al., 2023), and remain for future research.

Whereas most meta-analyses exhibited good quality, there is room for improvement, especially in reporting sample statistics and results as well as making data and analytic code openly available. We believe that the current developments in open science, which offer detailed guidelines and tools for ensuring high methodological quality and reproducibility, can lead to a transparent and reproducible research process, establish trustworthiness and reliability, and have an impact on psychological research on TSRs.

6 Study 2

The Relation Between Executive Functions and Math Intelligence in Preschool Children: A Systematic Review and Meta-Analysis Model

Valentin Emslander[a] and Ronny Scherer[b, c]

[a] Luxembourg Centre for Educational Testing (LUCET) at the University of Luxembourg,
Faculty of Humanities, Education and Social Sciences, University of Luxembourg, Luxembourg

[b] Centre for Educational Measurement at the University of Oslo (CEMO),
Faculty of Educational Sciences, University of Oslo, Norway

[c] Centre for Research on Equality in Education (CREATE),
Faculty of Educational Sciences, University of Oslo, Norway

Published in *Psychological Bulletin* [2]

http://dx.doi.org/10.1037/bul0000369

An older version of this study was pre-registered at
https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42017076291
The data, code, and material are available at
https://osf.io/t67qa/?view_only=5ee953125f584687ae61af98b292464b.

[2] The published article is based on my master's thesis (Emslander, 2020) and uses additional data and more elaborate methodological approaches. The numbering of headings, tables, and figures has been adjusted to align with the structure of the present work. The Appendix is integrated in the general Appendix at the end of this thesis.

Abstract

Executive functions (EFs) are key skills underlying other cognitive skills that are relevant to learning and everyday life. Although a plethora of evidence suggests a positive relation between the three EF subdimensions inhibition, shifting, and updating, and math skills for schoolchildren and adults, the findings on the magnitude of and possible variations in this relation are inconclusive for preschool children and several narrow math skills (i.e., math intelligence). Therefore, the present meta-analysis aimed to (a) synthesize the relation between EFs and math intelligence (an aggregate of math skills) in preschool children; (b) examine which study, sample, and measurement characteristics moderate this relation; and (c) test the joint effects of EFs on math intelligence. Utilizing data extracted from 47 studies (363 effect sizes, 30,481 participants) from 2000 to 2021, we found that, overall, EFs are significantly related to math intelligence ($\bar{r}$ = .34, 95% CI [.31, .37]), as are inhibition ($\bar{r}$ = .30, 95% CI [.25, .35]), shifting ($\bar{r}$ = .32, 95% CI [.25, .38]), and updating ($\bar{r}$ = .36, 95% CI [.31, .40]). Key measurement characteristics of EFs, but neither children's age nor gender, moderated this relation. These findings suggest a positive link between EFs and math intelligence in preschool children and emphasize the importance of measurement characteristics. We further examined the joint relations between EFs and math intelligence via meta-analytic structural equation modeling. Evaluating different models and representations of EFs, we did not find support for the expectation that the three EF subdimensions are differentially related to math intelligence.

Keywords: cognitive skills, executive functions, mathematics, meta-analysis, pre-school children

Public Significance Statement: Executive functions (EFs) are key to learning and children who score higher on EFs also show better scores in mathematics. This meta-analysis confirms that children who can better avoid (or *inhibit*) being distracted, *shift* easily between different tasks, or *update* the information they have just learned are also scoring high on math intelligence tests. However, this link between EFs and math intelligence should always be interpreted in light of the measurement characteristics of EFs, as suggested by the moderator analyses. We found evidence to support the idea that the three EFs (inhibition, shifting, and updating) are equally important for math intelligence.

6.1 Introduction

Executive functions (EFs)—a set of mental processes that regulate human cognition and behavior (Miyake et al., 2000; Miyake & Friedman, 2012)—and their three arguably most investigated subdimensions (response inhibition, mental set shifting, and updating of working memory; aka inhibition, shifting, and updating) are considered prerequisites for many cognitive skills, such as reading, and correlate with fluid intelligence (Cassidy et al., 2016; Diamond, 2013; Follmer, 2018). In addition to providing empirical evidence of the relation between EFs and these cognitive skills, researchers have repeatedly shown that EFs are linked to math skills, such as basic number knowledge, calculation, spatial skills, and mathematical reasoning, in schoolchildren and adults (see, e.g., Best et al., 2011; Cragg et al., 2017; Friso-van den Bos et al., 2013; Peng et al., 2016; Yeniad et al., 2013) and play a crucial role in the development of math skills (van der Ven, 2011; van der Ven et al., 2012). Therefore, it is critical to examine the preschool years and comprehensively investigate the constructs that have been found to contribute to the development of mathematical skills. For instance, development of an understanding of numbers begins before school entry (Passolunghi & Lanfranchi, 2012) and is related to EFs (Geary et al., 2019), suggesting that the preschool years are a formative period with rapid development of mathematical skills and EFs (Zelazo & Carlson, 2012). These developmental patterns might even differ between EF subdimensions in this age group (Diamond, 2013; Garon et al., 2008) as some EFs develop earlier than others. As preschool children are on the brink of structured schooling, their EFs constitute an important part of their school readiness and predict their academic success throughout their school careers (Blair, 2002; Diamond, 2013; Duncan et al., 2007). EFs and early mathematical skills are not only core areas of school readiness (the very goal of the preschool years) but also predict academic success later in the school career. Thus, both constructs are crucial to investigate in preschool children due to their age and preschool status.

Following the literature on the differentiation of cognitive skills over time, some researchers have argued that EFs and basic math skills actually measure the same underlying construct in children, as both are strongly linked to intelligence and can be integrated into the Cattel-Horn-Carroll (CHC) theory of intelligence (Ackerman et al., 2005; Allan et al., 2014; Fleming & Malone, 1983; Jewsbury et al., 2016; Roth et al., 2015). Specifically, Jewsbury et al. (2016) suggested subsuming inhibition and shifting under a general speed factor (*Gs*), and updating can be captured by a general memory factor (*Gsm*). Math skills have been categorized as quantitative knowledge (*Gq*), fluid intelligence (*Gf*), and visual processing (*Gv*; Schneider & McGrew, 2018; Uttal et al., 2013). However, other scholars have argued that EFs and math skills measure correlated but distinct constructs (see, e.g., Best et al., 2011; Cragg et al., 2017; Friedman et al., 2006; Peng et al., 2016; Yeniad et al., 2013). These perspectives differ in whether math skills are conceptualized as broader (e.g., students' grades or performance on teacher-constructed math tests) or narrower (e.g., only intelligence tests with math components). Typically, EFs and math skills in preschool children are not measured with standardized questionnaires that require the children to read and write (cf. the Behavior Rating Inventory of Executive Function [BRIEF]; Gioia et al., 2000), and researchers have to resort to innovative ways of assessing EFs and math skills. Thus, EFs might vary in their relation to math skills in terms of the measurement properties (e.g., Allan et al., 2014; Cortés Pascual et al., 2019) or study and sample characteristics (David, 2012; Friso-van den Bos et al., 2013; Peng et al., 2016). Investigating the influence of diverse measurement, sample, and study characteristics on the relation between EFs and math skills can provide valuable information to practitioners who want to streamline the assessment of these constructs and researchers who aim to better understand the nature of the relation.

The extant body of literature also reveals another issue. There is some disagreement about the magnitude of the joint relations of inhibition, shifting, and updating with math skills in preschool children. While some researchers found that all three are key predictors of math skills (e.g., Duncan et al., 2016; Purpura et al., 2017), others found that updating (Friso-van den Bos et al., 2013) or shifting (Jacob & Parkinson, 2015) is superior to the other EF subdimensions. With this equivocal discussion at hand, researchers should avoid jumping to conclusions about the differential effects of specific EFs, as strong claims call for even stronger evidence, and the field of EF research is no stranger to critical discussion (see, e.g., the discussion on the relation between working memory and intelligence; Ackerman et al., 2005; Beier & Ackerman, 2005; Kane et al., 2005; Oberauer et al., 2005). Concerning the joint and individual contributions of inhibition, shifting, and updating to math skills, some of these divergent findings may depend on the ways in which EFs are represented in measurement models, for instance, as single or multiple latent variables or composite factors (Camerota et al., 2020). The common practice of utilizing single, reflective latent variables has been questioned, especially when describing the relations between EFs and math skills. For instance, Nguyen et al. (2019) examined an alternative model to the single latent EF variable model and showed that relations with math skills could also operate via unique components of EFs. To integrate these different approaches, we performed meta-analytic structural equation modeling and examined the effects of EFs on math intelligence jointly rather than separately.

To examine the relations among the three EF subdimensions and narrow math skills, in the present study, we synthesized 363 effect sizes from 47 studies from 2000 to 2021. Most meta-analyses have investigated the relation between a specific EF and math skills (e.g., Allan et al., 2014; David, 2012; Peng et al., 2016) or the relation between EFs and general cognitive abilities in preschool children (Ackerman et al., 2005; Brydges et al., 2012; Jewsbury et al., 2016). However, the present meta-analysis combines these two approaches by focusing on math skills related to general cognitive abilities, that is, math intelligence and the three EF subdimensions. Furthermore, we identified which study, sample, and measurement characteristics may moderate these relations and examined their multivariate nature. To the best of our knowledge, this study is the first to investigate the relation between EFs and math intelligence using a multilevel and multivariate meta-analysis of preschool children.

6.2 Theoretical Framework

6.2.1 Executive Functions and Their Measurement

EFs are a set of general-purpose control mechanisms that regulate human cognition and behavior and are important for self-control and self-regulation (Miyake et al., 2000; Miyake & Friedman, 2012). Miyake and Friedman (2012) characterized these control mechanisms in their EF framework based on four conclusions drawn from the literature on EFs. First, the three EF subdimensions in their framework—inhibition, shifting, and updating—tend to be distinct in adults, showing a diversity factor structure with three distinct factors. Yet the three EFs are correlated, especially in preschool children, yielding a unitary factor for all three EFs (Wiebe et al., 2008). Second, twin studies have suggested that a large proportion of EFs are genetically inherited (Friedman et al., 2008). Third, measures of executive functions can be used to differentiate between clinical and nonclinical behaviors (Young et al., 2009). Fourth, longitudinal studies have shown that EFs are expressed in a stable way over the course of one's life (Mischel et al., 2011), although the structure of EFs develops over time and therefore differs between age groups. Other researchers have investigated the differences between emotionally arousing ("hot"; e.g., due to a punishment or reward) and emotionally neutral ("cool") EFs (Brock et al., 2009; Zelazo & Carlson, 2012). Their findings suggested that, when considered jointly, "hot" EFs are uniquely related to disruptive behavior in preschool children, and "cool" EFs are uniquely related to academic achievement (Brock et al., 2009; Willoughby et al., 2011). This distinction was further corroborated for inhibitory control by Allan et al.'s (2014) meta-analysis, hinting at a variety of executive function dimensions that can be distinguished beyond Miyake et al.'s (2000) work. Additionally, higher-level EFs have been examined, such as "planning," tapping into multiple more basic EF processes (Miyake & Friedman, 2012). For the purpose of this meta-analysis, we further describe the EF framework proposed by Miyake et al. (2000) and its three subdimensions (inhibition, shifting, and updating), the ways EFs are measured in children, and their implications for math skills.

Inhibition is the ability to deliberately override a dominant, automatic, or prepotent response when needed (Miyake et al., 2000; Miyake & Friedman, 2012). Inhibition tasks generally include a conflict between the child's automatic response and the correct response. For instance, in the head-toes-knees-shoulders task (McClelland et al., 2014), a child is instructed to touch their head when prompted to touch their toes and vice versa, as well as to touch their knees when prompted to touch their shoulders and vice versa. One of the dependent variables is the number of correct trials, and the prepotent response that needs to be inhibited is touching the stated body part (for a comprehensive review of EF tasks, see Garon et al., 2008). In math learning, this ability comes into play whenever the obvious answer to a question is not the correct one, and the task demands a more thorough approach (Graziano et al., 2016). When a child is trading cards, stickers, or marbles with a friend, for instance, the friend might ask how many items (i.e., marbles) the child has. A specific answer requires the child to avoid shouting out the answer that comes to mind first (e.g., "many, many marbles") and instead count the marbles one by one and then tell their friend ("I have four marbles"). In addition, for children working on a math question, inhibition is important to stay focused and avoid the urge to give up or be distracted by more attractive alternatives (Clements et al., 2016).

Shifting—also referred to as attention shifting or cognitive flexibility (Diamond, 2013)—is the ability to switch back and forth between mental sets or tasks (Miyake et al., 2000; Miyake & Friedman, 2012). Tasks that measure shifting generally have an element of novelty to them, be it changing rules or adjusting priorities. For example, in the dimensional change card sorting task (Zelazo, 2006), the child is presented with cards displaying various shapes in different colors. The child is asked to sort the cards by color, and after some time, they are asked to sort the cards by shape. The number of correct responses after the sorting rule is changed serves as the dependent variable. This change in sorting rules represents the novelty inherent in shifting tasks. In a math skill-testing situation, shifting is represented by the act of switching to a new question and then applying the appropriate rule to the new question rather than sticking with one rule, although the question demands otherwise. Shifting also comes into play when the solutions to math questions require children to change their perspectives. To be a successful marble trader, for instance, the child in the previous example must constantly switch from receiving marbles and adding them to their count to giving marbles to their friend and subtracting them.

Updating is the ability to constantly monitor and rapidly manipulate the content of working memory (Miyake et al., 2000; Miyake & Friedman, 2012). Although this definition is only slightly different from working memory itself (namely, the ability to store and process information at the same time; Baddeley, 1992), the two constructs can be distinguished regarding several points. In Baddeley's (1992) memory model, working memory is an overarching brain system with three components: the central executive, the phonological loop, and the visual sketch. Both updating and working memory tasks go beyond merely holding content in working memory (Baddeley, 1992) and add the aspect of manipulation (Lehto, 1996), while updating tasks introduce an aspect of monitoring and replacing old pieces of information with new ones. Consequently, working memory can be assessed with a backward digit span. The child is asked to repeat a list of digits (or words) backward, with the number of correct responses in a row representing the dependent variable. This task requires the child to listen to and remember the list of digits (holding them in their working memory) while simultaneously reversing the order of the digits (processing the content of working memory). To make this an updating task, the child can be asked to use only the third last digit, which adds the aspect of monitoring and manipulating, rather than repeating all digits backward. In the present study, we refer to working memory and updating as *updating* due to the similarities in their assessment and considerable ambiguity in primary studies. In the mathematics context, updating is crucial for solving multistep problems. Using the marble trading example, the child needs to remember the preceding steps or preliminary results (e.g., "I have two marbles") to decide on the next step and then update their previous results mentally (Harvey & Miller, 2017). The latter step could be to add one marble to the count (e.g., "My friend gave me one marble. So, I have three now") without having to count the marbles again.

Despite the distinction between multiple EF subdimensions, this distinction may not be clear-cut for complex EF tasks. In fact, such tasks may require multiple EF processes or processes other than executive functioning. Friedman et al. (2008) described this issue of "task impurity" using an example from the Wisconsin Card Sorting Test (WCST): While some WCST tasks require shifting due to changing sorting rules, the tests may also require perceptual and motor skills. Moreover, EF and math tasks may share design and assessment features, such as relying on numbers or operations (i.e., "method overlap"). Therefore, the relations between EFs and other constructs could be biased. To circumvent such bias, latent variable models have been proposed to account for common and unique variations among EF tasks and processes (Camerota et al., 2020).

Several models that describe the structure of EF assessments have been developed and reported in primary studies. These models include, but are not limited to, latent variable models with a single latent EF variable, multiple-correlated latent EF variables, or multiple latent EF variables with a general EF factor and some specific factors (Camerota et al., 2020; Friedman & Miyake, 2017). The question of the unity or diversity of executive functions is a key question in EF research and has initiated many empirical studies that have explored the fit of appropriate measurement models (Karr et al., 2018). However, such measurement models may vary among age groups, and most evidence supports the representation of EFs by a single latent variable for young and preschool-age children (Wiebe et al., 2008; Willoughby et al., 2012). At the same time, some evidence also points to the representation of EFs by a two-dimensional model (Karr et al., 2018; Lerner & Lonigan, 2014). Nonetheless, over time, the various EF processes may become more differentiated in later childhood (Brydges et al., 2014; Lerner & Lonigan, 2014).

6.2.2 Math Intelligence and Its Measurement

In the present study, we conceptualized math skills as part of intelligence due to their inherent fluid and crystallized aspects (Wechsler, 2003; Woodcock et al., 2001). Some theoretical frameworks subsume math skills in subfacets of intelligence. For instance, the CHC theory of intelligence categorizes math skills mainly under the narrow stratum I of fluid intelligence (*Gf*) as quantitative reasoning—that is, the inductive and deductive reasoning abilities involving mathematical relations and properties—and under stratum I quantitative knowledge (*Gq*), which includes mathematical knowledge and achievement (Schneider & McGrew, 2018). Mathematics skills related to geometry and space may also involve visual processing (*Gv*) generally and spatial reasoning specifically (Uttal et al., 2013). This conceptualization of math skills as part of general intelligence, which we refer to as "math intelligence" (e.g., Dweck, 2014; Martens et al., 2006; Rattan et al., 2012), is based on two elements. First, the present sample was comprised of preschool children who had not yet been exposed to formal schooling; thus, school achievement scores or teacher grades were not included. Second, we reasoned that intelligence and EFs are related (Friedman et al., 2006). Although much research attention has been paid to examining the relations between EFs and domain-general intelligence components (Ackerman et al., 2005), little is known about the relation between EFs and math-specific intelligence components.

Further, we differentiate between three subdimensions of math intelligence tasks: calculation and reasoning, basic number knowledge, and spatial tasks. Calculation and reasoning, which encompass numerical operations and applied mathematical problems, can be measured by, for instance, the Applied Problems subscale of the Woodcock-Johnson Test of Cognitive Ability (version III; Woodcock et al., 2001). This is a widely used intelligence test validated for individuals ages 2 to older than 90. The Applied Problem scale includes subitizing, simple subtraction, and calculation tasks, among others. Basic number knowledge, which comprises counting and cardinality, can be measured with simple tasks in which children are asked to count for as long as they can. The highest number to which the children can correctly count serves as the dependent variable. Spatial tasks involve aspects of geometry, shapes, and algorithms. Examples include shape recognition or shape-matching tasks in which children must match a probe with a target shape (e.g., Child Math Assessment; Starkey et al., 2004). In the present study, spatial tasks were later merged with the “other” category due to the small number of available effect sizes.

The three subdimensions of math intelligence were derived from previous theory and research. From a theoretical perspective, they reflect the three aspects of the CHC theory of intelligence adapted to math intelligence in preschool-age children: Calculation and reasoning and basic number knowledge represent fluid intelligence (*Gf*) and quantitative knowledge (*Gq*), while spatial tasks represent math-related visual processing (*Gv*) in preschool children (Schneider & McGrew, 2018). We merged the calculation and reasoning categories into one subdimension following previous factor-analytic research, which indicated that mathematical operations (i.e., calculation) and reasoning load on quantitative knowledge (*Gq;* Parkin & Beaujean, 2012). Friso-van den Bos et al. (2013) specified similar categories (simple arithmetic; counting, and basic understanding of numerical concepts; geometry, shapes, and algorithms) but included additional categories that did not apply to preschool children (e.g., advanced arithmetic or teacher rating). Peng et al. (2016) drew on similar but more differentiated categories of math tasks (i.e., basic number knowledge, whole-number calculations, fractions, geometry, algebra, and word-problem solving). Thus, we derived the math intelligence construct and its three subdimensions in the present study from previous theory and research.

6.2.3 Relations Between Executive Functions and Math Skills

EFs contribute to the process of learning and performing math skills. Not only are there applied settings where EFs and math skills are related, as described previously, but there is also strong evidence that preschool children’s EFs can predict their math skills later (Bull et al., 2008; Clements et al., 2016; Cragg & Gilmore, 2014; McClelland et al., 2014). For researchers and educators, it is crucial to better understand this relationship in preschool children in order to pave the way for success in their (academic) careers (Ancker & Kaufman, 2007; Duncan et al., 2007). In addition, measures of EFs can be deployed to identify children who need more help learning mathematical skills (Clark et al., 2010).

Because EFs are crucial for learning and performing math skills, a lack of EFs can predict difficulties in math skills. This has been suggested by experimental dual-task studies, such as those focused on updating (for a review, see Raghubar et al., 2010). To adequately assist children who are struggling with math skills early on, it is important to clarify the structure of the EF construct and the potential differential contribution of the three EF subdimensions. In contrast to most of the evidence suggesting a unitary EF construct in preschool children (Wiebe et al., 2008), several researchers have shown results that support the notion of distinct EF subdimensions in that age group (Carlson, 2005; Espy et al., 2001). Thus, it is unclear whether EFs are best represented as one construct or as several subdimensions. However, this knowledge could help educators and parents assist struggling children to thrive with tailored instruction later in school (e.g., by reducing updating demands, as in Case et al., 1996) or could help parents provide support and strengthen EFs in a playful way (see, e.g., Hutchison & Phillips, 2018). However, there seems to be no strong evidence of far transfer from, for instance, updating to reading comprehension and arithmetic (Melby-Lervåg & Hulme, 2013) or between EFs (Kassai et al., 2019). Furthermore, although it might be possible to train EFs, the benefits of such training seem to vanish quickly (Takacs & Kassai, 2019). From an economic perspective, knowledge of the EF subdimensions is relevant for the construction of measures for EFs and math skills. If the constructs cannot be differentiated, they could be measured with one instrument, saving time and financial resources as well as reducing the effort required of children. Examining the relation between EFs and math skills will inform not only theorization but also the current assessment practices for EFs in preschool educational contexts.

The unity and diversity of EFs have implications for describing the EF–math intelligence relationship. Different representations of EFs can result in different inferences drawn from this relationship. Nguyen et al. (2019) noted that there is a lack of agreement about the theoretical assumptions underlying the link between EFs and math skills. Specifically, the authors observed the common practice of representing EFs as a single latent variable or a composite score (see also Camerota et al., 2020). With this representation, the relation between EFs and math skills operates through a variable that captures what is common among the manifest EF indicators (Borsboom et al., 2003). Such a model ultimately provides information about the extent to which a unitary EF construct explains variations in the measures of skills in mathematics and sheds light on the joint contribution to math skills. However, if researchers are interested in the unique effects of EFs, the EF–math relations could be described via the residuals of the manifest EF indicators or multiple latent EF variables (Gignac & Kretzschmar, 2017; Nguyen et al., 2019). Finally, models describing the direct relations between multiple manifest or latent EF variables and math skills have the potential to unravel whether differential or uniform effects exist (Arán Filippetti & Richaud, 2017). Overall, the joint relations between multiple EFs and math intelligence can be modeled and interpreted in several ways.

6.2.4 Previous Meta-Analyses

We identified eight meta-analyses that investigated the relation between math skills and EFs as a whole, as well as the specific subdimensions (inhibition, shifting, and updating). This body of research revealed the need to bring together multiple EF subdimensions, a focus on age groups, and conceptualizations of math skills. Table 9 presents an overview of the previous meta-analyses.

In general, studies have found moderate correlations between all three subdimensions of EF and math skills. Allan et al. (2014) found a moderate average correlation ($\bar{r}$ = .34) between inhibition and the development of academic skills, including math skills, in their meta-analysis of preschool children. For shifting, Yeniad et al.'s (2013) meta-analysis yielded a correlation of $\bar{r}$ = .26 between shifting and math skills (operationalized as math performance) in children ages 4–14 years. Five meta-analyses (Cortés Pascual et al., 2019; David, 2012; Friso-van den Bos et al., 2013; Peng et al., 2016; Swanson & Jerman, 2006; see Table 9) focused on the relation between updating and math skills and found slightly higher correlations than those reported previously for inhibition and shifting. Two meta-analyses (David, 2012; Swanson & Jerman, 2006) found medium to large group differences in updating between children and young adults with and without math difficulties in favor of the group without math difficulties. Extending the sample to participants between 3 and 52 years of age, Peng et al. (2016) found a correlation of $\bar{r}$ = .35 between updating (operationalized as working memory) and math skills. Recently, Cortés Pascual et al. (2019) examined the links between composite EF and updating scores and math performance in children ages 6–12 years, finding similar average correlations ($\bar{r}$ = .37) for both links (Table 9).

***Table 9**. Prior Meta-Analytic Findings on the Relation between Executive Functions and Math Skills in Children*

Reference	Description and Method	Significant moderators (selected)	Age range	Examined relation	k_S	$\bar{r}$	*95% CI*
Allan et al. (2014)	Meta-analysis of the link between inhibition and academic skills (encompassing literacy & math)	• EF task type (hot vs. cold) • Test mode (behavioral vs. rating)	2.5–6.5	Inhibition & the development of math as a school subject	75	.34	[.29, .39]
Cortés Pascual et al. (2019)	Meta-analysis of the link of WM and an EF composite score with math performance	• EF subdimensions • Gender composition	6–12	EF composite & math performance	18	.37	[.30, .42]
			6–12	WM & math performance	11	.37	[.29, .45]
David (2012)	Meta-analysis comparing groups with and without math learning difficulties on three aspects of WM	• WM task type (numerical vs. non-numerical) • Age	9–21	Phonological loop d = -0.36	12	–	[-0.58, -0.14]
			9–21	Visuospatial sketchpad d = -0.59	9	–	[-0.87, -0.31]
			9–21	Central executive d = -0.93	17	–	[-1.53, -0.33]
Friso-van den Bos et al. (2013)	Multilevel meta-analysis of the link of inhibition, shifting, and updating with math skills	• Math skills task type • WM task type • Age	4–12	Inhibition & math skills	29	.27	–
			4–12	Shifting & math skills	18	.28	–
			4–12	Visuospatial updating & math skills	21	.34	–
			4–12	Verbal updating & math skills	85	.38	
Jacob & Parkinson (2015)	Meta-analysis looking at cross-sectional and longitudinal data on the EF subdimensions, EF composite, and three age groups accounting for dependencies by clustering standard errors by lead authors	• EF subdimension	3–5	EF composite & math achievement	26	.29	[.23, .36]
			6–11	EF composite & math achievement	34	.35	[.28, .41]
			12–18	EF composite & math achievement	8	.33	[.25, .42]
			3–18	EF composite & math achievement	60	.31	[.26, .37]
			3–18	Inhibition & math achievement	33	.31	[.25, .38]
			3–18	Shifting & math achievement	17	.34	[.24, .44]
			3–18	Updating & math achievement	40	.31	[.22, .39]
Peng et al. (2016)	Meta-analysis of the WM-math-skills-link, accounting for dependencies of effect sizes with robust standard error estimation	• WM type (e.g., verbal, numerical) • Types of math skills	3–52	WM & math skills	110	.35	[.32, .37]
Swanson & Jerman (2006)	Multilevel meta-analysis comparing memory performance of students with and without math disabilities	• Type of cognitive measure • Age	6–13	Verbal WM d = -0.70	43[a]	–	[-0.79, -0.61]
			6–13	Visuospatial WM d = -0.63	13[a]	–	[-0.77, -0.48]
Yeniad et al. (2013)	Meta-analysis of the links between shifting, math performance, and intelligence	• None, due to the small number of included studies	4–14	Shifting & math performance	18	.26	[.15–.35]
			4–14	Shifting & Intelligence	11	.30	[.18–.41]

Note. Larger positive effect sizes indicate a closer relationship between EFs and math intelligence. Common meta-analysis refers to a meta-analysis which takes only one effect size per study into account. k_S = Number of included studies; Age range = Mean age of the youngest and the oldest sample within the meta-analysis in years; EF = Executive function; $\bar{r}$ = weighted mean correlation; d = Cohen's d; WM = working memory. [a] = Number of included effect sizes (*not* studies).

Two meta-analyses examined the connections between math skills and all three EF subdimensions separately. Friso-van den Bos et al. (2013) examined the relation between math skills and inhibition ($\bar{r} = .27$), shifting ($\bar{r} = .28$), and updating (visuospatial: $\bar{r} = .34$; verbal: $\bar{r} = .38$) in samples of 4- to 12-year-olds. Jacob and Parkinson (2015) investigated the link between the three EF subdimensions and math skills (indicated by math achievement) in preschool children and schoolchildren, as well as in adolescents (ages 3–5 years, 6–11 years, and 12–18 years, respectively). Their results showed moderate correlations across the age groups and EF subdimensions, with an overall correlation of $\bar{r} = .31$. However, Jacob and Parkinson (2015) did not examine inhibition, shifting, and updating separately across the three age groups. Thus far, only this meta-analysis and one other have focused on preschool children (Allan et al., 2014; Jacob & Parkinson, 2015). All the meta-analyses broadly conceptualized math skills, including a wide range of subdimensions, measures, and tasks. Two meta-analyses showed substantial heterogeneity in correlations due to variations in math skills and tasks (Friso-van den Bos et al., 2013; Peng et al., 2016). In addition, key aspects of the unique and joint effects of EFs on math skills have not yet been investigated meta-analytically.

6.2.5 Possible Moderators

Previous meta-analyses of the relations between EFs and math skills have suggested a wide array of possible moderators, focusing on differences in age and measurement (see Table 9). In the present meta-analysis, we focused on study, sample, and measurement characteristics as three key sources of heterogeneity (for an extensive list of all coded moderators, see the Coding of Studies section and the codebook in Supplemental Material S1). As in any meta-analysis, study characteristics, such as variables indicating the context of the primary study or whether a primary study was published, can reveal research trends and design issues (see Ferguson & Heene, 2012). Sample characteristics, such as gender composition, country of origin, average age, and whether the children were in kindergarten, preschool, or still at home, have been found to explain substantial heterogeneity (see Friso-van den Bos et al., 2013; Yeniad et al., 2013). In particular, differences in age may lead to divergent findings because the development of EFs and math skills evolves rapidly before children enter school (Allan et al., 2014; Garon et al., 2008). David (2012) and Swanson and Jerman (2006) found that the links between updating and math skills are stronger in younger children than in older ones, and Friso-van den Bos et al. (2013) reported similar findings for shifting. We coded the authors' descriptions of the preschool status of the respective samples. As for gender composition, some studies found no effect on the relation between EFs and math skills (Bull et al., 2008; Clark et al., 2010). In a meta-analysis of children ages 6–12 years, however, Cortés Pascual et al. (2019) found gender composition significantly moderates the relationship of updating and an EF composite with math performance. At the same time, most studies in the present sample did not explicitly examine the possible influence of gender, leaving a research gap in the literature on preschool children. Additionally, there is some evidence that children from different countries and continents vary in terms of performance speed and problem-solving strategies, which might hint at differences in instruction and number language (Imbo & LeFevre, 2009). Furthermore, studies have reported that socioeconomic status is related to EFs (Blair, 2010) and influences their relation to math skills (Noble et al., 2007; Riggs et al., 2006). Finally, measurement characteristics are especially important in the context of preschool children because children at that age generally are not yet able to read or write. Therefore, commonly used measures based on reading and writing (e.g., paper-and-pencil word problem-solving tasks) cannot be applied. As a result, a wide array of alternative measures of inhibition, shifting, and updating have been used, introducing heterogeneity to the pool of measurements for preschool children. Some previous meta-analyses have supported the possible differential relation between different types of EF tasks and math skills (see Allan et al., 2014; Jacob & Parkinson, 2015). Table S6.1 presents a summary and examples of all types of EF tasks examined in the present meta-analysis. Regarding the measurement characteristics of math skills, different task types (e.g., basic number knowledge tasks vs. numerical reasoning tasks) have been proposed (see Allan et al., 2014; Peng et al., 2016) and have been found to moderate the relation between updating and math skills (Friso-van den Bos et al., 2013). Despite this evidence, several key characteristics have been tested to only a limited extent, including whether using a verbal, paper-and-pencil, behavioral, or computer-based test of EFs or math skills makes a difference (see Allan et al., 2014); whether the constructs were tested in a group setting or individually; and whether the test was a performance test or a third-person rating. In the present meta-analysis, we investigated the possible moderating effects of such characteristics on the relation between measures of EFs and math intelligence within and between studies. In the Coding of Studies section, we provide a list of all moderators.

6.3 The Present Meta-Analysis

The present meta-analysis was aimed at elucidating the relation between EFs and math skills—the latter of which is conceptualized as a facet of intelligence—in preschool children. The meta-analysis also aims to quantify variations within and between studies through multilevel meta-analysis, explaining these variations based on the study, sample, and measurement characteristics, and to test the amount of variation in math skills that can be jointly explained by inhibition, shifting, and updating. To answer the research questions, we followed the steps of the Meta-Analysis Reporting Standards (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008), and we utilized advanced meta-analytic approaches, including multilevel meta-analysis and meta-analytic structural equation modeling. Specifically, the present meta-analysis addresses three research questions:

RQ4: To what extent are EFs (represented by a composite and by the three subdimensions of inhibition, shifting, and updating) and math skills (conceptualized as math intelligence) related in preschool children? (Overall correlations)

RQ5: To what extent do these relations vary within and between studies, and which sample, study, and measurement characteristics explain this variation? (Heterogeneity and moderators)

RQ6: To what extent do the three subdimensions of EFs (i.e., inhibition, shifting, and updating) differ in their ability to explain variations in math intelligence, and how much variation do they jointly explain? (Model testing)

6.4 Methods

6.4.1 Literature Search

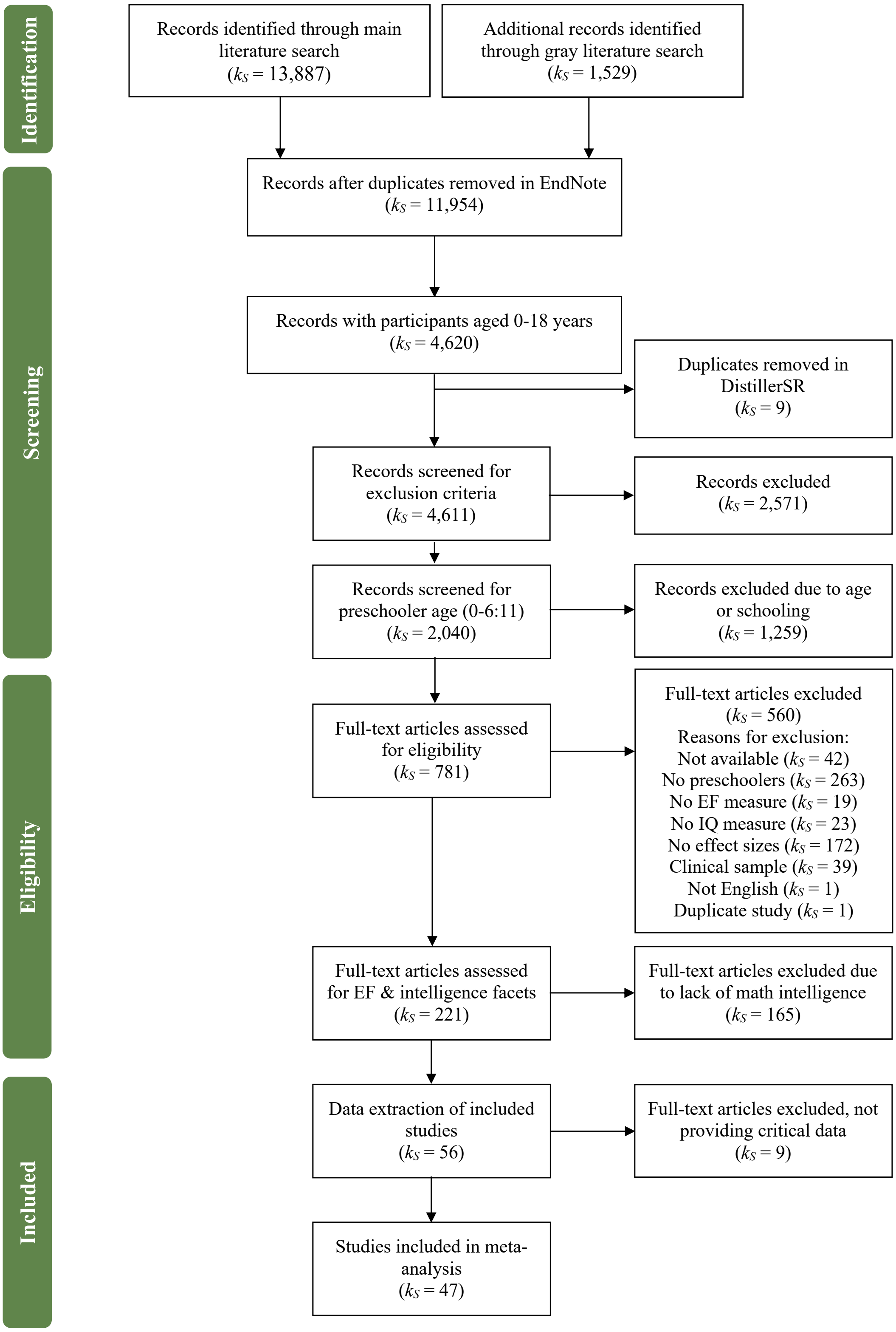
We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. The literature search and parts of the study selection were part of a broader project conducted by the authors that focused on the interrelations between EFs and different facets of intelligence in the general population. For this reason, the first steps of the literature search, screening, and coding procedure included the broad construct of intelligence. The literature search focused on published peer-reviewed articles and dissertations listed in Medline, Embase, ERIC, PsycINFO, ISI Web of Science (Core Collection), and ProQuest Dissertations and Theses. To mitigate potential publication bias, we also searched for gray literature in Google Scholar, OpenGrey, PsyArXiv, and ResearchGate. We performed an initial search of these databases in November 2018 and updated the search in March 2019 and in February–March 2021. The two updates yielded about 1,800 additional publications. We adjusted the search terms to fulfill the requirements of the different databases. The details of these searches, the search terms, and the full search strategy are in Supplemental Material S2. All searches were limited to results from between January 1, 2000, and the dates above; English-language publications; and studies with human subjects

(Boyle et al., 2018). The start date was chosen to ensure that the studies utilized the core EF dimensions defined in Miyake et al.'s (2000) seminal paper. More practical reasons for this start date were to minimize cohort effects and keep the number of studies manageable. Given that math skills are considered an element of intelligence and cognate reasoning skills, we first based the search terms on EFs, intelligence, and reasoning:

> ((executive (function* or dysfunction* or control* or abilit*)) OR (dysexecutive (function* or dysfunction* or control* or abilit*)) OR (cognitive control)) OR (short term memor*) OR (working memor*) OR (updating) OR (Multitasking Behavior) OR (Inhibition) OR (Shifting) OR (Switching) OR ((cognitive or mental) flexibilit*)) AND ((Intelligence) OR (IQ) OR ((Intelligence or IQ) (test* or measure* or examination* or tool* or score* or scoring or scale* or instrument* or assessment* or rating* or evaluat* or questionnaire*)) OR (reasoning) OR (Problem Solving) OR (Decision Making)).

The first step of the search (the main literature search) yielded 13,887 publications, and the second step (gray literature search) yielded 1,529 publications. We removed duplicates and all publications with participants older than 18 years to ensure a focus on children and youths (Bramer et al., 2016), resulting in 4,620 publications. We then transferred these publications to DistillerSR (Evidence Partners, 2021) and deduplicated them once more, resulting in 4,611 publications for further screening. The literature search, deduplication, and screening procedure for all references are summarized in Figure 12. As part of the screening, we extracted relevant publications that focused on math intelligence.

Figure 12. *Flowchart Summarizing the Literature Search and Study Selection*



Identification
Screening
Eligibility
Included
Records identified through main literature search (k_S = 13,887)
Additional records identified through gray literature search (k_S = 1,529)
Records after duplicates removed in EndNote (k_S = 11,954)
Records with participants aged 0-18 years (k_S = 4,620)
Duplicates removed in DistillerSR (k_S = 9)
Records screened for exclusion criteria (k_S = 4,611)
Records excluded (k_S = 2,571)
Records screened for preschooler age (0-6:11) (k_S = 2,040)
Records excluded due to age or schooling (k_S = 1,259)
Full-text articles assessed for eligibility (k_S = 781)
Full-text articles excluded (k_S = 560)
Reasons for exclusion:
Not available (k_S = 42)
No preschoolers (k_S = 263)
No EF measure (k_S = 19)
No IQ measure (k_S = 23)
No effect sizes (k_S = 172)
Clinical sample (k_S = 39)
Not English (k_S = 1)
Duplicate study (k_S = 1)
Full-text articles assessed for EF & intelligence facets (k_S = 221)
Full-text articles excluded due to lack of math intelligence (k_S = 165)
Data extraction of included studies (k_S = 56)
Full-text articles excluded, not providing critical data (k_S = 9)
Studies included in meta-analysis (k_S = 47)

Note. k_S = Number of studies.

6.4.2 Study Selection

The screening procedure comprised three steps—initial screening of titles and abstracts, screening for preschool children, and full text screening (see Figure 12). Table S6.2 in the Supplemental Material presents detailed descriptions of all three screening steps. After following the three steps, we included all studies that fulfilled the following five criteria: (a) The study had to be empirical and report original research findings. (b) The mean age of the sample had to be at or less than 6 years and 11 months, with more than half of the sample not in primary school yet. We chose this high cut-off age to possibly include children from countries such as Germany, where children tend to start first grade in the year they turn 7 years old. In combination with including only samples of which the majority was not yet in school, we tried to create an age-inclusive but predominantly preschool overall sample. (c) The study had to contain at least one sample in which the majority of participants were healthy and not diagnosed with a disorder or medical condition. This criterion maximizes the generalizability of the findings to the general public, as recent meta-analyses have found impaired EFs in children with conditions such as type 1 diabetes mellitus, fetal alcohol spectrum disorder, or high-functioning autism spectrum disorder (Broadley et al., 2017; Kingdon et al., 2016; Lai et al., 2017). (d) The facet of intelligence and at least one EF had to be measured. (e) The zero-order correlations of this relation had to be reported, or sufficient information to compute such an effect size had to be given. We did not include partial correlations or the results of multivariate analyses, as they do not purely represent the relation between EFs and facets of intelligence and, therefore, are not equivalent to zero-order correlations (Lipsey & Wilson, 2001).

We excluded studies during the three screenings if they fulfilled one or more of the following seven exclusion criteria: (a) The study was published before January 1, 2000. (b) The language of reporting was not English. (c) The subjects were nonhuman. (d) The results of the same sample were included in another study. (e) The abstract, full text, and secondary sources reporting the results of the study were unavailable. (f) Intelligence was operationalized as anything other than math intelligence, numerical intelligence, figural intelligence, verbal intelligence, or literacy skills (e.g., emotional intelligence, theory of mind). (g) Results were reported only for samples with a medical condition or a disorder or for samples with a mean age older than 6 years and 11 months.

A total of 221 studies were eligible for the next screening step after we applied the inclusion and exclusion criteria. The screening was conducted by one of the authors together with another graduate coder who had undergone training with a screening manual in 2020.

After the update in 2021, the coding author worked with the DistillerSR (Evidence Partners, 2021) artificial intelligence module (for an introduction, see Baguss, 2020). Between 20% and 30% of the studies in each screening step were double-screened. Any disagreement was settled through discussion among the authors, and the interrater reliability of the coding (κ) ranged from 93% to 98%.

After this screening, we conducted a finer-grained screening to identify which studies examined EFs and math intelligence in preschool children. Table 10 presents an evidence gap map of the frequency of intercorrelations between EF subdimensions and facets of intelligence. As we aimed to sort the studies by EF and intelligence facets, we differentiated among four categories of EFs: inhibition, shifting, updating, and a composite of multiple EFs. Intelligence measures were divided into four categories (subcategories are listed in parentheses): standardized intelligence tests (figurative, verbal, or numerical), math intelligence tests (standardized math test or math-related test), literacy skills tests (reading comprehension, reading fluency, receptive vocabulary, productive vocabulary, language fluency, letter identification, or language comprehension), and working memory tests. Of the 221 studies included in the facet screening, 56 studies reported results for at least one EF and math intelligence and therefore met the inclusion criterion for the present meta-analysis.

Table 10. *Evidence Gap Map of the Number of Studies Measuring EF Subdimensions and Intelligence Facets in Preschool Children*

Intelligence	Inhibition	Shifting	Updating	EF composite	**EF sum**
Standardized IQ test	58	37	48	39	102
Figurative intelligence	41	28	36	29	77
Verbal intelligence	32	17	23	22	57
Numerical intelligence	8	3	6	1	9
Math intelligence	39	21	34	18	56
Standardized math test	32	16	28	18	48
Math-related test	11	9	10	2	14
Literacy skills	86	53	60	40	124
Reading comprehension	8	5	7	5	15
Reading fluency	6	4	4	1	8
Receptive vocabulary	60	37	38	31	89
Productive vocabulary	24	15	15	10	31
Language fluency	11	9	10	4	16
Letter identification	22	10	15	11	32
Language comprehension	15	9	12	3	20
Working memory	90	63	95	21	100
Intelligence sum	160	104	129	65	

Note. Darker shaded cells identify sufficient amounts of research to conduct a meta-analysis. Lighter cells indicate a need for more primary studies. Numbers are based on k_S = 221 studies with preschool children identified in the screening process

6.4.3 Coding of Studies

Three independent coders extracted data from the full texts and coded data on the study, sample, and measurement characteristics for moderator analyses beyond the statistical relation between EFs and math intelligence. The following study characteristics were coded: publication year (2000–2021), publication status (published study vs. gray literature), type of publication (journal article, conference paper, dissertation, or preprint), study background (whether the research question of the study was educational vs. clinical), and study design (whether the data were longitudinal vs. cross-sectional). The coded sample characteristics were children’s mean age (in years), gender composition (percentage of girls in the sample), percentage of school students in the sample (samples with 50% or more school students were excluded), country of sample origin (e.g., the United States, Spain, and China), continent of sample origin (North America vs. Australia, Asia, and Europe), sample size, preschool status of the majority of the children in the sample (prekindergarten, kindergarten, preschool), and socioeconomic status of the sample (low, low to medium, medium, medium to high, high). We coded socioeconomic status based on either an explicit indication (e.g., “families reported a medium-high socio-economic level”; Cueli et al., 2020, p. 239) or, if that was not available, a combination of factors (e.g., parents' education, eligibility for free meals, household income in comparison to the national average) rated by two coders.

For the measurement characteristics of math intelligence, we coded the number of items used to assess math intelligence, math intelligence subdimension (basic number knowledge, calculation and reasoning, spatial, composite), mode of math intelligence testing (paper-and-pencil, verbal, behavioral, verbal and paper-and-pencil, computer-based), whether math intelligence was tested with a performance-based test or a third-party rating, whether math intelligence was tested in a group setting or individually, and the internal consistency of the math intelligence measure (Cronbach's α). Measurement characteristics of EFs included the number of items used to assess EFs, EF subdimensions (inhibition, shifting, updating, or composite of at least two of the other subdimensions), type of EF task (e.g., Stroop task, Simon task, dimensional change task, span task; see Table S6.1 for descriptions and examples of all 15 task types), mode of EF testing (verbal, behavioral, apparatus-based, computer-based), whether EFs were tested with a performance-based test or a third-party rating, whether EFs were tested in a group setting or in an individual setting, and the Cronbach's α value of the EF measure. We considered the study characteristics and sample characteristics moderators at the study level and the measurement characteristics moderators at the effect size level. The three coders, one of the authors and two graduate coders, had undergone training with a coding manual. The two graduate coders showed excellent interrater reliability with the coding author (κ = 91% and 94%) for the 23 studies they had double coded. Disagreement was resolved through discussion after the interrater reliability estimation. The codebook, with a description of all coded variables and their possible values as well as all data, can be found in Supplemental Material S1.

Twenty-five studies did not report crucial information for the meta-analysis, such as the correlation coefficients between EFs and math intelligence. In such cases, we contacted the authors and asked them to provide the missing information. After 2 to 6 weeks and a reminder, 16 of 25 authors provided the missing information. Therefore, it was possible to include 47 studies. From these studies, we extracted 363 effect sizes for a total sample of 30,481 preschool children. The studies included three gray literature publications.

6.4.4 Data-Analytic Approaches

After conducting preliminary analyses for publication bias and influential cases, we applied multilevel meta-analysis, which has two major advantages over conventional meta-analysis, to address RQ1 and RQ2. First, this approach enabled us to include multiple effect sizes per study, in contrast to conventional meta-analyses, and enhanced statistical power (see Soveri et al., 2017). Second, it is possible to examine heterogeneity and moderator effects at the effect size and study levels All data sets as well as the general analytic code for RQ1 and RQ2 can be found in Supplemental Material S1 and S3 respectively. To address Research Question 3, we performed meta-analytic structural equation modeling (MASEM). Supplemental Material S1 and S4 represent the corresponding data set and the analytic code for the MASEM approach, respectively.

6.4.4.1 Publication Bias and Influential Cases

In psychological research, nonsignificant results are less likely to be published—a phenomenon called publication bias (Egger et al., 1997; Ferguson & Heene, 2012). To avoid replicating selectively published findings in the meta-analysis, we assessed the publication bias in the present data. We graphically inspected the 363 effect sizes of this meta-analysis via contour-enhanced funnel plots.

A contour-enhanced funnel plot depicts all effect sizes in comparison to their respective standard errors, and different shades of color indicate their level of statistical significance (Peters et al., 2008). In cases of publication bias, the plot should be skewed to one side at the base of the funnel. Otherwise, the plot should be symmetrical, and studies with larger standard errors should be evenly spread around the base of the funnel (Egger et al., 1997). We performed Begg’s rank correlation test to check for a significant association between the effect size estimates and their sampling variances (Begg & Mazumdar, 1994). Moreover, Fernández-Castilla et al. (2021) recently extended trim-and-fill analyses to multilevel meta-analysis and developed a procedure for obtaining estimates of the number of effect sizes that may have been suppressed due to selection bias (L_0^+ and R_0^+). Estimates above 2 (L_0^+) or 3 (R_0^+), respectively, could indicate the presence of selection bias.

Extending the baseline (correlated and hierarchical effects [CHE]) model, we performed the Precision Effect Test (PET) by estimating the moderator effects of the sampling standard errors, the funnel plot test by estimating the moderator effects of the study sample sizes, and the Precision Effect Estimate with Standard Error (PEESE) test by estimating the moderator effects of the sampling variances.

6.4.4.2 Multilevel Meta-Analyses

For the meta-analysis, we used the R packages metafor (Viechtbauer, 2010), metaSEM (Cheung, 2015b), dmetar (Harrer et al., 2019), robumeta (Fisher et al., 2017), and clubSandwich (Pustejovsky, 2021). To address all the research questions, we conducted the analyses with five data sets: the complete data set (encompassing the entirety of k_{ES} = 363 effect sizes), the inhibition data set (with $k_{ES} = 137$ effect sizes for inhibition as the EF), the shifting data set (with $k_{ES} = 96$ effect sizes for shifting as the EF), the updating data set (with $k_{ES} = 107$ effect sizes for updating as the EF), and the pooled data set with study-level data (i.e., one average effect size per study, $k_S = k_{ES} = 47$). To obtain the average effect sizes of the pooled data set, we aggregated the correlations between EFs and math intelligence to weighted-average, study-level correlations utilizing the aggregate()-function in the R package metafor. This aggregation was based on inverse-variance weighting and accounted for the dependencies between the multiple correlations within the studies (Viechtbauer, 2021). Specifically, we assumed a within-study correlation of $\rho = 0.3$ and specified a compound symmetric structure to represent this dependency.

As noted, most of the 47 studies provided more than one correlation and reported a relation between multiple EF tasks and multiple measures of math intelligence. Consequently, the effect sizes were dependent (Borenstein et al., 2009; Cheung, 2014). Specifically, due to the inclusion of multiple correlations within studies and the inclusion of multiple measures of EFs or math intelligence in the same study, the dependency structure may be correlational and hierarchical (Pustejovsky & Tipton, 2021). To account for these two forms of dependencies, we specified a three-level random-effects model with a constant sampling correlation (ρ) to obtain an estimate of the average correlation between EFs and math intelligence in preschool children and the respective variance components. Pustejovsky and Tipton (2021) referred to this model as the CHE model and emphasized that it is suitable in situations in which the data structure may be correlational and hierarchical, and where an informed estimate of the within-study correlation between effect sizes (ρ) can be obtained. Compared to multilevel meta-analysis (Van den Noortgate et al., 2013), the Level 1 sampling errors in the CHE model are assumed to be correlated. This assumption provides a suitable representation of the nature of the present meta-analytic data. Compared to robust variance estimation (Fernández-Castilla et al., 2020), the CHE model quantifies the variance components at the different levels of analysis and thus could shed light on the level at which heterogeneity exists or is explained by moderators.

In a review of the existing measurement models describing the structure of executive functions, Camerota et al. (2020) found support for average correlations between EF tasks of $\rho = 0.20$–0.40. Drawing from existing reviews and influential studies on the relations between EF tasks (Ackerman et al., 2005; Friedman & Miyake, 2017; Jewsbury et al., 2016) and the existing meta-analyses on the relations between mathematics skills and EFs (see Table 9), we assumed a constant sampling correlation of ρ = 0.30 for the complete, composite, and pooled data sets (as they include a broad range of EF and math tasks), and ρ = 0.40 for the inhibiton, shifting, and updating datasets (as they include EF tasks assessing only one EF subdimension). However, we examined the sensitivity of choosing these values, varying ρ between 0.20 and 0.60 and examining the effects on model parameters.

The CHE model without moderators represents a three-level random-effects model incorporating ρ as follows (Cheung, 2015; Pustejovsky & Tipton, 2021):

Level 1 (Sampling variation): $\hat{\theta}_{ij} = \theta_{ij} + \epsilon_{ij}$,

Level 2 (Within-study variation): $\theta_{ij} = \kappa_j + \zeta_{(2)ij}$,

Level 3 (Between-study variation): $\kappa_j = \beta_0 + \zeta_{(3)j}$,

where the estimator of the ith effect size, $\hat{\theta}_{ij}$, in the jth study is decomposed into the true effect size, θ_{ij}, and the residuals, ϵ_{ij}, at Level 1 with $\epsilon_{ij} \sim N(0, s_j^2)$ and $Cov(\epsilon_{hj}, \epsilon_{ij}) = \rho s_j^2$ for $h \neq i$ and s_j^2 representing the average known sampling variance of study j. At Level 2, the true effect size, θ_{ij}, is then decomposed into the average study-level effect size, κ_j, and some deviation of the effect size, θ_{ij}, from κ_j. The variance $\tau_{(2)}^2$ indicates the within-study heterogeneity, $\zeta_{(2)ij} \sim N(0, \tau_{(2)}^2)$. Similarly, the average effect size, κ_j, is decomposed into the average population effect, β_0, and the study-specific deviation to quantify the between-study heterogeneity, $\zeta_{(3)j} \sim N(0, \tau_{(3)}^2)$. This three-level approach allowed us to quantify three components of variance via restricted maximum likelihood estimation: variance between studies (Level 3), variance between the effect sizes within a study (Level 2), and sampling variance (Level 1; see Van den Noortgate et al., 2013).

To support our choice of the CHE model as the baseline model, we specified alternative models with random effects and a fixed-effects model and compared them to the CHE model. For these comparisons, we examined the following information criteria: Akaike's information criterion (AIC), the Bayesian information criterion (BIC), and the corrected AIC (AICc; Cavanaugh, 1997). Smaller values indicate a preference for a model.

To assess whether variations other than the sampling error existed, we calculated Cochran's Q (Cochran, 1950). A statistically insignificant Q value indicates homogeneity, whereas a statistically significant Q value indicates heterogeneity within the distribution of effect sizes (Ellis, 2010). To complement this overall heterogeneity test, we calculated the level-specific I^2 indices. The I^2 index quantifies the proportion of variance caused by heterogeneity, and it is categorized as low (25%), moderate (50%), and high (75%) heterogeneity (Higgins et al., 2003). The I^2 indices for levels 2 and 3 in a three-level random-effects meta-analysis are defined as follows (Cheung, 2014):

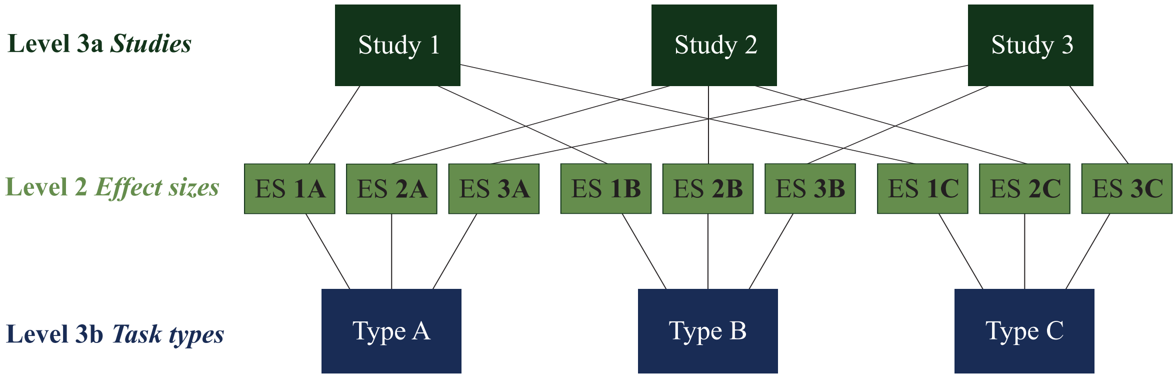
$$I^2_{(2)} = \frac{\hat{\tau}^2_{(2)}}{\hat{\tau}^2_{(2)} + \hat{\tau}^2_{(3)} + \tilde{v}} \qquad \text{and} \qquad I^2_{(3)} = \frac{\hat{\tau}^2_{(3)}}{\hat{\tau}^2_{(2)} + \hat{\tau}^2_{(3)} + \tilde{v}},$$

where $\hat{\tau}^2_{(2)}$ represents the Level 2 variance, $\hat{\tau}^2_{(3)}$ represents the Level 3 variance, and $\tilde{v}$ represents the within-study sampling variance (Cheung, 2015a). We supplemented these indices at the respective levels.

6.4.4.3 Moderator Analyses

To address the second research question, we examined possible moderator effects on the relation between EFs and math intelligence by entering the study, sample, and measurement characteristics into the CHE model. To facilitate interpretation, we mean-centered some of the continuous moderators, z-transformed the children's average age, arcsine-transformed proportional measures (e.g., gender composition coded as the percentage of girls in the sample; see Schwarzer et al., 2019), and dummy-coded categorical and ordinal moderators with more than two categories. The respective mixed-effects meta-regression models with the constant sampling correlation ρ and moderator variables $\mathbf{x}_{ij}$ represented a direct extension of the CHE model (Pustejovsky & Tipton, 2021): $\hat{\theta}_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \zeta_{(3)j} + \zeta_{(2)ij} + \epsilon_{ij}$, with $\epsilon_{ij} \sim N(0, s_j^2)$, $Cov(\epsilon_{hj}, \epsilon_{ij}) = \rho s_j^2$ for $h \neq i$, $\zeta_{(2)ij} \sim N(0, \tau^2_{(2)})$, and $\zeta_{(3)j} \sim N(0, \tau^2_{(3)})$. For moderators with many categories, we extended the CHE model by an additional level of analysis (CHE + model; Pustejovsky & Tipton, 2021)For instance, to assess the moderator effects of the EF task types, we specified a cross-classified random-effects model with the level of task type next to the effect size and study levels (Figure 13). This model allowed us to explicitly quantify the variance between EF task types.

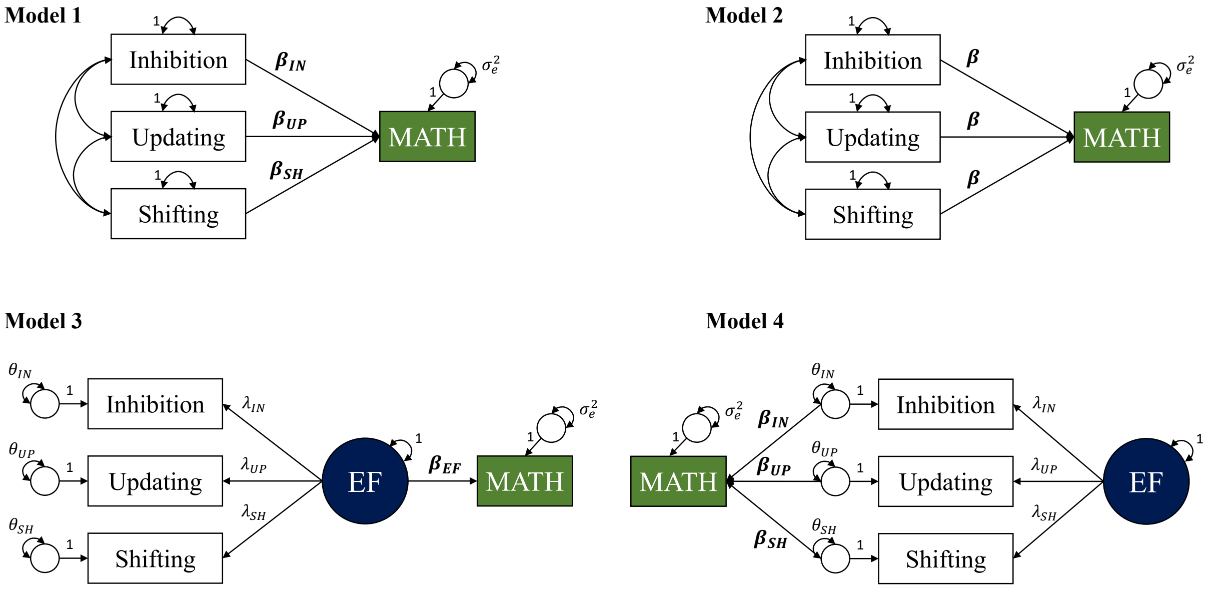
Figure 13. *Exemplary Structure of a Cross-Classified Model*



6.4.4.4 Meta-Analytic Structural Equation Modeling

To examine the joint effects of EFs on math intelligence and test for possible differential effects (RQ3), we performed correlation-based MASEM. To distinguish clearly between the three EF subdimensions (i.e., inhibiting, shifting, and updating), we excluded correlation matrices that contained only composite EF measures. Specifically, we tested four structural equation models (see Figure 14) and evaluated their goodness-of-fit to the meta-analytic data at the study level.

Figure 14. *Meta-Analytic Structural Equation Models (Models 1-4) Representing the Relations between Executive Functions and Math Intelligence*



Note. MATH = Math intelligence, IN = Inhibition, UP = Updating, SH = Shifting, EF = Latent variable representing general executive functions.

Model 1 represented a regression model in which math intelligence was predicted by the three EFs. This model quantified the overall variance explanation, allowing for differential effects of the EF subdimensions. Model 2 further constrained the regression coefficients in Model 1 to equality across the three executive functions. This model assumed equal effects on math intelligence (Marsh, Dowson, et al., 2004). Model 3 represented the three EF subdimensions with a latent variable (i.e., a general factor underlying the EFs; Miyake & Friedman, 2012) that predicted math intelligence. This model captured the covariation among the three EFs in a latent variable and quantified the explanation of joint variance (e.g., Scherer et al., 2018). As in Model 3, Model 4 assumed a latent variable of EFs but used the residuals of the EF subdimensions as predictors of math intelligence. Model 3 focused on a common latent EF variable as the primary source of variance in math intelligence ("common factor model"); Model 4 focused on the residuals of EFs as the sources of variation ("residual factors model"). In other words, these two models targeted common or unique effects on math intelligence (see also Nguyen et al., 2019). Notably, Model 4 estimated the unique effects after we controlled for the common latent EF variable, but not after we controlled for the effect of the common latent EF variable. In fact, adding a direct path between the latent EF variable and math intelligence resulted in a model that was not identified ($df_M = -1$). Following Nguyen et al.'s (2019) procedure, we specified a hybrid model that estimated the direct effect of the latent EF variable *and* the direct effect of one EF residual at a time on math intelligence. This model allowed us to include the latent variable and the residual effect and resulted in an identified model. We refer to these models (one for each EF residual) as *hybrid models*.[3] Supplemental Material S1 and S5 represent the corresponding data set and the analytic code for the hybrid models, respectively.

Correlation-based MASEM allows researchers to synthesize not only single correlations but also entire correlation matrices across primary studies to test multivariate hypotheses (Cheung, 2015a). Several procedures have been developed to perform correlation-based MASEM, and they benefit from these advantages (Sheng et al., 2016). In a review of these procedures, Scherer and Teo (2020) highlighted that they comprise two key elements: pooling of correlation matrices and structural equation modeling based on the pooled

[3] We also examined alternative modeling approaches to identify Model 4 with an additional direct path, such as constraining factor loadings (and ultimately residual variances) to equality or fixing the parameters of the EF measurement model to values that had been estimated from MASEM of a confirmatory factor analysis model with only the EF subdimensions (the two-step estimation approach). However, none of these approaches resulted in robust and trustworthy standard errors and, thus, were not considered for further analyses.

correlation matrix. Two-stage MASEM pools the correlation matrices utilizing maximum likelihood estimation and allows researchers to quantify the between-study heterogeneity of the correlation coefficients (i.e., random-effects modeling) in the first stage (Cheung & Cheung, 2016). In the second stage, the structural equation model is fitted to the pooled correlation matrix utilizing weighted least squares estimation (Cheung, 2015a) but without quantifying the heterogeneity of the model parameters. One-stage MASEM performs these two stages in one step and utilizes maximum likelihood estimation (Jak & Cheung, 2020). In the present meta-analysis, we performed two-stage MASEM to address the third research question and compared the resultant model parameters with those obtained by performing one-stage MASEM as a sensitivity test. At the time of writing, these two MASEM approaches had not been extended to accommodate hierarchical or correlation dependencies among correlation matrices (Jak & Cheung, 2020). Therefore, we estimated Models 1–4 based on the study-level correlation matrices, combining multiple correlations for the same pairs of variables within studies via Fisher's *z*-transformation (Borenstein et al., 2009). Thus, each primary study contributed only one correlation matrix. This data set and the corresponding analytic code can be found in Supplemental Material S1 and S4 respectively.

To evaluate the fit of the four models, we relied on common guidelines for the comparative fit index (CFI), the root mean square error of approximation (RMSEA), and the standardized root-mean-square residual (SRMR; acceptable fit was defined as CFI ≥ .95, RMSEA ≤ .08, SRMR ≤ .10; e.g., Hu & Bentler, 1999; Marsh et al., 2005). However, these guidelines have been validated only on a limited set of specific structural equation models; therefore, we did not apply them as "golden rules" (Marsh, Hau, et al., 2004). Model comparisons were performed via chi-square difference testing and differences in information criteria (Kline, 2015). We performed correlation-based MASEM in the R package metaSEM (Cheung, 2015b).

6.4.4.5 Sensitivity Analyses

To evaluate the sensitivity of the results, we examined the effects of (a) including vs. excluding the study with the largest sample (Nguyen & Duncan, 2019; more than 17,000 participants), (b) different choices of the constant sampling correlation ρ between 0.20 and 0.60, as previously noted (Fisher et al., 2017), (c) one-stage vs. two-stage MASEM (Jak & Cheung, 2020), and (d) unattenuated vs. attenuated correlations with the reliabilities of EF (α_{EF}) and math intelligence (α_{Math}) measures (Hunter & Schmidt, 2004). For (d), we considered only Cronbach's α values as reliability coefficients. If Cronbach's α was not provided in a primary study, we used the average reliabilities of all available EF or math intelligence measures in the data set. These average reliabilities were obtained with a random-effects meta-analysis of all available reliabilities for the respective constructs (see Supplementary Material S3).

6.4.4.6 Evaluation of Primary Study Quality

Evaluating the quality of the primary-level studies is an important step towards crafting an argument for the evidential value of the meta-analytic data (Johnson, 2021). In this meta-analysis, we extracted the following quality indicators from the primary studies: Publication status (*0 = gray, 1 = published*), overall sample size *N*, the reporting of the reliability coefficients for the study sample (*0 = not reported, 1 = reported*), the reliability coefficients of the executive functions and math intelligence assessments (Cronbach's Alpha), and the reporting of *p*-values associated with the EF-math intelligence correlations (*0 = not reported, 1 = reported*). To synthesize these indicators, we created a study quality index as an emergent composite variable via structural equation modelling. This model contained the EF-math intelligence correlation as an outcome variable that is predicted by the quality index (see Supplementary Material S7). To identify the model, the residual variance of the composite variable was fixed to zero, and the regression coefficient connecting study quality and the correlations was fixed to 1 (Henseler, 2021). Creating a composite quality index via SEM has several advantages: First, it allows meta-analysts to combine a diverse set of quality indicators, such as binary, ordinal, count, and continuous variables, into a single index. Second, it handles missing data in the quality indicators efficiently via multiple imputation or model-based procedures. Third, the quality indicators receive individual rather than equal weights. Fourth, it accounts for the hierarchical structure of the meta-analytic data (i.e., multiple samples or effect sizes per study). After estimating the SEM parameters in the R package lavaan (Rosseel, 2012), we extracted the resultant composite index and explored the extent to which it explained heterogeneity in the EF-math intelligence correlations.

6.4.5 Transparency and Openness

This review's protocol was not pre-registered. We followed the PRISMA reporting guidelines for our systematic review and meta-analysis. We share all Supplemental Material, analysis syntaxes, data, the codebook, and the documentation of the systematic search at https://osf.io/vn5pe/?view_only=d4a5f0670ca5469599d13828918957ba (Emslander & Scherer, 2022).

6.5 Results

6.5.1 Description of the Included Studies

The study, sample, and measurement characteristics of all ($k_S = 47$) studies included in the present meta-analysis are described in Table S6.3, and the frequencies and missing values of the categorial moderators are shown in Table S6.4 in the Supplemental Material. Thirty-eight of the 47 studies reported multiple effect sizes, ranging from 1 to 36 per study, with an average of 8. Overall, the present meta-analysis included 65 samples and $k_{ES} = 363$ effect sizes. These effect sizes stem from 44 published journal articles and three gray literature items: one dissertation (Ahmed, 2019), one preprint (Costa et al., 2021), and one conference proceeding (Duncan et al., 2016; see Table S6.3). Figure S6.1 in the Supplemental Material displays the number of effect sizes and studies by the country of sample origin. Most studies (66%) drew on samples from the United States (representing North America; none of the studies were based on Canadian samples), whereas 17% of the studies used samples from Europe, 13% drew from Asia, and 4% drew from Australia. All studies were conducted in an educational context. Most EF measures (98%) and math intelligence measures (99%) were administered in an individual setting, although one study measured EFs in a group setting (Ahmed, 2019), and another measured both constructs in group settings (Träff et al., 2020; see Table S6.4).

Figure S6.2 in the Supplemental Material presents the number of effect sizes by EF subdimension, math intelligence subdimension, EF task type, EF test mode, and math intelligence test mode. The largest number of effect sizes (38%) indicated the relation between inhibition and math intelligence ($k_{ES} = 137$), followed by the relation between updating and math intelligence ($k_{ES} = 107$) and the relation between shifting and math intelligence ($k_{ES} = 96$). Thus, the literature slightly emphasizes the correlation between inhibition and math intelligence. Regarding math intelligence, most effect sizes were associated with the correlations of EFs and basic number knowledge ($k_{ES} = 157$), calculation and reasoning ($k_{ES} = 105$), or a math intelligence composite ($k_{ES} = 93$). The tasks that were most frequently used to measure EFs were Stroop-like tasks ($k_{ES} = 79$) to measure inhibition, dimensional change tasks ($k_{ES} = 83$) to measure shifting, and difficult span tasks ($k_{ES} = 50$) to measure updating. These EF measures were mostly administered verbally ($k_{ES} = 120$) followed by apparatus-based ($k_{ES} = 93$) or computer-based ($k_{ES} = 71$) and behaviorally ($k_{ES} = 62$). Math measures were administered verbally even more often than EF measures ($k_{ES} = 273$; 77% of effect sizes).

Table 11 shows the descriptive statistics of the continuous moderators. The total sample of the present meta-analysis included 30,481 preschool children ranging in age from 3.0 to 6.6 years. Half of the children were girls (*M* = 50.4%, *SD* = 4.2%). Although 46 studies drew on samples of children who had not yet entered first grade, 15% of Mills et al.'s (2019) sample were first graders. All studies were published between 2007 and 2021 (see Table 11).

Table 11. *Descriptive Results of Continuous Moderators*

Continuous moderator	k_S	k_{ES}	*M*	*SD*	*Mdn*	*Min*	*Max*
Publication year	47	363	2016.21	1.13	2017	2007	2021
Gender composition	47	363	50.36	4.23	50	37.59	63.90
Age (in years)	46	355	4.93	0.84	5	3	6.58
Reliability of math intelligence measures	22	169	0.85	0.10	0.88	0.46	0.97
Number of math intelligence items	42	268	32.46	22.51	33	3	120
Reliability of EF measures	24	63	0.87	0.08	0.87	0.56	0.95
Number of EF items	39	237	35.52	40.30	20	3	240

Note. k_S = Number of included studies; k_{ES} = Number of effect sizes; Gender composition = Percentage of girls in the sample.

6.5.2 Preliminary Analyses

6.5.2.1 Identifying a Baseline Model

First, we specified and estimated the baseline model (i.e., the CHE model) with the respective constant sampling correlations. To support this model, we identified alternative meta-analytic models with different variance components. Table 12 displays the fit statistics for the complete data set. Based on these results, we found that the three-level random-effects model with a constant sampling correlation best represented the data (CHE in Table 12). Consequently, this model served as the baseline model for additional moderator analyses and for reporting an overall effect size. The CHE model also served as a baseline model for the inhibition, updating, and shifting data sets. Given the small sample sizes of the pooled and composite data sets, we chose the random-effects model with robust variance estimation and a constant sampling correlation of $\rho = 0.30$.

Table 12. *Model Fit Indices for the Baseline and Cross-Classified Models*

Variable	*df*	LogLik	*AIC*	*BIC*	*AIC*c
CHE model ($\tau^2_{(2)}$ and $\tau^2_{(3)}$ freely estimated, hierarchical nesting, *r* = rho = 0.3)	3	210.51	-415.01	-403.34	-414.94
HE model ($\tau^2_{(2)}$ and $\tau^2_{(3)}$ freely estimated, hierarchical nesting)	3	207.78	-409.57	-397.89	-409.50
RE model ($\tau^2_{(2)}$ = 0 and $\tau^2_{(3)}$ freely estimated)	2	-625.45	1254.91	1262.69	1254.94
RE model ($\tau^2_{(2)}$ freely estimated and $\tau^2_{(3)}$ = 0)	2	162.39	-320.79	-313.00	-320.75
FE model ($\tau^2_{(2)}$ = 0 and $\tau^2_{(3)}$ = 0)	1	-1364.99	2731.98	2735.88	2732.00
CHE+ model($\tau^2_{(2)}$ and $\tau^2_{(3)}$ freely estimated, hierarchical nesting, *r* = rho = 0.3)	4	257.67	-507.35	-491.78	-507.23
HE+ model ($\tau^2_{(2)}$ and $\tau^2_{(3)}$ freely estimated, hierarchical nesting)	4	251.49	-494.98	-479.41	-494.87

Note. LogLik = Log-Likelihood; *AIC* = Akaike information criterion; *BIC* = Bayesian information criterion; *AICc* = corrected Akaike information criterion; CHE model = three-level random-effects model with hierarchical nesting and constant sampling correlation; HE model = three-level random-effects model with hierarchical nesting; RE model = common random-effects model; FE model = fixed-effects model; CHE+ model = cross-classified four-level random-effects model with hierarchical nesting, constant sampling correlation, and executive function task type as an extra level; HE+ model = cross-classified four-level random-effects model with hierarchical nesting and executive function task type as an extra level.

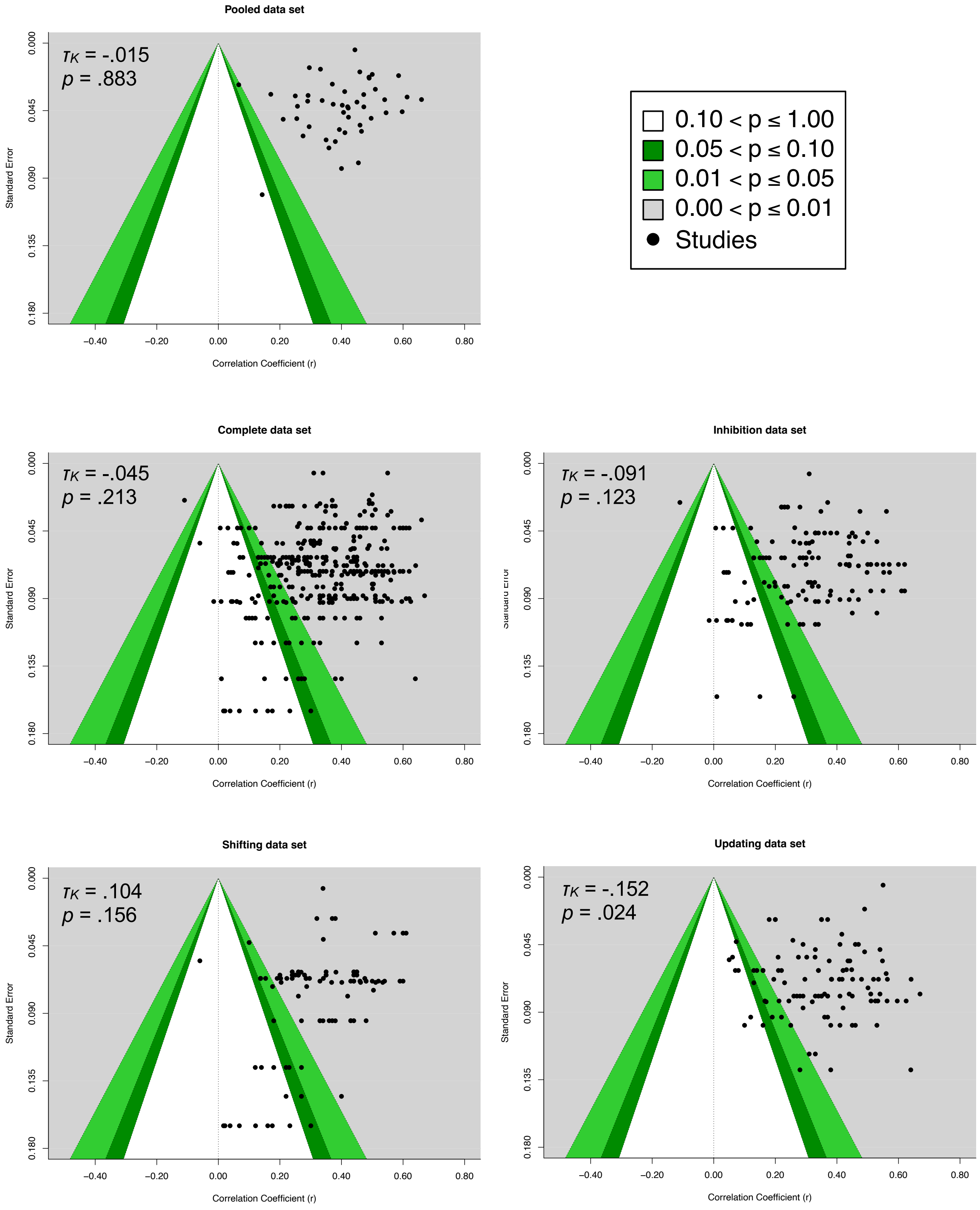
6.5.2.2 Detecting Influential Effect Sizes

Outlier analysis identified two of the 363 effect sizes (Ahmed et al., 2018; Mills et al., 2019) as influential. The confidence interval of these effect sizes exceeded the average correlation confidence interval and displayed considerable difference in fits values (DFFITS > 0.4 standard deviations), Cook's (1977) distance values (> 0.15), and covariance ratio (< 1), suggesting that these effect sizes should be removed for greater precision (Viechtbauer & Cheung, 2010). However, removing these two influential cases yielded a similar average correlation as the meta-analysis that included all effect sizes ($\bar{r} = .347$ vs. $\bar{r} = .341$), showed almost equal population variances ($\tau_{(2)} = .101$ and $\tau_{(3)} = .084$ vs. $\tau_{(2)} = .106$ and $\tau_{(3)} = .085$), and displayed only marginally different confidence intervals (95% CI [.32, .38] vs. [.31, .37]). As a result, we assumed that the meta-analysis was robust against influential cases. Furthermore, both effect sizes stemmed from studies drawing on samples larger than the median sample size of all 47 studies, exhibited good psychometric quality, and reported core study characteristics. Therefore, we refrained from excluding the two effect sizes.

6.5.2.3 Analyses of Publication Bias

The contour-enhanced funnel plots for the five data sets are shown in Figure 15. For the complete data, the precision estimate test indicated the presence of publication bias, because the sampling errors moderated the overall effect, $B = –13.6$, $SE = 0.6$, $p < .001$. Similarly, the PEESE test resulted in a statistically significant moderator effect of the sampling variances, $B = –69.5$, $SE = 3.7$, $p < .001$. The funnel plot test did not suggest any dependency between the effect sizes and the sample sizes, $B = 0.0$, $SE = 0.0$, $p = .58$. The two trim-and-fill estimates were zero; thus, no extra effect sizes were needed to counterbalance potential asymmetry in the funnel plot. However, these trim-and-fill results should be interpreted with caution, as the added value of this method is controversial (Duval & Tweedie, 2000; Peters et al., 2007). The funnel plots for the complete, inhibition, and shifting data sets yielded nonsignificant Kendall's τ values ranging from –.04 to –.10, providing no evidence of publication bias. The Kendall's τ value in the updating data set, –.15, was statistically significant and suggested some evidence of publication bias (see Figure 15). We did not adjust for publication bias in the subsequent models.

Figure 15. *Contour-Enhanced Funnel Plots of all Data Sets*



Note. Larger positive effect sizes indicate a closer relationship between EFs and math intelligence. Correlation coefficients on the x-axis are plotted against the standard errors on the y-axis for every effect size. τ_K = Kendall's τ value.

6.5.3 Main Analysis: Overall Correlations (RQ1)

Table 13 reports the results of the meta-analyses for the five data sets: the complete, pooled, inhibition, shifting, and updating data sets. To address RQ1, we calculated a three-level random-effects meta-analysis with the complete data set, which yielded a moderately positive average correlation ($\bar{r} = .34$) between EFs and math intelligence in preschool children. Similarly, the average correlation was $\bar{r} = .40$ for the pooled data set (see Table 13) and $\bar{r} = .47$ for all composite EF measures, testing at least two EF subdimensions. Figure S6.3 in the Supplemental Material displays a forest plot for the pooled dataset with effect sizes and their confidence intervals (see Supplemental Material S3).

Investigating the three EF subdimensions separately, we found a substantial average correlation between inhibition and math intelligence ($\bar{r} = .30$), shifting and math intelligence ($\bar{r} = .32$), and updating and math intelligence ($\bar{r} = .36$). However, given the overlapping confidence intervals, there was no evidence of statistically significant differences between the three average correlations (see Table 13). The complete and pooled data sets and the corresponding analytic codes are presented in Supplemental Material S1 and S3, respectively.

6.5.4 Heterogeneity and Moderator Analyses (RQ2)

6.5.4.1 Heterogeneity Indices

The effect sizes of the five data sets were highly heterogeneous ($p_Q < .001$; see Table 13 for the exact Q values). For the three-level random-effects meta-analysis with constant sampling correlation of the complete data set, the heterogeneity at the effect size level and the study level was 55% and 35% of the total variance, respectively, which was not due to sampling error. The univariate meta-analysis of the pooled data set showed high total heterogeneity (90%), as did the data set with only composite EF measures (81%). The effect sizes of the inhibition, shifting, and updating data sets also varied substantially within and between studies (see Table 13). The large heterogeneity throughout the data sets motivated subsequent moderator analyses, in accordance with Research Question 2.

***Table 13**. Results of the Meta-Analysis of the Relation of EFs and their Subdimensions with Math Intelligence*

Relation	Data set	$\bar{r}$	95% CI	*SE*	k_S	k_{ES}	*t*	$\tau_{(2)}$	$\tau_{(3)}$	*Q*	$I^2_{(2)}$	$I^2_{(3)}$
EF and math intelligence	complete	.34*	[.31, .37]	.02	47	363	21.28	.106	.085	3690.13*	.55	.35
EF and math intelligence	pooled	.40*	[.36, .44]	.02	47	47	21.70	–	.117	658.69*	–	.93
Inhibition and math intelligence	inhibition	.30*	[.25, .35]	.02	30	137	12.35	.089	.110	955.49*	.36	.54
Shifting and math intelligence	shifting	.32*	[.25, .38]	.03	20	96	9.94	.063	.124	362.17*	.18	.69
Updating and math intelligence	updating	.36*	[.31, .40]	.02	27	107	14.89	.105	.088	1032.50*	.52	.37

Note. Larger positive effect sizes indicate a closer relationship between EFs and math intelligence. $\bar{r}$ = Weighted average correlation, pooled from all effect sizes within the respective study; k_S = Number of included studies; k_{ES} = Number of effect sizes; $\tau_{(2;3)}$ = Heterogeneity at effect size (2) and study level (3), respectively; *Q* = Sum of squared deviations of each effect size's estimate from overall meta-analytic estimate; $I^2_{(2;3)}$ = Heterogeneity indices at effect size (2) and study level (3), respectively.
* $p < .001$

6.5.4.2 Continuous Moderators

To explain the heterogeneity (RQ2), we performed a moderator analysis of the complete data set, as well as separate analyses for the inhibition, shifting, and updating data sets. In the complete data set, we examined a total of seven continuous and 14 categorical moderators. Table 14 reports the results of the continuous moderators in the complete data set, and Table S6.5 reports the results for inhibition, shifting, and updating data sets separately. Publication year, gender composition, age, number of math items, EF reliability, and number of EF items did not moderate the relation between EFs and math intelligence. The only continuous characteristic that marginally statistically significantly moderated this relation was the reliability of the math intelligence measures (*B* = 0.30, *SE* = 0.15, 95% CI [–0.01, 0.60], *p* = .055). The higher the reliability of the math intelligence test, the higher the correlation with EFs tended to be (see Table 14). This moderating effect was exclusive to the effect size level ($R^2_{(2)}$= 3%; $R^2_{(3)}$= 0%).

Table 14. *Results of the Continuous Moderators*

Continuous moderator	k_S	k_{ES}	*B*	*SE*	95% CI	*p*
Publication year	47	363	0.00	0.01	[-0.03, 0.02]	.744
Gender composition	47	363	0.05	0.31	[-0.55, 0.65]	.882
Age (in years)	46	355	0.00	0.01	[-0.02, 0.03]	.758
Reliability of math intelligence measures	23	169	0.30	0.15	[-0.01, 0.60]	.055
Number of math intelligence items	42	268	0	0	[-0.00, 0.00]	.373
Reliability of EF measures	13	63	0.11	0.30	[-0.50, 0.71]	.730
Number of EF items	39	237	0	0	[-0.00, 0.00]	.840

Note. A significant *p*-value indicates that there is a statistical difference between the levels of the moderator. k_S = Number of included studies; k_{ES} = Number of effect sizes; Gender composition = Percentage of girls in the sample.

6.5.4.3 Categorical Moderators

Table 15 displays the results for categorical moderators in the complete data set, and Table S6.6 shows the results for the inhibition, shifting, and updating data sets separately. To ensure meaningful interpretation of the results, moderators with fewer than four effect sizes per category were excluded from the testing (see Bakermans-Kranenburg et al., 2003). This criterion was applied to publication type, study background, and math intelligence performance test vs. third-party rating. To facilitate interpretability, we reduced the categories of the moderators' mode of EF testing (to verbal, behavioral, apparatus-based, and computer-based) and socioeconomic status (to low, medium-low, medium-high, and high).

Table 15. *Results of the Categorical Moderators*

Moderator	k_S	k_{ES}	$\bar{r}$ *[95% CI]*	*SE*	$R^2_{(2)}$	$R^2_{(3)}$	*p*
Baseline ($I^2_{(2;3)}$: .55; .35)	47	363	.34 [.31, .37]	0.02	–	–	–
Country	47	363	–	–	.00	.30	.021*
United states	31	247	.36 [.33, .40]	0.02	–	–	–
Australia	2	20	.32 [.15, .49]	0.09	–	–	–
Spain	2	21	.28 [.11, .45]	0.09	–	–	–
Netherlands	2	11	.34 [.13, .54]	0.10	–	–	–
United Kingdom	2	22	.37 [.20, .55]	0.09	–	–	–
Sweden	2	9	.25 [.07, .42]	0.09	–	–	–
Japan	1	4	.47 [.25, .69]	0.11	–	–	–
China	3	14	.30 [.14, .46]	0.08	–	–	–
Turkey	1	9	.13 [-.16, .42]	0.15	–	–	–
Pakistan	1	6	.07 [-.14, .28]	0.11	–	–	–
Continent	47	363	–	–	.00	.07	.041*
North America	31	247	.36 [.33, .40]	0.02	–	–	–
Asia, Australia, Europe	16	116	.30 [.19, .40]	0.05	–	–	–
Publication status	47	363	–	–	.00	.09	.09
Preschool status	47	363	–	–	.00	.02	.31
Socioeconomic status	38	304	–	–	.00	.00	.77
Study design	47	363	–	–	.00	.00	.58
Math intelligence subdimension	47	355	–	–	.00	.08	.06
Mode of math intelligence testing	46	353	–	–	.00	.00	.48
Math performance testing	46	360	–	–	.00	.00	.46
Math group testing	46	360	–	–	.00	.00	.31
EF subdimension	47	363	–	–	.06	.24	< .001**
Inhibition	32	137	.30 [.26, .33]	0.02	–	–	–
Shifting	19	96	.32 [.25, .40]	0.04	–	–	–
Updating	27	107	.35 [.28, .41]	0.03	–	–	–
Composite	9	23	.45 [.35, .55]	0.05	–	–	–
EF task type	46	356	–	–	.24	.17	< .001**
Composite	6	11	.47 [.38, .56]	0.05	–	–	–
Stroop (Inhibition)	19	79	.24 [.05, .43]	0.10	–	–	–
Simon (Inhibition)	10	21	.40 [.20, .60]	0.10	–	–	–
Slow (Inhibition)	2	5	.26 [.03, .49]	0.12	–	–	–
Shape school (Inhibition)	1	4	.38 [.14, .62]	0.12	–	–	–
Delay (Inhibition)	5	9	.30 [.09, .51]	0.11	–	–	–
Flanker (Inhibition)	4	12	.31 [.10, .52]	0.11	–	–	–
Tap (Inhibition)	6	14	.44 [.23, .66]	0.11	–	–	–
Dimensional change (Shifting)	17	83	.31 [.12, .50]	0.10	–	–	–
Flexible selection (Shifting)	4	14	.37 [.16, .58]	0.11	–	–	–
Random generation (Updating)	2	8	.45 [.23, .67]	0.11	–	–	–
Easy span (Updating)	9	18	.30 [.10, .50]	0.10	–	–	–
Difficult span (Updating)	15	50	.39 [.20, .59]	0.10	–	–	–
Last word (Updating)	1	12	.14 [-.08, .35]	0.11	–	–	–
Search (Updating)	2	16	.26 [.04, .47]	0.11	–	–	–
Mode of EF testing	45	346	–	–	.03	.00	.03*
Verbal	23	120	.31 [.27, .35]	.02	–	–	–
Behavioral	22	62	.37 [.29, .46]	.04	–	–	–
Apparatus-based	20	93	.35 [.26, .43]	.04	–	–	–
Computer-based	13	71	.36 [.27, .46]	.05	–	–	–
EF performance testing	46	349	–	–	.00	.00	.79
EF group testing	45	347	–	–	.00	.00	.63

Note. Larger positive effect sizes indicate a closer relationship between EFs and math intelligence. A significant *p*-value indicates that there is a statistical difference between the levels of the moderator. I^2 *(2;3)* = Heterogeneity indices at levels two and three, respectively; k_S = Number of included studies; k_{ES} = Number of effect sizes; $\bar{r}$ = Weighted average correlation; R^2 *(2;3)* = Variance explanation at levels two and three, respectively.

* *p* < .05. ** *p* < .001

Most categorical characteristics (publication status, preschool status, socioeconomic status, study design, math intelligence subdimension, mode of math intelligence testing, group vs. individual math intelligence test, EF performance test vs. third-party rating, and group vs. individual EF test) did not show a statistically significant moderating effect on the relation between EFs and math intelligence (see Table 15). However, several categorical characteristics—country, continent, EF subdimension, mode of EF testing, and EF task type—statistically significantly moderated this relation for effect size and study level. The country where the study was conducted was identified as a statistically significant moderator at the study level ($R^2_{(2)}$= 0%; $R^2_{(3)}$= 30%). We found nonsignificant correlations in the two studies that drew on samples from Turkey (Söğüt et al., 2021) and Pakistan (Armstrong-Carter et al., 2020). All other countries yielded significant correlations, which were especially high in studies with samples from Japan (Fujisawa et al., 2019) and the United Kingdom (Blakey & Carroll, 2015; Costa et al., 2021). Closely related, the continent on which the study was conducted significantly moderated the relation between EFs and math intelligence almost exclusively at the study level ($R^2_{(2)}$= 0%; $R^2_{(3)}$= 7%). Specifically, studies conducted in the United States on the North American continent yielded higher correlations than studies conducted in countries on the Asian, Australian, and European continents (see Table 15). The EF subdimension was a significant moderator as well, explaining most variance at the study level ($R^2_{(2)}$= 6%; $R^2_{(3)}$= 24%). Inhibition, shifting, and updating correlated significantly with math intelligence. However, the EF composite showed the greatest relation. Examination of the confidence intervals revealed that the EF composite had a significantly stronger correlation with math intelligence than inhibition (see Table 15). The mode of EF testing significantly moderated the relation between EFs and math intelligence, with descriptively higher correlations for behavioral, apparatus-based, and computer-based assessments than for verbal assessments. This moderator explained only a small amount of variance between effect sizes and none between studies ($R^2_{(2)}$= 3%; $R^2_{(3)}$= 0%). The EF task type also showed a statistically significant influence on the relation between EFs and math intelligence and had a moderating effect at the effect size and study levels ($R^2_{(2)}$= 24%; $R^2_{(3)}$= 17%). All EF task types exhibited statistically significant correlations with math intelligence. The Simon and tap tasks revealed the largest correlations for inhibition, the flexible selection tasks for shifting, and the random generation and difficult span tasks for updating (see Table 15). For the data set with composite EF measures, there was a statistically significant difference in the effects between prekindergarten and kindergarten with higher correlations for pre-K children (*B* = –0.07, *SE* =

0.02, $p < .05$). For separate results of the categorical moderators of inhibition, shifting, and updating, please see Table S6.6. Overall, mostly the location of the sample origin and the characteristics of the EF measurement moderated the relation between EFs and math intelligence.

To further test the significant effects of the EF task types, we specified this moderator as another level of analysis in addition to the study level. The resultant cross-classified model, in which the EF task types represented a fourth analytic level, was favored over the three-level model with constant sampling correlation ($\Delta\chi^2[1] = 94.3$, $p < .001$) and resulted in a between-task variance of $\tau^2_{(4)} = .01$, 95% CI [.01, .02]. These results further indicated that the EF–math intelligence correlations varied across EF task types (see Table 12 and Supplemental Material S3).

6.5.5 Meta-Analytic Structural Equation Modeling (RQ3)

To test the hypothesized models that shed light on the extent to which the three EFs explain variation in math intelligence jointly and uniquely (RQ3), we performed a two-stage MASEM utilizing study-level data. The analytic code for the MASEM approach and the hybrid models can be found in Supplemental Material S4 and S5 respectively.

6.5.5.1 Stage 1: Pooling Correlation Matrices

Excluding the correlations between the composite scores of executive functions and math intelligence, we were able to include 38 studies and 120 correlation coefficients based on the data of 26,281 children in these analyses. Before pooling the correlation matrices across the primary studies, we tested them for positive definiteness, a key prerequisite for structural equation modeling (Cheung, 2015a; Kline, 2015). All matrices were positive definite and could be submitted to the pooling stage.

Next, we pooled the correlation matrices under a random-effects model via the tssem1()-function in the R package metaSEM. This model resulted in statistically significant between-study heterogeneity of the correlation matrices ($Q_E[114] = 1039.5$, $p < .001$). The corresponding pooled correlation matrix, between-study variances, and heterogeneity indices are shown in Table 16.

Table 16. *Pooled Correlation Matrix Under the Stage-1 Random-Effects Model (k_S = 23, N = 5402)*

Constructs	1.	2.	3.	4.
1. Math intelligence				
$\bar{r}$	1.000			
2. Inhibition				
$\bar{r}$	0.312	1.000		
95% CI	[0.245, 0.379]			
τ^2	0.020			
I^2	85.0%			
3. Shifting				
$\bar{r}$	0.381	0.294	1.000	
95% CI	[0.311, 0.451]	[0.217, 0.371]		
τ^2	0.007	0.007		
I^2	63.1%	60.2%		
4. Updating				
$\bar{r}$	0.379	0.279	0.245	1.000
95% CI	[0.323, 0.435]	[0.190, 0.367]	[0.178, 0.312]	
τ^2	0.009	0.021	0.001	
I^2	71.7%	84.4%	16.5%	

Note. Larger positive effect sizes indicate a closer relationship between EFs and math intelligence. The pooled correlation coefficients ($\bar{r}$) were based on the stage-1 random-effects model of the two-stage MASEM procedure. k_S = Number of included studies; *N* = Number of included students; 95% CI = Wald 95% confidence intervals, τ^2 = Between-study variation of the correlation coefficients, I^2 = Heterogeneity index.
** $p < .01$.

6.5.5.2 Stage 2: Structural Equation Modeling

Model 1 (Regression Model). Fitting Model 1 to the pooled correlation matrix via the tssem2()-function in metaSEM resulted in positive and statistically significant regression coefficients for inhibition ($\beta_{Inhibition}$ = 0.19, 95% CI [0.14, 0.24]), shifting ($\beta_{Shifting}$ = 0.19, 95% CI [0.12, 0.26]), and updating ($\beta_{Updating}$ = 0.26, 95% CI [0.20, 0.33]). With a residual variance of σ_e^2 = 0.79 (95% CI [0.75, 0.83]), about 21% of the variance in math intelligence can be explained. Dominance analyses indicated that updating contributed to explaining the overall variance with 10%, while inhibition and shifting contributed less (about 6% each). Although this result may indicate that inhibition and shifting have the lowest explanatory power, the differences in the average contribution among the three EFs were not substantial (see Supplemental Material S4). Model 1 was just identified and exhibited a perfect fit to the data.

Model 2 (Regression Model With Equal Regression Coefficients). Model 2 assumed equal regression coefficients for the EFs and exhibited a good fit to the data: $\chi^2(2)$ = 2.9, p = .23, CFI = 0.999, RMSEA = 0.004, SRMR = 0.018. The overall regression coefficient was β = 0.21 (95% CI [0.19, 0.24]), and the residual variance was σ_e^2 = 0.80 (95% CI [0.76, 0.83]). Overall, about 20% of the math intelligence variance was explained by this model. Given that Model 1 was exactly identified, the chi-square difference test statistic corresponded to the chi-square value of Model 2 (χ^2[2] = 3.6, p = .16) and indicated that the equality of the regression coefficients could be assumed. Thus, for the sample of primary studies, there was no evidence of differential effects of the three EFs on math intelligence (RQ3).

Model 3 (Structural Equation Model With a Latent EF Variable). Fitting Model 3 to the pooled correlation matrix, we obtained fit indices that indicated a close-to-perfect fit to the data: $\chi^2(2)$ = 0.6, p = .75, CFI = 1.000, RMSEA = 0.000, SRMR = 0.008. Factor loadings were positive and statistically significant for inhibition ($\lambda_{Inhibition}$ = 0.47, 95% CI [0.41, 0.53]), shifting ($\lambda_{Shifting}$ = 0.47, 95% CI [0.42, 0.53]), and updating ($\lambda_{Updating}$ = 0.54, 95% CI [0.48, 0.60]). The overall regression coefficient was β = 0.65 (95% CI [0.58, 0.72]), and the residual variance was σ_e^2 = 0.58 (95% CI [0.48, 0.66]). Overall, about 42% of the math intelligence variance was explained in this model. Therefore, representing EFs with a latent variable and thus capturing the covariance among the three EFs explained substantially more variance in math intelligence than representing them as distinct but correlated variables (RQ3).

Model 4 (Structural Equation Model With Unique EF Effects). We further specified and estimated Model 4, a model proposed by Nguyen et al. (2019), which describes the unique effects of EFs on math intelligence after a common trait shared among the three subdimensions is controlled for. This model was just identified and exhibited a perfect fit to the meta-analytic data. As in Model 3, the factor loadings were positive and statistically significant for inhibition ($\lambda_{Inhibition}$ = 0.47, 95% CI [039, 0.57]), shifting ($\lambda_{Shifting}$ = 0.49, 95% CI [0.41, 0.57]), and updating ($\lambda_{Updating}$ = 0.51, 95% CI [0.42, 0.61]), and the three EF residuals explained 21% of the variance in math intelligence. Furthermore, the regression coefficients were positive and statistically significant for inhibition ($\beta_{Inhibition}$ = 0.19, 95% CI [0.14, 0.24]), shifting ($\beta_{Shifting}$ = 0.19, 95% CI [0.12, 0.26]), and updating ($\beta_{Updating}$ = 0.26, 95% CI [0.20, 0.33]), and matched those in Model 1. Similar to Model 2, constraining the regression coefficients to equality did not diminish the fit of Model 4 significantly, $\chi^2(2)$ = 2.9, p = .23. After their shared variation was controlled for, the unique effects did not differ significantly among the three EF subdimensions. Finally, Model 3 and Model 4 exhibited a somewhat similar fit to the data,

$\chi^2[2] = 0.6$, $p = .75$. Although both models may explain the joint relations between EF subdimensions and math intelligence, the unitary Model 3 fit slightly better and was more parsimonious, as it represented the data with a smaller number of parameters (Kline, 2015).[4]

Hybrid Models (Structural Equation Models With a Latent EF Variable and One Unique EF Effect). Finally, we estimated three hybrid models specifying the effects of the latent EF variable and one residual effect on math intelligence (Nguyen et al., 2019). These models exhibited a very good fit to the meta-analytic data (e.g., CFIs = 1.000, RMSEAs = 0.000, SRMRs ≤ 0.008; see Supplementary Material S5). In all models, the latent EF variable was positively and statistically significantly related to math intelligence (βs = 0.60–0.68, $p <$ .001); however, each of the three residual effects was statistically insignificant ($\beta_{Inhibition}$ = –0.02, 95% CI [–0.15, 0.08]; $\beta_{Shifting}$ = –0.03, 95% CI [–0.24, 0.10]; $\beta_{Updating}$ = 0.05, 95% CI [–0.06, 0.16]). Thus, in these models, the variation in math intelligence was primarily explained by the latent EF variable but not the EF residuals (between 39% and 45% in total).

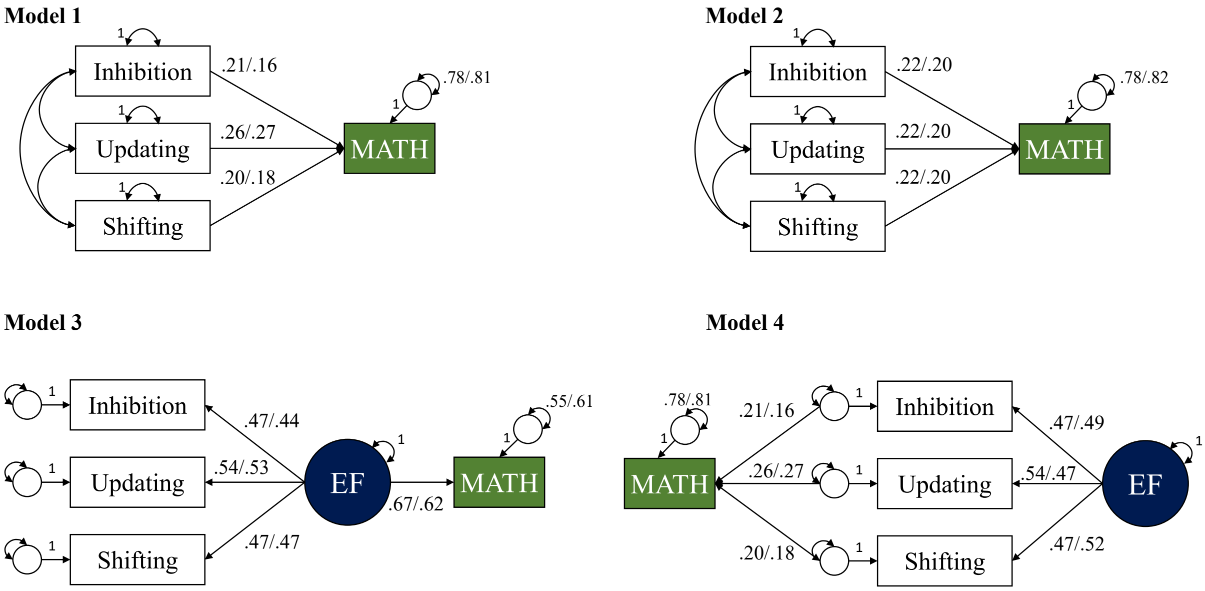
6.5.5.3 Subgroup Analysis

As the final step, we examined the extent to which the four models applied to two key groups of primary studies in the samples: studies with prekindergarten and kindergarten children (k_S = 19, N = 21,401) and studies with preschool children (k_S = 19, N = 4,880). Figure 16 shows the meta-analytic structural equation models with model parameters for these two subgroups. These models were based on two separate Stage 1 random-effects models. Although the variance explanations in Model 1 were similar (20% vs. 21%), the dominance of the EF subdimension updating was more pronounced in the studies that included preschool children. However, given the few studies in this subgroup, the differences in the regression coefficients were not statistically significant (Model 1 vs. 2; for the MASEM syntaxes, see Supplementary Material S4). Model 3 revealed a slightly higher variance explanation by the latent EF variable for the prekindergarten and kindergarten samples (45%) than for the preschool samples (39%). Model 4 supported the dominance of shifting in the preschool samples after we controlled for the latent EF variable. This model did not exhibit a fit superior to that of Model 3 for both study subgroups. Overall, the conclusions drawn for the total meta-analytic samples held for the two

[4] To further substantiate the unique effects of the EFs, we performed a Cholesky decomposition (Dang et al., 2015) and found support for the positive relation between inhibition and math intelligence (b = 0.30, 95% CI [0.25, 0.34]), the positive residual effect of shifting on math intelligence after partialling out inhibition (b = 0.23, 95% CI [0.17, 0.29]), and the unique residual effect of updating after partialling out inhibition and shifting (b = 0.24, 95% CI [0.18, 0.30]). This decomposition explained about 20% in variance and represented an approach for extracting unique relations while controlling for what is shared by the three EF subdimensions. The equality of the regression parameters remained ($\chi^2[2] = 3.4$, $p = .18$).

subgroups of primary studies and revealed a tendency toward more pronounced unique effects of shifting for preschool samples.

Figure 16. *Meta-Analytic Structural Equation Models for Prekindergarten/Kindergarten and Preschool Samples*



Note. Larger positive effect sizes indicate a closer relationship between EFs and math intelligence. MATH = Math intelligence, IN = Inhibition, UP = Updating, SH = Shifting, EF = Latent variable representing general executive functions. The coefficients for the pre-K and K samples are shown first, the coefficients for the preschool samples second. All coefficients were statistically significant ($p < .05$).

6.5.6 *Sensitivity Analyses*

6.5.6.1 Exclusion of a Large Primary Study

Table S6.7 shows the results of all sensitivity analyses. Comparing the results of the complete data set with and without the largest study (both rounded to $\bar{r} = .34$; Nguyen & Duncan, 2019), we found very similar correlations. Additionally, the confidence intervals of the two results were identical (95% CI [.31, .39]). After Nguyen and Duncan's (2019) study was excluded, the heterogeneity between effect sizes remained similar ($I^2_{(2)}$ = 55% vs. 53%; $I^2_{(3)}$ = 35% vs. 33%), as did the variance in the effect size and study levels ($\tau^2_{(2)}$ = 0.01 vs. 0.01; $\tau^2_{(3)}$ = 0.01 vs. 0.01; see Table S6.7 in the Supplemental Material S6).

6.5.6.2 Within-Study Correlation Between Effect Sizes

We further examined the extent to which the parameters of the baseline model were sensitive to the choice of the within-study correlation between effect size (ρ), varying its values between 0.2 and 0.6 (see Supplementary Material S3). Across all data sets, the estimates of the weighted-average effect sizes and the Level 2 and Level 3 variance components were not substantially affected by the choice of ρ. For instance, the effect sizes of the baseline model for the complete data were 0.3405 (ρ = 0.2), 0.3410 (ρ = 0.3), 0.3415 (ρ = 0.4), 0.3419 (ρ = 0.5), and 0.3422 (ρ = 0.6). The heterogeneity indices were similar, $I^2_{(2)}$ = 37.0% and $I^2_{(3)}$ = 53.4% (ρ = 0.2), $I^2_{(2)}$ = 35.0% and $I^2_{(3)}$ = 55.4% (ρ = 0.3), $I^2_{(2)}$ = 33.0% and $I^2_{(3)}$ = 57.4% (ρ = 0.4), $I^2_{(2)}$ = 31.2% and $I^2_{(3)}$ = 59.4% (ρ = 0.5), $I^2_{(2)}$ = 29.3% and $I^2_{(3)}$ = 61.4% (ρ = 0.6). Overall, the sensitivity to the choice of ρ was marginal.

6.5.6.3 MASEM Estimation

Estimating Models 1–4 via one-stage MASEM resulted in the same directions and sizes of effects as those resulting from the two-stage approach (see Supplementary Material S4). Marginal differences occurred in the estimated standard errors and goodness-of-fit indices. However, the conclusions drawn from the four models did not change. The model parameters were not sensitive to the MASEM estimation procedure.

6.5.6.4 Attenuation of Correlation Coefficients

Effect sizes were attenuated for their corresponding reliability or, if no reliability had been reported, for the average reliability of the EF composite ($\alpha_{composite}$ = .86), inhibition ($\alpha_{inhibition}$ = .84), shifting ($\alpha_{shifting}$ = .81), updating ($\alpha_{updating}$ = .90), and math intelligence (α_{math} = .84) measures. Comparing the average correlation between EFs and math intelligence with unattenuated and attenuated effect sizes revealed a higher correlation with attenuated effect sizes ($\bar{r}$ = .34 vs. $\bar{r}$ = .40). The average correlation remained statistically significant, of moderate size, and positive, meaning that its overall interpretation did not change. In addition, the confidence intervals of the results with unattenuated and attenuated effect sizes overlapped (95% CI [.31, .37] vs. [.36, .44]). Descriptive comparisons with the other data sets yielded similar results. It should be noted that the difference between unattenuated and attenuated average correlations ($\bar{r}$ = .40 vs. $\bar{r}$ = .49) for the relation between EFs and math intelligence in the pooled data set was statistically significant (see Table S6.7). As we averaged all available Cronbach's α values for this attenuation, we refrained from interpreting this difference in average correlations to avoid overgeneralizing the unreliability of EF measures. Due to the very similar average correlations, identical confidence intervals, and comparable heterogeneity, we decided to use unattenuated effect sizes, including the three effect sizes reported by Nguyen et al. (2019). Overall, these findings indicate that the meta-analysis was only marginally sensitive to the specified conditions.

6.5.6.5 Influence of Primary Study Quality

The composite index of primary study quality ranged between -0.02 and 0.25 and deviated from a normal distribution at the two tails (skewness = 0.31, kurtosis = 0.87). On average, the quality index was 0.12 (*SD* = 0.05) with a median of 0.12. Using the quality index as a moderator, we found that it explained about 29.1% of the within-study heterogeneity, yet no between-study heterogeneity. Primary study quality moderated the EF-math intelligence correlation significantly (*F*[1, 361] = 20.2, *p* < .001), and the respective effect was positive (*B* = 0.05, *SE* = 0.01, 95% CI [0.02, 0.08]). Hence, primary studies with higher quality tended to report larger correlations between EFs and math intelligence. After controlling for study quality and considering the cluster-robust standard errors, the effects of all other moderators remained and were similar in size and direction (see Supplementary Material S7).

6.6 Discussion

6.6.1 Summary of Key Findings

To contribute to the debate on the relations among cognitive skills in preschool children, we addressed three research questions by performing a meta-analysis of a total of 363 effect sizes from 47 studies. First, we synthesized the relations between EFs and math intelligence in preschool children, differentiating among three EF subdimensions: inhibition, shifting, and updating (RQ1). Second, we identified moderators that explain heterogeneity within and between studies (RQ2). Table 17 displays the key results for RQ1 and RQ2. Third, we examined the differential contributions of inhibition, shifting, and updating to the explanation of variance in math intelligence (RQ3). Overall, the meta-analysis provided evidence of the relation between math intelligence and EFs as a composite as well as separate subdimensions. Moreover, we found that several study, sample, and measurement characteristics moderated this relationship and explained variations within and between studies. Finally, there was no evidence of differential relations between math intelligence and inhibition, shifting, and updating. After the relation between a latent variable representing the EFs and math intelligence was controlled for, none of the EF residuals were related to math intelligence.

Table 17*. Summary of the Main Findings of RQ1 and RQ2*

Relation (data set)	*RQ1: Correlations* $\bar{r}$ (95% CI)	*RQ2: Significant Moderators*
EF and math intelligence (complete)	.34 [.31, .37]	▪ Country: Samples from the United Kingdom and Japan showed large effects, samples from Turkey and Pakistan small, non-significant ones ▪ *Continent*: Larger effect for American samples ▪ *EF Subdimension*: Order of effects, Composite > Updating > Shifting > Inhibition ▪ *EF task type*: Largest effects for Composite, Tap (inhibition), Simon (inhibition), Random generation (updating), and Difficult span (updating) tasks ▪ *Mode of EF testing*: Order of effects, verbal < behavioral ≈ apparatus-based ≈ computer-based testing
Inhibition and math intelligence (inhibition)	.30 [.25, .35]	▪ *Reliability of math intelligence measures*: Math intelligence measures with greater reliability showed closer link to inhibition ▪ *Continent:* Larger effect for American samples ▪ *Inhibition task type*: Largest effects for Simon, Shape school, and Tap tasks ▪ *Mode of inhibition testing*: Order of effects, verbal < behavioral = apparatus-based = computer-based testing
Shifting and math intelligence (shifting)	.32 [.25, .38]	▪ *Reliability of math intelligence measures*: Math intelligence measures with greater reliability showed closer link to shifting ▪ *Publication status*: Larger effect for published studies
Updating and math intelligence (updating)	.36 [.31, .40]	▪ *Math intelligence subdimension*: Order of effects, Basic number knowledge < Calculation & Reasoning < Composite ▪ *Updating task type*: Largest effects for Random generation and Difficult span tasks

Note. Larger positive effect sizes indicate a closer relationship between EFs and math intelligence. $\bar{r}$ = Weighted average correlation

6.6.2 Overall Correlations (RQ1)

The results of the meta-analysis indicated a moderate relation between EFs and math intelligence ($\bar{r} = .34$, $p < .001$; see Table 9). Investigated separately, the three EF subdimensions—inhibition, shifting, and updating—each showed moderately statistically significant relations to math intelligence ($\bar{r} = .30$, $\bar{r} = .32$, and $\bar{r} = .36$, respectively; $p < .001$). Answering Research Question 1, these results indicate that EFs and their subdimensions are substantially related to math intelligence in preschool children similarly to other age groups (see Best et al., 2011; Cragg et al., 2017; Friso-van den Bos et al., 2013; Peng et al., 2016; Yeniad et al., 2013). From a conceptual perspective, this finding corroborates existing frameworks that integrate EFs and math skills (Diamond, 2013; Miyake et al., 2000). As they share some, but not all, variations, EFs and math intelligence can be differentiated but might have several processes in common (Kovacs & Conway, 2016) and be based on a similar collection of cognitive and metacognitive skills, which aid children in solving EF and math intelligence tasks (Van Der Maas et al., 2006). Similar to the assumptions underlying the CHC theory, math intelligence and EFs may represent distinct but related constructs that share some common skills. Jewsbury et al. (2016) argued that EFs and facets of intelligence represent different facets of general cognition. In practice, this implies that assessing one of the two constructs does not make assessing the other redundant, although the performance on one may predict the performance on the other to some extent (van Aken et al., 2019). Thus, reducing the testing burden for young test-takers by focusing on only one construct may come at the cost of insufficient construct coverage. Additionally, this implies that educators should promote EFs and math intelligence in young children, as training in one may not necessarily transfer to the other (Webb et al., 2018). However, EFs might not be causally related to academic achievement, as suggested by a fixed-effect analysis of longitudinal data on the link between inhibition and several achievement measures (Willoughby et al., 2012). Additionally, Willoughby et al. (2019) only found a small within-person association between shifting and working memory and academic achievement from kindergarten to second grade. The transferability of EF training has also been questioned meta-analytically (Melby-Lervåg & Hulme, 2013). More research is needed to test the causality and, thus, the transferability of EF trainings to evaluate the economic sense of a more general use of EFs in teaching (see, e.g., Vaughn et al., 2012).

Overall, previous meta-analytic research that included mostly schoolchildren and adolescents reported somewhat similar relations between EFs as a whole and math skills. For instance, Cortés Pascual et al. (2019) observed an effect ($\bar{r}$ = .37) in primary school children (6–12 years old) that was comparable to the present results. This might imply that there is no drastic decrease in the strength of the relation between EFs and math intelligence from preschool to primary school. After integrating the results for the separate EF subdimensions, however, we observed a slight trend of decreasing relations between math intelligence and inhibition with older age. The relation between inhibition and math skills seemed to be stronger in meta-analyses with younger children and weaker with older children, with correlations ranging from $\bar{r}$ = .34 (Allan et al., 2014; 2.5–6.5 years) to the present result of $\bar{r}$ = .30 (95% CI [.24, .36]; 3–6.5 years) to $\bar{r}$ = .27 (Friso-van den Bos et al., 2013; 4–12 years). The relation between shifting and math skills, while yielding considerably lower average correlations, showed more variation over all the meta-analyses we reviewed. We found an average correlation of $\bar{r}$ = .32 for the preschool sample, which is descriptively larger than those reported by Yeniad et al. (2013; $\bar{r}$ = .26; 4–14 years) and Friso-van den Bos et al. (2013; $\bar{r}$ = .28; 4–12) and comparable to that reported by Jacob and Parkinson (2015; $\bar{r}$ = .34, 3–18 years). The link between updating and math intelligence did not vary substantially, and the correlation of $\bar{r}$ = .36 identified in the present study was similar to those reported in previous meta-analyses ($\bar{r}$ = .37, Cortés Pascual et al., 2019; $\bar{r}$ = .34 and .38, Friso-van den Bos et al., 2013; $\bar{r}$ = .35, Peng et al., 2016; see Table 9). Given these similarities, our narrower conceptualization of math skills as a facet of intelligence did not result in substantially different relations compared to other conceptualizations of these skills.

As noted previously, one key issue challenging the interpretation of the EF–math intelligence relation is task impurity (Nguyen et al., 2019). We addressed this issue in two ways. First, effect sizes based on complex EF tasks that required children to engage in multiple processes rather than a single EF process were meta-analyzed separately as part of the "composite" data set. In this way, possible bias due to task impurity may have been reduced in the meta-analyses of the data specific to the EF subdimensions. Second, as a common practice, we represented EFs with a latent variable (Model 3) to single out what was common across the tasks measuring the three EF subdimensions and what was considered a measurement error (Camerota et al., 2020). However, Nguyen et al. (2019) argued that this representation may not account for possible task impurity in relation to mathematics skills. Thus, we followed their recommendation and examined the EF–math intelligence relation via the residuals, that is, the latent variables describing what is unique to each EF subdimension (Model 4). In this way, common variation that might be due to overlapping EF processes required in the tasks may be controlled for by the single latent EF variable. Although these approaches could counter possible task impurity bias to some extent, alternative latent variable models allowing for cross-loadings or task-specific factors may construct explicitly overlapping EF processes at the level of tasks. To specify and estimate such models, meta-analytic correlations among EF tasks are needed (Scherer & Teo, 2020). Moreover, accounting for the overlap between EF processes does not address the possible impurity due to the processes involved in EF tasks outside executive functioning (Friedman et al., 2008). In this sense, some processes involved in EFs and math intelligence may also be shared. A more detailed differentiation of math intelligence into its subdimensions and a data set providing a sufficiently large sample of correlations at the level of subdimensions are needed to control for this form of task impurity via latent variables.

6.6.3 Heterogeneity and Moderators (RQ2)

We found substantial heterogeneity in the data between and within studies. Five study and measurement characteristics could explain significant amounts of this heterogeneity in the relation between EFs and math intelligence: (a) the country and (b) the continent on which the study was conducted; (c) the EF subdimension; (d) the EF task type; and (e) the mode of EF testing. For the relation between inhibition and math intelligence, heterogeneity was explained by the reliability of the math intelligence and EF measures, the continent, the inhibition task type, and the mode of inhibition testing. Only the reliability of math intelligence measures and publication status moderated the correlation with shifting, and only the math intelligence subdimension and the updating task type explained the variability in the correlation with updating (Table 9). Regarding RQ2, these findings stress the importance of measurement characteristics (compared to study and sample characteristics) as moderators of the relation between EFs and math intelligence in preschool children (Table 9). At the same time, they contrast some of the moderator effects identified in previous meta-analyses with a different conceptualization of math skills.

6.6.3.1 Study and Sample Characteristics

In the present meta-analysis, the only sample characteristics moderating the relation between EFs and math intelligence were the country and the continent on which the primary studies had been conducted. These results were most likely driven by two studies, drawing on samples from Turkey (Söğüt et al., 2021) and Pakistan (Armstrong-Carter et al., 2020), which found nonsignificant correlations between EFs and math intelligence. This finding aligns somewhat with the notion that the involvement of EFs in math intelligence differs between children from different countries and perhaps cultures (Friso-van den Bos et al., 2013). However, we refrain from further interpreting cultural differences in these skills as the only reason for these effects. To the best of our knowledge, such cultural effects have yet to be examined in detail. Moreover, 31 of the 47 primary studies included in the present meta-analysis were conducted in the United States, limiting the possible inferences that could be drawn regarding country differences.

Several promising moderators did not yield statistically significant results. For example, children's age did not statistically significantly explain any heterogeneity in the present meta-analysis. Similarly, Cortés Pascual et al. (2019) found that age does not significantly moderate the EF–math relation in children between 6 and 12 years of age. One explanation for these results could be that the age range was restricted in the present meta-analysis, as well as in previous ones (i.e., Allan et al., 2014; Cortés Pascual et al., 2019; Friso-van den Bos et al., 2013; Yeniad et al., 2013). This result contrasts with that of David's (2012) meta-analysis, in which age was identified as a significant moderator. However, we can observe the previous trend of younger children exhibiting stronger relations of inhibition with math intelligence than older children when reviewing previous meta-analyses. In line with previous studies (Bull et al., 2008; Clark et al., 2010), but contrary to one of eight prior meta-analyses (Cortés Pascual et al., 2019), the gender composition of the samples did not show any moderator effects, probably due to the small variation in this characteristic between studies. These results are very plausible, as even in the clearest case of gender composition in education—namely, single-sex education—girls and boys are likely to show large differences other than interests (Eliot, 2013).

6.6.3.2 Executive Function Measurement Characteristics

Three characteristics of the EF measurement moderated the relation between EFs and math intelligence in preschool children. First, this relation varied slightly between the EF subdimensions under investigation, such that the correlation was descriptively weaker for inhibition, moderate for shifting, and stronger for updating (see RQ2). Similar differences in the correlations with math skills between the EF subdimensions have been found in previous meta-analyses (Cortés Pascual et al., 2019; Friso-van den Bos et al., 2013; Jacob & Parkinson, 2015). In the moderator analysis, we also examined a composite EF score that summarizes all EF measures, examining at least two of the three EF subdimensions. With a high correlation with math intelligence, the composite EF is likely to be the driver of this moderation effect. The relation between this score and math intelligence was stronger than the relations for each subdimension individually. This might be because tests yielding a composite score for EFs tend to be similar in design to tests of general cognitive ability and, therefore, may overlap substantially with math intelligence tests (see, e.g., Yeniad et al., 2013). Additionally, these findings are in line with the notion that the differentiation among the three EF subdimensions might be greater in the later years of childhood, explaining the small differences between subdimensions in the present preschool sample.

Second, the EF task type significantly explained heterogeneity within and between studies. The extent of the explained heterogeneity was so great that a cross-classified random-effect model in which EF task type was the fourth level suggested substantial variation across the EF task types. The present meta-analysis categorized more than 70 EF tasks into 15 distinguishable categories. The strongest relations to math intelligence were found for the composite task type (which measures at least two EF subdimensions), tap and Simon tasks (which measure inhibition), and random generation and difficult span tasks (which measure updating). These results are not surprising; the composite, random generation, and difficult span tasks are very close in structure to math intelligence tasks and require individuals to handle numbers, especially in updating tasks such as the backward number span task (Peng et al., 2016). This is another example of task impurity (Friedman et al., 2008), which describes the common phenomenon that one task requires multiple EF subdimensions, making a clear-cut interpretation more difficult. Overall, the results reflect the vast variety of EF tasks used in the field and simultaneously draw attention to inconsistencies related to measuring EFs at an age at which children cannot read.

Third, the mode of EF testing moderated the relationship between EFs and math intelligence at the effect size level, with very small descriptive differences among verbal, behavioral, apparatus-based, and computer-based testing. Looking more closely into the single EF subdimensions, we found that this moderation effect occurred only in inhibition tasks, and not in shifting or updating tasks. Therefore, inhibition drives the overall moderator effect. The higher effects for behavioral, apparatus-based, and computer-based testing might also be due to their overlap with common actions involving math intelligence. The source for this speculation is the similarities between the behavioral tap task and counting on one's fingers, for instance, or the computer-based shape school task and preschool geometry puzzles. Although this finding might raise the question of how testing mode-general vs. mode-specific inhibition is, we should not overinterpret it, as the mode of EF testing explained only a very small amount of variance between effect sizes and none between studies. Previous meta-analyses had somewhat similar findings for working memory domains (David, 2012; Friso-van den Bos et al., 2013; Peng et al., 2016). Overall, this indicates the disadvantage of verbal (inhibition) testing compared to the other modes of testing for preschool children.

6.6.4 Model Testing: Explaining Variations in Math Intelligence (RQ3)

Utilizing the analytic framework of correlation-based MASEM, we substantiated the evidence that the EF subdimensions statistically significantly and jointly explain variation in math intelligence via testing and comparing several structural equation models (Models 1–4). Examining the joint effects via multiple regression (Model 1), we found that all EFs exhibited statistically significant relations with math intelligence and that these relations did not differ among the EF subdimensions (Model 2). Representing EFs as a single latent variable (Model 3), we obtained a strong and positive effect on math intelligence, an effect much larger than the EF factor loadings. An alternative interpretation of this finding might be that math intelligence tests are better single measures of general EF than any single EF measure, because the common variance of inhibition, shifting, and update, can be better explained by math intelligence than with any one of the EF subdimensions. Further testing with a broader set of measures representing the constructs more comprehensively is needed to corroborate this finding conceptually.

Finally, we examined the unique effects via the EF residuals (Model 4) and found positive and statistically significant path coefficients that did not differ between the three EF subdimensions. Overall, the variance explanations in math intelligence were substantial, ranging from 21% to 42%. Of the three EFs, updating descriptively, but not statistically significantly, explained more variance than inhibition and shifting. This trend may point to possible overlaps in the processes or assessment methods involved in updating and math intelligence tasks and supports previous findings (Bull & Lee, 2014). Despite the interpretation of these overlaps as method effects or commonalities among the processes involved in EFs, another interpretation lies in the domain specificity of EFs, bringing together these two elements. Specifically, Peng et al. (2018) sought evidence of the domain specificity of working memory dimensions (numeric, verbal, and visuospatial) and argued that working memory operates through domain knowledge, skills, and procedures in task-specific situations. This perspective aligns with the situational model by Doebel (2020), who considered the development of EFs to be a set of skills activating specific knowledge, beliefs, values, norms, and preferences. Thus, EFs are directed at meeting the demands and goals of specific tasks. In this sense, the shared EF process and specific task demands cannot be strictly separated.

From a substantive perspective, several findings are worth noting. First, the evidence of uniform relations with math intelligence if the three EF subdimensions are considered jointly may indicate the unitary (rather than multifaceted) nature of the EF construct. Gonzalez et al. (2020) argued that uniform relations to external criteria are key to establishing that constructs may represent the same underlying processes or traits. Although the present results allow for this conclusion, this criterion is not sufficient to establish that the three EFs are the same. Additional evidence of the factor structure of EF measures across samples, contexts, and time is needed to substantiate this conclusion (Marsh et al., 2019). However, to generate such meta-analytic evidence, item- or subscale-level correlation matrices are needed, and the diversity of EF assessments across studies limits the possibility of such research syntheses (Karr et al., 2018). The present meta-analysis revealed small-to-moderate correlations among the three EFs. At the same time, there is some existing evidence of the unitary nature of EFs for preschool children (Garon et al., 2008; Nelson et al., 2016). These observations highlight that relations to external criteria (i.e., math intelligence) may not represent the only source of information for or against the unity of the three EFs.

Second, the MASEM results did not support the differential relations to math intelligence identified in experimental and longitudinal studies of EFs (Cragg & Gilmore, 2014; Jacob & Parkinson, 2015). One possible reason for these divergent results may be the way in which math skills were operationalized in the present study. In contrast to measures of school performance and educational achievement, we conceptualized math skills within the intelligence framework, defining them as math intelligence. In this sense, the uniform relations between EFs and math intelligence align with mounting evidence of substantial relations between EFs and general intelligence (Jewsbury et al., 2016). The strength of these relations is comparable to that reported for EFs and other intelligence measures (Friedman & Miyake, 2017). Thus, to meaningfully interpret the EF–math link, researchers must carefully consider the conceptualization of math skills.

Third, Model 3 revealed a strong relation between the single latent EF variable and math intelligence, with a regression coefficient that was higher than the factor loadings of the EF subdimensions. Reviewing the data from a smaller set of primary studies, Nguyen et al. (2019) obtained the same results as we did in the present meta-analysis. This result may have several explanations and implications. From a substantive perspective, the results may highlight that the processes common to the three EFs and the processes captured by the math intelligence assessment tasks overlap substantially (Kovacs & Conway, 2016). In other words, the two constructs may be based on a common set of cognitive and metacognitive or mutually reinforcing skills that support children in performing the respective assessment tasks (Van Der Maas et al., 2006). Similar to the notion of "*g*" in describing what is considered common across intelligence tests, these overlapping processes may also be considered indicators of a general factor underlying EFs and math intelligence (Webb et al., 2018). Identifying what is common across the EF subdimensions and math intelligence could aid in understanding possible transfer effects between the two constructs (Melby-Lervåg et al., 2016). From a measurement perspective, the high structural parameter connecting the single latent EF variable and math intelligence may also be due to a misrepresentation of the EF construct. Despite a very good fit to the meta-analytic data, the reflective measurement model in Model 3 may not represent the true EF structure. In fact, Rhemtulla et al. (2020) showed that common factor models can introduce severe bias to the structural parameters in structural equation models, for instance, if the true model is a formative rather than a reflective measurement model. Similarly, Camerota et al. (2020) argued for considering alternative representations of EFs, for instance, as composites.

Fourth, after what was shared among the EF subdimensions was controlled for, the residuals were statistically significantly and positively related to math intelligence in Model 4. This finding is in line with some of the results reported by Nguyen et al. (2019). Of the nine empirical studies the authors reviewed, two supported Model 4 and showed the same pattern we observed in the meta-analysis. The positive relations between the EF residuals and the measure of math intelligence indicate that unique effects might exist. However, these unique effects did not differ among the three EF residuals, possibly due to the limited number of studies included. This finding calls into question the extent to which the possible unique processes underlying the three EF subdimensions are substantively unique. Although there was a slight preference for Model 3 over Model 4 in the pooled meta-analytic data set (Model 3 was more parsimonious and showed a descriptively better fit than Model 4), both models described the data well and shed light on different aspects of the EF–math intelligence relation.

To further test for the uniqueness of the EF effects, we also adopted Nguyen et al.'s (2019) hybrid models, in which the relationships between the EF residuals and math intelligence were controlled for the effect of the latent EF variable—that is, what is common to all three EF subdimensions. These models supported the dominance of the latent EF variable, as none of the residual effects existed. Similar to Nguyen et al.'s (2019) observations, "latent EF was a very robust predictor of math" (p. 282), and this latent variable accounted largely for the EF–math intelligence relationship.

Fifth, we observed a tendency toward more differentiated relations among the EF subdimensions and math intelligence for the preschool samples than for the prekindergarten and kindergarten samples. Although age ranges in preschool, prekindergarten, and kindergarten differ between countries, this result may inform the discussion on the differentiation of EFs and other cognitive skills over time (Lerner & Lonigan, 2014). However, a much broader age range is needed in future meta-analyses to investigate the possible breaking points at which differentiation occurs.

From a methodological perspective, the MASEM approach offered a powerful procedure for jointly describing and examining the relations among the four constructs and exploring the fit of different models. This allowed us to obtain additional evidence of the relations between EFs and math intelligence beyond univariate relations (Cheung, 2015a). In addition, we were able to evaluate the fit of structural equation models representing different assumptions (Cheung & Cheung, 2016). Specifically, in the meta-analysis, we followed extant modeling approaches, assuming either correlated but distinct EFs or a unitary EF construct (Friedman & Miyake, 2017). However, as noted previously, correlations among EF and math intelligence indicators at the level of tasks or task paradigms (Lehtonen et al., 2018) could shed more light on the structure of EFs and the extent to which math intelligence may be incorporated into this structure. Based on such meta-analytic data, finer-grained models could be tested, such as models that assume a general factor underlies EFs and math intelligence and specific factors representing their unique components (Friedman et al., 2008). In this way, the relationship between EFs and math intelligence could be described factor-analytically and may inform existing factor models that incorporate EFs and other intelligence measures (Jewsbury et al., 2016; Webb et al., 2018). However, the preference for one or another factor model depends on the assignment of the indicators to the respective latent variables (Ackerman et al., 2005). Therefore, establishing clear EF assessment frameworks and examining the sensitivity toward the choice of alternative frameworks are key (Oberauer et al., 2005).

Finally, although the choice of the four models and the hybrid models we tested via MASEM was based on different theoretical assumptions on the EF–math intelligence relationship, we could not support the preference for one model over the others. In this sense, all four models were useful for examining the joint relations between the EF subdimensions and math intelligence from different perspectives. Models 1 and 2 focused on direct joint relations and their equality across EF subdimensions; Models 3 and 4 focused on the effect of a single EF variable or, respectively, the unique effects of the EF subdimensions. Although these models are commonly specified in EF research and exhibited a very good fit to the meta-analytic data, they may oversimplify the complex relationship between EFs and math skills and be subject to task impurity bias (Camerota et al., 2020; Nguyen et al., 2019). Therefore, we encourage researchers in the field to evaluate multiple models rather than a single model for describing the EF–math relationship, consider alternative representations of the respective EF and mathematics constructs, such as composite scores or formative measurement models

(Rhemtulla et al., 2020), and examine in greater detail the processes that are common or unique to EFs and math skills (Cragg & Gilmore, 2014; Kovacs & Conway, 2016).

6.6.5 Limitations and Future Directions

Several limitations of the meta-analysis are worth noting. First, there are several limitations inherent in the inclusion/exclusion criteria and the language restriction. Thus, the results are limited to preschool children without a diagnosed medical condition or disorder. Thus, the results do not generalize across all possible samples of preschoolers and do not allow one to test specific disadvantages in EFs and math intelligence due to medical conditions or disorders. Including clinical studies might have been especially interesting, because EFs are impaired in children with specific disorders (Broadley et al., 2017; Kingdon et al., 2016; Lai et al., 2017). As we restricted our search to English records, a mono-language bias might be present, which limits the generalizability of our results (Johnson, 2021). Future meta-analytic research should, therefore, test the robustness of the meta-analytic evidence across multiple languages. Further, we excluded studies published before the year 2000, potentially missing otherwise eligible studies and their contribution to evidence formation in the field.

Second, 66% of all included studies used American samples. Although 34% of the included studies examined samples from Asia, Australia, or Europe, the study pool did not include samples from South America or Africa. Therefore, further evidence of samples from a broader spectrum of countries is needed to be able to generalize the present findings further and test for cultural differences in cognitive processes (Imbo & LeFevre, 2009). Third, the study pool was too small to investigate all the moderators of interest. This was partially due to the strict inclusion and exclusion criteria applied during the study selection. For example, we had to exclude eligible clinical studies that did not report crucial results for the healthy control group. Therefore, we encourage researchers to extend and update this meta-analytic sample to examine additional moderators of the EF–math link. Fourth, the categorization of EF task types in the meta-analysis represented a compromise between precision and manageability due to the large variety of possible categorizations in the extant EF literature (Baggetta & Alexander, 2016; Garon et al., 2008). In line with Miyake et al. (2000), we did not explicitly distinguish between tasks measuring hot and cool EFs, which might have explained further variations between task types (Brock et al., 2009). In the same vein, we could not distinguish clearly between pure working memory tasks and updating tasks and thus subsumed them under the updating category. Future meta-analyses should try to examine these two constructs separately if the reporting precision of primary studies allows. To account for the methodological overlap between EF and math intelligence tasks, the content of a task (e.g., whether a span task uses numbers, letters, or pictures as stimuli) should be coded and examined as a methodological moderator in future research. To further enhance precision, n-back tasks and difficult span tasks could be investigated separately in accordance with meta-analytic findings (Redick & Lindsey, 2013) if enough effect sizes are available. Generally, variation between task definitions is a well-known phenomenon in EF research and has led to divergent findings in the relations between EFs and other cognitive abilities (Ackerman et al., 2005; Friso-van den Bos et al., 2013; Lehtonen et al., 2018). A clear and consistent framework of EF tasks for different age groups and comprehensive reporting in primary studies might be the much-needed solutions to this problem. Although it was not possible in this meta-analysis due to the small number of primary studies providing intercorrelations among math intelligence tasks or subdimensions, we argue that a further differentiation of the math intelligence construct, especially in Models 1–4, could provide more detailed evidence of the connections between specific EFs and specific math skills. As several meta-analyses have demonstrated (Hjetland et al., 2020; Scherer et al., 2020; Yang et al., 2021), hypotheses regarding the connections between multifaceted constructs could be tested with a range of factor models, such as single-, correlated-, or nested-factor models. To obtain such evidence, however, a more detailed reporting of primary studies, especially the correlation matrices containing not only the correlations among EF subdimensions and correlations between EF subdimensions and math skills but also the correlations among math skills, is needed. In the future, finer-grained reporting could further allow researchers to tease apart the multiple facets of EFs and math intelligence, as well as partial out the influence of general intelligence, which could not have been fully achieved with the present data.

A comprehensive framework of EFs and academic skills could be established in the future to streamline EF assessment and account for several trends in the literature. Namely, such a framework should aim to explain (a) the very similar correlations across EF subdimensions (see Table 9), (b) an age trend of the EF–academics relationship increasing/decreasing with age for different EF subdimensions, (c) very small training and transferability effects (Kassai et al., 2019; Melby-Lervåg et al., 2016; Melby-Lervåg & Hulme, 2013; Takacs & Kassai, 2019), and (d) the domain specificity of EFs in task-specific situations (Doebel, 2020; Peng et al., 2018). Such a framework would introduce exciting new hypotheses to test and take the field forward in leaps and bounds.

Similar to the present study, new research helps to create a solid basis for this comprehensive framework. Spiegel et al. (2021), for instance, recently examined the relationship between EFs and reading, language, and mathematics in primary school children.

The authors' findings suggest that a simple age trend—that EF subdimensions diverge, but their link to academic skills strengthens with age—is very unlikely. Instead, the predictive strength of specific EF subdimensions seems to depend on the developmental stage and the exact academic skill being measured. Combined with insights from Peng and Kievit's (2020) review of the mutual effects of cognitive functions and academic skills, these findings provide further grounds for a comprehensive framework of EFs and academic skills in the future.

6.7 Conclusions

The present meta-analysis provides a comprehensive overview of the literature on the relation between EFs and math intelligence in preschool children from 2000 to 2021. The present findings suggest that EFs, represented by a composite as well as three subdimensions, are positively and significantly related to math intelligence in preschool children. This relation testifies to the overlap in some skills and measures and, ultimately, the involvement of EFs in solving math intelligence tasks, and vice versa. To some degree, the performance on one construct measure could be used to predict the performance on the other. Nevertheless, the evidence presented in this meta-analysis does not suggest that assessing one of the two constructs may make assessment of the other redundant. In addition to the positive correlations, there is substantial heterogeneity within and between studies, suggesting that these effect sizes are reproducible across studies. As measurement characteristics, rather than sample or study characteristics, primarily explain parts of this heterogeneity, we highlight the importance of considering the psychometric quality of EFs and math intelligence assessments when interpreting their correlation. When considered jointly, the relations between the EF subdimensions and math intelligence were similar for the study-level data. This finding does not provide evidence of the differential relations between a single EF and math intelligence after the other EFs are controlled for. Further research is needed to establish a comprehensive framework of EF task types to streamline the EF assessments clarifying, for instance, the impact of age on the relation between EFs and math intelligence.

7 Study 3

Value-added Scores Show Limited Stability over Time in Primary School

Valentin Emslander[a], Jessica Levy[a], Ronny Scherer[b, c], Antoine Fischbach[a]

[a] *Luxembourg Centre for Educational Testing (LUCET) at the University of Luxembourg, Faculty of Humanities, Education and Social Sciences, University of Luxembourg, Luxembourg*

[b] *Centre for Educational Measurement at the University of Oslo (CEMO), Faculty of Educational Sciences, University of Oslo, Norway*

[c] *Centre for Research on Equality in Education (CREATE), Faculty of Educational Sciences, University of Oslo, Norway*

Published in *PLOS ONE*[5]

https://doi.org/10.1371/journal.pone.0279255

The minimal dataset and codebook are available at https://osf.io/t67qa/?view_only=5ee953125f584687ae61af98b292464b.

[5] The numbering of headings, tables and figures has been adjusted to align with the structure of the present work. The Appendix is integrated general Appendix at the end of this thesis.

Abstract

Value-added (VA) models are used for accountability purposes and quantify the value a teacher or a school adds to their students' achievement. If VA scores lack stability over time and vary across outcome domains (e.g., mathematics and language learning), their use for high-stakes decision making is in question and could have detrimental real-life implications: teachers could lose their jobs, or a school might receive less funding. However, school-level stability over time and variation across domains have rarely been studied together. In the present study, we examined the stability of VA scores over time for mathematics and language learning, drawing on representative, large-scale, and longitudinal data from two cohorts of standardized achievement tests in Luxembourg (N = 7,016 students in 151 schools). We found that only 34-38% of the schools showed stable VA scores over time with moderate rank correlations of VA scores from 2017 to 2019 of r = .34 for mathematics and r = .37 for language learning. Although they showed insufficient stability over time for high-stakes decision making, school VA scores could be employed to identify teaching or school practices that are genuinely effective—especially in heterogeneous student populations.

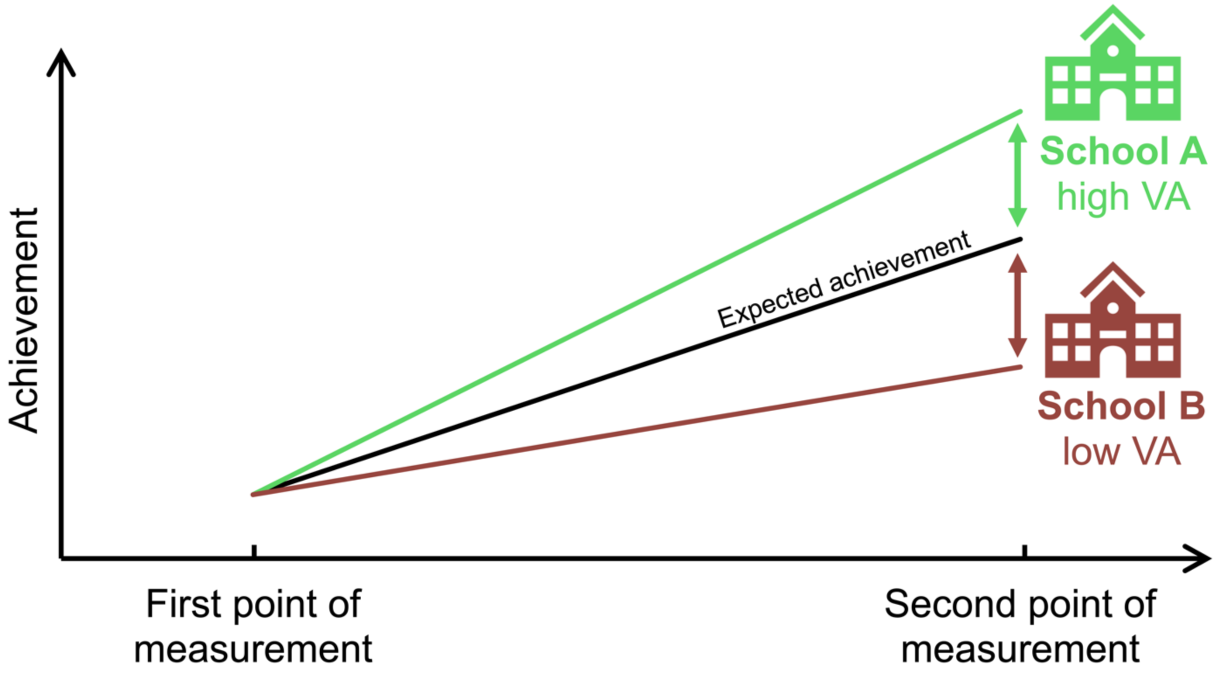
Keywords: Value-added modeling, school effectiveness, longitudinal data, primary school

Public Significance Statement: This study suggests that school value-added scores are only moderately stable over time. Therefore, using value-added scores for high-stakes decisions (e.g., about how to allocate school funds) is ill-advised. Value-added scores might be best suited for finding effective schools, learning from their practices, and improving school effectiveness.

7.1 Introduction

Can the effectiveness of a teacher or a school be quantified with a single number? Researchers in the field of value-added (VA) models may make this exact argument (Chetty et al., 2014; Kane et al., 2013). VA models are used to calculate a score that represents the learning gains a student has received through their teacher or school. The VA score quantifies the difference between the expected achievement of students with similar background characteristics and their actual achievement (Sanders et al., 1997). Positive VA scores signify higher-than-expected achievement, given the student's background characteristics (e.g. socioeconomic status [SES], language, or prior achievement), whereas negative scores imply lower-than-expected achievement. Attempting to make a fair comparison between schools, these student VA scores can be averaged per school (or teacher) and indicate the value a school adds to its students (Braun, 2005; Tymms, 1999, p. 27) independent of their background. Figure 17 illustrates such a comparison for one school with a high VA score and one school with a low score. In this example, the students from the two schools had comparable starting characteristics (e.g., prior achievement) and were thus expected to perform similarly from a statistical perspective. However, the students from School A performed better than what was statistically expected for a comparable group of students, indicating that this school offered *added value* to its students' achievements. School A would thus receive a high VA score, whereas School B would receive a low VA score.

Figure 17. *Illustration of high and low VA schools performing above and below what is expected of them from one measurement point to another*



Note. Double-headed arrows signify the VA score as the difference between a school's expected achievement and its actual achievement.

Because this approach seems reasonable for quantifying school effectiveness and teacher effectiveness, VA modeling was adopted for accountability purposes in U.S. schools in the late 90s, and its application has since resulted in a surge in high-stakes decision making in educational settings (Aslantas, 2020). In such VA-based assessment systems (e.g., Education Value-Added Assessment System, EVAAS), changes in the VA scores of teachers or schools are used to reward highly effective teachers, by offering them a tenured position, and schools, by allocating more funding. At the same time, teachers with low VA scores might face in extremis unemployment, and schools with low VA scores might need to operate on a tighter budget (Goldhaber & Hansen, 2010; Sass, 2008). This allocation of resources is based on the notion that differences in a student's VA score from one year to another are due solely to this student's teacher or school.

Parallel to the rise in applications of VA models, research interest in evaluating the actual performance and precision of VA models has bloomed (Aslantas, 2020; Levy et al., 2019). Due to a lack of consensus on how to calculate VA scores, and because VA scores differ greatly in their accuracy depending on the exact model used to calculate them (Aslantas, 2020; Levy et al., 2019, 2020), a school's VA score could differ from one year to the next. Hence, a school with a high VA score one year may receive a low VA score the next year without any actual change happening at that school. VA scores may change because of variation in teaching or school effectiveness but also due to different errors in the measurement and construction of VA scores (Loeb & Candelaria, 2012). Due to this instability, researchers and policymakers have argued that VA scores are not suitable for high-stakes decision making (Amrein-Beardsley, 2014; Gorard et al., 2013), and their use should be restricted to informing teachers and schools about how they can improve their schools. To utilize VA scores for high-stakes decision making (e.g., teachers' tenure or allocation of funding), these scores need to be highly stable over time. However, findings on this school-level stability is mixed, with some studies indicating generally high stability in school VA scores over time (Ferrão, 2012; Thomas et al., 2007) and others reporting instability (e.g., Gorard et al., 2013; Perry, 2016).

Furthermore, there is no consensus on which independent or dependent variables should be used in VA models beyond prior achievement (Everson, 2017; Levy et al., 2019), and the stability of VA scores could vary between different outcome domains (e.g., between language and mathematics). If VA scores lack stability *over time*, we cannot be sure whether these changes are due to actual changes in teachers and schools or measurement issues, such as error or choice of models. If VA scores lack stability *across outcome domains*, differences could be driven by different teacher and school practices in the two domains but also by variation in the measurement of the two domains so that they cannot be meaningfully compared. If VA scores lack stability *over time and across outcome domains*, we would not be able to attribute VA score variation to changes in effectiveness over time or to educational differences between domains. Thus, the use of VA scores as the primary information for high-stakes decision making is in question, and the inferences drawn from them could be compromised. To shed light on the "stability problems" of VA models (Aslantas, 2020), we estimated VA scores for two cohorts of students from 151 primary schools and examined the stability of these scores over time and across the most frequently used outcome domains: mathematics and language learning. In our large-scale data set, we included all eligible primary schools in Luxembourg and thus worked with population data. We chose to investigate students at the beginning of their school careers because younger students have been found to show greater response to interventions than older students (Heckman, 2008). Our study extends the body of research on VA score stability by considering multiple domains and primary schools, adding a rich set of background variables to our analysis, and using state-of-the-art VA models and covariate combinations (Emslander, Levy, Scherer, et al., 2021; Levy et al., 2020, 2022).

7.2 Theoretical framework

7.2.1 Value-added models and their use

Effectiveness is often difficult to compare across schools because no two schools are alike in the composition of students' language background, SES, or prior achievement. To solve this difficulty in comparing schools, researchers have drawn on VA scores that control for student composition factors (e.g., students' language background, SES, and prior achievement) and single out the "net effect" of school effectiveness (Driessen et al., 2016). VA scores can be calculated in different ways in terms of variables and choice of statistical models (Levy et al., 2019, 2020; Tekwe et al., 2004). However, the underlying idea, which originated from economics (Hanushek, 1971), is the same for all of these statistical models: When controlling for all available background variables and prior achievement, all gains in achievement that are left are likely to be due to teacher or school effectiveness (for an in-depth literature review of research on teacher and school VA scores, please see (Everson, 2017; Koedel et al., 2015; Levy et al., 2019).

The idea underlying VA scores can be expressed in two simple statistical steps: Equation 1 shows the first step, in which the expected achievement $\hat{y}$ is estimated for every student i in school j as a function *f* of their initial characteristics x_{ij} at an earlier time point (e.g., prior achievement). This function *f* is usually a linear regression or a multilevel model (Kurtz, 2018; Levy et al., 2019).

$$\hat{y}_{ij} = f(x_{ij}) \tag{1}$$

Equation 2 shows the second step, in which a VA score is estimated for each school j by calculating the mean difference (i.e., residuals) between the expected achievement $\hat{y}$ and the actual achievement y for all n students in this specific school j. This is equal to the average error term e of all students i in school j.

$$VA_j = \frac{\sum_i^j (y_{ij} - \hat{y}_{ij})}{n_{ij}} = \frac{\sum_i^j e_{ij}}{n_{ij}} \tag{2}$$

A high VA score indicates that students in school j achieved above what was expected of them, as was the case in the example of School A in Figure 17. A low VA score indicates that students in school j achieved below what was expected statistically. The idea is to find the "true" school effect—namely, the value a school adds to its students' achievement—by statistically controlling for everything that cannot be changed by a school. Following this idea, everything that is left can be seen as the true school effect (i.e., the residuals; Equation 2).

Therefore, it is important to ensure a high level of quality in the initial prediction step (i.e., Equation 1).

Researchers have explored different uses of VA models to render issues of educational effectiveness visible. One use is to employ VA models for high-stakes decisions, such as decisions about teachers' salaries or tenure or a school's funding (Hanushek, 2019). In the US, where VA research is flourishing (Everson, 2017; Levy et al., 2019), VA models are applied to policymaking in large parts of the country (Amrein-Beardsley & Holloway, 2017; Kurtz, 2018). In other parts of the world, VA models are used as well, for example, in Italy, Portugal, Brazil, the United Kingdom, and the Netherlands (Agasisti & Minaya, 2021; Ferrão, 2014; Perry, 2016; Timmermans et al., 2015). Another practice is to use VA models to identify high-performing schools and the factors that determine their success (Emslander, Levy, & Fischbach, 2022).

The use of VA scores for accountability purposes is based on highly debated scientific findings and a mixed research literature. While using VA scores is a significant improvement over using only achievement tests to compare school effectiveness (Perry, 2016), researchers still debate crucial aspects of VA scores. Some issues of VA models have already been identified and tackled, such as creating a consensus on how to best estimate VA scores (Everson, 2017; Levy et al., 2019, 2020). The most widely used models are the multilevel and linear regression models, with the former outperforming the latter (Emslander, Levy, Scherer, et al., 2021; Levy et al., 2022). However, the stability of VA scores over time in primary schools still needs to be examined to ensure that VA scores are informative for this educational level.

7.2.2 Stability of value-added scores over time

Let us consider one of this year's top-performing primary schools: Do we expect this school to perform as well next year and in two years? We would probably expect this school's VA scores to be stable over time because we can assume that a school's VA scores are susceptible to outside events only to a small extent.

Research on school-level stability over time is still scarce and has produced mixed results. Table 18 gives an overview of prior research on the stability of school VA scores over time with the included variables and samples. On the one hand, some studies have supported the stability of VA scores over time: Ferrão (2012), for example, found a moderate level of stability in VA scores over two years in Portugal with 65% of scores remaining in the same quartile of VA rankings. She recommended using VA scores for school improvement, especially in countries with high retention rates. Similarly, Thomas et al. (2007) investigated changes in VA scores of English secondary schools over ten years. They found that only one school (which had a low VA score) out of 16 schools was able to meaningfully raise its VA score across a period of more than four consecutive years, whereas most other schools' VA scores improved only over two consecutive years before stagnating or declining again. On the other hand, several researchers found a lack of VA score stability in their data. In their study of all secondary schools in England, Gorard et al. (2013) found that none of the schools showed consistently high VA scores across five consecutive years. They interpreted their finding as evidence that VA scores were unstable over time and should not be used in practice until the reliability of the scores could be improved. Perry (2016) found moderate to large correlations between VA scores over one ($r = .59$-.61), two ($r = .45$-.46), and three years ($r = .35$) and showed that student characteristics, such as English as an additional language and eligibility for free school meals, could explain 11% and 35% of the variance in VA scores in primary and secondary school, respectively. Despite this moderate to large stability but given the dependence on student characteristics, Perry (2016) and more recent research from the UK (Leckie & Goldstein, 2019) recommended avoiding school VA scores as a basis for policymaking or other high-stakes decisions.

Table 18. *Overview of Prior Research on the Stability of School Value-Added Scores Over Time, their Included Variables, Samples, and Conclusions*

Reference	Included Variables	Sample	Findings & Conclusion
Ferrão (2012)	• Achievement in mathematics or reading and prior achievement • Student characteristics: SES, gender, self-declaration of race/skin color, kindergarten attendance, SEN, attendance of mixed class, grade repetition • School characteristics: composition SES, type of school governance	• Portugal: 45 primary and 14 elementary and lower secondary schools with two cohorts spanning grades 1 to 8 were researched over two years each • Over 4 years (2005–2008)	• Moderate level of stability over two consecutive years with 65% of scores remaining in the same quartile of VA rankings. ⇒ Use VA scores for school improvement, especially in countries with high retention rates
Gorard et al. (2013)	• Achievement in English, mathematics, and science • Student characteristics: gender, SEN, ethnicity, free school meals, first language, school changes, age, IC, IDACI • School characteristics: Variance of student achievement within a school	• England: All secondary schools with VA scores (*n* = 2,897) • Over 5 years (2006–2010) • High VA: confidence interval does not include the mean VA score	• No school showed consistently high VA scores across five consecutive years ⇒ VA scores are unstable over time and should not be used in practice ⇒ The missing data problem must be solved
Thomas et al. (2007)	• General Certificate of Secondary Education points of students, prior achievement at age 11 with the Cognitive Abilities Test • Student characteristics: age, gender, ethnicity, free school meals, poverty, SEN, school changes • School characteristics: none, as gender composition was about 50% in all schools	• England: Ten consecutive cohorts of 16-year-old secondary school students (*n* = 134) in one large school district • Over 10 years (1993–2002)	• One out of 16 schools was able to meaningfully raise its VA score across a period of more than four consecutive years • Most other schools' VA scores improved only over two consecutive years before stagnating or declining again ⇒ Low VA schools are more likely to improve
Perry (2016)	• Achievement at ages 7 and 11, averaged for each cohort • Student characteristics were aggregated on the school-level • School characteristics: gender (%), SEN (%), English as an additional language (%), free school meals (%), child looked after status (%), number of students in a cohort, coverage (inclusion in the measure)	• England: nearly all primary and secondary state-schools • Over 4 years (2011–2014)	• Moderate to large correlations between VA scores over one (r = .59-.61), two (r = .45-.46), and three years (r = .35) in primary schools • Student characteristics could explain substantial variance in VA scores ⇒ Avoid school VA scores as a basis for policy-making or other high-stakes decisions

Note. *n* = number of schools in the sample; SES = socio-economic status; SEN = Special Educational Needs; VA = Value-Added; IC = students who have been 'In Care' at any time while being at this school; IDACI = Income Deprivation Affecting Children Index measuring deprivation based on student postcode. Please see the respective original study for more details.

The extant literature has discussed some reasons for the instability of VA scores. Perry (2016) hypothesized that the most variance could be found within rather than between schools. The small amount of between-school variance might therefore not be very informative (Wiliam, 2010). As noted above, Perry (2016) also considered variables outside of the control of teachers and schools as potential reasons for instability. This is a common phenomenon in longitudinal studies such that the actual change in the school hardly drives variation, but rather, the variability is driven by outside events, such as a successful team-building intervention or the admission of several new students. Changes within one cohort over multiple years could also simply be due to the maturation of the students, whose cognitive abilities (Zelazo & Carlson, 2012) and peer relationships (Hardy et al., 2002) develop rapidly before and during school. But teachers can also change such that professional development, positive feedback, and experience could have a positive influence on the VA scores of the entire school, whereas critical, personal life events, or additional responsibilities could have a negative impact (Agasisti & Minaya, 2021).

Measurement error and regression to the mean might also be drivers of change in VA scores (Perry 2016). Regression to the mean is most prevalent in extreme groups (high and low VA scores), which might make these two groups most prone to variations. This effect was mostly ignored in VA research (Smith & Smith, 2005). Looking at the extreme groups of schools with high and low VA scores, some of these schools might be in this group only due to the inevitable measurement error in achievement scores (Ferrão & Goldstein, 2009). Such schools that were misplaced accidentally due to this measurement error would likely be closer to the mean at the next measurement point, implicating unwanted variance in VA scores over time (Smith & Smith, 2005).

A solution to the stability issues in VA scores might be, for example, to control for a rich set of background variables and to estimate the VA scores with the same models over time (Aslantas, 2020). Model choice and included variables are highly influential in estimating VA scores (Dumay et al., 2014; Levy et al., 2020; Perry, 2016) and should therefore be held constant when looking at the stability of VA scores.

7.2.3 Different outcome domains

While there seems to be some consensus about which variables should be included when estimating VA scores (Levy et al., 2019), less is known about the impact of different outcome domains. Only a few studies have looked at stability across diverse measures of the same achievement domain (Papay, 2011) or even contrasted two different achievement domains (Ferrão & Couto, 2013). At the student level, in accordance with the reciprocal internal/external frame of reference model (Niepel et al., 2014), we might expect differences between mathematics and language achievement to be large. This model describes the link between, for example, a student's achievement in mathematics or language with their academic self-concept in the other domain in a longitudinal setting. However, this effect might average out when considering the teacher or school level on which the VA scores are summarized.

Comparing VA scores and students' regular achievement measures helps evaluate the VA scores' validity, which is crucial for the credible use of VA scores in policymaking or other kinds of high-stakes decision making. Naturally, policymakers would like schools with a high VA score to do well not only on mathematics tests but also on language tests and vice versa. Thus, a comparison of different outcome domains is indicated to investigate whether VA scores are equally meaningful for the domains of mathematics and language.

Prior research on both school and teacher VA scores—two closely related research areas—has focused on the comparison of correlations between VA scores with different outcome domains or different measures of the same outcome domain (Ferrão & Couto, 2013; Lockwood et al., 2007; Papay, 2011; Sass, 2008). On one end of the spectrum, Ferrão and Couto (2013) found that school VA scores had strong correlations with students' mathematics and Portuguese performance as the outcome domains in Grades 2 through 5, ranging from *r* = .43 to .70. Looking at teacher VA scores, Sass (2008) investigated their stability over time and across two test instruments and found a correlation of *r* = .48 between VA scores of two different achievement tests. He also found that this correlation was higher than the correlation of VA scores over time (*r* = .27). At the same time, he acknowledged the substantial difference between the two outcome measures, which he attributed to differences in testing material, potential ceiling effects, and differences in pressure to perform, as one measure was a high-stakes test and one was a rather low-stakes test (Sass, 2008). Lockwood et al. (Lockwood et al., 2007) found large correlational differences between two subscales from the same mathematics test on the teacher level. They attributed large parts of the variation in the VA scores to these different measures in a middle school sample, suggesting that VA scores are very sensitive to the choice of outcome measure. These results were later replicated by Papay (2011), thus corroborating the conclusion that the choice of outcome measures can have a larger influence on the stability of VA scores than the model specifications. Going beyond Lockwood et al.'s (Lockwood et al., 2007) original study, Papay (2011) also explored differences between three comprehensive reading achievement measures as outcome variables. He found correlations of *r* = .15 to .58 between VA scores with these three different outcome measures. These results indicate some comparability between the measures but not enough for high-stakes decisions. In a real-life situation, almost half of all teachers would have had different salaries if the outcome measures underlying their VA scores changed, as Papay (2011) demonstrated. Taken together, these findings demonstrate the limited stability of VA scores across different outcome domains.

Focusing on the stability of VA scores between different tests of the same outcome domain, prior research has left one question open: What is the stability of VA scores between not only different outcome tests (e.g., two different mathematics subtests) but between different outcome domains (i.e., mathematics and language)? We would expect that a school with a high VA score is not only helping its students excel in one domain (e.g., mathematics) but is also facilitating learning in another domain (e.g., language; Papay, 2011). Stated differently, a VA score that shows great variability in its predictive power across different outcome domains would not be stable across domains and would thus not be helpful. Then again, VA scores that were well-aligned no matter whether mathematics or language was used as the outcome domain would provide an argument for the validity of school VA scores, especially if the VA scores in the two domains showed similar levels of stability.

7.2.4 International use of value-added scores

In the US and Europe, VA scores have experienced increasing research interest (Hanushek, 1971; Levy et al., 2019). With the No Child Left Behind Act (No Child Left Behind Act, 2002) and the Race to the Top Act (Race to the Top Act, 2011), VA models have been applied for policymaking in large parts of the US (Amrein-Beardsley & Holloway, 2017; Kurtz, 2018). Parallel to their increase in use, VA scores have also experienced critical resistance from researchers and teachers due to methodological flaws and their real-life implications (Amrein-Beardsley, 2014; Collins, 2014). After 2009, several unsuccessful lawsuits had been filed against the use of VA scores in decisions about teachers' remuneration or tenure, leaving many educators with smaller pay or without a contract at all after their VA-based assessment. In 2017, in the school district of Houston, TX in the US, however, the federal court ruled it unconstitutional to terminate a teacher's contract on the basis of undisclosed VA score data (Paige & Amrein-Beardsley, 2020). Overall, VA scores have found most of their use in the US, which has led to several lawsuits against their use in high-stakes decisions in education policy.

Parallel to the development of the "Tennessee Value-Added Assessment System" in the US (TVAAS; Sanders & Horn, 1994), for instance, the French ministry of education introduced VA scores to be used in school evaluations as well ("Indicateurs de valeur ajoutée"; Duclos & Murat, 2014; MEN-DEP., 1994). As discussed above, VA scores are used in the United Kingdom, where other variables in addition to prior achievement are used to estimate the VA scores (Perry, 2016). In this way, student characteristics such as ethnicity, SES, or gender could also inform VA scores and increase their fairness. Here, VA scores are usually used as an approximation of a school's effectiveness and to provide school performance rankings (Gorard et al., 2013; Perry, 2016). In the Netherlands, VA scores are used to identify disadvantaged primary schools that are at risk of underperforming so that targeted interventions can be offered to them (Timmermans et al., 2015). Further, Ferrão (2014) reviewed research on school-level VA scores in Brazil and Portugal.

In a nutshell, most countries outside the US avoid using VA scores for high-stakes decisions. Nonetheless, they apply VA scores in order to inform such decisions as one of multiple tools that can be used to estimate a facet of school effectiveness. Such examples introduce VA scores as a means of identifying high-performing teachers or schools to learn from them or their low-performing counterparts in order to support such lower performers in a less punitive and more appreciative way.

7.2.5 Unsolved issues in value-added research and the Luxembourgish context

Prior research has presented several unsolved issues and open questions on the use of VA scores, for example: How stable are school VA scores over time? How stable are they across outcome domains? These questions are crucial for highly diverse educational contexts, because they can profit the most from reliable and valid VA scores. For example, diversity in the student population can arise from diverse languages spoken at home, a migration background, or a family's SES. This diversity leads to different preconditions for learning mathematics and new languages (or even the language of instruction) and thus shapes students' school careers (Hadjar & Backes, 2021) and school completion (Ferrão, 2022). Consequently, VA scores improve by including relevant background information, such as the languages spoken at home, alongside prior achievement.

The Grand Duchy of Luxembourg provides one such highly diverse educational context with a multilingual student body, leading to gaps in students' achievement that widen with age (Sonnleitner et al., 2021). This multilingualism is reflected in the fact that in 2020, only 43% of all students in Grades 1 and 3 spoke Luxembourgish or German at home (Fischbach et al., 2021). Alongside students who come from a socioeconomically disadvantaged family or attend a secondary school in the lower tracks, students who do not speak the language of instruction are specifically challenged to do well in school (Lenz et al., 2021).

Another specificity of the Luxembourgish primary schools is that they operate in four learning cycles, spanning two years each. Cycle 1 starts with two years of preschool before Cycle 2 spans the first and second years of school. These two-year cycles continue until Cycle

4 ends in Grade 6. Within one cycle, students typically have the same class teacher until they progress to the next cycle, where they get a new class teacher. Thus, longitudinal studies conducted in Luxembourg, such as the Luxembourg school monitoring programme (LUCET, 2023), are usually conducted every other year to correspond to the structure of the learning cycles. In this way, similar results within one learning cycle, for example, due to having the same class teacher, are not overinterpreted.

In a context as diverse as Luxembourg, VA scores could provide a much fairer measure of school effectiveness than averaging the standardized achievement of all the students in one school. However, VA scores can only be used in these contexts if they exhibit a sufficient level of stability across time and across different outcome domains. In other words, they need to exhibit satisfactory reliability and validity. Otherwise, VA scores would fail to flexibly adjust to the constantly evolving language and school landscape (Kirsch & Seele, 2022) and should thereby not be used in high-stakes decision making.

7.2.6 Relevance of value-added score stability

The relevance of the stability of VA scores is directly linked to their reliability and validity. If they are stable over time in multiple outcome domains (touching on their reliability and validity, respectively), they could be a great additional tool for evaluating school effectiveness. This would be especially helpful in highly diverse educational contexts, where controlling for students' backgrounds makes the effectiveness estimate much more meaningful. This effect might be smaller in rather homogeneous contexts that lack variation in students' and schools' backgrounds.

If VA scores prove to be unstable over time and in different outcome domains, however, their use should be cautioned. A larger variety in VA scores that cannot be explained by the predictor variables (i.e., noise in the data) makes VA scores less reliable and less valid. They would be less reliable if the same school with unchanged circumstances places high one year and low in some other year without apparent real-life change occurring between these years. Similarly, VA scores would be less valid if they had vastly different results for different outcome domains (i.e., mathematics and language) without an actual difference in achievement, leading to a lack of real-life explanatory power. Therefore, without sufficient explanatory power to set apart high- and low-VA schools and without the stability to demonstrate a reliable estimation of school effectiveness, VA scores would not be suited for making high-stakes decisions.

In a concrete application example, at an unlucky school in the US whose VA score plummets due to measurement error or changing circumstances, several teachers might lose their jobs. Across the big pond, in the United Kingdom, a school with an unstable VA score might move to the top of the effectiveness ranking undeservedly, attracting more students without having improved its school climate or bettered its general level of instructional quality. Such unwarranted fluctuations need to be considered, as their real-life implications can be radical and, in some cases, harmful. To avoid disadvantageous outcomes from the application of VA scores, they need to either (a) show sufficient stability across time and outcome domains (touching on their validity and reliability) or (b) be abolished as a means of measuring a school's effectiveness for high-stakes decisions and should be used only for informative purposes.

7.3 The present study

With the present study, we examine the stability of school VA scores over time and quantify differences in stability between two outcome domains (i.e., mathematics and language achievement). Specifically, we address the following research question: *How stable are school VA scores over time and across outcome domains?*

To answer this research question, we first calculated VA scores at the school level, ranked all the schools accordingly, and estimated the correlations of the ranks over time to arrive at a stability estimate. To calculate the VA scores, we used the same selection of covariates across the two outcome domains. To choose the covariates, we reviewed models of school learning (e.g., Haertel et al., 1983; Wang et al., 1993), prerequisites, and correlates of students' achievement as well as recent findings on the selection of covariates in school VA models (Levy et al., 2022). We followed the approach specified by Levy et al. (2022), who found that including prior mathematics achievement, prior language achievement, and covariates related to students' sociodemographic and sociocultural backgrounds (i.e., socioeconomic status of the parents, languages spoken at home, migration status, and sex) in multilevel school VA models could help leverage between-school differences in student intake and in the resulting school VA scores. By examining the stability of school VA scores and comparing different outcome domains, we seek to contribute to the existing debate on the stability and use of VA scores in educational effectiveness.

7.4 Method

7.4.1 Participants

The present study drew on longitudinal large-scale data from the Luxembourg School Monitoring Programme Épreuves Standardisées (ÉpStan; LUCET, 2023). The ÉpStan assesses all students in Grades 1, 3, 5, 7, and 9 in Luxembourg at the beginning of every school year. By doing so, every student who follows the usual path through school is tested every other year. Data on the students are collected in three main areas: academic competencies (in mathematics and languages), motivation and emotion, and background variables (e.g., language background and SES). In the present study, we used data from the cohorts of Grade 1 students in the years 2015 and 2017 to inform our VA scores for the Grade 3 students in the years 2017 and 2019 and form a large longitudinal data set with school VA scores at two time points (2017 and 2019).

The final data set comprised *N* = 7,016 students, nested in 151 primary schools in Luxembourg. Students were included if they participated in Grade 1 and two years later in Grade 3. We thus excluded data from students who (a) had missed data collection in Grade 3 (e.g., due to sickness or grade repetition) or (b) were enrolled at a different school in Grade 3.

The ÉpStan received approval from the national committee for data protection and has a proper legal basis. Current ethical standards were respected at all times (American Psychological Association, 2017). The participating students and their parents or legal guardians were duly informed before the data were collected and had the opportunity to opt out. All data were pseudonymized with a so-called “Trusted Third Party” (for more information, see LUCET, 2021, 2023), in accordance with the European General Data Protection Regulation, to ensure the privacy of students and their families. In the present study, we used an anonymized data set.

7.4.2 Measures

To calculate the VA scores, we included measures of prior academic achievement and measures of an outcome domain (i.e., mathematics achievement and language achievement) all with high psychometric quality. To further inform the VA scores, we also included measures of sociodemographic and sociocultural background variables.

7.4.2.1 Academic achievement

Academic achievement plays two crucial roles in VA modeling: Whereas prior academic achievement is used as a covariate, current academic achievement is irreplaceable as an outcome variable. We used measures of mathematics and language achievement for Grade 1 simultaneously to calculate all VA scores. For the outcome measures in Grade 3, however, we report all results separately, once for mathematics and once for language achievement. Our choice to incorporate both mathematics and language achievement into our analyses was further corroborated by meta-analytic evidence of their mutual relationship (for a recent meta-analysis on the mutual relationship between language and mathematics, see Peng et al., 2020).

In the first months of Grades 1 and 3, students' mathematics and language achievement scores were collected with standardized achievement tests. Expert groups consisting of teachers, content specialists in teaching and learning, and psychometricians developed these tests to ensure the content validity of these tests (Fischbach et al., 2014), based on the Luxembourgish national curriculum standards (Ministry of National Education, Children and Youth, 2011). On the day of testing, the students completed the achievement tests in their own classrooms in a paper-and-pencil format. Most items were designed as closed questions and scaled by a unidimensional Rasch model (Fischbach et al., 2014; Nagy & Neumann, 2010; Wu et al., 2007). Warm's Mean Weighted Likelihood Estimates were used (WLE (Warm, 1989, p. 19)) to indicate student achievement. We used these WLE values and their standard errors to calculate the reliability of all the achievement scores in the R package TAM version 3.3.10 (Robitzsch et al., 2019). Table 19 shows these reliability coefficients for both mathematics and language in Grades 1 and 3 in 2015-2017 and 2017-2019.

Table 19. *Reliability Coefficients for the Achievement Scores for the 2015-2017 and 2017-2019 Data Sets*

Variable	2015-2017	2017-2019
Mathematics achievement in Grade 1	.84	.85
Language achievement in Grade 1	.71	.73
Mathematics achievement in Grade 3	.93	.93
Language achievement in Grade 3	.83	.83

For mathematics achievement, students completed a test in Grade 1 in Luxembourgish because the language of instruction in preschool is Luxembourgish. Whereas Luxembourgish can be described as a language cognate to German (Dalby, 1999), politically and culturally it is a language of its own. The students answered items from three domains of mathematics competence: "numbers and operations," "space and shape," and "size and measurement" (see, for a more comprehensive explanation https://epstan.lu/en/assessed-competences-21/). In Grade 3, the students took the mathematics tests in German because the students had been taught in German during Grades 1 and 2. Again, the students answered items from three mathematics competence domains: "numbers and operations", "space and form", and the novel area of "quantities and measures" (see, for a more comprehensive explanation https://epstan.lu/en/assessed-competences-31/, LUCET, 2019).

For language achievement, students completed a test in Grade 1 in Luxembourgish because the language of instruction in preschool is Luxembourgish. The standardized language tests consisted of "listening comprehension" and "early literacy comprehension." To assess listening comprehension in the two language competence domains "identifying and applying information presented in a text" and "construing information and activating listening strategies," the students listened to different kinds of texts in an audio recording. For early literacy comprehension, the students were tested on three competence domains, namely, "phonological awareness," "visual discrimination," and "understanding of the alphabetic principle" (see https://epstan.lu/en/assessed-competences-21/). We averaged the scores for listening and reading comprehension as prior language achievement in the VA model to have one single score for language achievement.

In Grade 3, the students took the listening and reading comprehension language achievement tests in German, which had been the language of instruction during Grades 1 and 2. The listening comprehension test consisted of two domains of competence: "identifying and applying information presented in a text" and "construing information and activating listening strategies." For reading comprehension, the students were again tested on two competence domains, namely, "identifying and applying information presented in a text" and "construing information and activating reading strategies/techniques" (see https://epstan.lu/en/assessed-competences-31/; LUCET, 2019). Analogous to prior language achievement in Grade 1, we calculated a mean score across listening and reading comprehension in the German language, resulting in a single dependent variable for language achievement.

In the present study, we conducted secondary analyses of archived data sets and had only limited information on the psychometric quality of the achievement tests. However, as these data sets formed the basis of political and practical decisions in Luxembourg, the psychometric quality of the tests had been optimized in that regard (Fischbach et al., 2014; Martin et al., 2015). More specifically, as mentioned above, several expert panels developed all the items for the domain-specific achievement tests to ensure their content validity. After pilot testing, all items underwent psychometric quality checks concerning their empirical fit to the Rasch model—that is, the model that was used to generate WLE estimates to represent students' domain-specific achievement in Grades 1 and 3. To additionally ensure adequate testing between student cohorts at the same grade level (e.g., between Grade 1 in 2017 and Grade 1 in 2019), all test items were examined for differential item functioning. These psychometric quality checks were complemented by analyses of convergent and discriminant validity. Finally, the students' domain-specific achievement scores, as represented by the WLE scores, demonstrated reliability coefficients between .70 and .90, which were considered sufficient (Schmitt, 1996).

7.4.2.2 Sociodemographic and sociocultural background variables

All parents with a child in Grade 1 were asked to complete a questionnaire on their child's sociodemographic and sociocultural background. Table 20 shows the descriptive data for each of the two samples (2015-2017 and 2017-2019) and for the entire sample of 7,016 students. The sociodemographic and sociocultural distributions were similar in both data sets. Parents identified their occupation from a list of categories that were based on the ISCO (International Standard Classification of Occupations) classification. To approximate the parents' SES, an average value of all occupational categories was computed on the basis of the validated ISEI (International Socio-Economic Index of occupational status, see Ganzeboom, 2010) scale. For the Grade 1 sample, parents reported a mean ISEI value of 50.3 for the sample from 2015 and 49.7 for the sample from 2017. These mean values were only slightly above the average ISEI for all OECD countries of 48.8 from the first PISA tests in 2000 (OECD & UNESCO Institute for Statistics, 2003).

Table 20. *Descriptive Data of the two Samples Used in the Present Study*

Years of the sample	*N*	% of female students	% of students speaking Luxembourgish with at least one parent at home	Mean (SD) SES	Migration status (% native)
2015-2017	3443	50%	50%	50.3 (15.6)	48%
2017-2019	3573	49%	48%	49.7 (16.0)	47%

Note. SES = Socioeconomic status as measured by HISEI (Highest International Socio-Economic Index of Occupational Status).

To indicate a child's migration status, parents further specified where they and their child were born. This was translated into three categories of migration status: "native," "first generation," and "second generation." We created a dummy variable for migration status with "native" as the reference category. In the 2015-2017 sample, 48% of the students had a "native" migration status, and 47% of the students in the 2017-2019 sample had this status. To complement their parents' answers to the questionnaire, the first graders also answered a questionnaire on the language(s) they spoke with their parents. Not speaking any Luxembourgish at home could be considered a disadvantage for students in Grade 1 because both the testing and the preschool instruction were in Luxembourgish. Therefore, we coded the language(s) spoken at home as a dummy variable to distinguish between students who spoke Luxembourgish with at least one parent (reference category) and those who did not speak any Luxembourgish at home. In the 2015-2017 sample, 50% of the students indicated that they spoke Luxembourgish with at least one parent, and 48% in the 2017-2019 sample did so. Furthermore, a data set from the Ministry of National Education, Students, and Youth provided information on the students' sex, with 50% (2015-2017) and 49% (2017-2019) girls in the samples.

7.4.3 Data analysis

All analyses were performed in R version 3.6.1 (R Core Team, 2021). After preparing the data sets and imputing the missing data, we estimated the VA scores for the schools. Part of the data preparation and the estimation of school VA scores were analogous to the study by Levy et al. (2022), who used different data from the school monitoring programme (LUCET, 2021) from the years 2014 and 2016.

7.4.3.1 Data preparation

Because a criterion for inclusion in the study was that students participated in the Grade 3 achievement tests, there were no missing data in the achievement data in Grade 3. To impute missing data on the covariates, we used multiple multilevel imputation with 20 imputations, 50,000 burn-in iterations, and 5,000 iterations between imputations using the R packages *mitml* version 0.3-7 (Grund et al., 2019) and *jomo* version 2.6-9 (Quartagno & Carpenter, 2019). The S1 Syntax and S2 Syntax show the R code for the data imputation for the data from 2015 – 2017 and 2017 – 2019, respectively.

7.4.3.2 Estimation of school value-added scores

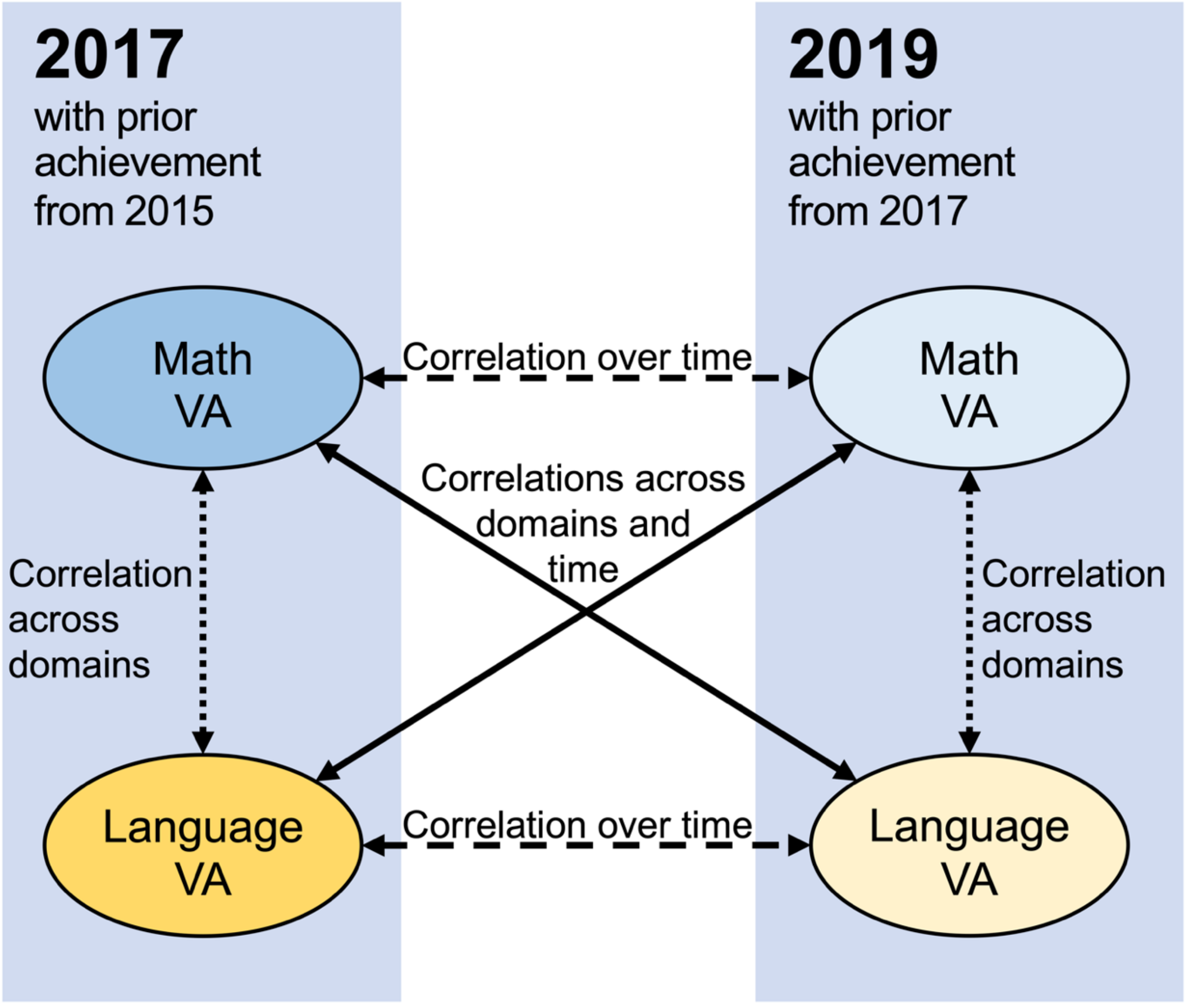
We estimated the random effects within each school to obtain VA scores via Equations 1 and 2 (Ferrão & Goldstein, 2009). More specifically, these were the Level 1 residuals from the multilevel model, averaged within a school. In other words, all student-level VA scores were averaged into one VA score per school. To estimate the model, we used the log-likelihood as the estimator. We applied the *lmer* and *ranef* functions from the R package *lme4* (Bates et al., 2015) and defined the multilevel model as follows:

```
Achievement_in_Grade_3 ~ Prior_Math_Achievement +
Prior_Language_Achievement + SES + migration_status +
language_spoken_at_home + sex + (1|school_ID)
```

In this model, the outcome variable represents either mathematics or language achievement in Grade 3. Prior achievement in mathematics and in language, SES, migration status, language spoken at home, and sex are covariates, thus being statistically accounted for.

To address our research question on the stability of VA scores across time and outcome domains, we ranked the schools by their VA scores from highest to lowest in one year. Then, we created three indicators of VA score stability. First, we estimated correlations between the two outcome domains for 2017 and 2019, respectively. Second, we checked for the stability of the school VA scores within one outcome domain over two years by calculating one correlation between the VA mathematics score in 2017 and the respective VA mathematics score in 2019, and we did the same for language. Third, we estimated the correlations between the VA mathematics scores in 2017 with the VA language scores in 2019, and vice versa (i.e., the correlation between language as an outcome domain in 2017 and mathematics as an outcome domain in 2019). Figure 18 graphically represents these three types of correlational indicators. The resultant correlation coefficients will be interpreted as indicators of VA score stability across time and domains (Loeb & Candelaria, 2012; Papay, 2011).

Figure 18. *Graphical Representation of the VA Score Intercorrelations across Time and Domains*


2017
with prior achievement from 2015
2019
with prior achievement from 2017
Math VA
Math VA
Language VA
Language VA
Correlation over time
Correlation over time
Correlations across domains and time
Correlation across domains
Correlation across domains

Note. Double-headed arrows signify the correlations we calculated.

On the basis of these findings, we investigated the total number and percentage of schools that showed a stable or unstable VA score rank with mathematics as an outcome domain compared with language as an outcome domain. For this purpose, we consulted commonly used benchmarks (Marzano & Toth, 2013) and decided to use four levels of VA scores to indicate a school's effectiveness:

- High VA scores are in the top 25% (highly effective schools)
- Upper medium VA scores are between the 50th and 75th percentiles (moderately to highly effective schools)
- Lower medium VA scores are between the 25th and 50th percentiles (moderately effective schools to schools that might need improvement)
- Low VA scores are in the lowest 25% (schools that need improvement)

We defined schools with a stable VA score as schools that remained in the same VA rank quartile in 2017 and 2019. We defined schools with an unstable VA score as schools that were in different quartiles in 2017 and 2019. The S3 Syntax shows the R code we used to analyze the correlations between the VA scores. The S4 Table shows the covariance table of VA scores with different outcome domains (mathematics and language) over time. The S5 Dataset shows the minimal dataset and codebook of school's VA quartile ranking across time and domains.

7.5 Results

7.5.1 Stability of value-added scores

We found positive correlations between school VA scores in both the mathematics and language outcome domains and across the two years of testing. Table 21 depicts the correlations for the school VA scores in the different outcome domains (mathematics and language) across time. More specifically, the school VA scores in mathematics and language within the same years were moderately to highly correlated with correlation coefficients of *r* = .59 in 2017 and *r* = .47 in 2019. The correlations of the school VA scores within one outcome domain across the two years were smaller but still moderate. We found that the school VA mathematics scores from 2017 and 2019 were correlated at *r* = .34. Similarly, the school VA language scores showed a correlation of *r* = .37 across the two years. The correlations of the school VA scores across both the domains and time were smaller than the other correlations. The correlation between the school VA language scores in 2017 and the school VA mathematics scores two years later was *r* = .22. The correlation between the school VA mathematics scores in 2017 and the school VA language scores two years later was *r* = .32. Overall, we found moderate correlations across outcome domains but only small correlations over time in primary school for the school VA scores.

Table 21. *Correlation Table of VA Scores in Different Outcome Domains (Mathematics and Language) over Time*

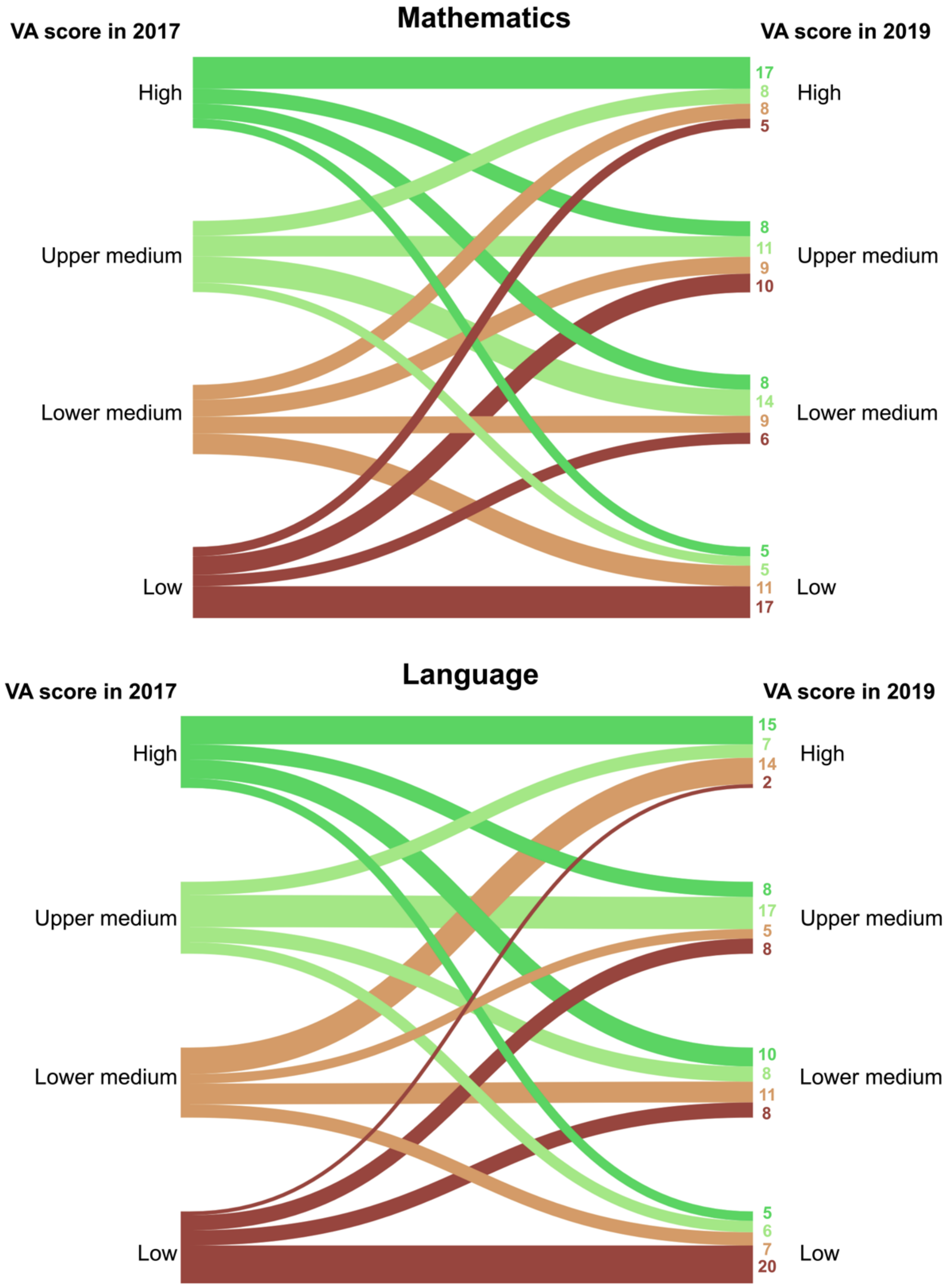
	2015-2017 Mathematics	2015-2017 Language	2017-2019 Mathematics	2017-2019 Language
2015-2017 Mathematics	-			
2015-2017 Language	.59	-		
2017-2019 Mathematics	.34	.22	-	
2017-2019 Language	.32	.37	.47	-

Note. Correlations were calculated on the basis of *n* = 7,016 elementary schools students' VA scores.

7.5.2 Prevalence of schools with stable and unstable VA scores

Figure 19 shows a transition diagram illustrating the number of schools that changed or remained in their VA score rank quartile. Looking at the school VA mathematics scores, 54 out of 151 schools had the same rank in both 2017 and 2019. This was roughly one third of the schools with mathematics as an outcome domain that remained stable over the two years (approx. 35.8%). There were more schools in the highest and lowest ranks that remained stable (i.e., 34 schools) than in the two middle ranks (i.e., 20 schools). While most schools (i.e., 97 schools) changed one or two ranks up or down, 10 schools were classified as having a high VA score in one year and a low VA score two years later, or vice versa. When language was the outcome domain, the results were similar. We found that 63 out of 151 had a stable VA score over time. Again, most schools had a different rank in 2019 than the rank that was based on their VA score in 2017. Two schools had moved from the lowest to the highest rank, whereas 5 schools had gone from the highest to the lowest rank over the two years.

Figure 19. *Transition Diagrams Indicating Changes in the Schools' Value-Added Score Ranking Quartiles from 2017 to 2019.*



Note. A school's position in the VA score ranking is indicated on either side of the transition diagram. High VA scores are in the top 25%, Upper medium VA scores are between the 50th and 75th percentiles, Lower medium VA scores are between the 25th and 50th percentiles, and low VA scores are in the lowest 25%.

Overall, the results for mathematics and language were comparable. We also found that only about one third of the primary schools exhibited stable VA scores across the two years. The other two thirds fluctuated, with some schools changing their VA score rank position substantially.

7.6 Discussion

VA models are used for accountability purposes in education and quantify the value a teacher or a school adds to their students' achievement. For this purpose, these models predict achievement over time and attempt to control for factors that cannot be influenced by schools or teachers (i.e., sociodemographic and sociocultural background). Following this logic, what is left must be due to differences in teachers or schools (Braun, 2005). To contribute to the debate about the stability of VA scores over time and across outcome domains, we drew on representative longitudinal data from two cohorts of standardized achievement tests administered to a total of 7,016 students attending 151 primary schools in Luxembourg. First, we calculated correlations between the VA scores within and across time and outcome domains. Additionally, we investigated the total numbers and percentages of schools that showed a stable versus unstable VA score across time in mathematics and language as the outcome domains.

7.6.1 Stability of value-added scores across time and domains

In our sample of primary schools, we found moderate correlations of $r = .34$ for mathematics and $r = .37$ for language as the outcome domains across a two-year period. These correlations are far from perfect (i.e., $r = 1$), moderate in size, and can thus be considered to show instability in VA scores over time. If VA scores worked perfectly, this instability could be explained by actual improvement or decline in the schools' quality, but this is probably not the case because several other factors influence VA scores. Instead, researchers have pointed to the multiple sources of error and bias in VA scores introduced by, for example, variation on the student level and measurement issues (Gorard et al., 2013; Perry, 2016; Sass, 2008). Thus, the instability in VA scores we found should be attributed primarily to these disturbances rather than to changes in school effectiveness. The correlations we found across a two-year period in primary school were even somewhat smaller than those found in a secondary-school sample by Perry (2016), who cautioned against the use of VA scores as a basis for high-stakes decision making and policymaking. We also share the interpretation expressed by Gorard et al. (2013), who warned against the use of VA scores until consensus has been reached on how to handle missing data, which models to use, and until the predictive power of VA scores for school effectiveness has been unambiguously shown.

While we found large correlations across mathematics and language within the same years—2017 and 2019, respectively—we found only small correlations across domains and across years. Correlations were larger across outcome domains in the same year than they were for the same outcome domain across two years' time. Stated differently, time has a greater effect on VA stability than different outcome domains do. This could either indicate greater changes in school effectiveness between years than between domains or simply show less noise in data collection within the same year than over time due to similarity in assessments even in different domains. This interpretation is in line with research by Thomas et al. (2007), who identified that time as a factor introduces considerable instability. The larger correlations within the same year across outcome domains replicate prior findings (Ferrão & Couto, 2013; Sass, 2008).

The reasons for the instability in VA scores can be manifold. Of course, schools improving through effective leadership or successful teacher development introduce the kind of change and variation that is desired in VA scores (Agasisti & Minaya, 2021). As discussed above, however, school VA scores might vary for statistical reasons, such as greater variability within than between schools, measurement error, and regression to the mean. However, external variables might influence the stability of school VA as well, such as changes in a student cohort due to maturation, (adverse) life events of a class or an entire school, or changes in teachers or teacher behavior due to professional devolvement or personal circumstances, for example (Agasisti & Minaya, 2021; Loeb & Candelaria, 2012; Perry, 2016; Sass, 2008). However, most of these sources of variance are outside the control of the teachers and the schools but can introduce instability in VA scores.

Given our correlational findings and the multiple sources of instability, we conclude that VA scores have insufficient predictive power for the value a school adds to its students over time. Thus, these scores cannot be considered stable enough over time to be used for high-stakes decisions (Gorard et al., 2013; Perry, 2016). As VA scores are a highly political topic and their use is controversially discussed, they should not be used as the sole indicator for a schools effectiveness on a public policy level (Amrein-Beardsley & Holloway, 2019; Conaway & Goldhaber, 2020). In the public interest, however, schools with stable VA scores can be used in research contexts for informative purposes, including learning from schools with stable high VA scores.

7.6.2 Differences between outcome domains

We found that 34-38% of the schools showed stable VA scores from Grade 1 to Grade 3 by remaining in the same VA score ranking quartile in 2017 and 2019. These results were similar across outcome domains. Considering the similarity in the correlations between mathematics and language over time, the differences in the VA scores between the outcome domains can be considered small (Cohen, 1992).

While 34-38% of the schools remained in the same VA score ranking quartile, all other schools had VA scores that changed quartiles between Grade 1 and Grade 3. As multiple sources of variance may be influencing the stability of VA scores, the changes in the rankings of these schools might not be due to their actual improvement or deterioration but may have been driven by statistical or external factors. Thus, about two thirds of the schools were at risk of falling victim to unwarranted consequences. Especially for the 10 schools that changed from the highest to the lowest VA ranks or vice versa, this instability could have dramatic consequences if VA scores were applied for accountability reasons, such as decisions about funding or the closing of schools. With no major advances in the stability of VA scores by adding richer sets of background variables or more advanced statistical methods, on its own, “Value-added is [still] of little value” (Gorard, 2006) to public educational policy.

While we did not find similar numbers of educational units that remained the same as Ferrão (2012) did, we share her sentiment that VA scores could be used as indicators of school improvement. At the same time, we argue for the complementary use of VA scores, paired with observational, qualitative, and other kinds of data to obtain a broader picture of school effectiveness (Emslander, Levy, & Fischbach, 2022).

7.6.3 Limitations and future directions

The present study has potential limitations. First, how to treat missing data is an important question in research in general, especially in VA research, as small changes in data handling might have large impacts and might sway real-life decisions. In the present study, we included only cases that had complete achievement test data in the two outcome domains and then imputed the other missing values as described above rather than using case-wise deletion methods. By excluding the incomplete achievement tests, we might have missed relevant cases, potentially introducing bias in the VA score. Future research should investigate the sensitivity of VA scores across different ways of handling missing data (Levy et al., 2022).

One specificity of the Luxembourgish school system is its structure in two-year learning cycles. During these two years, it is expected that the same teachers teach the students, as is customary in Italy for example (Minaya & Agasisti, 2019). Therefore, we did not investigate students in two consecutive years but in the first year of two consecutive learning cycles. In research, most VA scores are estimated from one year to the next. Had we focused on students in two consecutive years, however, with the same teachers, in the same classroom, and less time between the measurement points, VA score stability might have been higher, as the two measurement points would be more similar (Perry, 2016).

The two-year learning cycles can be prolonged by one year, so a child remains in the same learning cycle not for two but for three years, which is a form of repeating a grade. Retention is quite common in Luxembourg, leading to a large number of students participating in the ÉpStan in Grade 1 in 2017 but not in Grade 3 in 2019, for example. These excluded students tend to have a lower socioeconomic status, are likely to show lower achievement, and are less likely to speak the language of instruction with their parents than the included students. The practice of looking at students with regular educational pathways is in line with the common practice of estimating VA scores (Goldhaber & Hansen, 2013; McCaffrey & Lockwood, 2011), and VA scores can especially be used for constructive purposes in educational settings with high retention rates (Ferrão, 2012).

In the present study, we focused on primary school Grades 1 and 3 in Luxembourg, a highly diverse and multilingual school context compared with other countries. While our findings add further evidence of the instability of VA scores to the literature, we should not extrapolate our findings from Grades 1 and 3 to other grades or blindly take the results from Luxembourg and apply them to samples in vastly different educational settings. Perry (2016) found that even in the same Grade 4 and Grade 6 cohorts, for example, VA scores had a correlation of only *r* = .24. And, as could be expected, cohorts in two consecutive years are more likely to be similar than cohorts that are several years apart (Perry, 2016). Further, VA scores were suggested to show greater variability between different cohorts than within one cohort (Minaya & Agasisti, 2019; Newton et al., 2010; Sass, 2008).

Future research could tackle the stability of VA scores and the implications of stability in these scores in three distinct ways: (a) by investigating fundamental research questions on the parameters that influence the stability of VA scores, (b) by conducting a systematic review of both published and gray literature on the stability of VA scores over time, and (c) by looking at the real-life implications that (un)stable VA scores have for schools. To tackle open basic research questions, future research could extend the present study to replicate our findings in a less diverse and rather monolingual educational context. These are just a few directions future research could take, for which prior research has already provided some evidence. Findings from these endeavors could then be used to evaluate the specificity of the present study—in other words, whether Luxembourg, whose strength lies in a heterogeneous student population, is either significantly different from other countries or quite comparable to other places in the world.

Looking at the real-life implications for schools of (un)stable VA scores and their productive use outside of high-stakes decision making is another area that future research should explore. It will be informative to investigate differences between schools with stable high VA scores and those with stable low or moderate VA scores and learn about effective pedagogical strategies (Emslander, Levy, & Fischbach, 2022; Ferrão, 2012). Such a constructive use of VA scores can help create parsimonious samples, identify and appraise effective schooling, and aid schools that are in need of support. These could be future applications of VA scores and the present findings on VA score stability.

7.7 Conclusion

The present study provided evidence for the moderate stability of primary schools' VA scores. Only 34-38% of the schools showed stable VA scores across two years with moderate correlations of r = .34 for mathematics and r = .37 for language achievement as the outcome domains. The number of stable schools did not differ greatly between mathematics and language. Real-life implications for schools may be consequential with only about one third of schools having a stable VA score over time. This finding indicates that VA scores should not be used as the only measure for purposes of accountability. Thus, both public and private educational services should refrain from using VA scores as the (sole) metric to rank schools in their effectiveness and policymakers need to consider the present controversy about school VA scores. Complementary sources of data to make appropriate educational decisions are strongly recommended, as school VA scores do not seem to be stable enough over time and theoretical assumptions about the VA scores don't seem to hold in practice (Conaway & Goldhaber, 2020; Scherrer, 2011). Nonetheless, we argue that VA models could be employed to find genuinely effective teaching or school practices—especially in heterogeneous student populations, such as Luxembourg, in which educational disparities are already an important topic in primary school (Hoffmann et al., 2018). Here, VA scores could help researchers look past these disparities and investigate the schools with stable positive VA scores and learn from them.

8 Study 4

Instructional Quality and School Climate in Luxembourg's *Écoles Fondamentales*: Findings from the SIVA Study

Valentin Emslander[a], Cassie Rosa[a], Sverre Berg Ofstad[b], Jessica Levy[a], and Antoine Fischbach[a]

[a] Luxembourg Centre for Educational Testing (LUCET) at the University of Luxembourg, Faculty of Humanities, Education and Social Sciences, University of Luxembourg, Luxembourg

[b] Centre for Educational Measurement at the University of Oslo (CEMO), Faculty of Educational Sciences, University of Oslo, Norway

Submitted to *National Education Report Luxembourg* [6]

Currently *In-Principle Accepted*

This study was preregistered and all materials can be found at https://doi.org/10.17605/OSF.IO/X3C48

[6] The numbering of headings, tables and figures has been adjusted to align with the structure of the present work. The Appendix is integrated general Appendix at the end of this thesis. Additional methods and results from the SIVA project will be presented. They had been excluded from publication in the National Education Report Luxembourg for brevity.

Abstract

In such a diverse context as Luxembourg, educational inequalities can arise from the language spoken at home, a migration background, or a family's socioeconomic status. This diversity leads to different preconditions for learning maths and languages (e.g. the language of instruction) in school and thus shapes the school careers of students (Hadjar & Backes, 2021).

The purpose of the systematic identification of high value added in educational contexts (SIVA) project was to answer the questions (1) what highly effective schools are doing 'right' and (2) what other schools can learn from them to alleviate inequalities. In collaboration with the Observatoire National de l'Enfance, de la Jeunesse et de la Qualité Scolaire – Section Qualité Scolaire (OEJQS), we qualitatively and quantitatively investigated the differences between schools with stable high value added (VA) scores and those with stable medium or low VA scores from multiple perspectives. VA is a statistical regression method usually used to estimate school effectiveness considering diverse student backgrounds and preconditions.

We identified 16 schools with stable high, medium, or low VA scores in two cohorts. In these16 *écoles fondamentales*, we collected data on, among others, their pedagogical strategies, student backgrounds, and school climate in 49 classrooms, involving 511 Grade 2 students, and 191 teachers as well as parents, school presidents and regional directors. This diverse sample unintentionally mirrored Luxembourg's general population of *écoles fondamentales* in terms of gender balance, class size, socioeconomic status, and prior achievement.

The data collection focuses on variables that are central in school effectiveness and student success models such as teaching quality, school climate, and school organization (e.g., Hattie, 2008; Helmke et al., 2008; Klieme et al., 2001). We further investigated specificities about the Luxembourgish school system, which are not represented in international school learning models, such as the two-year learning cycles, the multilingual school setting, or the diverse student population.

When examining instructional quality and school climate, the study unveiled generally positive perceptions from students. They reported strong teacher-student relationships, although classroom management scored lower than other aspects of instructional quality, namely cognitive activation and teacher support. We observed a dynamic multilingual environment within classrooms, especially during informal interactions. Although German was predominantly used for instruction, teachers often switched to other languages to accommodate students' diverse linguistic backgrounds, emphasizing adaptability. This is important to foster a conducive learning environment and strengthen teacher-student relationships.

Despite the variation in value added (VA) scores between schools, the study did not find significant differences on most variables. This implies that other factors contribute significantly to school success and instructional quality in Luxembourg's *écoles fondamentales,* and further research is needed.

Keywords: Value-Added Modeling, School Effectiveness, Primary School

8.1 Educational Inequalities in Luxembourg

In such a diverse context as Luxembourg, educational inequalities can arise from various factors. For example, the language spoken at home, the migration background, or the family's socio-economic status can impact a child's school success. Also, the educational landscape in Luxembourg is unique due to its linguistic diversity, with three languages of instruction in the traditional public education system: Luxembourgish, German, and French. However, only about 32 % of *écoles fondamentales* students speak Luxembourgish—a language of primary instruction—with at least one of their parents as their first language (MENJE – Ministère de l'Éducation nationale, de l'Enfance et de la Jeunesse & SCRIPT, 2022). This linguistic diversity means that students bring different language backgrounds to Luxembourg's *écoles fondamentales*, where they will be taught in Luxembourgish in preschool and German in subsequent grades. Consequently, these different language backgrounds result in varying preconditions for school success, which impact not only language learning but also subjects like mathematics and thus shape students' school careers (Hadjar & Backes, 2021). With the growing language diversity and educational inequality, one would expect Luxembourg to perform increasingly poorly in international comparisons. However, the data suggests that Luxembourg remains stable in its test results in international comparison studies (Weis et al., 2020). This stability suggests that there must be factors driving school success against the odds, such as effective strategies to tackle educational inequalities. If we could identify these educational strategies, they could be a starting point for all schools to learn from. To find such factors is precisely the goal of our project, "Systematic Identification of High Value-Added in Educational Contexts" (SIVA). Concretely, we wanted to find out what some *écoles fondamentales* in Luxembourg could be doing "right" in dealing with (linguistic) diversity and what other schools might learn from them.

8.2 The SIVA Project: Identifying Effective Schools

In the SIVA project, we aimed to explore how to mitigate educational inequalities in Luxembourg's *écoles fondamentales*. With this project, we had three key objectives: (1) to identify highly effective schools and compare them with other schools, (2) to collect quantitative and qualitative data in these schools to discover what highly effective schools are doing differently, and (3) to identify practices that other schools can learn to reduce educational disparities. Collaborating with the *Observatoire National de l'Enfance, de la Jeunesse, et de la Qualité Scolaire – Section Qualité Scolaire (OEJQS)*, the project team set out to find the literal needle in the haystack of what makes the *écoles fondamentales* in Luxembourg effectively navigate diversity and bring all students to success. To identify especially effective schools, we used a statistical method called "value-added" (VA) analysis, which aims to identify how much "value" schools can "add" to the achievement of their students. VA analysis estimates school effectiveness while considering diverse student backgrounds, prior achievement, and other preconditions. The idea behind these scores is that there are several student factors, such as the students' home language, that the school cannot influence. With the help of VA scores, we can account for such factors and approximate the school's influence on students' achievements. To find a school's VA score, we can average all the students' VA scores within a school. This VA score says how well a school helps students thrive independent of their backgrounds. VA scores are widely applied for accountability reasons, for instance, in the US, to compare the value different teachers or schools add to their students' achievements (Amrein-Beardsley & Holloway, 2019). In our project, we used them as a starting point to analyze effective strategies in the school system. We thus selected *écoles fondamentales* with stable high value-added scores and compared them with those achieving stable medium or low VA scores.

8.3 Methods and Data Collection

To select the schools, we drew on two representative longitudinal data sets of students who participated in the *Épreuves standardisées* (ÉpStan) in Grade 1 in 2014 or 2016 and then again in Grade 3 in 2016 and 2018 two years later. The dataset included the students' math and language achievement and their background variables, such as SES and home language. In a prior study, we found that the stability of VA scores was rather low when only considering math *or* language as an outcome variable (for a review of research on VA-score instability, see Emslander et al., 2022). It is not possible to infer the cause of this instability with correlational data - it might reflect real changes but could also be due to issues with reliability or measurement error. Thus, for the SIVA project, we decided to use schools that had a stable VA score in *both* math and language. That way, we increased the reliability of selecting schools with a high, medium, or low VA score. As such, we identified 16 schools with stable high, medium, or low VA scores, across both subject domains, over two years as our sample.

After identifying the 16 schools, we collected data in all their 2nd-grade classrooms from January to March 2022. We collected data from several perspectives, combining quantitative questionnaire data with qualitative classroom observations. We collected quantitative questionnaire data from students, their parents and guardians, classroom and subject teachers, school presidents, and regional directors. Additionally, we collected qualitative observational data by visiting each classroom for a one-hour math lesson. Giving all stakeholders a voice, we aimed to find the literal needle in the haystack that drives possible differences between the schools.

Generally, we collected data on several aspects of instructional quality, school climate, and language use, among others. During the structured classroom observations, we observed which languages and pedagogical strategies the teachers used. We focused on three essential concepts of instructional quality (after Klieme et al., 2001; Praetorius et al., 2018):

- *student support*: Teachers promoting positive interactions among students and being attentive to students' needs to build strong teacher-student relationships.
- *cognitive activation*: Teachers engaging students in challenging tasks and building on what the students already know.
- *classroom management*: Teachers creating a well-structured working environment with minimal disruption so students know what to do and can concentrate on the learning activities.

Ultimately, the SIVA project draws from multiple data sources, including observations in 49 classrooms in 16 schools and questionnaires completed by 511 2nd-grade students, 410 of their parents, 191 of their classroom and subject teachers, 14 school presidents, and 13 regional directors. Comparing the SIVA sample with the entire population of about 150 *écoles fondamentales* in Luxembourg, we unintentionally made a close-to-representative choice as our sample is quite comparable to the general population on several variables, such as on the schools' gender balance, measures of socio-economic status, and prior achievement. Also, the 16 included schools were geographically evenly dispersed throughout the Grand Duchy, with more schools in the center and the south, where most people live.

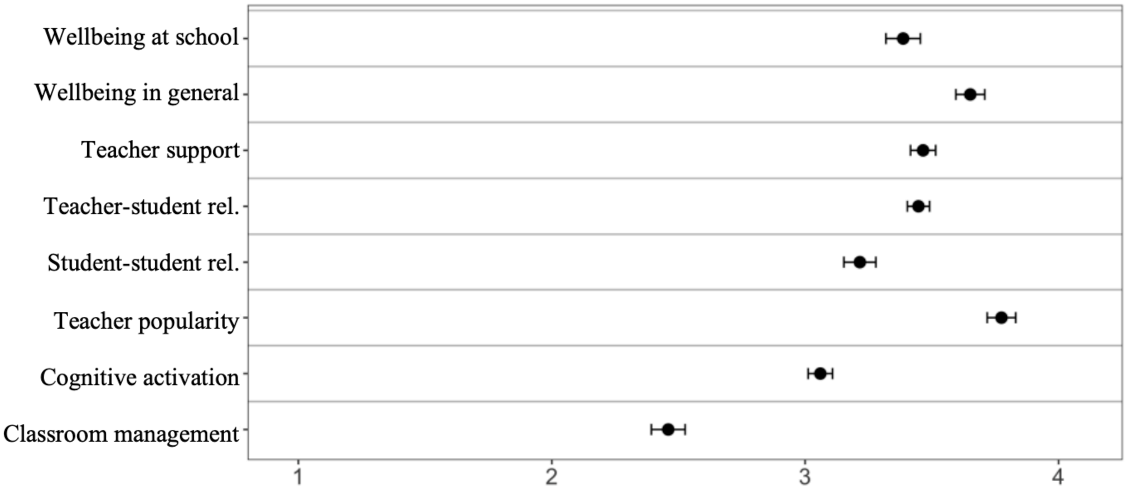
Although data collection for the SIVA project was somewhat invasive and took a lot of effort during the COVID-19 pandemic, we were met with much support from participants. This allowed us to gather rich qualitative and quantitative data from multiple perspectives. While our sample closely resembles key characteristic of the full population of cycle 2.2, we could only assess a select subsample of schools. In this chapter, we will focus on the quantitative data from the students and teachers because they are the most important actors in the classroom. We additionally highlight the more qualitative classroom observations and open-text questions to complement the quantitative aspects of learning and help to shed light on the classroom processes.

8.4 Findings and Implications

8.4.1 Instructional Quality and School Climate in Luxembourg's Écoles Fondamentales

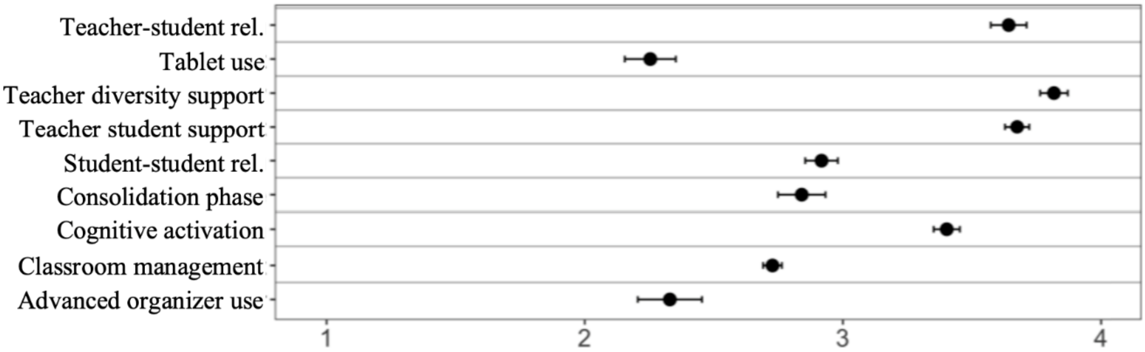
In Figure 20, we summarize the questionnaire results from the perspective of the 511 students. These results indicate that the students are experiencing a sense of well-being. While their relationships with fellow students are good, the students report better relationships with their teachers, and the teachers enjoy a high level of popularity among students. Whereas it is still discussed what influence teacher-student relationships might have in 2nd grade, teacher popularity is likely to positively influence the students' ratings of teacher-student relationships (Aleamoni, 1999; Emslander et al., 2023). The only aspect that the children perceive as somewhat lower, compared to other strategies of instructional quality, is classroom management, which is followed by cognitive activation and teacher support. This pattern is also found in German studies (e.g., Fauth et al., 2014).

Figure 20. *Student questionnaire results on selected variables (n = 511)*



The perspective of the 191 teachers presents a more mixed picture (see Figure 21). On the positive side, the teachers hold a favorable view of their relationships with students and their own ability to support their students, regardless of their background. However, the teachers perceive a lack of a digitalization strategy, particularly regarding the use of tablets. Additionally, similar to the students' perspective, they find the relationships among students to be more challenging than between students and teachers. Furthermore, teachers report that strategic consolidation phases, in which students "save" what they have learned, are not very common in their lessons. The teachers reported making even lesser use of advanced organizers, thus not presenting the structure of the learning unit to the students ahead of time. They also identify room for improvement in classroom management strategies but less in cognitive activation and student support. This pattern of classroom management being the least developed of the three teaching strategies—with cognitive activation in the middle and student support as the most developed strategy of instructional quality—aligns with the observations made by the students in our sample and from other studies (Fauth et al., 2014).

Figure 21. *Teacher questionnaire results on selected variables (n = 191)*



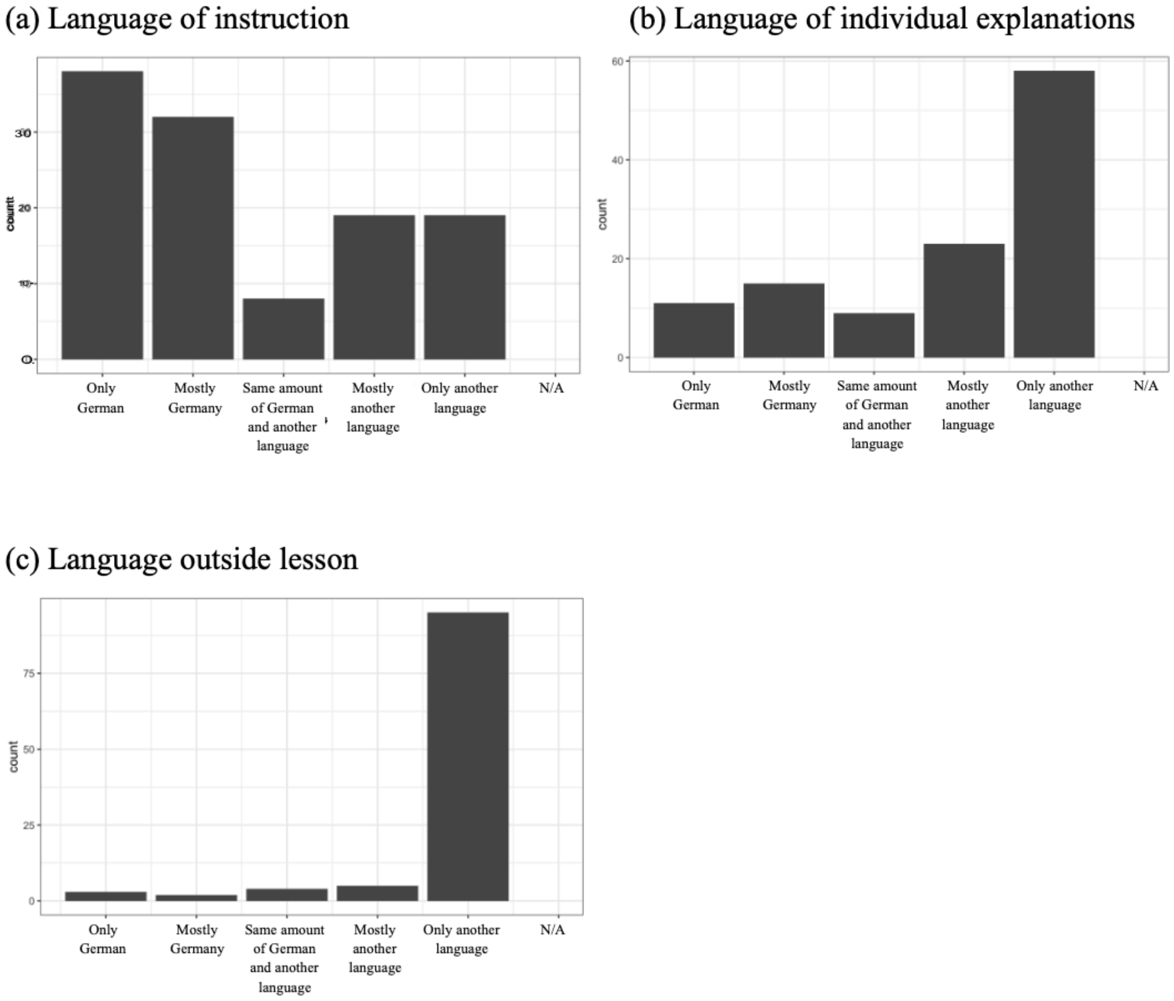
Surprisingly, we found no significant differences between schools with high-, medium-, and low-VA scores. This means that the pattern of the results was the same among schools with all types of VA scores, indicating that none of the described factors seem to be linked to differences in school VA scores. In other words, we did not find any pedagogical strategies or other aspects that set schools with a high VA score clearly apart from other *écoles fondamentales.* Thus, we did not find the figurative needle in the haystack of the quantitative data we investigated.

8.4.2 What we Observed During the Standardized Classroom Observations

The more qualitative part of our study involved classroom observations conducted by two educational experts during one-hour math lessons. We averaged their ratings, which showed high agreement. We observed the language(s) spoken by the teacher during the lessons, during individual explanations, and during breaks or before and after the lessons (see Figure 22). Whereas the language of instruction was either “only German” or “mostly German” in most cases, this changed in more informal settings. More than half of the time, when the teacher explained something to an individual student, they used a language other than German. This language diversity reflects the students' multilingual backgrounds and the teachers’ willingness to adapt flexibly to their students’ language skills or preferences. The language diversity was even higher in informal chatter before and after the lessons or during breaks: About 85% of the time, the students and teachers spoke another language than German. The most popular non-

German language was Luxembourgish, followed by French. A few conversations between a teacher and a student were in Portuguese, Italian, or Bosnian.

Figure 22. *Observed language use*



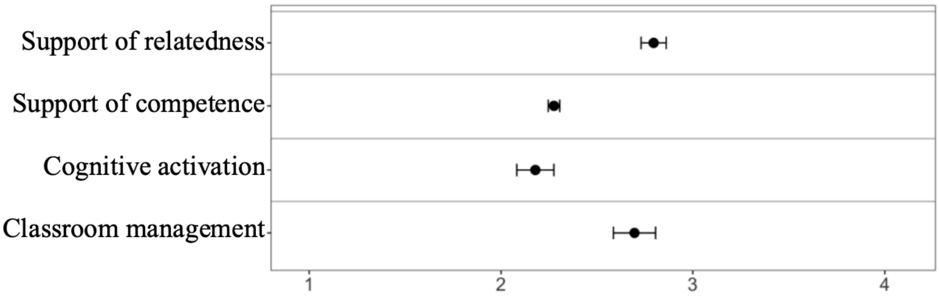
(a) Language of instruction

(b) Language of individual explanations

(c) Language outside lesson

Figure 23 sums up selected variables from the classroom observations. Similar to the students' and teachers' findings, classroom management and cognitive activation appeared less well-developed in the observed lessons. However, the order of importance of the strategies for instructional quality was different. From the classroom observers' perspective, cognitive activation was the least developed aspect, followed by the two other strategies of student support—supportive teacher-student relationships and the teacher's support competence—and classroom management (see Figure 23).

Figure 23. *Classroom observations questionnaire results on continuous variables (n = 49 classrooms observed with two observers each)*



When comparing schools with a high, medium, or low VA score, we encountered some interesting, yet mostly statistically insignificant, findings. High VA schools demonstrated a descriptively higher tendency to use German as the language of instruction. However, in cases where teachers in high VA schools employ other languages, teachers in schools with a high VA score speak the home language of the students. This adaptivity demonstrates a strategic utilization of language diversity among the students. Additionally, it is important to have a fixed language of instruction while remaining flexible and responsive to the linguistic needs of the students. At the same time, it seems that *code-switching* is an asset to the students. *Code-switching* refers to speaking multiple languages in the same situation, such as when a teacher explains a question in Luxembourgish and German (Rampton, 2017). This lively-discussed theory argues that using the students' home language makes them feel safe and valued, which are the prerequisites for deep learning and understanding in diverse societies (Creese & Blackledge, 2015; Lin, 2013). The results of our study support this theory somewhat by indicating that strategic use of *code-switching* might be associated with a higher VA score in a school.

Interestingly, while we observed no significant differences in the quantitative tools assessing pedagogical strategies, we could observe some substantial differences in the qualitative ratings of language use. These differences hint at the importance of conducting classroom observations during on-site school visits to capture unique teaching practices and identify differences between schools. It is vital to acknowledge the complementarity of methods, with structured observations being particularly valuable in the school context.

As another source of qualitative data, we investigated the open texts the teachers wrote in the questionnaires. Here, three teachers emphasized the need for smaller classes and more teaching personnel to create an effective working environment. Another three teachers urged for a faster digitalization of teaching and learning materials—they mentioned the lack of tablets and computers—for their students to support teaching 21st-century skills and practical media use. Two teachers further expressed their positive experiences with multilingual teaching and *code-switching* in everyday school life, and another two teachers shared their positive experiences with supportive colleagues, describing teaching as a great job due to the immediate positive feedback. Overall, these answers mirror what we found in the questionnaire results.

8.4.3 Conclusion and Outlook

Two key take-home messages emerged from the SIVA project. Firstly, our analyses revealed generally positive perceptions of educational quality among students and teachers in Luxembourg. Key findings emphasize the significance of teacher-student relationships while highlighting opportunities for improvement in classroom management strategies. Intriguingly, differences in VA scores did not yield significant variations in most measures, suggesting the presence of other influential factors in school performance. Our analyses of open-text responses and observations suggested a faster digitalization and the active use of multiple languages, namely *code-switching*, as important strategies to investigate in future research. In the classroom observations, we witnessed many teachers using German as the language of instruction but switching to a student's home language for individual explanations and during breaks.

Secondly, our study showed the importance of incorporating qualitative alongside quantitative data in educational research because it enables a deeper understanding of teaching strategies and school environments. Whereas quantitative analyses, particularly VA scores, help identify overall trends and differences, it is through qualitative observations that we discover subtle nuances and intricacies. Thus, by combining quantitative and qualitative approaches, researchers can comprehensively understand educational contexts and uncover valuable insights to inform effective educational strategies and decision-making. Thus, it is crucial to perform structured classroom observations or teacher interviews to go beyond looking at inputs and outcomes of the educational system and understand the learning processes taking place directly in the classroom.

These correlational findings could suggest implications for educational decision-makers in Luxembourg. Fostering already successful strategies, such as creating positive teacher-student relationships, and addressing classroom management challenges could enhance educational quality. The importance of such actions is supported in the international literature (Emslander et al., 2023). Motivating teachers to stick to the instruction language and, for additional explanations, use their students' home language might help create a welcoming and effective learning environment. Our research also advocates for continued exploration of additional variables potentially influencing educational quality further, such as digitalization and language use. A longitudinal perspective may provide further insights into the evolution of educational dynamics in Luxembourg and could guide future efforts to refine the education system based on causal results. In conclusion, the SIVA project offers a correlative, multifaceted perspective on educational quality in Luxembourg, highlighting the importance of embracing multilingualism in the proper context and integrating quantitative and qualitative research methods.

8.4.4 Bulletpoints

1. The SIVA project encompassed a multi-perspective data collection in 16 *écoles fondamentales*, 49 classrooms, involving 511 2nd-grade students, 191 teachers, as well as parents, school, presidents, and regional directors. This diverse sample unintentionally mirrored Luxembourg's general population of *écoles fondamentales* regarding gender balance, class size, socio-economic status, and prior achievement.

2. When examining instructional quality and school climate, the study unveiled generally positive perceptions of students and teachers in quantitative and qualitative data we collected through questionnaires and observations. Students and teachers reported strong teacher-student relationships, although classroom management strategies were less frequent than other instructional quality aspects, namely, cognitive activation and teacher support.

3. We observed a dynamic multilingual classroom environment, especially during informal interactions. While teachers used predominantly German for instruction, they often switched to other languages to accommodate students' diverse linguistic backgrounds, emphasizing adaptability. This flexibility might be important to foster a conducive learning atmosphere and strengthen teacher-student relationships.

4. Despite the variation in value-added (VA) scores among schools, the study found no statistically significant differences in questionnaire responses from students or teachers and only small statistical differences during structured classroom observations. This lack of differences implies that other factors contribute to the differences in school performance and instructional quality in Luxembourg's *écoles fondamentales* and that classroom observations are crucial.

5. The study underscores the essential synergy between quantitative and qualitative research approaches to understand the teaching and learning processes in schools better. It highlights the importance of structured classroom observations and on-site school visits, calling for a greater focus on the teacher in Luxembourg's educational research landscape also, e.g., through structured teacher interviews.

8.4.5 Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the authors used GPT-3.5 from OpenAI (2023), Grammarly (https://www.grammarly.com), and Writefull (https://www.writefull.com) for proofreading purposes and to ensure linguistic precision, good style, and readability. After using these tools, the authors reviewed and edited the text as needed and take full responsibility for the content and wording of the publication.

.

8.5 Additional Methods to Study 4 [7]

8.5.1 School Selection and Sample Size

In study 3, we found that the stability of VA scores was relatively low when only considering math or language as an outcome variable. Therefore, for Study 4, we decided to use schools with a stable VA score in *math and language*. In this way, we increase the reliability of selecting schools with a high, medium, or low VA score that is stable over time and across subject domains.

We defined a *stable* school as one with a VA score in the same quartile of the VA ranking in 2016 and 2018. As such, the target sample size is defined by the underlying data, resulting in a total of 16 primary schools with stable high (*n* = 5), medium (*n* = 4), or low (*n* = 7) VA scores. In these schools, we focused on Grade 2 (cycle 2.2), and—after a statistical power analysis—we aimed for a sample of at least 258 students. With the program G*Power, we conducted a power analysis (Faul et al., 2007, 2009). Given a significance level of .05 and a power of .90, we should be able to detect small correlational effects of .02 with a sample of 258 children. Given our 16 statistically selected schools and considering dropouts, we expected a sample size between 320 and 420 students and roughly the same number of parents and teachers. The exact student sample is defined by the number of Grade 2 students (1) whose parents have agreed to their child's participation and (2) who were in school the day of the test. The exact teacher and school president (principal) sample is defined by the number of teachers at the school, teaching in grades 1 (cycle 2.1) through 6 (cycle 4.2), who have agreed to participate and send us their questionnaire until the end of March 2022. The exact parent sample is defined by the number of Grade 2 students' parents who have agreed to participate and send us their questionnaire by the end of March 2022. Concretely, within the 16 selected schools, we were able to collect data on 49 classrooms with a total of 511 students, their 410 parents, 191 classroom and subject teachers, 14 school presidents, and questionnaires from regional directors in 13 of those schools.

8.5.2 Measured Variables

We chose a wide range of potentially explanatory variables to get a well-rounded overview of variables that could explain the differences between schools with a stable high,

[7] The additional methods and results for study 4 are adapted from the unpublished third and final SIVA project report to the OEJQS: Emslander, V., Rosa, C., Ofstad, S. B., Levy, J., & Fischbach, A. (2024). *Systematic Identification of High Value-Added in Educational Contexts: SIVA* (Report No. 3). Observatoire National de l'Enfance, de la Jeunesse et de la Qualité Scolaire (section Qualité Scolaire).
The study was preregistered and all materials can be found at https://doi.org/10.17605/OSF.IO/X3C48

medium, or low VA score. The Theoretical Considerations section details the selection of variables based on a narrative literature review. We collected data on several aspects of instructional quality, school climate, TSR, language use, and more. For a comprehensive list of all variables and their measurement in specific target groups, please see the Scale Handbook and the respective questionnaires, as well as the observation sheet in the 2nd SIVA report, freely accessible at https://doi.org/10.17605/OSF.IO/X3C48. In the following, general explanations for the decision of the assessed constructs will be given. The tools for classroom observations and questionnaires rely on well-established international references but were adapted to the particularities of the Luxembourgish school system. For example, terms commonly used in Luxembourg replace international terms (e.g., “school president” replaces “school principal”), and questionnaires were translated into the most common language for wide accessibility (e.g., the parent questionnaire was available in German, Portuguese, English, and French).

To later calculate inter-rater agreement, all observations were conducted by two trained observers—one from the *Observatoire national de l’enfance, de la jeunesse et de la qualité scolaire - Section Qualité Scolaire* and one from the *Luxembourg Centre for Educational Testing*. We observed one math lesson per classroom with standardized observation sheets. All observations were made in the morning between 7 and 12 a.m. to have some stability in the circadian rhythm. The observation sheets were designed based on the TBD model of Klieme et al. (2001), which is why we observed the constructs of cognitive activation, student support, and classroom management, based on the observations conducted by Praetorius et al. (2018). Other constructs observed are instructional and pedagogical strategies, such as language use, the implementation of rituals in the classroom, and the application of new media. All observers were trained in a workshop before the observation.

The student questionnaires were provided on paper in German, with a standardized protocol for the assessment, including translations in Luxembourgish, French, and Portuguese for further specifications. Their maximum duration was one school lesson (50 minutes). The questions were based on the TBD model by Klieme et al. (2001) and were adapted from existing questionnaires for similar age groups (e.g., Fauth et al., 2014). Furthermore, further constructs such as school climate (e.g., Bear et al., 2011), well-being (Weber et al., 2013), and other motivational/emotional variables (e.g., Peixoto et al., 2015) were assessed.

Questionnaires for teachers, school presidents, and regional directors were provided in German and took less than 30 minutes to complete. Teachers were asked for information on their background (e.g., career pathways) as well as on the same constructs of the Klieme et al. (2001) model (oriented at Fauth et al., 2014). In addition, constructs such as collective teacher efficacy (Skaalvik & Skaalvik, 2007) and school climate (e.g., Bear et al., 2015; Leff et al., 2011) were assessed. School presidents and regional directors were also asked about their perception of collective teacher efficacy and school climate. Furthermore, organizational aspects were assessed, such as collaboration or communication within the school and with the *maison relais*.

The questionnaires for parents (max. 20 minutes) were provided in German, Portuguese, English, and French. It assessed background information on students' educational pathways of the child (e.g., grade retention), as well as parental involvement (e.g., Georgiou & Tourva, 2007) and the parents' perception of school climate (e.g., Bear et al., 2015).

All participants/their legal guardians were informed about the study, including how to contact the authors for any questions they may have. In the consent form, it was explained which data would be collected and that participants could withdraw at any time. Participants gave their informed consent before participating. Legal guardians gave their child's informed consent in advance for their child's participation.

8.5.3 Quality Assurance

Before data collection, we registered the SIVA project in the University of Luxembourg's General Data Protection Regulation (GDPR, https://eugdpr.org/the-regulation/) registry and the Bridge Builder procedure for educational research in Luxembourg (https://bridgebuilder.lu). All data were pseudonymized in accordance with European GDPR guidelines (https://eugdpr.org/the-regulation/). Pseudonymization of the collected data was warranted with the help of a trusted third party (https://www.itrust.lu). The Ethics Committee of the University of Luxembourg approved the SIVA project on November 16, 2021 (https://wwwen.uni.lu/research/researchers_research/ethics_policies_and_committees). Thus, the project adheres to the American Psychological Association Ethical Principles & Code of Conduct (https://www.apa.org/ethics/code/). Finally, we preregistered the SIVA project, its data collection procedure, materials, and data analyses on the online repository of the Open Science Framework at the Centre for Open Science (OSF.io; https://doi.org/10.17605/OSF.IO/X3C48; Emslander et al., 2022). OSF is an online tool that facilitates collaboration on large projects and makes them publicly accessible for broad dissemination. The latter aligns with the open science movement, which strives to make research replicable and reliable and combat scientific misconduct.

Data Preparation

After data collection, participants sent their questionnaires to the Luxembourg Centre for Educational Testing, where a team of research and development specialists digitalized them and extracted the data. The digitalized data were then formatted into six datasets, one for each of the six groups of students, teachers, parents, school presidents, regional directors, and classroom observations. These datasets were checked by IT experts for obvious errors and then by the research team for errors in the questionnaires or minor typing mistakes.

To allow for the calculation and interpretation of the results, the data had to be prepared by (1) inversing, (2) scale creation, and (3) rescaling several items. First, we inverted the answer options to negatively worded items. That is, a high agreement with the statement "I don't like to go to school," for example, signifies a positive attitude towards school. This was a prerequisite for the creation of the scale in the second step. Here, we combined several items that assess the same constructs into one scale by calculating the average of these items. For example, all items assessing teacher support were averaged to build a single indicator for teacher support. Further, we rescaled several measures to be comparable with the others on visual displays. Teacher age, for example, ranges between 25 and 70, whereas most other items were answered on a five-point Likert scale from 0 to 4. To make both kinds of measures accessible and readily understandable in the same graph, we truncated the age scale, for instance, so it fits the other scales better.

8.5.4 Data Analysis

In this final SIVA project report, we followed a four-step process to analyze the current data and present the results. First, we examined the sample statistics to check the generalizability of our sample across the Grand Duchy. Second, we calculated and reported means and standard deviations for all variables in all schools to gain a general overview of all constructs. These results are the first insights into whether the constructs functioned as expected. Third, we divided the sample into three groups of schools with high, medium, or low VA scores, respectively. Between these three groups, we want to determine whether descriptive differences in their background variables exist. Additionally, we focus on the differences between schools in the measured variables of the SIVA project. Going beyond the background variables, we aimed to determine what might drive the difference between schools with a stable high, medium, or low VA score—a potential "needle in the haystack." Here, we examine group differences in a graphical way to make the group differences easily visible. Fourth, we investigated the responses to open-text questions in our surveys and observation records, aiming to cluster them for a deeper understanding of the school climate and developmental needs of the schools.

8.5.5 *Additional Results*

In this additional results section, we delve into supplementary findings of the SIVA project that were omitted from the National Education Report Luxembourg for brevity. This section is structured along the steps of the statistical analyses:

8.5.5.1 **SIVA Sample Statistics:** We present descriptive statistics for the entire SIVA sample to determine its representativeness for Luxembourg. Examining key characteristics and demographics of the sample provides a context for subsequent analyses.

8.5.5.2 **Means and Standard Deviations for Constructs of Interest:** This section explores the means and standard deviations of the constructs of interest identified in the narrative literature review.

8.5.5.3 **Differences Between High, Medium, and Low VA Schools:** To further analyze our findings, we compare the descriptive statistics across high, medium, and low VA groups. Additionally, we investigated the differences between high, medium, and low VA schools in educational psychological variables. This analysis also offers insights into the potential implications of our findings for educational practices.

8.5.5.4 **Analysis of the Open Text Fields**: As we acknowledge the significance of incorporating qualitative data, this section delves into the responses for the open-text questions in our questionnaires and the observation sheet. Our specific focus is on whether these responses can be clustered to provide a more comprehensive understanding of the perspective of teachers and school presidents. This analysis underscores the importance of considering qualitative perspectives in educational decision-making processes.

8.5.5.5 **Conclusion:** In the concluding section, we summarize additional findings, implications, and recommendations from our SIVA project. We discuss the significance of our research in filling existing gaps in educational research and the potential implications for educational practices.

Throughout the results, we emphasize the need for methodological rigor, the potential for future research, and the importance of using our findings as a foundation for informed decision-making in education.

8.5.5.1 SIVA Sample Statistics

Table 22 displays the exact number of schools, classrooms, students, parents, teachers, school presidents, and regional directors from whom we collected data. We have identified 16 primary schools in Luxembourg with a stable high, medium, or low VA score over two years, with math and language as an outcome variable. These were the basis of our data collection.

Within these 16 schools, we were able to collect data on 49 classrooms with a total of 511 students. This number exceeded our expectations, and the a priori calculated minimum sample size of 258. The number of participating parents also exceeded our expectations, with a total of 410 parents participating in the SIVA data collection. This number of responses equals a participation rate of more than 80%. Further, 191 classroom teachers and subject teachers participated, which roughly equates to one classroom teacher and about three subject teachers per class filling out our survey. Of the 16 schools, 14 school presidents participated. In 13 of these schools, we have completed questionnaires from regional directors.

With this sample, we will be able to statistically detect small to moderate effects in students and parents. Both groups were organically oversampled. According to our prior calculation of statistical power, we can be confident in uncovering a small effect if it exists in a sample of Luxembourgish students in Grade 2 (cycle 2.2).

Table 22. *Sample sizes of all participant groups*

	High VA	Medium VA	Low VA	Total
Schools	5	4	7	**16**
Classrooms	14	11	24	**49**
Students	129	131	251	**511**
Parents	85	100	211	**410**
Teachers	68	39	92	**191**
School Presidents	5	3	6	**14**
Regional Directors	4	3	6	**13**

Note. Not all regions were involved with the SIVA study, but the 13 regional directors represent all regions where we collected data for the SIVA project. VA = value added.

As shown in Table 22, we collected somewhat similar amounts of data in all three VA groups. However, as we identified seven schools with a low VA score and only four with a medium and five with a high VA score, the schools with a low VA score are somewhat overrepresented. Because schools with low VA scores were also somewhat larger, about half of the classroom observation data and about half of the student, parent, and teacher questionnaire data were collected in these schools.

Table 23 shows summary statistics on the background variables. Comparing the SIVA sample with the entire sample of about 150 primary schools in Luxembourg, we unintentionally made a close-to-representative choice, as our sample is quite comparable to the general population. There are only a few slight variations. When looking at class sizes, for example, we have slightly more students per class in the SIVA sample than in the general population. Further, we have 5.4% more students in the SIVA sample who do not speak Luxembourgish at home.

Table 23. *Descriptive statistics of high, medium, and low VA schools*

	High VA	Medium VA	Low VA	SIVA sample	Entire Population
Number of schools	5	4	7	**16**	**146**
Average number of students	24	21	30	**26**	**24**
Percentage female	44.1%	51.8%	52.1%	**49,8%**	**49.9%**
Percentage speaks no Luxembourgish at home	53.4%	44.7%	58.9%	**56,1%**	**50.7%**
Mean HISEI-score (SD)	48 (15)	52 (14)	44 (15)	**47 (15)**	**50 (15)**
Mean math achievement in grade 1 (SD)	0.14 (1.34)	0.13 (1.06)	0.04 (1.21)	**0,09 (1,22)**	**0.27 (1.24)**
Mean language achievement in grade 1 (SD)	0.16 (1.09)	0.17 (0.98)	-0.02 (0.94)	**0,07 (1,00)**	**0.22 (0.97)**
Mean math achievement in grade 3 (SD)	0.01 (1.10)	0.19 (0.92)	0.06 (0.91)	**0,07 (0,97)**	**0.19 (0.99)**
Mean language achievement in grade 3 (SD)	0.05 (1.35)	0.26 (1.24)	-0.24 (1.25)	**-0,05 (1,29)**	**0.13 (1.27)**

Note. Here, high, medium, and low VA schools are compared with the entire population of primary schools in Luxembourg. VA = value added, SD = standard deviation (a higher value shows greater dispersion around the mean), HISEI = highest available *international socioeconomic index of occupational status* (ISEI; Ganzeboom, 2010).

Although the differences in the achievement mean between the SIVA sample and the entire population might seem meaningfully large, a look at the standard deviations (SD, a measure of variance around the mean) reveals that these deviations are negligible.

These descriptive statistics on the SIVA sample suggest that we will be able to present results on key constructs for Luxembourg's Grade 2 (cycle 2.2) children from a sample that is close to representative and covers roughly 10% of a Luxembourgish birth cohort. The large sample is also due to a relatively high response rate among parents, almost as high as the response rates in ÉpStan (LUCET, 2023), in which participation is mandatory. To further corroborate the value of the SIVA sample, the schools we identified as having a stable high, medium, or low VA score happen to be well-distributed across the country.

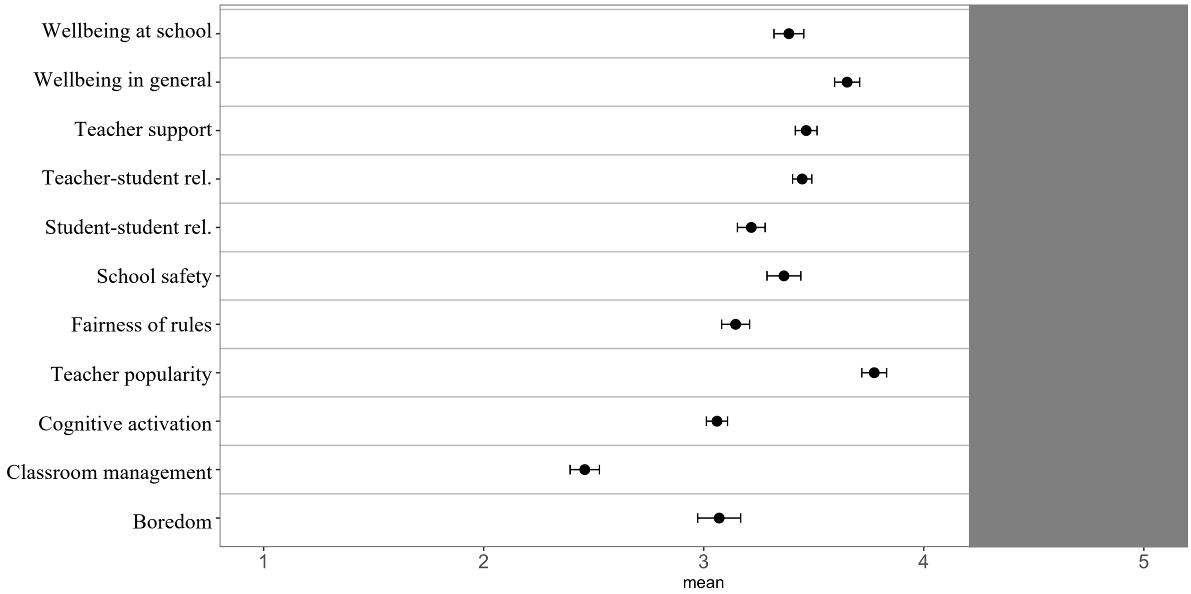
In conclusion, the present sample is close to a representative sample. It will provide reliable insights into the happenings inside Grade 2 classrooms and the primary schools in Luxembourg from 6 different perspectives. In addition to classroom observations, students, teachers, parents, school presidents, and regional directors filled out questionnaires on their views on the classroom and the school. With this wide array of different angles and a mixed-methods database, we will now take a first look at the results of the SIVA study.

8.5.5.2 Means and Standard Deviations (SD) for the Constructs of Interest

In the following, we describe the mean values of the SIVA sample and compare them with international research, such as in Germany. We concentrate on analyzing the dispersion (standard deviation) and overall means. Although our sample is not intentionally stratified, it offers a good representation of the Grand Duchy, allowing us to highlight several blind spots. We will focus on a selection of meaningful variables to preserve the readability of this manuscript but also report null results.

From the students' perspective, the mean scores provide a generally positive image. A summary of the selected variables from the student's point of view is shown in Figure 24. It suggests that students are doing well at school and are experiencing a high sense of well-being. Although their relationships with fellow students are good, the students report having even better relationships with their teachers. While it is still discussed what influence teacher-student relationships might have at that age, teacher popularity could affect student ratings of their relationship with their teacher (Aleamoni, 1999; Emslander, Holzberger, Fischbach, et al., 2023). The students seem to like their teachers very much. The only aspect that children perceive as notably lower is classroom management, compared to other dimensions of instructional quality, followed by cognitive activation and teacher support. This pattern is also shown in studies from Germany. For example, Fauth et al. (2014) also found the smallest mean value for classroom management, followed by cognitive activation and supportive climate in their study focusing on science education with *n* = 1,556 students in 89 classes. It is crucial to remember that most variables were rated on a 4-point scale and the fifth point only refers to a select few variables denoted with a blue asterisk in the respective figure.

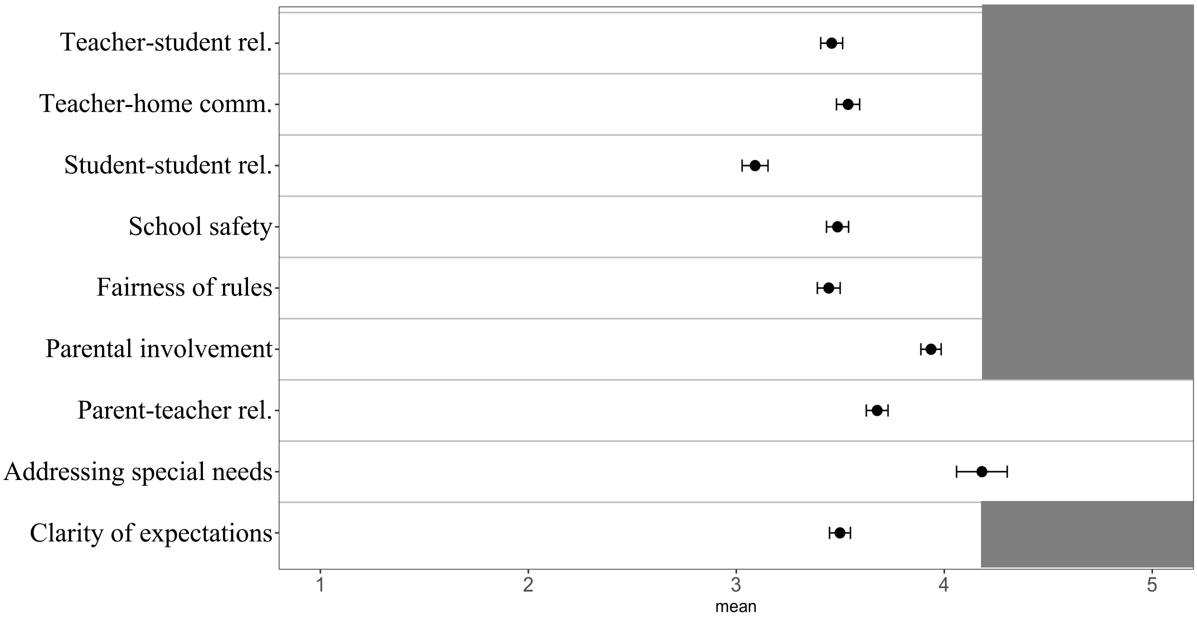
Figure 24. Student questionnaire results on continuous variables (n = 511)



Note. The items/scales were answered on a 4-point Likert scale from 1 to 4 and thus are truncated with a gray bar. The 5-point depiction only helps compare the results with the other groups.

Analyzing the variables from the parents' perspective in Figure 25, we observe an overall positive picture, similar to the students' perspective. Parents, too, have a considerably more critical view of the relationship between students than between students and their teachers. As primary school is a time of important cognitive and socioemotional development (Emslander & Scherer, 2022), children are relearning how to engage with their peers throughout this stage of development. As a result, students notice that children may not show the same levels of empathy as adults. Children gradually eschew egocentric worldviews throughout their school careers (Kesselring & Müller, 2011). The figure also shows that parents express high satisfaction with their involvement in their children's education, recognizing its significance for their academic success. Additionally, parents believe that schools are adapted to addressing special needs, showing a favorable opinion of the school's support systems.

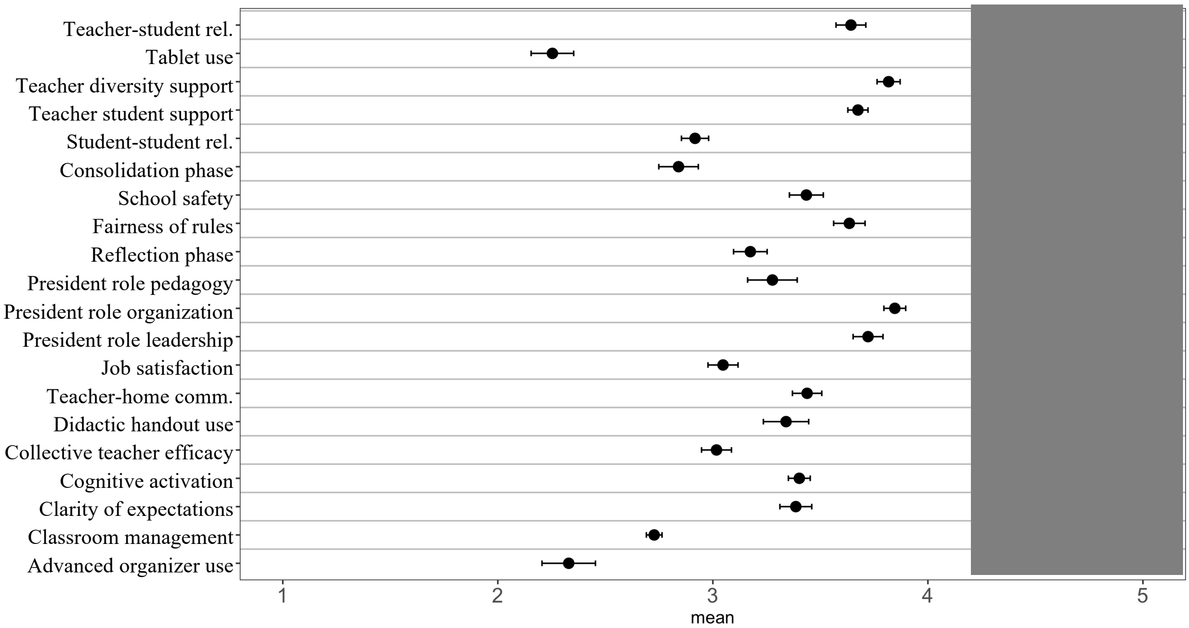
Figure 25. Parent questionnaire results on continuous variables (n = 396)



Note. Most items/scales were answered on a 4-point Likert scale from 1 to 4 and thus are truncated with a gray bar. The 5-point only applies to the items/scales that are not grayed out.

The perspective of the roughly 200 teachers presents a more mixed picture (Figure 26). On the positive side, teachers have a favorable view of their relationships with students and their ability to support them. They also acknowledge the importance of organizational and leadership tasks carried out by the school administration. The role of the school president is mostly seen as providing organizational support and taking over leadership responsibilities. However, the teachers also express some negative views. They perceive a lack of progress in digitalization, particularly regarding the use of tablets. In addition, they find the relationships between students more challenging than between students and teachers, similar to the perspectives of the students and parents. Furthermore, they report that reconciliation phases, in which students “save” what they have learned, are less frequent in their lessons. Teachers reported making even less use of advanced organizers, thus not presenting the structure of the lesson or learning unit to the students ahead of time. Like students, teachers also identify room for improvement in classroom management, but less so in cognitive activation and student support. This pattern of classroom management being the least developed area—with cognitive activation in the middle and student support as the most developed area of instructional quality—again aligns with the observations made by the students in our sample and from other studies (Fauth et al., 2014).

Figure 26. Teacher questionnaire results on continuous variables (n = 200)

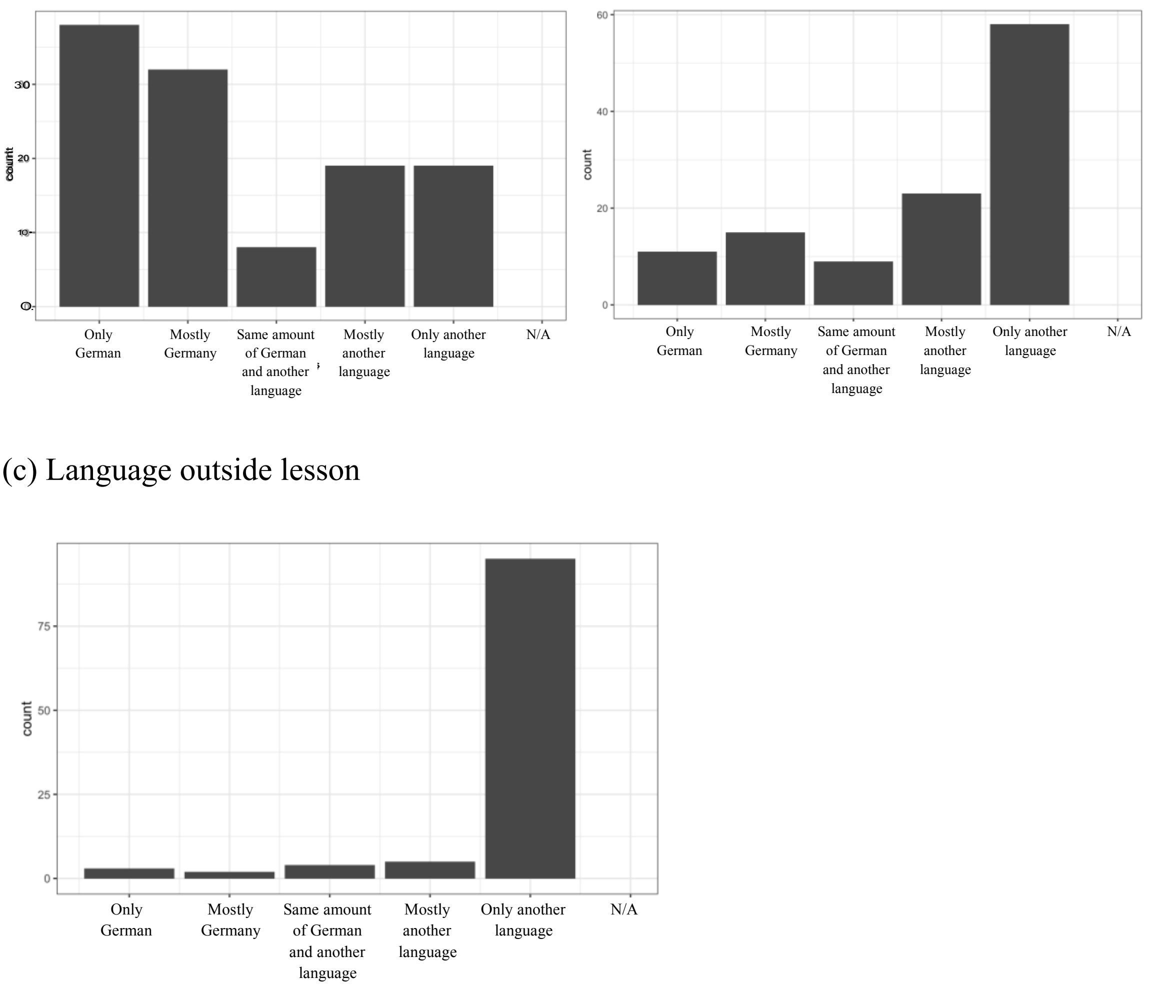


Note. The items/scales were answered on a 4-point Likert scale from 1 to 4 and thus are truncated with a gray bar. The 5-point depiction only helps compare the results with the other groups.

One of the more qualitative parts of the SIVA data collection is the perspective gained from classroom observations. Regarding the languages in the breaks and between learning phases, the classroom environment was highly multilingual. We observed the languages spoken by the teacher and the students during the lessons as the language of instruction and explanation, as well as the languages spoken before and after the classes or during breaks. Figure 27 shows the results of this active multilingualism in the classroom. Although the language of instruction was usually “only German” or “mostly German,” this changed in more informal settings. More than half of the time, when the teacher explained something to an individual student, they used exclusively a language other than German. It reflects both the multilingual background of the students and the willingness of the teachers to adapt flexibly to the language skills and preferences of their students.

In informal chat before and after lessons or during breaks, teachers and students again spoke languages other than German more often. Approximately 85% of the time, the students and teachers spoke another language than German. The additional languages spoken were mostly Luxembourgish, followed by French. Much fewer conversions were made in Portuguese, Italian, or Bosnian, with mainly only one or two teachers speaking these languages with their students.

***Figure 27**. Observed language use*



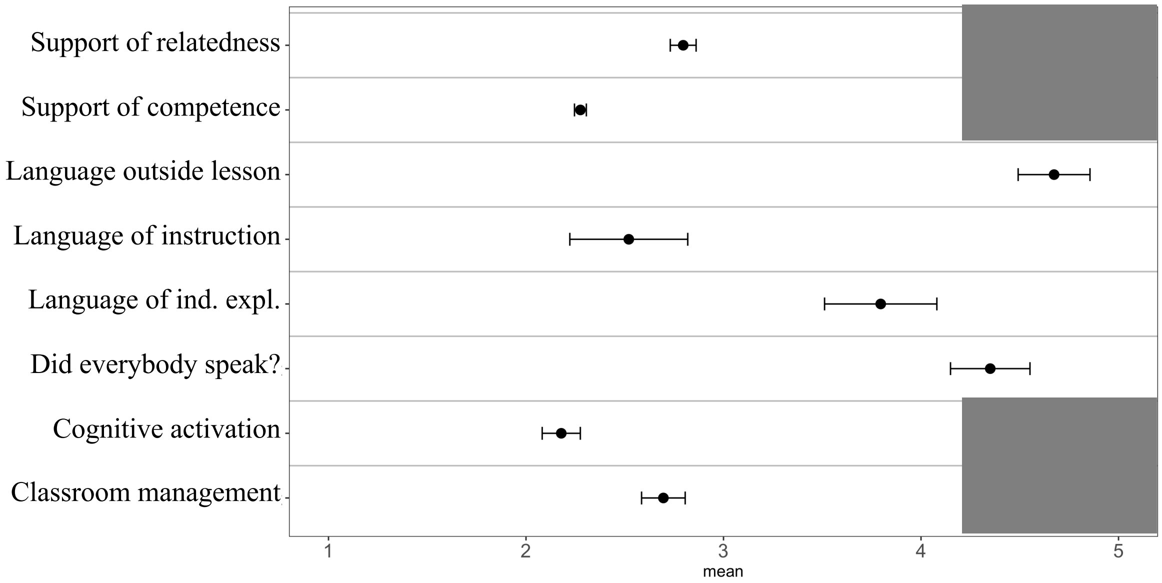
(a) Language of instruction

(b) Language of individual explanations

(c) Language outside lesson

Figure 28 summarizes the language variables, showing the use of any language other than German and selected continuous variables from the classroom observations. In particular, in most cases, all students had the opportunity to contribute and participate in the lessons. Similar to the previous findings, classroom management and cognitive activation appeared less developed in this context. However, the order of the instructional quality indicators differs from the previous perspectives. From the classroom observers' perspective, teachers used cognitive activation fewer than the other two dimensions of student support—i.e., supportive teacher-student relationships and the teacher's support competence—and classroom management.

Figure 28. *Results of the classroom observation questionnaire on continuous variables (n = 49 classrooms observed with two observers each)*



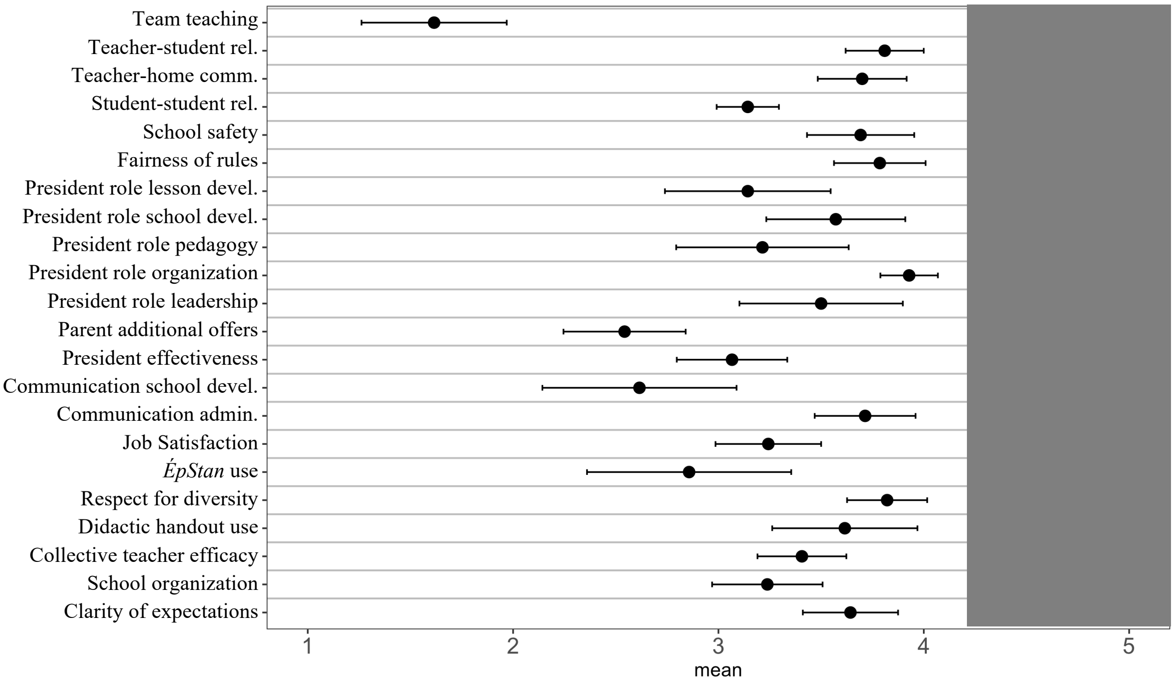
Note. Most items/scales were answered on a 4-point Likert scale from 1 to 4 and thus are truncated with a gray bar. The 5-point only applies to the items/scales that are not grayed out.

Moving on to the perspectives of the 14 school presidents, we identified several key points and summarized them in Figure 29. Firstly, team teaching does not seem very prevalent, suggesting that teachers had separate classes rather than joining classes to teach as a team. Regarding the presidents' roles, organizational and administrative communication tasks were seen as the most important, followed by school development and leadership, with pedagogy ranking as the least important task. Interestingly, teachers mostly share this perspective, and we can identify some agreement between these two groups. Additionally, school development and communication appear to be regarded as relatively unimportant or not part of the profile of a school president. When asked about their use of the ÉpStan results, the school presidents indicated that they frequently used didactic handouts in their lessons but also that there was a lack of active engagement with the results of the school monitoring. Another area where the school presidents identified areas for improvement was parental participation and additional offers from the school. For instance, only a few of the school presidents indicated having a network of experts (e.g., counselors, therapists, psychologists) at hand to refer parents. Further, none of the school presidents fully agreed that they had a concept for parent support centered on learning support. At the same time, 13 out of 14 school presidents agreed that their teachers showed high support for the needs of parents.

The overall assessment is relatively favorable for team teaching, school development and communication, parental participation, and utilizing ÉpStan results aside. For example, school presidents highly rated aspects of school climate: teacher-student relationships, teacher-home communication, school safety, fairness of rules, respect for diversity, and clarity of expectations. These points paint a positive picture of how teachers, parents, and students work together. Again, the relationship between students and their peers is a bit of a caveat.

We also find high agreement by looking at the school presidents' job satisfaction and collective teacher efficacy. As such, 13 of 14 school presidents feel free to introduce new ideas and learning techniques in their lessons. Although a sense of job satisfaction is an important protective factor for teacher dropout and other adverse outcomes (Yaniv, 1982), self-efficacy is a crucial factor in predicting student learning in Hattie's (2008) overview of meta-analyses.

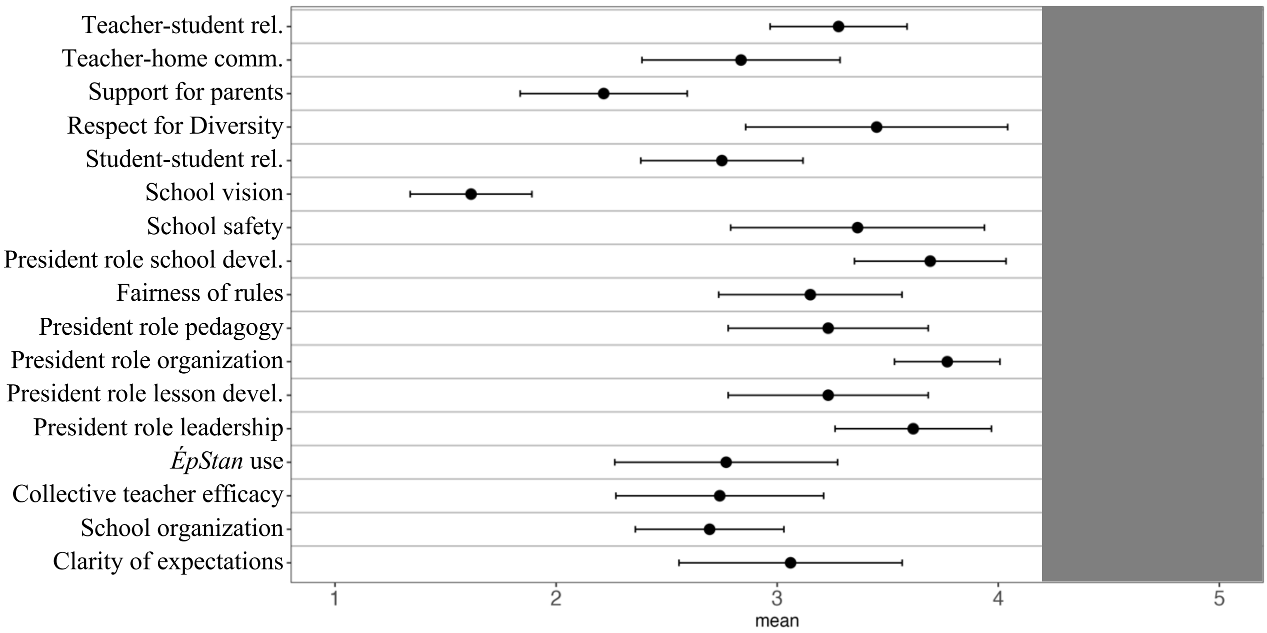
***Figure 29**. School president questionnaire results on continuous variables (n = 14)*



Note. The items/scales were answered on a 4-point Likert scale from 1 to 4 and thus are truncated with a gray bar. The 5-point depiction only helps compare the results with the other groups.

Figure 30 summarizes the regional directors' perspective. The regional directors indicated that their school does not have a vision or greater plan they work towards. Also, they perceived the support for the parents to be somewhat lower than other variables. Regional directors rated teacher-student relationships and support for diversity rather highly, and we find a similar picture for school development. Both school organization and leadership seemed to play an essential role in the schools we investigated.

Figure 30. Results of the regional director questionnaire on continuous variables (n = 13)



Note. The items/scales were answered on a 4-point Likert scale from 1 to 4 and thus are truncated with a gray bar. The 5-point depiction only helps compare the results with the other groups.

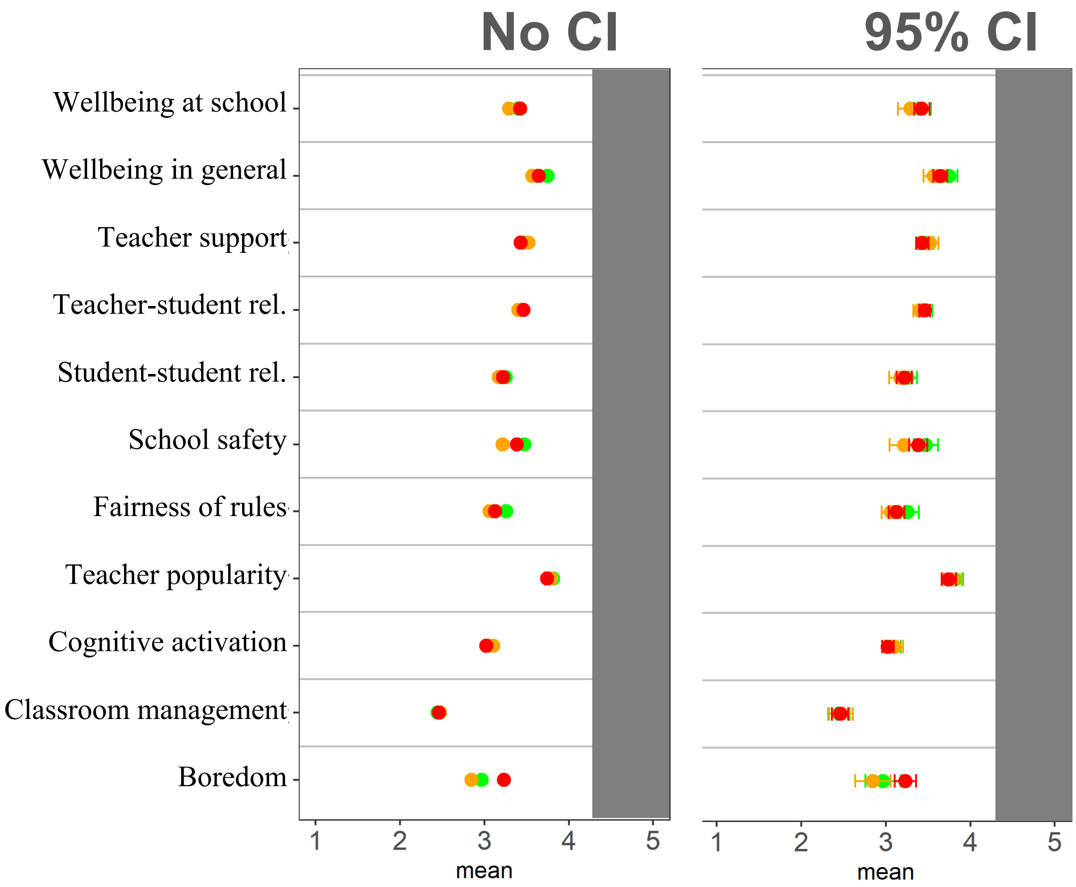
8.5.5.3 Differences Between High, Medium, and Low VA Schools on Background Variables

In the following, we examine the differences between schools with high, medium and low VA scores. We have a close-to-representative sample of schools across Luxembourg and somewhat balanced groups for the analysis, allowing for effective and meaningful comparisons. However, schools with a low VA score are over-represented. Still, when considering the descriptive statistics, we find that the schools are well-balanced across various factors. For example, the gender distribution, the proportion of Luxembourgish-speaking students, and academic performance are comparable in all groups. This alignment with expectations is precisely what the VA metric aims to capture (i.e., rather than the differences in student and school background). This result suggests that our sampling method, with the help of VA scores, has worked as expected.

To check for statistically significant differences between the three school groups, we focus on the confidence intervals. When the confidence intervals overlap, the difference between the groups is statistically nonsignificant. The two groups show statistically significant differences only when the confidence intervals are separate. In the left row of Figure 31, we have plotted only the mean values for each of the three VA groups without including confidence intervals. On the right, we have used a commonly employed confidence interval, which represents the range within which the true value is expected to fall with a 95% probability.

On initial examination, we observe a notable pattern: Boredom appears more prominent in low-VA schools. Upon closer inspection, the confidence interval of the high-VA group overlaps with both the low and the medium VA group. Consequently, only the low-VA group expressed less boredom in mathematics than the medium-VA group. Boredom in school is closely associated with lower motivation and lower academic achievement (Waack, 2018). This academic emotion could be very influential, correlated with other critical school-related factors, such as teacher support (Karagiannidis et al., 2015; King et al., 2012; Peixoto et al., 2015) and could drive the differences between the three groups.

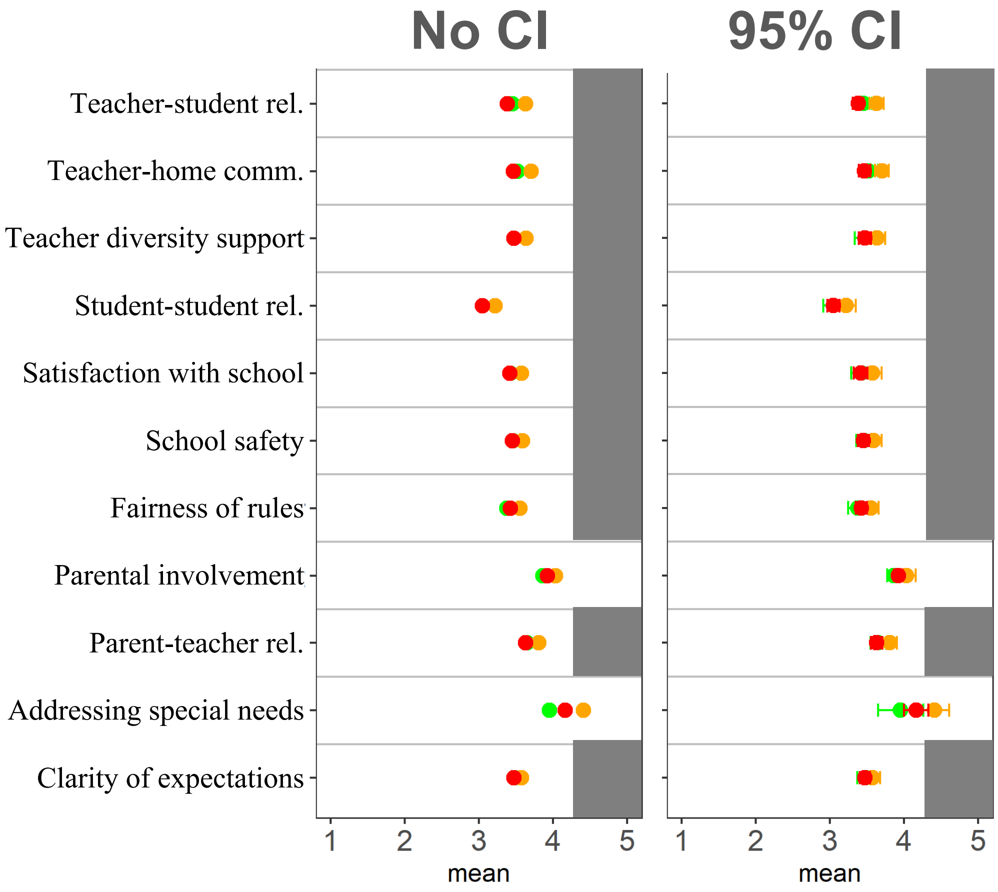
Figure 31. *Group difference in student questionnaire results on continuous variables (n_{highVA} = 85; $n_{mediumVA}$ = 100; n_{lowVA} = 211)*



Note. The items/scales were answered on a 4-point Likert scale from 1 to 4 and thus are truncated with a gray bar. The 5-point depiction only helps compare the results with the other groups. Schools with a high VA score are displayed in green, with a medium VA score in orange and a low VA score in red. CI = confidence interval.

Among the parents in Figure 32, we find an equally large agreement between the groups. In terms of content, there are only insignificant differences in how well parents perceive the school's actions to respond to the potential special educational needs of children. This response to special needs in educational settings was another promising factor that could explain the differences in the VA group, as they greatly affect students' success and well-being (Hattie, 2008). However, there are no statistically significant group differences from the parents' perspective.

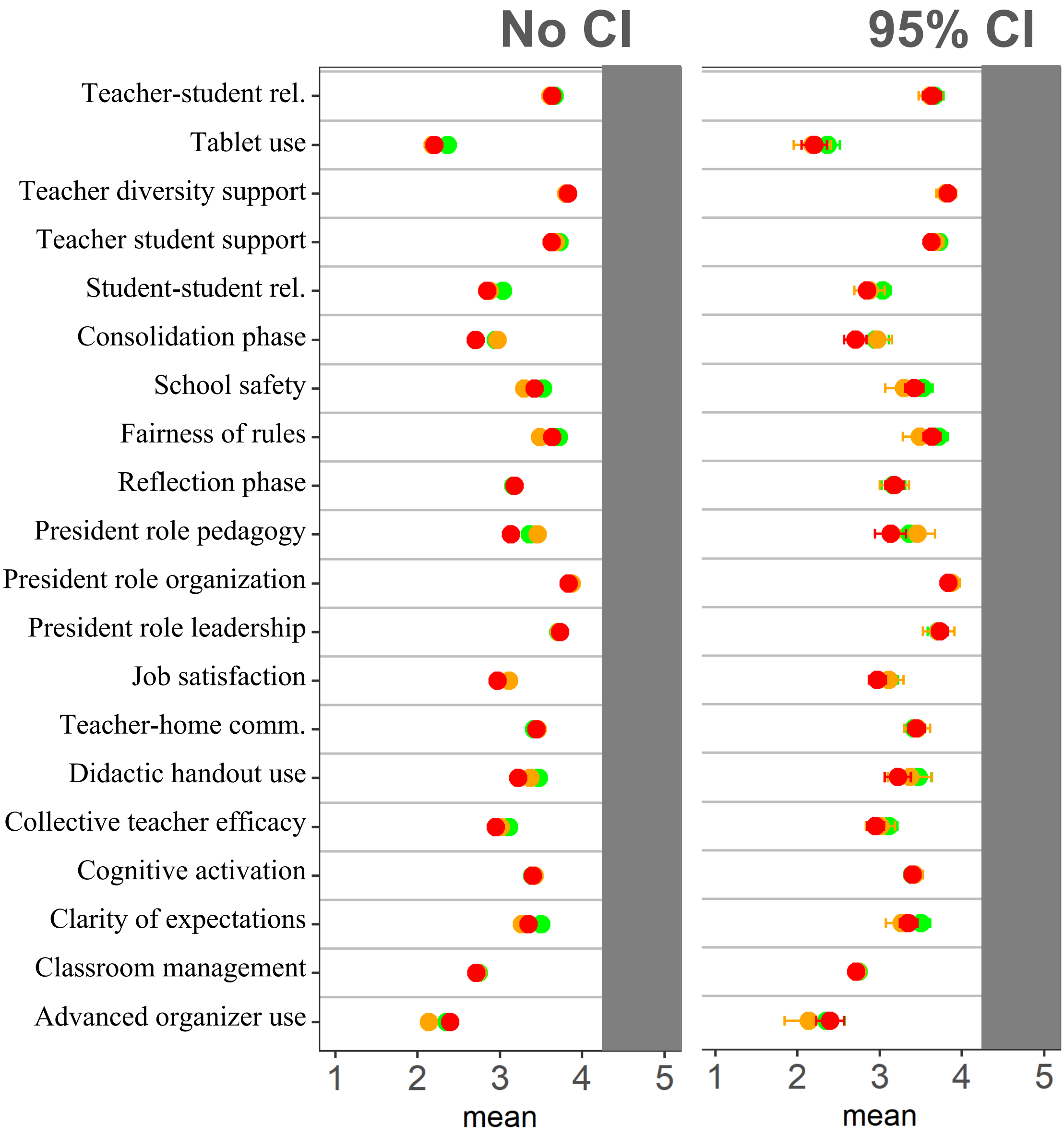
Figure 32. *Group difference in the results of the parent questionnaire on continuous variables (n_{highVA} = 85; $n_{mediumVA}$ = 100; n_{lowVA} = 211)*



Note. Most items/scales were answered on a 4-point Likert scale from 1 to 4 and thus are truncated with a gray bar. The 5-point only applies to the items/scales that are not grayed out. Schools with a high VA score are displayed in green, with a medium VA score in orange and a low VA score in red. CI = confidence interval.

In the analysis of the teachers' responses in Figure 33, we found no significant group differences between high-, medium-, and low-VA schools. However, descriptively, teachers in low VA schools perceive the importance of the regional director of pedagogy to be relatively lower than the other groups. Descriptively, more teachers in high VA schools use didactic handouts than in the other two groups. Additionally, high VA school teachers tend to provide students with clearer requirements. However, it is important to note that these differences are minor and statistically insignificant.

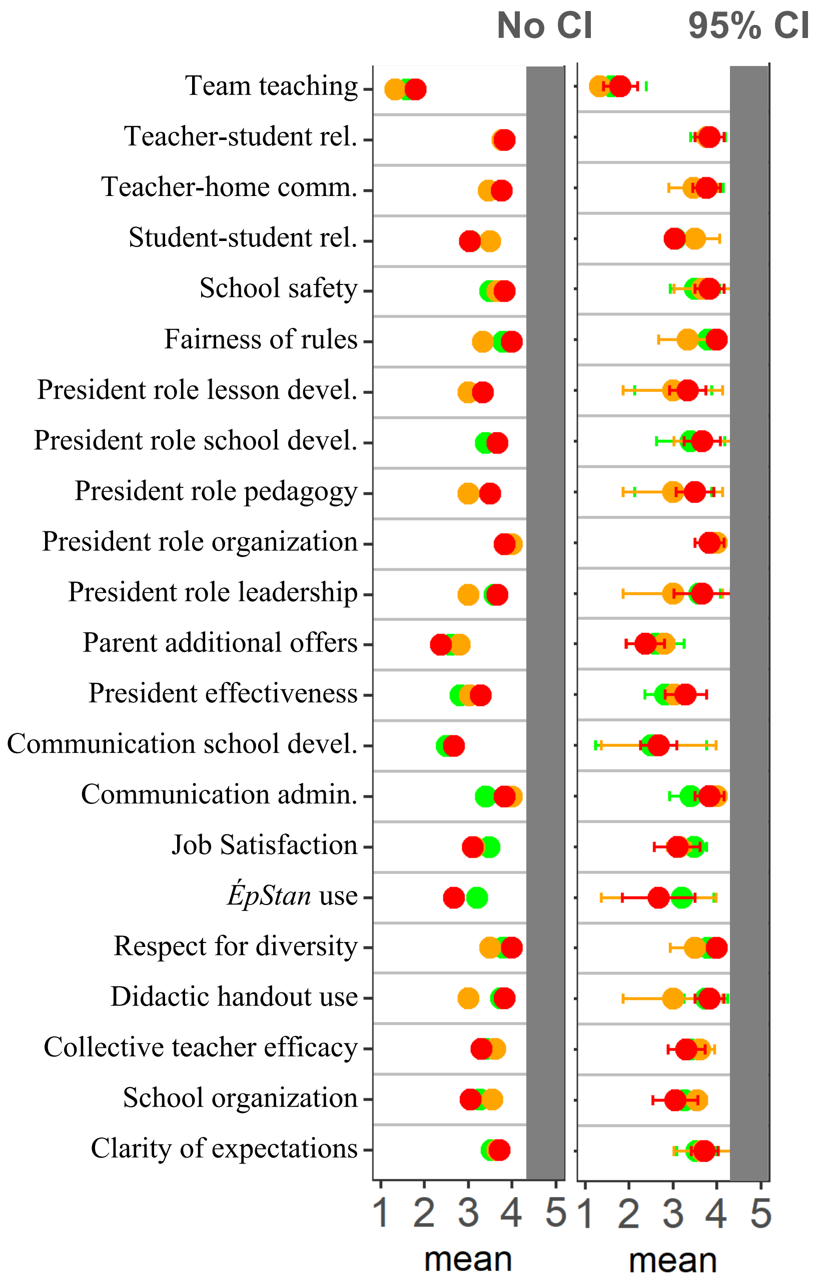
Figure 33. *Group difference in teacher questionnaire results on continuous variables (n_{highVA} = 68; $n_{mediumVA}$ = 39; n_{lowVA} = 92)*



Note. The items/scales were answered on a 4-point Likert scale from 1 to 4 and thus are truncated with a gray bar. The 5-point depiction only helps compare the results with the other groups. Schools with a high VA score are displayed in green, with a medium VA score in orange and a low VA score in red. CI = confidence interval.

Descriptively, school presidents showed slightly lower levels of job satisfaction and less engagement with the ÉpStan in low VA schools than in high VA schools, as seen in Figure 34. Again, there are no statistically significant differences between the three groups.

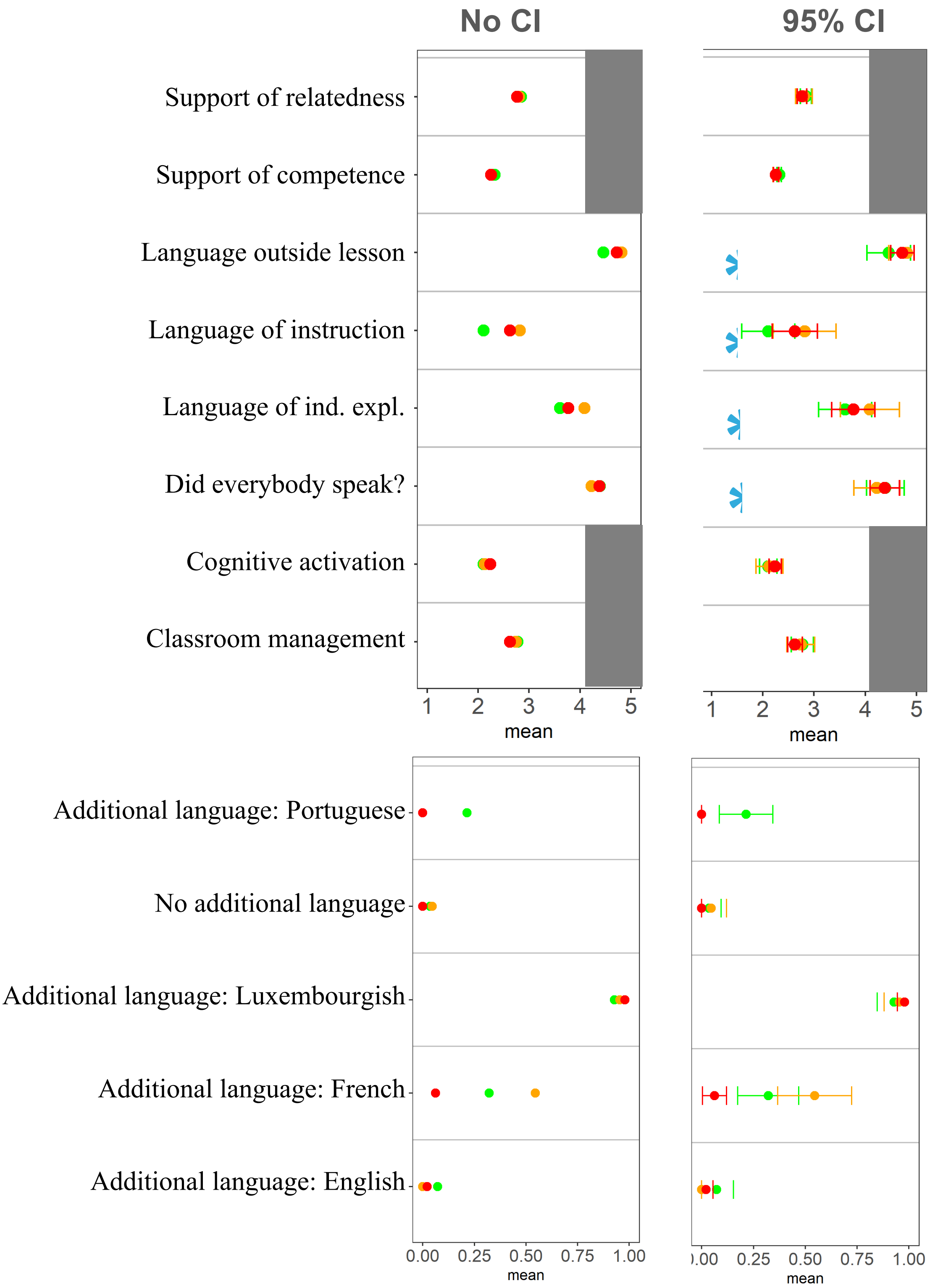
Figure 34. *Group difference in the results of the school president questionnaire on continuous variables (n_{highVA} = 5; $n_{mediumVA}$ = 3; n_{lowVA} = 6)*



Note. The items/scales were answered on a 4-point Likert scale from 1 to 4 and thus are truncated with a gray bar. The 5-point depiction only helps compare the results with the other groups. Schools with a high VA score are displayed in green, with a medium VA score in orange and a low VA score in red. CI = confidence interval.

In our classroom observations in Figure 35, we had some potentially interesting—yet again statistically insignificant—findings. High VA schools demonstrate a descriptively higher tendency to use German as the language of instruction and employ other languages to give individual explanations or during breaks. These results and their implications for *code-switching* (Rampton, 2017) and qualitative research are discussed above.

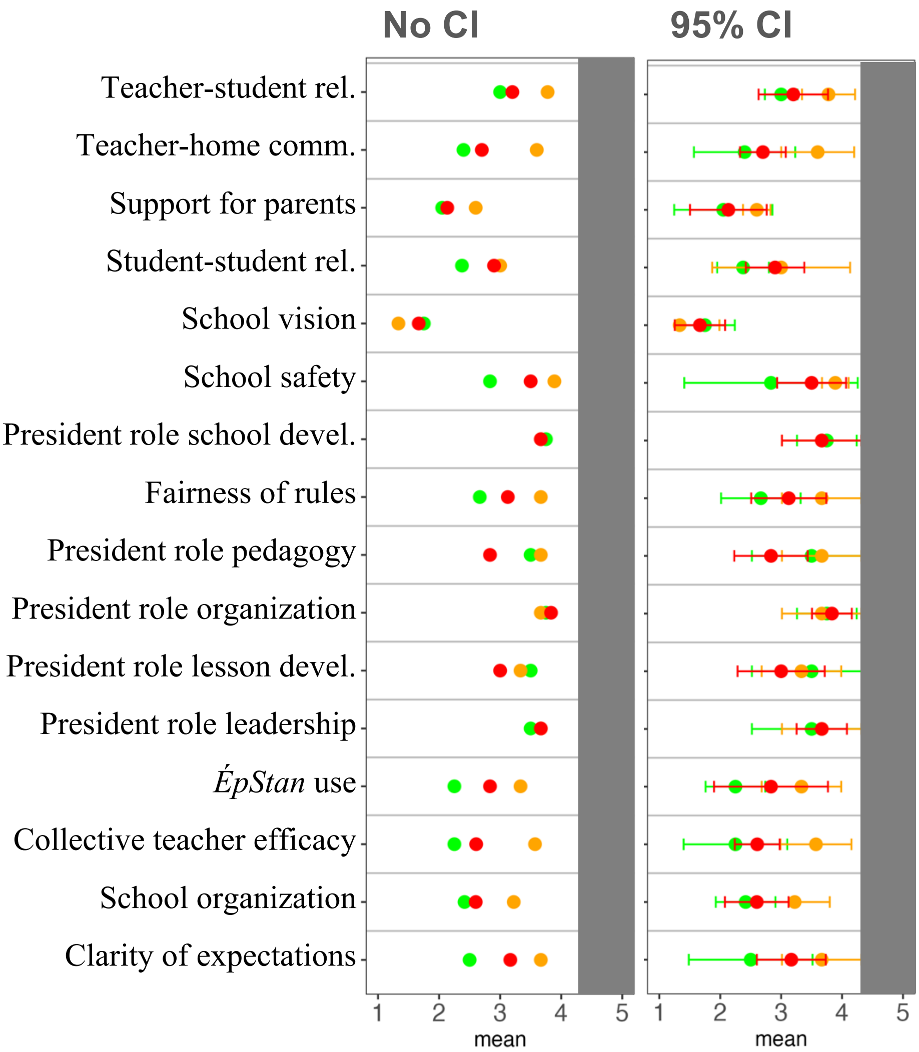
Figure 35. *Group difference in classroom observations results on continuous variables (n_{highVA} = 14; $n_{mediumVA}$ = 11; n_{lowVA} = 24 classrooms observed with two observers each)*



Note. Most items/scales were answered on a 4-point Likert scale from 1 to 4 and thus are truncated with a gray bar. The 5-point only applies to the items/scales that are not grayed out. Schools with a high VA score are displayed in green, with a medium VA score in orange and a low VA score in red. CI = confidence interval.

We find only marginal differences between the three school groups when inspecting the results from the perspective of the regional directors in Figure 36. Although an observation could be that the regional directors who responded to high VA schools indicated descriptively less favorable scores on most variables, it is likely that these slight differences stem from the small unequal sample sizes of $n_{highVA} = 4$, $n_{mediumVA} = 3$, and $n_{lowVA} = 6$, respectively.

Figure 36. *Group difference in regional directors' results on continuous variables ($n_{highVA} = 4$; $n_{mediumVA} = 3$; $n_{lowVA} = 6$ schools)*



Note. The items/scales were answered on a 4-point Likert scale from 1 to 4 and thus are truncated with a gray bar. The 5-point depiction only helps compare the results with the other groups. Schools with a high VA score are displayed in green, with a medium VA score in orange and a low VA score in red. CI = confidence interval

8.5.5.4 Evaluation of Open Text Fields

Moving beyond quantitative measures, we integrated substantial qualitative data collection into the SIVA project. On the one hand, we conducted classroom observations with two observers in 49 mathematics classrooms. On the other hand, we provided regional directors, school presidents, and teachers with open-text questions to share their thoughts about their schools. In this section, we offer a brief, additional overview of the responses provided for the open-text questions in our questionnaires and the observation sheet. By categorizing these responses, we aim to provide insight into the qualitative data.

During classroom observations, the observers noted many age-appropriate and engaging learning activities. In more than half of the 49 observed lessons, at least one observer noticed that children were encouraged to move around the classroom, participate in hands-on tasks involving measurement tools, or participate in activating activities such as dancing and clapping. Furthermore, observers noted that teachers interacted with the children in multiple languages. In addition to Luxembourgish, German, English, French, and Portuguese, the teachers welcomed responses in Italian, Bosnian, Ukrainian, Arabic, and Scandinavian languages. This highlights the active use of code-switching in Grade 2 (cycle 2.2) classrooms (Rampton, 2017).

The regional directors and the school presidents indicated that they have two main focus areas in their plans to develop their schools. Firstly, most of them saw a need to strengthen the use of and competencies in Information and Communication Technologies (ICT) among both teachers and students. Second, many of the regional directors and school presidents recognized the need to further develop the language competencies of their students. A handful of respondents also suggested combining these two issues by promoting language development in children through ICT.

In addition to the main issues of ICT and language development, several school presidents were also concerned about building a better school climate, emphasizing that social interaction is a focal point in day-to-day school life. Regional directors and presidents offered ten distinct school mottos when asked for their school's motto. They all centered around promoting respect and tolerance at the school to create a safe school climate. Only three schools included social-emotional aspects such as having fun while learning, enjoying learning, or laughing together.

In addition, we asked the teachers about the learning materials they used in their lessons and whether they wanted to share any additional information. Most teachers used school books and worksheets in their classes regarding the learning materials. Several teachers indicated that they created their learning materials, such as worksheets, rather than using materials from a book. Six teachers mentioned that they primarily used Internet learning materials and often integrated ICT into their lessons.

While we did not find systematic differences in these responses between schools with high, medium, and low VA scores, future research could apply finer-grained coding and artificial intelligence to search for differential patterns in these open responses. However, we agree on pressing issues from several perspectives: the need for ICT and a clear language development strategy in schools. These results underscore the importance of qualitative data in educational research and decision-making processes.

8.5.5.5 Additional Conclusions

The SIVA project is an ambitious, multi-method, multi-perspective, and explorative scientific quest to find the literal needle of what creates high VA in the haystack of grade 2 classrooms throughout Luxembourg. Given the ultimately statistically nonsignificant results between different VA groups, we can but conclude that, despite these complex exploratory research efforts, the metaphorical needle has not revealed itself (yet).

Still, we contributed considerable scientific evidence through the SIVA data collection. By collecting the SIVA data, we created a multi-perspective, multi-method, and (coincidentally) quasi-representative dataset on teaching quality, school climate, TSR, language use, and many other variables in Luxembourg’s primary schools. Researching educational processes in classrooms and schools, as opposed to research on educational input and/or output, can be considered one of the frontiers in the quest for high-quality education. However, such studies tend to be invasive by design and invasive study designs tend to deter practitioners and stakeholders alike. The SIVA project showed that such an invasive study can be carried out and that communication with and appreciation of all stakeholders is vital.

In addition, our research advocates for continued exploration of the identified variables in Luxembourg, the consideration of additional variables, especially at the teacher level, and interventions to bolster areas needing improvement. A longitudinal perspective may provide insight into the evolution of educational dynamics in Luxembourg and could guide future efforts to refine the primary education system.

Whereas the exploratory nature and the COVID-19 pandemic posed several challenges, the project could be realized due to the support of all parties involved. We successfully identified suitable schools, gained participating schools' consent, and effectively collected complex data during the exceptionally COVID-19 pandemic challenging teachers and students.

Looking forward, the SIVA project has several implications for future studies. Incorporating SIVA learnings on using VA scores and mixed-methods and multi-perspective data collection in diverse multilingual student populations in future research can provide a more comprehensive understanding of educational dynamics. The potential to implement an intervention study as a follow-up project or realize more research ideas with the rich data we have collected is worth considering. Finally, once the communication embargo on the collected data is lifted, the rich dataset can be shared using open science practices, and several further paper projects can be realized.

9 General Discussion

The diverse student population of Luxembourg faces challenges within a trilingual education system, creating language barriers and exacerbating socioeconomic disparities. These barriers and disparities ultimately create educational inequality in Luxembourg. Despite the apparent challenges, student achievement research surprisingly contradicts assumptions of declining performance (Weis et al., 2018). In fact, Luxembourg maintains a stable educational standing in international comparisons, suggesting effective strategies for managing linguistic diversity and mitigating educational inequalities. Such effective educational strategies will likely affect the youngest students (Heckman, 2008). Identifying and understanding these effective strategies in high-performing primary schools could provide valuable insights for other primary schools aiming to support their students from diverse socioeconomic and language backgrounds. In the first research strand, the present doctoral thesis aims to shed light on social and cognitive learning processes in the international literature. More specifically, we used meta-analytic methods to achieve the following:

1. Delve into the social learning processes and investigate the role of TSRs for diverse student outcomes in the meta-analytic literature (see Study 1).
2. Dive deeper into essential cognitive skills of EFs in mathematics to get a fuller picture of the cognitive learning process in the preschool literature (see Study 2).

In a second research strand, the thesis seeks to identify successful educational psychological strategies within the *SIVA* project, specifically in Luxembourg. Within the second research strand, we use VA methods to accomplish the following two aims:

3. Test the stability of Luxembourgish primary schools’ VA scores, given their students' socioeconomic and linguistic backgrounds (see Study 3).
4. Compare educational psychological variables between such schools with stable high, medium, or low VA scores in a multi-perspective, mixed-methods data collection (see Study 4).

We made five methodological decisions to address these four research aims, as outlined in more detail in the introductory section. Namely, we (1) focused on pre- and primary education students, (2) applied research-synthesis methods in Research Strand 1, (3) used VA scores for informative purposes (Levy, 2020) in Research Strand 2, (4) recognized multiple perspective quantitative and qualitative methods in Research Strand 2, and (5) followed current open science methods throughout all four studies. Under the umbrella of these five methodological considerations, we designed and described a narrative literature review as the theoretical considerations of this thesis. For the first research strand, we conducted a SOMA (Study 1) and a meta-analytic study (Study 2). For the second research strand, we conducted a cross-sectional VA study with Luxembourgish large-scale data (Study 3) and a cross-sectional, multi-perspective, and multi-methods study—the SIVA study (Study 4). The results of the four studies will be discussed below.

9.1 Summary of Results

The results are structured along the narrative literature review and the four studies in the present doctoral thesis. The narrative literature review and the four studies each have essential functions. With the narrative review as preparatory work, we aimed to gain an overview of potential drivers of school success in diverse student populations. With Study 1 and Study 2, we detail the possible characteristics of such driving factors. In Study 3, we map educational system of Luxembourg to find out where a search for such success factors might be most effective. Finally, equipped with insights from our prior research, we search for the well-defined drivers for students' success in carefully selected primary schools in Study 4.

9.1.1 Narrative Literature Review as Preparatory Work

This narrative literature review of this thesis motivated Study 1 and Study 2 as the Research Strand 1. Further, these methodological considerations helped identify a suitable theoretical foundation and relevant variables to assess in the SIVA project (i.e., Study 3 and Study 4). As such, we examined influential educational effectiveness models for overlapping constructs in a narrative literature review. We found the TBD*,* described by Klieme et al. (2001), to be the smallest common denominator. Thus, we decided to collect data on these three dimensions in the SIVA project: cognitive activation, student support, and classroom management. We further found school climate to be a second core model for the project and used its definition by Wang and Degol (2016). Here, we emphasized the importance of TSRs as one of the constructs included in both models of instructional quality and school climate. Alongside other relevant variables, we found several sources of evidence for an association between general cognitive functions and school success (e.g., Cortés Pascual et al., 2019; Follmer, 2018; Jacob & Parkinson, 2015).

Consequently, the first research strand further investigated TSRs and EFs. Study 1, in particular, explored the wide variety of student outcomes associated with TSR. As an student-factor of school success and a prerequisite for learning, we analyzed the EFs' connection to domain-specific skills in Study 2. Further, the SIVA project is constructed around the TBD and school climate, focusing on TSR. Further, we included language use measures, the school president's role, and assessments of other Luxembourg-specific variables. In conclusion, we used a theory-focused narrative literature review to build the theoretical foundation for the four studies in this thesis and concretely deduct relevant variables to assess in the SIVA data collection (Study 4).

9.1.2 Study 1 Emphasizes the Importance of TSR for Student Outcomes

In our narrative literature review, we identified TSR, as a facet of school climate, to be a highly relevant social learning process and aimed to broaden our understanding of this construct. Thus, Study 1 describes a systematic review of meta-analyses and several SOMAs. The study synthesized meta-analyses on the link between TSRs and student outcomes for the first research question. We found substantial correlations between TSRs and student outcomes such as academic achievement, academic emotions, appropriate student behavior, behavior problems, EFs and self-control, motivation, school belonging and engagement, and student well-being. As we cannot draw causal conclusions from this correlational design, our results do not hint at the direction of these associations.

All examined student outcomes did not statistically differ in their association with TSRs. This similarity in correlations suggests that TSRs are equally important for several student outcomes. These results were independent of whether TSRs were negative or positive. Similarly, negative and positive student outcomes showed statistically equal associations with TSRs. Thus, we can assume that both negative and positive TSRs are essential variables for student outcomes. At the same time, this finding means that intervention programs or teacher education should equally address how to build positive TSRs and avoid negative ones. Similarly, both negative and positive student outcomes should be investigated in future TSR research.

To find moderators of the TSR-outcome link, we applied two approaches: first, a review of moderator results of prior meta-analytic research, and then a second-order moderator analysis. Reviewing previous meta-analyses, we found that age, gender, and informant (student-, peer-, or teacher-assessments) were most frequently assessed as moderators. The included studies reported partially contrary moderator effects of these variables. These inconsistencies are likely because the meta-analyses differed between student outcomes and the TSR facets. Our analyses suggested student grade level and social minority status as moderators. Concretely, this would mean that meta-analyses encompassing older students and students with a social minority status show stronger links between TSRs and student outcomes. The moderation by grade level was surprising because developmental theories suggest younger students have closer relationships with their teachers, implying the greater significance of TSRs in younger years (e.g., Bowlby, 1982; Bronfenbrenner & Morris, 2007). To accompany our explanation in Study 1, novel theories should include this counter-intuitive finding and aim to fully unravel this age effect.

To critically appraise our data, we investigated the methodological quality of the included meta-analyses. A sum score of the 16 assessed items did not correlate with the effect sizes. This result indicates no general issue with methodological quality in the included meta-analyses, as their results are not confounded with study quality. Looking at the 16 assessed items individually offered more nuanced findings and interpretations. While most meta-analyses reported their inclusion and exclusion criteria clearly, and more than half included unpublished literature to mitigate publication bias, there were some shortcomings. For example, only three recent meta-analyses shared their data, and two made their analytical code publicly available in line with recommended open science practices. Further, only six of the meta-analyses reported using at least two people to screen and code their included studies. This potentially introduces considerable bias in two-thirds of the included meta-analyses. While the reporting has improved over time, probably due to the traction gained by the open science movement (Open Science Collaboration, 2015), these findings conclude that future meta-analytic research would benefit from adhering to methodological guidelines and open science practices.

Based these findings, we can conclude that TSRs are essential to student success in diverse student populations. Our findings corroborate that TSRs are equally strongly related to several fundamental student outcomes. Further, we found that TSRs are important for boys and girls of any age. Meta-analyses show high quality but should adhere to open science standards, as presented in this doctoral thesis.

9.1.3 Study 2 Provides Evidence on the EF-Math Intelligence Link in Preschool Children

EFs were identified as significant drivers of school success during our narrative literature search (e.g., Cortés Pascual et al., 2019; Follmer, 2018). Additionally, Study 1 describes EFs and self-control as one of the meta-analytically researched correlates of TSR, further highlighting their relevance in school success and student development (see Vandenbroucke et al., 2018). Thus, Study 2 encompasses several meta-analyses and a MASEM study on the link between EFs and mathematical skills in preschool children. For the first research question, we found meta-analytic evidence for an association between math intelligence (defined as math skills assessed through intelligence tests) and EFs as a composite and as separate subdimensions: inhibition, shifting, and updating. We found several study, sample, and measurement characteristics to moderate this association. This means the EF-math intelligence link differed depending on the constructs' measurements. We found no evidence of differential relations between math intelligence and inhibition, shifting, or updating with the MASEM approach.

Our meta-analyses in Study 2 offer a broad overview of the literature from 2000 to 2021 on the EFs-math link in preschool children. In addition to the positive correlations between the two constructs, we found several statistically significant moderators. Those moderators were mostly measurement characteristics explaining heterogeneity within and between studies. Therefore, it is important to consider the psychometric quality of EFs and math intelligence assessments when designing studies and interpreting their correlation. The structural equation modeling findings suggest that EFs are still best theorized as one composite in preschool children rather than three separate cognitive processes. While the SIVA project had no means of testing EFs in the data collection of Study 4, it is crucial to acknowledge their importance for school success (Study 2) and school climate and TSRs (Study 1).

9.1.4 Study 3 Shows Instability of VA Scores in Luxembourg Primary Schools

With Study 3, we dive into the second research strand of the thesis, focusing on VA methods within the SIVA project. We aimed to test the VA method to identify highly effective primary schools in order to compare them with other groups of schools in Study 4. More concretely, we tested the stability of VA scores over time and across subject domains. We focused on mathematics and language as outcome domains. Overall, we found moderate stability in primary schools' VA scores. Across the two-year timeframe, only 34-38% of the schools showed stable VA scores. These VA scores correlated moderately with the outcome domains of mathematics and language achievement at $r = .34$ and $r = .37$, respectively.

With only about one-third of Luxembourg's primary schools showing a stable VA score over time, about two-thirds of schools fluctuate substantially in their scores. This instability should warn policymakers and researchers against proposing consequential real-life implications solely based on VA scores. Hence, VA scores should be accompanied by other valid and reliable measures of school success when used for high-stakes decision-making. VA scores do not seem stable over time or across subject domains. Thus, our findings suggest that theoretical assumptions about the power and potential use cases of VA scores might be wrong (Conaway & Goldhaber, 2020; Scherrer, 2011). Still, VA scores could be used to find genuinely effective schools and teaching practices when focusing on both stable schools over time and across subject domains.

As suggested by prior research, this application might be especially indicated in diverse student populations where educational inequalities are already an essential topic in primary school (Hoffmann et al., 2018). Additionally, VA scores might be especially effective in educational systems with high rates of grade repetition (Ferrão, 2012). Diverse student populations and high retention rates are characteristics of Luxembourg's primary schools. Consequently, we can use VA scores to look past these disparities, identify the schools with stable positive VA scores, and see what they do differently than other schools with a stable medium or low VA score.

9.1.5 Study 4 Finds Home-Language Acknowledgement as One of Few School Differences Between VA Scores Groups

In Study 4, we combined the knowledge and evidence gathered from the theoretical considerations, Research Strand 1, and Study 3 on VA to address the fourth research of this thesis, which stood at the core of the SIVA project. As suggested in Study 3, we used VA scores to identify schools with stable positive VA scores. We compared them to primary schools with stable medium or low VA scores, intending to learn from their effective strategies to address educational inequalities. We used two longitudinal data sets from the ÉpStan with students in Grade 1 in 2014 or 2016 and then again in Grade 3 in 2016 and 2018, respectively. We identified 16 schools with stable high, medium, or low VA scores across time and mathematics and language as outcome domains. In these 16 schools, we collected data on the variables we had identified as relevant for educational effectiveness and overcoming educational inequalities. The data stem from six sources, including questionnaires for students, parents, teachers, school presidents, regional directors, and classroom observations.

While most variables (e.g., the TBD or TSRs) did not differ between the schools with a stable high, medium, or low VA score, two key take-home messages emerged from the SIVA data collection in Study 4. As a first take-home message, we find generally positive perceptions of educational quality in Luxembourg's primary schools from the six included perspectives. Especially teacher-student relationships are viewed positively, for example, while classroom management could be improved. During the classroom observations, we collected data on the active use of multiple languages. This *code-switching* reflects the diverse linguistic dynamics within Grade 2 classrooms (Lin, 2013; Rampton, 2017). While primarily using German as the language of instruction, switching to a student's home language for individual explanations or during breaks seemed more prevalent in schools with a high VA score. As such, we cautiously conclude that it might be advisable to use one language of instruction but also acknowledge and value the students' home language to help students follow the lesson and to create a welcoming learning environment.

As a second take-home message, the SIVA project highlights the advantages of combining qualitative and quantitative measures in educational research. This mixed-methods approach enables a thorough understanding of teaching practices and school environments. Also, incorporating several perspectives can enrich the evidence we draw from data collections. These conclusions can be drawn because it was through quantitative analyses that general positive trends were identified. Still, through qualitative observations, we detected the differences in language use between the three school-VA groups. Thus, classroom observations are a central addition to standard quantitative questionnaire approaches to assess an educational system or examine learning processes in the classroom.

In conclusion, Study 4 provides multi-perspective, mixed-methods insights into educational effectiveness in Luxembourg. Most variables, while theoretically corroborated, did not differ between schools depending on their VA scores. However, these statistically nonsignificant results hold no less value but equally contribute to a general scientific evidence base. The reason for these nonsignificant findings might be that other influential factors drive differences in VA scores. Alternatively, VA scores could be even less suited to use in school appraisal than the literature suggests. Finally, the statistically significant findings underline the benefits of supporting students in their home language in the proper context and integrating quantitative and qualitative research methods.

9.2 Contributions to Theory, Methodology, and Practice

The findings of the present thesis contribute to theory building, provide methodological advances, and can be used to develop effective practices. Thus, the following sections will resume the results and elaborate on theoretical, methodological, and practical advances through this thesis. Each section will first present the advances made in the first research strand and then in the second research strand.

9.2.1 Theoretical Contributions

The research presented in this thesis could be a stepping stone for advancing educational psychological theories in research. Through the narrative literature review, we found the TBD (Klieme et al., 2001) to be the culmination of several models of educational effectiveness. This model's centrality aligns with prior research (Praetorius & Charalambous, 2018). Thus, cognitive activation, student support, and classroom management can be seen as a common denominator of the most prevalent scientific models of educational effectiveness. Paired with the variables within the school climate model (Wang & Degol, 2016) and system-, sample-, or location-specific variables, the TBD represent some of the essential constructs of learning and instruction and, thus, educational psychology as a whole. In conclusion, the TBD (Klieme et al., 2001) and school climate models (Wang & Degol, 2016) are core theories of educational psychology. They can be applied to gain a broad overview of an educational system and its pedagogical psychological strategies to address educational inequalities.

The above-discussed theories already hint at the importance of positive relationships between teachers and their students for effective instruction and a welcoming school climate. The findings from Study 1 further corroborate the centrality of both positive and negative TSRs for various student outcomes. Several moderators were shown not to influence the TSR-outcome links, which means that the strength of the TSR-outcome link is robust against different ways of measurement and is generalizable over all students independent of age and gender. These latter findings corroborate the gender similarity hypotheses, which gained increasing support, for example, through the seminal work of Hyde (2014) and Else-Quest et al. (2010). Further, the links between positive and negative TSRs and positive and negative student outcomes were statistically equal. Thus, further theory building should include positive and negative TSRs equally (e.g., Bowlby, 1982; Bronfenbrenner & Morris, 2007).

While our overview of previous international research on TSRs can likely be generalized to Luxembourg's student population, we examined this central construct specifically in the Luxembourgish school in Study 4. The very positive TSRs found in Study 4 further suggest the importance of TSRs as an essential aspect of effective teaching, especially in Luxemburg's primary schools.

The findings from Study 2 contributed evidence to the theory of EF development. This evidence is based on the result that preschool students have not shown differential links to mathematics skills for the three subdimensions: inhibition, shifting, and updating. Although the age range we examined was restricted to preschool, prekindergarten, and kindergarten children, this result may inform the discussion on the differentiation of EFs and other cognitive skills over time (Brydges et al., 2012; Lerner & Lonigan, 2014). Additionally, we could show that EFs and mathematical skills are distinct from one another and, while correlated, cannot be subsumed under the same cognitive skill in preschool children (e.g., Ackerman et al., 2005; Allan et al., 2014; Jewsbury et al., 2016). As such, these findings are in line with the common notion that cognitive functions differentiate further and become more independent of one another when children develop (Brydges et al., 2014). These theoretical contributions help hone models of cognitive development in pre- and primary school children, informing the development of age-appropriate teaching in Luxembourg and internationally. Theoretical advances of Study 3 and Study 4 are closely linked to practical contributions and will, thus, be discussed in the Practical Advances section.

Finally, by publishing statistically nonsignificant findings throughout the four empirical works in this thesis, we counteract publication bias. This bias is prevalent in psychological science and hinders theory building by omitting large parts of robust evidence (Ferguson & Heene, 2012). This omitted evidence usually pertains to nonsignificant results or findings that would undermine currently popular theories. Thus, by reporting and publishing such null results, the present thesis aids the creation of a realistic evidence basis for educational psychological research and supports advancing an overall scientific body of knowledge.

9.2.2 Methodological Contributions

The current thesis presented several possibilities for methodological advances by combining and applying a wide array of state-of-the-art research and statistical methods. The narrative review introduces the concept of searching for overlap in theoretical models to identify the foundation of a family of theories. In the present thesis, the narrative literature review looked at models of educational effectiveness to find which variables they have in common. Consequently, these common variables served as the foundation for the presented studies and the core of the SIVA data collection. This demonstrates one effective use case for a narrative review as the basis for empirical work. Researchers searching for relevant constructs to assess in their field might, thus, resort to narrative reviews to identify the overlap of influential theories.

Study 1 introduced several small but helpful additions to the second-order meta-analytic workflow. For example, we used a graphic depiction of the sample constellation (Figure 4). While one might argue this is an unnecessary addition as the sample can easily be described in writing, the Figure 4 effectively communicates the exact sample sizes while breaking down the concept of a multi-level clustered sample. For another example, we provided a second-order meta-analytic forest plot (see Figure 8) to give an overview of the aggregated meta-analytic effect sizes. Such a graph helps the readers to easily notice the magnitude and relation of the study results concerning each outcome. While we have not tested these methods against other possible approaches, Study 1 could inspire other researchers to use such practical figures to effectively communicate their sample constellation or meta-analytic findings.

While being one of the first studies to report MASEM in the context of EFs (cf., Lin & Powell, 2022; Shen et al., 2022), the main methodological contribution of Study 2 lies elsewhere. In a novel combination of methods, we used the MASEM approach to construct a quality score for critical appraisal of the included primary studies. Concretely, we built one composite score of previously coded study quality indicators such as sample size and the reliability of the applied measures. This composite score was then weighted by the moderator effects of the coded indicators. Hence, this approach has several advantages over simply using the sum of binary indicators as a quality score (see Wedderhoff & Bosnjak, 2020). For example, with the MASEM method, we could give greater weight to those quality indicators that had a greater influence on the EF-math intelligence link. Further, we were able to combine not only binary indicators with each other but all sorts of categorical and continuous indicators into one score. Using a MASEM approach to combine primary study quality indicators sparked an ongoing research project, which will be explained in more detail in the Future Research Directions section.

Aligning with Levy (2020), our results from Study 3 suggest to use VA scores constructively rather to punish supposedly low-performing schools. Such a constructive use was to explore the VA scores' ability to make a fair comparison between schools, measuring the value they add to their students independently of their language or socioeconomic background. This technique seemed to be especially promising for Luxembourg's primary schools, with their diverse and multilingual student population. Further, this approach was tested in Study 4, which could be interpreted as a proof of concept for constructively using VA scores. It is likely to be the first study applying VA scores to identify groups of schools for a kind of extreme-group data collection. Such a focused sample of only schools with a stable VA score over time and across subjects combines the advantages of conserving resources while comparing high- and low-performing schools fairly. The resources saved by reducing the breadth of data in the data collection can consequently go into the depth of the data collection. In short, using VA scores to select schools for a resource-efficient data collection has been applied in the SIVA project with some success. In our case, this approach led to a close-to-representative sample (see Study 4). However, as we did not find group differences in most variables, future research needs to test further using VA scores to choose a sample that maximizes variation between extreme groups.

Notably, the studies included in this thesis ensure the accessibility of all these smaller and larger methodological advances by following open science principles. By publicly sharing materials for all four studies, other researchers can use our methods for their research projects. For example, the 6-perspective mixed-methods research approach with all the different questionnaires in up to four languages can be found online. Further, we have preregistered Study 1, Study 2, and Study 4. Study 1, Study 2, and Study 3 have links to their publicly accessible data (but not Study 4 due to data protection concerns and contractual obligations).

This openness in research allows for easy replication and reproducibility of results, which has already led to new collaborations and follow-up projects building upon the present thesis. In conclusion, following the open science principles whenever feasible could increase the impact of a body of research, such as the present thesis.

9.2.3 Practical Contributions

After looking into the theoretical and methodological advances the present thesis suggests, the following section will discuss potential advances for educational practice. As such, Study 1 showed similar associations of TSRs with a range of positive and negative student outcomes and reported several statistically nonsignificant moderator analyses. These findings have two main implications for teaching practice. First, TSRs are equally relevant for most student outcomes, and second, TSRs show associations for all students, largely independent of their age or gender. Thus, teachers should not only promote positive TSRs but also avoid negative ones. Whereas one must be cautious about interpreting these correlational findings with any causal direction, it seems warranted to appeal to the teachers' sensitivity to create positive and prevent negative TSRs. In a multilingual context, such Luxembourg, teachers could show an interest in a student's home language to build a positive relationship. Other meta-analytic research has already tested the effectiveness of interventions to improve TSR, giving an overview of what teachers could do to proactively improve the TSRs in their classrooms (see Kincade et al., 2020; Korpershoek et al., 2016). Such programs could already be helpful in teacher education, so preservice teachers commence their careers with their minds set on facilitating positive TSRs and counteracting negative ones.

As for EFs, in Study 2, we found that they were correlated but distinct from mathematical skills already in preschool children. If all cognitive skills in young children were indistinguishable, we could measure them all with one test, and fostering one such skill would also contribute to the increase of the other (see Jewsbury et al., 2016). However, our results from Study 2 suggest that assessing one of the two skills does not make assessing the other redundant. Also, training one of the skills is unlikely to transfer to the other (e.g., Melby-Lervåg et al., 2016; Webb et al., 2018). Therefore, we can neither reduce the testing burden for young test-takers by focusing on only one construct nor just train one of the skills and have the other one improve. This knowledge of developmental stages can help the teachers assess which tasks would be age-appropriate for their students and support them accordingly.

Study 3 showed that VA scores in primary school are relatively unstable over time and across subject domains. This finding holds, even when constructing the VA scores based on state-of-the-art VA modeling methods using prevalent covariates and applying multi-level modeling (Emslander, Levy, Scherer, et al., 2021; Levy et al., 2020, 2022). Thus, Study 3 concluded that the instability of VA scores should deter policymakers from using VA scores in high-stakes educational decision-making. This warning against the high-stakes use of VA scores aligns with prior research (Ferrão, 2012; Floden, 2012; Leckie & Goldstein, 2019; Levy, 2020). Hence, we strongly discourage policymakers from applying VA scores to decide on a teacher's tenure or a school's funding. Instead, Study 4 describes a way to use VA scores constructively.

Study 4 used VA scores to identify target schools and compare them to previously defined educational psychological variables. However, in this study, we found primarily null results, which might have two reasons. First, the constructs may not have varied between the schools with high, medium, and low VA groups enough and were not the constructs driving school differences. Thus, it could be that other variables drive the differences between the schools with high and low VA scores. These would be variables that were not identified through our theory-driven approach. Second, an alternative reason might be that VA scores hold even less meaning than expected. Conversely, when using this VA approach, we identified a nearly representative sample regarding several background variables (see Table 23). More research is needed to find conclusions from these nonsignificant findings. Still, such null results further question the use of VA scores for educational decision-making.

Even though most comparisons in study 4 did not reach statistical significance, some did variables showed differences between the school-VA groups. Following these findings, we can deduct concrete implications for the Luxembourgish school system and for teaching a diverse and multilingual student population. One of the findings was that language use and appreciation are important and should be thoughtfully tailored to the students' needs. As mentioned above, our data suggest that teachers should keep to one language of instruction in primary school for the Luxembourgish context. The question whether monolingual instruction in German is best for the diverse and multilingual Luxembourgish student population is beyond the scope of the present thesis but is already addressed in ongoing research projects (Luxembourg Centre for Educational Testing (LUCET) & Service de Coordination de la Recherche et de l'Innovation pédagogiques et technologiques (SCRIPT), 2023; SCRIPT, 2023). At the same time, teachers should acknowledge and value the student's home language whenever possible. This practice could help students follow the lesson despite insufficient German proficiency. Thereby, the teacher can create a positive school climate and TSRs through cognitive activation and student support. As for the school climate, students feel acknowledged and safe in an environment that recognizes their multilingual identity. On a personal level, this could improve TSR. Similarly, students could feel more support from their teacher, and the teacher has a broader set of options to activate the child’s prior knowledge using their home language (Creese & Blackledge, 2015; Lin, 2013; Rampton, 2017) and tailor-made solutions for multilingual children could be explored. As discussed in the Introduction, such findings could also be tested in other countries with a multi-lingual educational system, to specifically support their students facing language barriers.

In conclusion, our results suggest that teachers should value students' home language to overcome language barriers whenever possible while keeping to the language of instruction for formal explanations.

In sum, the research of the present thesis holds implications for developing scientific theory, research methodology, policy, and teaching practice. However, this work has limitations, which require further investigation, as discussed below.

9.3 Limitations and Directions for Future Research

While each of the four included studies discusses strengths, limitations, and future research ideas above, there are additional general limitations worth noting. The following three sections will focus on limitations and future research in the two research strands and open science practices. The three sections will also suggest future research to alleviate the discussed limitations. The present thesis has already sparked ongoing research, which the following sections will present together with future directions.

9.3.1 Limitations and Future Research with Research-Synthesis Methods

The narrative literature review, as well as Study 1 and Study 2, have discussed some general limitations of research synthesis methods. The following sections will raise several critical points and suggest possible solutions through future studies concerning Research Strand 1.

9.3.1.1 A Systematic Review and MASEM of the TBD

As a first limitation, the initial literature search was not conducted systematically. While narrative literature reviews have their advantages in summarizing and overviewing theory (Baumeister & Leary, 1997; Rother, 2007), it is not a systematic or reproducible approach in the sense that we had (1) no standardized search term, (2) no predefined inclusion and exclusion criteria, and (3) no sophisticated inter-rater system established. However, we used a somewhat standardized search, excluded ill-fitting literature, and applied a four-eyes principle, discussing every inclusion and exclusion within the project team. Still, future research could explore the overlap of educational effectiveness and school success models in a systematic, reproducible way. Also, a meta-analytic exploration of the TBD (Klieme et al., 2001) would hold merit. Concretely, a MASEM approach with a single item as the smallest investigation unit could substantially broaden our current understanding of the model. Additionally, comparing the perspectives of students, teachers, and observers would be a powerful tool to create robust evidence on the interplay of cognitive activation, student support, and classroom management.

9.3.1.2 A Comprehensive Theory of TSRs

The large wealth of TSR theories is somewhat scattered in the literature and would profit from comprehensive streamlining (e.g., Wentzel, 2022). Combining the theoretical approach of a narrative literature search with the breadth of a systematic review of meta-analyses, we could address another issue in Study 1. The wealth of theories concerning TSRs is comprehensive but scattered, with several well-received theories describing parallel but independent processes. One way to combine theories into one overarching TSR model would be to reuse the search from Study 1. In this set of meta-analyses, we could find the most prevalent models, identify their theoretic overlap and distinctive features, and combine the theories into one comprehensive TSR model. Such an overarching framework would help researchers gain a quick overview of the most prevalent TSR theories. Finally, as called for by recent efforts in educational psychological research (Decristan & Dumont, 2023; Greene, 2022; Praetorius & Charalambous, 2023), a joint TSR model would aid the theory-building of teaching and TSR.

9.3.1.3 Up-to-Date Guidelines for SOMAs

Having discussed shortcomings of the narrative literature review, systematic research synthesis methods also have potential disadvantages. Aside from the high level of abstraction, widely used and up-to-date guidelines for conducting a systematic review of meta-analyses in times of open science seem to be missing. Thus, Study 1 relies on guidelines that have not been updated in the past ten years since the open science movement has gained traction (i.e., Cooper & Koenka, 2012; Schmidt & Oh, 2013). Generally, poor use of open science principles and reporting standards may lead to considerable missing information and a lack of reporting in (reviews of) meta-analyses. Thus, future research must follow open science practices to share their data, code, and materials for a better scientific workflow. A growing number of researchers are applying open science practices (Federer et al., 2018; Kohrs et al., 2023; Nosek & Lindsay, 2018) and recent guidelines help meta-analysts adopt these practices, too (Moreau & Gamble, 2022). Such overviews of and guidelines for using open science in reviews of meta-analyses and SOMAs are missing. To contribute to this discussion, a systematic review of

SOMAs has already been started (i.e., Scherer & Emslander, 2023). This “meta-meta-meta-analysis” aims to take stock of good practices in reviews of meta-analyses and give recommendations on conducting such a third-order review. More specifically, we plan to review the methods used for the quantitative synthesis, explanation of heterogeneity, primary-meta-analysis overlap detection, publication bias assessment, and critical appraisal of methodological quality in the included meta-analyses (Scherer & Emslander, 2023a). In conclusion, our preregistered project aims to present a comprehensive overview of best practices in SOMA conduction and identify needs for further development.

9.3.1.4 Practical Guidance for Primary-Study Quality Assessment

One of these needs in SOMAs, but also systematic reviews in general, are guidelines to appraise primary study quality systematically. An increasing number of top-tier journals require authors to provide such a quality appraisal for transparency (Johnson, 2021). There is little guidance to assess methodological quality in a comprehensive and streamlined procedure. Thus, we have developed a tutorial to guide researchers through critical primary-study appraisal in a preregistered project. With the help of illustrative examples and open code and materials, our preprint guides meta-analysts through this process (Scherer & Emslander, 2022b, 2022a, 2023b). Evaluating primary-study quality in research syntheses indirectly supports the open science idea by making evidence accessible and reproducible, such as Study 1 and Study 2. They also provide means to assess their rigor and quality.

9.3.1.5 Developmental Trajectory of EFs

On a less methodological note, in Study 2, we did not find an age effect on the EF-math intelligence link. This finding was likely due to the small age range within our preschool sample. However, when examining the prior meta-analytic research on EFs and math skills, we found evidence for the theorized gradually increasing divergence between cognitive skills (Friedman & Miyake, 2017; Karr et al., 2018). Such a developmental trajectory of EFs becoming increasingly distinct and showing decreasing links to other cognitive skills could be tested in a systematic review of meta-analyses and SOMA. This supposed age effect could explain the divergent findings of prior meta-analyses on the EF-math intelligence link. Such a research effort could directly build on the evidence provided by the overview of previous research in Study 2 (see Table 9).

Future research should investigate the possible breaking points at which differentiation occurs with a broad age range of participants. Additionally, a SOMA on EFs could help establish a comprehensive framework of EF task types to streamline the EF assessments. As one of the challenges in Study 2 was identifying which task types pertain to which EF(s), such a SOMA could additionally address this need for a clear framework to categorize EF tasks. What would advance the field considerably is a SOMA of the link between EFs and mathematical skills focusing on age trajectories and creating an EF task framework.

9.3.1.6 Gender Differences in Meta-Analyses and the Lack Thereof

A disadvantage of research synthesis methods is their high level of abstraction. Effects that are prevalent in primary research might not be detectable in (reviews of) meta-analyses. One such effect pertains to gender differences. We found no gender differences in our research syntheses reported in Study 1 and Study 2. This might be because gender is usually only coded as a percentage of male/female students in the sample, and no separate results for different gender groups are reported. Thus, meta-analyses restrict their between-study heterogeneity, and gender moderator effects are more unlikely to be detected (Craig Aulisi et al., 2023). However, an alternative, more substantiative explanation is that we found no gender differences because there might be none. The gender similarity hypothesis would argue exactly that (Else-Quest et al., 2010; Hyde, 2014). Thus, we should check the gender similarity hypothesis in psychological meta-analytic research in general. While prior research explored this field about nine years ago (Zell et al., 2015), there has been a surge in meta-analytic research since. Thus, a SOMA on gender similarities and differences in psychological science has been preregistered to update the meta-analytic evidence on the gender similarity hypothesis (see Meyer, Emslander, et al., 2023).

9.3.2 Limitations and Future Research with VA Methods

Several general limitations of using VA scores pertain to Research Strand 2 and the SIVA study. The following section will raise issues such as the SIVA study's cross-sectional design, the unused data from the SIVA project, and how to tackle education inequalities in Luxembourg. These shortcomings, however, might be solved with future research concerning the research strand the following sections discuss.

9.3.2.1 Use of VA scores

As previously discussed, the use of VA scores for high-stakes decision-making is discouraged by the current findings from Study 3. However, we used VA scores constructively to select schools to test in Study 4. Despite creating an extreme-group design with the help of VA scores by comparing schools with stable high, medium, and low VA scores, we found little differences between the school groups. While finding null results has scientific value, too, we should discuss two possible reasons for this lack of variation between schools with high, medium, and low VA scores. The first explanation might be that other variables we have not assessed drive the difference between the schools. Possibly, the variables we assessed showed to little variance between schools either due to measurement error or because other variables are more informational in this particular school system. However, an alternative second explanation would be that VA scores are too unstable to use them as we did. This latter explanation should be examined in future research with the SIVA data and other datasets to provide evidence on the potential and limits of VA score use in research. Still, the multi-perspective, mixed-methods dataset of the SIVA project offers a unique insight into a diverse multilingual student population, and several additional research questions could be addressed with its help, as suggested below.

9.3.2.2 Further Use of Data from the SIVA Project

As Study 4 is just a first glimpse into the data and findings from the SIVA project due to space restrictions of the National Educational Report in which parts of Study 4 will be published, there still lies considerable potential in the SIVA data. As such, several future projects could be realized without combining the project data with other data sources. First, we have not used the full potential of the data on the TBD. Here, preliminary results already showed a similar pattern of cognitive activation, student support, and classroom management as in German students of a similar age (Fauth et al., 2014). A more in-depth examination could reveal further similarities and differences in an international comparison.

Additionally, the six perspectives of assessment have not yet been valorized. Here, a test of stakeholder agreement would provide novel insights into the accordance of different perspectives. This research could contribute to the discussion around the validity of teachers' self-, student, and parent ratings, for example, on instructional quality. Prior research has focused on the agreement of fewer perspectives (e.g., Wagner et al., 2016), which could be expanded with the help of the SIVA data.

The more qualitative open-text fields should be explored as an exploratory approach. First analyses have brought interesting insights from the teachers’ perspective to light (see Study 4). Potentially, with the help of large language models such as ChatGPT (OpenAI, 2023), future research could aim to find patterns, group differences, or common themes in this qualitative data. This would further strengthen the mixed-methods side of our study. These were just three ideas for valorizing further the SIVA project's multi-perspective and mixed-methods data. Further possible extensions of the data and connected research questions will be discussed in the following.

9.3.2.3 Possible Extensions to the SIVA Dataset

As typical of cross-sectional data, the SIVA study's core data collection in Study 4cannot address causal or longitudinal research questions. However, before the SIVA data collection, we acquired the consent of the parents to match the data from the SIVA project with the results from the ÉpStan from when the students were in Grade 1 the year before the data collection, and then from Grade 3 in the year after. This data matching would allow for investigating future longitudinal developments of these three years. As such, we can use the three measurement points to conduct profile analyses and examine questions of educational trajectories in our sample of students.

Similarly, the measures used in the SIVA data collection could be scaled up and applied in the ÉpStan to reach an entire cohort of Luxembourgish primary school students. While such a large extension would only be feasible for a few SIVA variables, it would provide a dataset spanning the entire population of interest. With such a dataset, all project aims and the question about the usefulness of VA scores could be finally clarified. Thus, multiple options exist to expand the SIVA dataset further and address longitudinal and potentially causal research questions.

9.3.2.4 Understanding Learning Processes from the Teachers' Perspective

The SIVA project provides a unique insight into the processes at Luxembourgish primary schools. One of its novel aspects is the inclusion of the teachers', regional directors', and school presidents' perspectives, which have not been the focus of prior research. Their perspectives are especially valuable, as these practitioners have a high level of expertise and deep insights in their specific school setting. One way to further strengthen the influence of the teachers' and school leaders' views on research could be for Luxembourg to participate in the *Teaching and Learning International Survey* (TALIS). Since 2008, this ongoing large-scale survey has focused on effective instruction and supporting institutional conditions to aid student learning (Ainley & Carstens, 2018). The TALIS allows linking their results with those from the PISA study to better assess the learning-enhancing processes. Alternatively, the ÉpStan could be extended by adding a teacher questionnaire, as many of the correlates of student achievement connect to the teacher (Hattie, 2023). This further corroborates a teacher's unique influence on their students' learning and would warrant strengthening their perspective in school research. With such a teacher questionnaire, the national educational monitoring could go beyond measuring the inputs (such as student backgrounds and school funding) and outputs of the educational system (student achievement) and investigate the learning processes and variables in the classroom. With the associations between TSRs and many of the students' emotional, psychological, cognitive, and other variables found in Study 1, we have further reason to emphasize the teachers' role in addressing educational inequalities in Luxembourg.

9.3.3 Additional Future Research Directions in Open Science

Whereas we followed open science principles whenever feasible for the four studies in the present thesis, there are some limitations to the adherence to open science practices. First, we could not share the data from the SIVA study, specifically Study 4, due to data protection concerns. Concretely, we wanted to avoid making the schools with high, medium, and low VA scores identifiable due to the political implications this might have. Additionally, we collected sensitive data on minors, which cannot be shared.

Future research could aim to develop an easy-to-follow tutorial on adhering to open science practices while protecting the participants' data. This advance could include creating a participant-consent form ensuring both the privacy of the participants and the sharing of the dataset or suggestions for a mock dataset with only non-sensitive data. Such resources must be easily accessible for all researchers without much extra time investment. With such guidelines, even more researchers might adopt the open science process.

More far-reaching actions to advance open science could be integrating it into university teaching and providing personnel support to follow open science practices. Concerning teaching, open science seems to being gradually more included in current curricula and teaching practices (Kohrs et al., 2023; Scheffel et al., 2023). Regarding structural support, funding agencies that require researchers to follow open science practices should also offer financial support for positions such as research facilitators or data stewards.

Generally, research following open science practices should produce more robust and reproducible findings while reporting more null results. In turn, meta-analyses striving to mitigate publication bias and following open science practices could also show similar benefits. Whether there is such a difference in meta-analyses in the psychological science will be tested in our project on the gender differences hypothesis (see Meyer, Emslander, et al., 2023). More precisely, we will code several indicators of adherence to open science practices to test whether meta-analyses applying these practices differ from those that do not. Assessing the impact of open science in meta-analytic research could help identify the current position of the field and the general attitude of meta-analysts toward open science (e.g., Abele-Brehm et al., 2019).

While the present thesis can only touch upon a few of the advances and challenges of open science, the field seems to be improving quite rapidly, and open science practices are on the advance in many parts of psychological research practice (e.g., Nosek & Lindsay, 2018). As such, not only are research practices evolving towards the inclusion of open science practices (e.g., Quintana, 2015), but also peer-reviewing (e.g., Morey et al., 2016), funding (e.g., Smaldino et al., 2019), and staff selection are urged to adopt the open science perspective (e.g., Cagan, 2013). These developments show two aspects of the current state of (educational) psychological science. First, the research community must still go a way to ensure robust, reproducible, and transparent research. Second, however, many active steps have been taken in the right direction.

10 General Conclusion

While educational inequalities may rise, Luxembourg's primary schools must have found ways to alleviate them. To identify such strategies, the present thesis explored factors driving school success in diverse students in two research strands. Research Strand 1 examined general social and cognitive drivers of school success with research synthesis methods in the international literature. Research Strand 2 used VA methods to identify effective educational psychological practices against educational inequalities, specifically in Luxembourg's primary schools within the SIVA project.

Our findings from Study 1 and Study 2 within the first research strand led to several conclusions. As a general social driver of school success, teachers should strengthen positive and avoid negative TSR. TSRs are an essential social learning process, as they are linked to crucial student variables such as academic achievement, emotions, behavior, motivation, well-being, or cognitive functions. They seem to work similarly for most students and across measurement methods. Interventions to improve TSRs (Korpershoek et al., 2016; Vandenbroucke et al., 2018) could help teachers build a strong relationship with their students. As for a general cognitive process driving school success, EFs are linked to math skills in preschoolers. Still, researchers must assess EFs and math skills separately, as they are clearly distinct already in preschool children (see Ackerman et al., 2005; Jewsbury et al., 2016). However, the three subdimensions of EFs (inhibition, shifting, and updating) are relatively indistinguishable at preschool age. Thus, future research must investigate the exact age trajectory when these three EFs are differentiating to further inform educational researchers, decision-makers, and practitioners.

Study 3 and Study 4 within the second research strand suggest the following conclusions. Specifically for Luxembourg, VA scores are too unstable for high-stakes educational decision-making (e.g., Levy, 2020). However, when using VA scores constructively, we found almost no differences between schools with high, medium, or low VA scores. Aligning with prior European research, the use of VA score should generally be cautioned in educational systems with diverse student populations. Still, the findings from Study 4 suggest primary school teachers adaptively acknowledge and value their students' home language to support their students by overcoming language barriers and cognitively activate, while tendentially sticking to one language of general instruction. Thus, home-language inclusion could be one of the drivers of school differences and thus help reduce educational inequities. Classroom observations and questionnaires for teachers, as the main actors in the classroom, are vital when assessing learning processes. Thus, Luxembourg's educational quality authority or national school monitoring should consider including such aspects. This more fine-grained assessment of the educational system will support identifying factors that help alleviate educational inequalities in diverse student populations.

As such, the present thesis has addressed factors driving school success in diverse students with meta-analytic and VA methods. All studies used open science practices and shared nonsignificant results with the aim to bring the field forward. Following this thesis, future research should adhere to open science practices to create a transparent, reproducible, and openly accessible evidence base.

Building on these results, future research should integrate the different frameworks of TSRs into an overarching model. Furthermore, researchers should make use of meta-analytic methods to identify which pedagogical psychological strategies work best for diverse student populations and whether such variables work equally for different genders and age groups. Further, the thesis urges decision-makers to avoid VA scores for accountability purposes and wait for further evidence on VA scores' (in)stability in different countries and age groups. Several follow-up research projects have been already commenced. Additionally, further research should aim to include here identified essential variables of educational psychology—instructional quality, TSR, EFs, school climate, and Luxembourg-specific variables—in a large-scale context. This way, the research community could find additional drivers of school success to support diverse student populations and help them thrive against the odds.

11 References

Abele-Brehm, A. E., Gollwitzer, M., Steinberg, U., & Schönbrodt, F. D. (2019). Attitudes Toward Open Science and Public Data Sharing. *Social Psychology*, *50*(4), 252–260. https://doi.org/10.1027/1864-9335/a000384

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working Memory and Intelligence: The Same or Different Constructs? *Psychological Bulletin*, *131*(1), 30–60. https://doi.org/10.1037/0033-2909.131.1.30

Agasisti, T., & Minaya, V. (2021). Precision and stability of schools' value-added estimates: Evidence for Italian primary schools. *Applied Economics Letters*, *28*(7), 541–545. https://doi.org/10.1080/13504851.2020.1763242

Ahmed, S. (2019). *Measuring Executive Function During Early Childhood: The Utility of Direct Assessments, Teacher Ratings, and Group-Based Tasks*. http://hdl.handle.net/2027.42/151502

Ahmed, S., Tang, S., Waters, N. E., & Davis-Kean, P. (2018). *Executive Function and Academic Achievement: Longitudinal Relations from Early Childhood to Adolescence*. https://doi.org/10.31234/osf.io/xd5jy

Ainley, J., & Carstens, R. (2018). *Teaching and Learning International Survey (TALIS) 2018 Conceptual Framework*. OECD. https://doi.org/10.1787/799337c2-en

Ainsworth, M. D. S., & Bell, S. M. (1970). Attachment, Exploration, and Separation: Illustrated by the Behavior of One-Year-Olds in a Strange Situation. *Child Development*, *41*(1), 49. https://doi.org/10.2307/1127388

Ainsworth, M. D. S., Blehar, M. C., Waters, E., & Wall, S. (2014). *Patterns of Attachment* (0 ed.). Psychology Press. https://doi.org/10.4324/9781315802428

Aleamoni, L. M. (1999). Student Rating Myths Versus Research Facts from 1924 to 1998. *Journal of Personnel Evaluation in Education*, *13*(2), 153–166. https://doi.org/10.1023/A:1008168421283

Ali, S., Khaleque, A., & Rohner, R. P. (2015). Influence of Perceived Teacher Acceptance and Parental Acceptance on Youth's Psychological Adjustment and School Conduct: A Cross-Cultural Meta-Analysis. *Cross-Cultural Research*, *49*(2), 204–224. https://doi.org/10.1177/1069397114552769

Allan, D. M., Allan, N. P., Lerner, M. D., Farrington, A. L., & Lonigan, C. J. (2015). Identifying unique components of preschool children's self-regulatory skills using executive function tasks and continuous performance tests. *Early Childhood Research Quarterly*, *32*(3), 40–50. https://doi.org/10.1016/j.ecresq.2015.02.001

Allan, N. P., Hume, L. E., Allan, D. M., Farrington, A. L., & Lonigan, C. J. (2014). Relations between inhibitory control and the development of academic skills in preschool and kindergarten: A meta-analysis. *Developmental Psychology*, *50*(10), 2368–2379. https://doi.org/10.1037/a0037493

Allen, K., Kern, M. L., Vella-Brodrick, D., Hattie, J., & Waters, L. (2018). What Schools Need to Know About Fostering School Belonging: A Meta-analysis. *Educational Psychology Review*, *30*(1), 1–34. https://doi.org/10.1007/s10648-016-9389-8

Alp Christ, A., Capon-Sieber, V., Grob, U., & Praetorius, A.-K. (2022). Learning processes and their mediating role between teaching quality and student achievement: A systematic review. *Studies in Educational Evaluation*, *75*, 101209. https://doi.org/10.1016/j.stueduc.2022.101209

American Psychological Association. (n.d.). *Diversity*. APA Dictionary of Psychology. Retrieved January 27, 2024, from https://dictionary.apa.org/diversity

American Psychological Association. (2017). *Ethical principles of psychologists and code of conduct*. https://www.apa.org/ethics/code/

Amrein-Beardsley, A. (2014). *Rethinking value-added models in education: Critical perspectives on tests and assessment-based accountability*. Routledge.

Amrein-Beardsley, A., & Holloway, J. (2017). Value-added models for teacher evaluation and accountability: Commonsense assumptions. *Educational Policy*, *33*, 516–542. https://doi.org/10.1177/0895904817719519

Amrein-Beardsley, A., & Holloway, J. (2019). Value-Added Models for Teacher Evaluation and Accountability: Commonsense Assumptions. *Educational Policy*, *33*(3), 516–542. https://doi.org/10.1177/0895904817719519

Amrein-Beardsley, A., Ryan Lavery, M., Holloway, J., Pivovarova, M., & L. Hahs-Vaughn, D. (2023). Evaluating the validity evidence surrounding use of value-added models to evaluate teachers: A systematic review. *Education Policy Analysis Archives*, *31*. https://doi.org/10.14507/epaa.31.8201

Ancker, J. S., & Kaufman, D. (2007). Rethinking health numeracy: A multidisciplinary literature review. *Journal of the American Medical Informatics Association*, *14*(6), 713–721. https://doi.org/10.1197/jamia.M2464

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist*, *63*(9), 839–851. https://doi.org/10.1037/0003-066X.63.9.839

Arán Filippetti, V., & Richaud, M. C. (2017). A structural equation modeling of executive functions, IQ and mathematical skills in primary students: Differential effects on number production, mental calculus and arithmetical problems. *Child Neuropsychology*, *23*(7), 864–888. https://doi.org/10.1080/09297049.2016.1199665

Armstrong-Carter, E., Finch, J. E., Siyal, S., Yousafzai, A. K., & Obradović, J. (2020). Biological sensitivity to context in Pakistani preschoolers: Hair cortisol and family wealth are interactively associated with girls' cognitive skills. *Developmental Psychobiology*, *62*(8), 1046–1061. https://doi.org/10.1002/dev.21981

Aslantas, I. (2020). The Stability Problem of Value-added Models in Teacher Effectiveness Estimations: A Systematic Review Study. *Imagining Better Education: Conference Proceedings 2019*, 1–14. http://dro.dur.ac.uk/31546/1/31546.pdf

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556–559. https://doi.org/10.1126/science.1736359

Baggetta, P., & Alexander, P. A. (2016). Conceptualization and operationalization of executive function. *Mind, Brain, and Education*, *10*(1), 10–33. https://doi.org/10.1111/mbe.12100

Baguss, J. (2020, June 20). Artificial Intelligence Solutions—One of These Things is Not Like the Other. *Evidence Partners Blog*. https://blog.evidencepartners.com/artificial-intelligence-solutions-one-of-these-things-is-not-like-the-other

Bakermans-Kranenburg, M. J., van IJzendoorn, M. H., & Juffer, F. (2003). Less is more: Meta-analyses of sensitivity and attachment interventions in early childhood. *Psychological Bulletin*, *129*(2), 195–215. https://doi.org/10.1037/0033-2909.129.2.195

Bardach, L., & Klassen, R. M. (2020). Smart teachers, successful students? A systematic review of the literature on teachers' cognitive abilities and teacher effectiveness. *Educational Research Review*, *30*, 100312. https://doi.org/10.1016/j.edurev.2020.100312

Barnes, M. A., Clemens, N. H., Fall, A.-M., Roberts, G., Klein, A., Starkey, P., McCandliss, B., Zucker, T., & Flynn, K. (2020). Cognitive predictors of difficulties in math and reading in pre-kindergarten children at high risk for learning disabilities. *Journal of Educational Psychology*, *112*(4), 685–700. https://doi.org/10.1037/edu0000404

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Baumeister, R. F., & Leary, M. R. (1997). Writing Narrative Literature Reviews. *Review of General Psychology*, *1*(3), 311–320. https://doi.org/10.1037/1089-2680.1.3.311

Bear, G. G., Gaskins, C., Blank, J., & Chen, F. F. (2011). Delaware School Climate Survey—Student: Its factor structure, concurrent validity, and reliability. *Journal of School Psychology*, *49*(2), 157–174. https://doi.org/10.1016/j.jsp.2011.01.001

Bear, G. G., Yang, C., & Pasipanodya, E. (2015). Assessing School Climate: Validation of a Brief Measure of the Perceptions of Parents. *Journal of Psychoeducational Assessment*, *33*(2), 115–129. https://doi.org/10.1177/0734282914545748

Begg, C. B., & Mazumdar, M. (1994). Operating Characteristics of a Rank Correlation Test for Publication Bias. *Biometrics*, *50*(4), 1088. https://doi.org/10.2307/2533446

Beier, M. E., & Ackerman, P. L. (2005). Working Memory and Intelligence: Different Constructs. Reply to Oberauer et al. (2005) and Kane et al. (2005). *Psychological Bulletin*, *131*(1), 72–75. https://doi.org/10.1037/0033-2909.131.1.72

Best, J. R., Miller, P. H., & Naglieri, J. A. (2011). Relations between executive function and academic achievement from ages 5 to 17 in a large, representative national sample. *Learning and Individual Differences*, *21*(4), 327–336. https://doi.org/10.1016/j.lindif.2011.01.007

Blair, C. (2002). School readiness: Integrating cognition and emotion in a neurobiological conceptualization of children’s functioning at school entry. *American Psychologist*, *57*(2), 111–127. https://doi.org/10.1037/0003-066X.57.2.111

Blair, C. (2010). Stress and the Development of Self-Regulation in Context: Stress and the Development of Self-Regulation. *Child Development Perspectives*, *4*(3), 181–188. https://doi.org/10.1111/j.1750-8606.2010.00145.x

Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development*, *78*(2), 647–663. Embase. https://doi.org/10.1111/j.1467-8624.2007.01019.x

Blakey, E., & Carroll, D. J. (2015). A Short Executive Function Training Program Improves Preschoolers’ Working Memory. *Frontiers in Psychology*, *6*(8), 1827. https://doi.org/10.3389/fpsyg.2015.01827

Blankson, A. N., Gudmundson, J. A., & Kondeh, M. (2019). Cognitive predictors of kindergarten achievement in African American children. *Journal of Educational Psychology*, *111*(7), 1273–1283. https://doi.org/10.1037/edu0000346

Bonjean, D. (2018). *The Bologna Process and the European Higher Education Area*. Education and Training - European Commission. https://education.ec.europa.eu/education-levels/higher-education/inclusive-and-connected-higher-education/bologna-process

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley & Sons. ps://doi.org/10.1002/9780470743386

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203–219. https://doi.org/10.1037/0033-295X.110.2.203

Bosman, R. J., Roorda, D. L., Van Der Veen, I., & Koomen, H. M. Y. (2018). Teacher-student relationship quality from kindergarten to sixth grade and students’ school adjustment: A person-centered approach. *Journal of School Psychology*, *68*, 177–194. https://doi.org/10.1016/j.jsp.2018.03.006

Bowlby, J. (1982). *Attachment and loss: Vol. 1. Attachment.* (2nd ed., Vol. 1). Basic Books.

Boyle, K., Felling, R., Yiu, A., Battarjee, W., Schwartz, J. M., Salorio, C., & Bembea, M. M. (2018). Neurologic Outcomes After Extracorporeal Membrane Oxygenation: A Systematic Review. *Pediatric Critical Care Medicine*, *19*(8), 760–766. https://doi.org/10.1097/PCC.0000000000001612

Brady, A. C., Griffin, M. M., Lewis, A. R., Fong, C. J., & Robinson, D. H. (2023). How Scientific Is Educational Psychology Research? The Increasing Trend of Squeezing Causality and Recommendations from Non-intervention Studies. *Educational Psychology Review*, *35*(1), 37. https://doi.org/10.1007/s10648-023-09759-9

Bramer, W. M., Giustini, D., de Jonge, G. B., Holland, L., & Bekhuis, T. (2016). De-duplication of database search results for systematic reviews in EndNote. *Journal of the Medical Library Association: JMLA*, *104*(3), 240. https://doi.org/10.3163/1536-5050.104.3.014

Braun, H. (2005). *Using Student Progress to Evaluate Teachers: A Primer on Value-Added Models.* (Policy Information Perspective). Educational Testing Service.

Broadley, M. M., White, M. J., & Andrew, B. (2017). A Systematic Review and Meta-analysis of Executive Function Performance in Type 1 Diabetes Mellitus: *Psychosomatic Medicine*, *79*(6), 684–696. https://doi.org/10.1097/PSY.0000000000000460

Brock, L. L., Rimm-Kaufman, S. E., Nathanson, L., & Grimm, K. J. (2009). The contributions of 'hot' and 'cool' executive function to children's academic achievement, learning-related behaviors, and engagement in kindergarten. *Early Childhood Research Quarterly*, *24*(3), 337–349. https://doi.org/10.1016/j.ecresq.2009.06.001

Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by nature and design*. Harvard university press.

Bronfenbrenner, U., & Ceci, S. J. (1994). Nature-nuture reconceptualized in developmental perspective: A bioecological model. *Psychological Review*, *101*(4), 568–586. https://doi.org/10.1037/0033-295X.101.4.568

Bronfenbrenner, U., & Morris, P. A. (2007). The bioecological model of human development. *Handbook of Child Psychology*, *1*. https://doi.org/10.1002/9780470147658.chpsy0114

Brophy, J. (2000). *Teaching. Educational Practices Series—1*. International Bureau of Education, P. https://eric.ed.gov/?id=ED440066

Brydges, C. R., Fox, A. M., Reid, C. L., & Anderson, M. (2014). The differentiation of executive functions in middle and late childhood: A longitudinal latent-variable analysis. *Intelligence*, *47*, 34–43. https://doi.org/10.1016/j.intell.2014.08.010

Brydges, C. R., Reid, C. L., Fox, A. M., & Anderson, M. (2012). A unitary executive function predicts intelligence in children. *Intelligence*, *40*(5), 458–469. https://doi.org/10.1016/j.intell.2012.05.006

Bull, R., Espy, K. A., & Wiebe, S. A. (2008). Short-term memory, working memory, and executive functioning in preschoolers: Longitudinal predictors of mathematical achievement at age 7 years. *Developmental Neuropsychology*, *33*(3), 205–228. https://doi.org/10.1080/87565640801982312

Bull, R., Espy, K. A., Wiebe, S. A., Sheffield, T. D., & Nelson, J. M. (2011). Using confirmatory factor analysis to understand executive control in preschool children: Sources of variation in emergent mathematic achievement. *Developmental Science*, *14*(4), 679–692. https://doi.org/10.1111/j.1467-7687.2010.01012.x

Bull, R., & Lee, K. (2014). Executive Functioning and Mathematics Achievement. *Child Development Perspectives*, *8*(1), 36–41. https://doi.org/10.1111/cdep.12059

Cagan, R. (2013). San Francisco Declaration on Research Assessment. *Disease Models & Mechanisms*, dmm.012955. https://doi.org/10.1242/dmm.012955

Cai, D., Zhang, L., Li, Y., Wei, W., & Georgiou, G. K. (2018). The role of approximate number system in different mathematics skills across grades. *Frontiers in Psychology*, *9*, 1733. https://doi.org/10.3389/fpsyg.2018.01733

Cameron, C. E., Brock, L. L., Murrah, W. M., Bell, L. H., Worzalla, S. L., Grissmer, D., & Morrison, F. J. (2012). Fine motor skills and executive function both contribute to kindergarten achievement. *Child Development*, *83*(4), 1229–1244. https://doi.org/10.1111/j.1467-8624.2012.01768.x

Cameron, J. J., & Stinson, D. A. (2019). Gender (mis)measurement: Guidelines for respecting gender diversity in psychological research. *Social and Personality Psychology Compass*, *13*(11). https://doi.org/10.1111/spc3.12506

Camerota, M., Willoughby, M. T., & Blair, C. B. (2020). Measurement models for studying child executive functioning: Questioning the status quo. *Developmental Psychology*, *56*(12), 2236–2245. https://doi.org/10.1037/dev0001127

Campos, D. G., Cheung, M. W.-L., & Scherer, R. (2023). A primer on synthesizing individual participant data obtained from complex sampling surveys: A two-stage IPD meta-analysis approach. *Psychological Methods*. https://doi.org/10.1037/met0000539

Carlson, S. M. (2005). Developmentally Sensitive Measures of Executive Function in Preschool Children. *Developmental Neuropsychology*, *28*(2), 595–616. https://doi.org/10.1207/s15326942dn2802_3

Case, R., Okamoto, Y., Griffin, S., McKeough, A., Bleiker, C., Henderson, B., Stephenson, K. M., Siegler, R. S., & Keating, D. P. (1996). The role of central conceptual structures in the development of children's thought. *Monographs of the Society for Research in Child Development*, i–295.

Cassidy, A. R., White, M. T., DeMaso, D. R., Newburger, J. W., & Bellinger, D. C. (2016). Processing speed, executive function, and academic achievement in children with dextro-transposition of the great arteries: Testing a longitudinal developmental cascade model. *Neuropsychology*, *30*(7), 874–885. https://doi.org/10.1037/neu0000289

Cavanaugh, J. E. (1997). Unifying the derivations for the Akaike and corrected Akaike information criteria. *Statistics & Probability Letters*, *33*(2), 201–208. https://doi.org/10.1016/S0167-7152(96)00128-9

Chambers, C. (2017). *The Seven Deadly Sins of Psychology: A Manifesto for Reforming the Culture of Scientific Practice*. Princeton University Press. https://doi.org/10.1515/9781400884940

Chamizo-Nieto, M. T., Arrivillaga, C., Rey, L., & Extremera, N. (2021). The Role of Emotional Intelligence, the Teacher-Student Relationship, and Flourishing on Academic Performance in Adolescents: A Moderated Mediation Study. *Frontiers in Psychology*, *12*, 695067. https://doi.org/10.3389/fpsyg.2021.695067

Cherne, J. L. (2008). *Effects of Praise on Student Behavior in the Classroom* [PhD Thesis]. University of Minnesota.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*(9), Article 9. https://doi.org/10.1257/aer.104.9.2593

Cheung, M. W.-L. (2014). Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. *Psychological Methods*, *19*(2), 211. https://doi.org/10.1037/a0032968

Cheung, M. W.-L. (2015a). *Meta-analysis: A structural equation modeling approach*. John Wiley & Sons, Inc.

Cheung, M. W.-L. (2015b). metaSEM: An R package for meta-analysis using structural equation modeling. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.01521

Cheung, M. W.-L., & Cheung, S. F. (2016). Random-effects models for meta-analytic structural equation modeling: Review, issues, and illustrations. *Research Synthesis Methods*, *7*(2), 140–155. https://doi.org/10.1002/jrsm.1166

Chu, F. W., vanMarle, K., Hoard, M. K., Nugent, L., Scofield, J. E., & Geary, D. C. (2019). Preschool deficits in cardinal knowledge and executive function contribute to longer-term mathematical learning disability. *Journal of Experimental Child Psychology*, *188*, 104668. https://doi.org/10.1016/j.jecp.2019.104668

Chu, P. S., Saucier, D. A., & Hafner, E. (2010). Meta-Analysis of the Relationships Between Social Support and Well-Being in Children and Adolescents. *Journal of Social and Clinical Psychology*, *29*(6), 624–645. https://doi.org/10.1521/jscp.2010.29.6.624

Clark, C. A. C., Pritchard, V. E., & Woodward, L. J. (2010). Preschool Executive Functioning Abilities Predict Early Mathematics Achievement. *Developmental Psychology*, *46*(5), 1176–1191. https://doi.org/10.1037/a0019672

Clark, C. A., Sheffield, T. D., Wiebe, S. A., & Espy, K. A. (2013). Longitudinal associations between executive control and developing mathematical competence in preschool boys and girls. *Child Development*, *84*(2), 662–677. https://doi.org/10.1111/j.1467-8624.2012.01854.x

Clearing House Unterricht. (2023, July 19). *Clearing House Unterricht | TUM School of Education*. https://www.clearinghouse.edu.tum.de/

Clements, D. H., Sarama, J., & Germeroth, C. (2016). Learning executive function and early mathematics: Directions of causal relations. *Early Childhood Research Quarterly*, *36*, 79–90. https://doi.org/10.1016/j.ecresq.2015.12.009

Cochran, W. G. (1950). The comparison of percentages in matched samples. *Biometrika*, *37*(3/4), 256–266. https://doi.org/10.2307/2332378

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Academic Press.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Colling, J. (2022). *Need for Cognition in Secondary Education: The Development of Need for Cognition, its Relation to Academic Achievement in Different Learning Environments and its Interaction with Theoretically Related Constructs* [Doctoral Thesis]. University of Luxembourg.

Collins, C. (2014). Houston, we have a problem: Teachers find no value in the SAS education value-added assessment system (EVAAS®). *Education Policy Analysis Archives*, *22*, 1–39. https://doi.org/10.14507/epaa.v22.1594

Conaway, C., & Goldhaber, D. (2020). Appropriate Standards of Evidence for Education Policy Decision Making. *Education Finance and Policy*, *15*(2), 383–396. https://doi.org/10.1162/edfp_a_00301

Cook, R. D. (1977). Detection of Influential Observation in Linear Regression. *Technometrics*, *19*(1), 15–18. https://doi.org/10.1080/00401706.1977.10489493

Cooper, H. (2017). *Research synthesis and meta-analysis: A step-by-step approach* (Fifth Edition). SAGE.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2019). *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation.

Cooper, H., & Koenka, A. C. (2012). The overview of reviews: Unique challenges and opportunities when research syntheses are the principal elements of new integrative scholarship. *American Psychologist*, *67*(6), 446–462. https://doi.org/10.1037/a0027119

Cornelius-White, J. (2007). Learner-Centered Teacher-Student Relationships Are Effective: A Meta-Analysis. *Review of Educational Research*, *77*(1), 113–143. https://doi.org/10.3102/003465430298563

Cortés Pascual, A., Moyano Muñoz, N., & Quílez Robres, A. (2019). The Relationship Between Executive Functions and Academic Performance in Primary Education: Review and Meta-Analysis. *Frontiers in Psychology*, *10*, 1582. https://doi.org/10.3389/fpsyg.2019.01582

Corwin Visible Learning Plus. (2023, June 30). *Visible Learning—Teacher-student relationships Details*. Visible Learning MetaX. https://www.visiblelearningmetax.com/influences/view/teacher-student_relationships

Costa, H. M., Partanen, P., & Van Herwegen, J. (2021). *The role of Working Memory, Processing Speed and Approximate Number System abilities in low maths achievement among preschoolers* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/b6p8d

Cragg, L., & Gilmore, C. (2014). Skills underlying mathematics: The role of executive function in the development of mathematics proficiency. *Trends in Neuroscience and Education*, *3*(2), 63–68. https://doi.org/10.1016/j.tine.2013.12.001

Cragg, L., Keeble, S., Richardson, S., Roome, H. E., & Gilmore, C. (2017). Direct and indirect influences of executive functions on mathematics achievement. *Cognition*, *162*, 12–26. https://doi.org/10.1016/j.cognition.2017.01.014

Craig Aulisi, L., Markell-Goldstein, H. M., Cortina, J. M., Wong, C. M., Lei, X., & Foroughi, C. K. (2023). Detecting gender as a moderator in meta-analysis: The problem of restricted between-study variance. *Psychological Methods*. https://doi.org/10.1037/met0000603

Creese, A., & Blackledge, A. (2015). Translanguaging and Identity in Educational Settings. *Annual Review of Applied Linguistics*, *35*, 20–35. https://doi.org/10.1017/S0267190514000233

Cueli, M., Areces, D., García, T., Alves, R. A., & González-Castro, P. (2020). Attention, inhibitory control and early mathematical skills in preschool students. *Psicothema*, *32*(2), 237–244. https://doi.org/10.7334/psicothema2019.225

Dalby, D. (1999). *The linguasphere register of the world's languages and speech communities /* (1–2). Linguasphere Press.

David, C. V. (2012). Working memory deficits in Math learning difficulties: A meta-analysis. *International Journal of Developmental Disabilities*, *58*(2), 67–84. https://doi.org/10.1179/2047387711Y.0000000007

Deater-Deckard, K., & Dodge, K. A. (1997). Externalizing Behavior Problems and Discipline Revisited: Nonlinear Effects and Variation by Culture, Context, and Gender. *Psychological Inquiry*, *8*(3), 161–175. https://doi.org/10.1207/s15327965pli0803_1

Decristan, J., & Dumont, H. (2023). *Adaptivity – a core principle of cognitively stimulating instruction—Call for papers—Learning and Instruction—Journal—Elsevier*. https://www.journals.elsevier.com/learning-and-instruction/call-for-papers/journals.elsevier.com/learning-and-instruction/call-for-papers/undefined

Diamond, A. (2013). Executive Functions. *Annual Review of Psychology*, *64*(1), 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Diamond, A., & Lee, K. (2011). Interventions Shown to Aid Executive Function Development in Children 4 to 12 Years Old. *Science*, *333*(6045), 959–964. https://doi.org/10.1126/science.1204529

Diener, E., Northcott, R., Zyphur, M. J., & West, S. G. (2022). Beyond Experiments. *Perspectives on Psychological Science*, *17*(4), 1101–1119. https://doi.org/10.1177/17456916211037670

Ditton, H. (2000). Qualitätskontrolle und Qualitätssicherung in Schule und Unterricht. Ein Überblick zum Stand der empirischen Forschung. *Qualität Und Qualitätssicherung Im Bildungsbereich; Schule, Sozialpädagogik, Hochschule.*, 73–92.

Ditton, H., & Arnoldt, B. (2004). Schülerbefragungen zum Fachunterricht–Feedback an Lehrkräfte. *Empirische Pädagogik*, *18*(1), 115–139.

Doebel, S. (2020). Rethinking Executive Function and Its Development. *Perspectives on Psychological Science*, *15*(4), 942–956. https://doi.org/10.1177/1745691620904771

Doidge, M. C. (2014). *Factors influencing Grade 7 teachers' implementation of outcomes-based approaches in the national curriculum when teaching 'human reproduction'—CORE* [Doctoral Thesis, University of the Witwatersrand]. https://core.ac.uk/display/39674018?utm_source=pdf&utm_medium=banner&utm_campaign=pdf-decoration-v1

Driessen, G., Agirdag, O., & Merry, M. S. (2016). The gross and net effects of primary school denomination on pupil performance. *Educational Review*, *68*(4), 466–480. https://doi.org/10.1080/00131911.2015.1135880

Duclos, M., & Murat, F. (2014). Comment évaluer la performance des lycées? Un point sur la méthodologie des IVAL (Indicateurs de valeur ajoutée des lycées). *Éducation et Formations*, *85*, 73–84.

Dumas, D., & Edelsbrunner, P. (2023). How to Make Recommendations for Educational Practice from Correlational Data Using Structural Equation Models. *Educational Psychology Review*, *35*(2), 48. https://doi.org/10.1007/s10648-023-09770-0

Dumay, X., Coe, R., & Anumendem, D. N. (2014). Stability over time of different methods of estimating school performance. *School Effectiveness and School Improvement*, *25*(1), 64–82. https://doi.org/10.1080/09243453.2012.759599

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K., & Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, *43*(6), 1428–1446. https://doi.org/10.1037/0012-1649.43.6.1428

Duncan, R., Nguyen, T., Miao, A., McClelland, M., & Bailey, D. (2016). *Executive Function and Mathematics Achievement: Are Effects Construct- and Time-General or Specific?* (ED567239). ERIC. https://eric.ed.gov/?id=ED567239

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463. https://doi.org/10.1111/j.0006-341X.2000.00455.x

Dweck, C. S. (2014). *Mindsets and math/science achievement*.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *Bmj*, *315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Egger, M., Smith, G. D., & Sterne, J. A. C. (2001). Uses and abuses of meta-analysis. *Clinical Medicine*, *1*(6), 478–484. https://doi.org/10.7861/clinmedicine.1-6-478

Eliot, L. (2013). Single-Sex Education and the Brain. *Sex Roles*, *69*(7–8), 363–381. https://doi.org/10.1007/s11199-011-0037-y

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press.

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin*, *136*(1), 103–127. https://doi.org/10.1037/a0018053

Emslander, V., Alzen, J., Ofstad, S. B., Rosa, C., & Fischbach, A. (2023). *Systematic identification of high value-added in educational contexts (SIVA)(Report No. 3)*.

Emslander, V., Holzberger, D., Fischbach, A., & Scherer, R. (2023, January 3). *Lehrer-Schüler-Beziehungen und ihre Korrelate: Ein systematisches Review von Meta-Analysen [Teacher-Student Relationships and their Correlates: A systematic Review of Meta-Analyses]*. In Emslander, V., & Holzberger, D. (Eds.) (2023, February 28 – March 2). Lehrer-Schüler-Beziehungen: Von der Generalisierbarkeit positiver

Befunde [Teacher-Student Relationships: Of the Generalizability of Positive Results]. Symposium at the 10th Conference of the Society for Empirical Educational Research (GEBF), Essen, Germany. http://hdl.handle.net/10993/54530

Emslander, V., Holzberger, D., Ofstad, S., Fischbach, A., & Scherer, R. (2023). *Teacher-Student Relationships and Student Outcomes: A Systematic Review of Meta-Analyses and Second-Order Meta-Analysis* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/qxntb

Emslander, V., Levy, J., & Fischbach, A. (2020). *Systematic Identification of High Value-Added in Educational Contexts (SIVA)(Report No. 1).*

Emslander, V., Levy, J., & Fischbach, A. (2021). *Systematic Identification of High Value-Added in Educational Contexts (SIVA)(Report No. 2).*

Emslander, V., Levy, J., & Fischbach, A. (2022). *Systematic Identification of High "Value-Added" in Educational Contexts (SIVA)*. https://doi.org/10.17605/OSF.IO/X3C48

Emslander, V., Levy, J., Scherer, R., Brunner, M., & Fischbach, A. (2021, September 15). *Stability of Value-Added Models: Comparing Classical and Machine Learning Approaches*. PAEPSY 2021 Tagung der Fachgruppe Pädagogische Psychologie, virtual conference. http://hdl.handle.net/10993/48087

Emslander, V., Levy, J., Scherer, R., & Fischbach, A. (2022). Value-added scores show limited stability over time in primary school. *PLOS ONE*, *17*(12), e0279255. https://doi.org/10.1371/journal.pone.0279255

Emslander, V., & Scherer, R. (2022). The relation between executive functions and math intelligence in preschool children: A systematic review and meta-analysis. *Psychological Bulletin*, *148*(5–6), 337–369. https://doi.org/10.1037/bul0000369

Endedijk, H. M., Breeman, L. D., van Lissa, C. J., Hendrickx, M. M. H. G., den Boer, L., & Mainhard, T. (2022). The Teacher's Invisible Hand: A Meta-Analysis of the Relevance of Teacher–Student Relationship Quality for Peer Relationships and the Contribution of Student Behavior. *Review of Educational Research*, *92*(3). https://doi.org/10.3102/00346543211051428

Espy, K. A., Kaufmann, P. M., Glisky, M. L., & McDiarmid, M. D. (2001). New Procedures to Assess Executive Functions in Preschool Children*. *The Clinical Neuropsychologist*, *15*(1), 46–58. https://doi.org/10.1076/clin.15.1.46.1908

European Commission. (2018). *OSPP-REC: Open Science Policy Platform Recommendations.* Publications Office. https://data.europa.eu/doi/10.2777/958647

Everson, K. C. (2017). Value-Added Modeling and Educational Accountability: Are We Answering the Real Questions? *Review of Educational Research*, *87*(1), 35–70. https://doi.org/10.3102/0034654316637199

Eysenbach, G. (2006). Citation Advantage of Open Access Articles. *PLoS Biology*, *4*(5), e157. https://doi.org/10.1371/journal.pbio.0040157

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction*, *29*, 1–9. https://doi.org/10.1016/j.learninstruc.2013.07.001

Federer, L. M., Belter, C. W., Joubert, D. J., Livinski, A., Lu, Y.-L., Snyders, L. N., & Thompson, H. (2018). Data sharing in PLOS ONE: An analysis of Data Availability Statements. *PLOS ONE*, *13*(5), e0194768. https://doi.org/10.1371/journal.pone.0194768

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, *7*(6), 555–561. https://doi.org/10.1177/1745691612459059

Fernández-Castilla, B., Declercq, L., Jamshidi, L., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2021). Detecting Selection Bias in Meta-Analyses with Multiple Outcomes: A Simulation Study. *The Journal of Experimental Education*, *89*(1), 125–144. https://doi.org/10.1080/00220973.2019.1582470

Fernández-Castilla, B., Jamshidi, L., Declercq, L., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2020). The application of meta-analytic (multi-level) models with multiple random effects: A systematic review. *Behavior Research Methods*, *52*(5), 2031–2052. https://doi.org/10.3758/s13428-020-01373-9

Ferrão, M. E. (2012). On the stability of value added indicators. *Quality & Quantity*, *46*(2), 627–637. https://doi.org/10.1007/s11135-010-9417-6

Ferrão, M. E. (2014). School effectiveness research findings in the Portuguese speaking countries: Brazil and Portugal. *Educational Research for Policy and Practice*, *13*(1), 3–24. https://doi.org/10.1007/s10671-013-9151-7

Ferrão, M. E. (2022). The evaluation of students' progression in lower secondary education in Brazil: Exploring the path for equity. *Studies in Educational Evaluation*, *75*, 101220. https://doi.org/10.1016/j.stueduc.2022.101220

Ferrão, M. E., & Couto, A. (2013). Indicador de valor acrescentado e tópicos sobre consistência e estabilidade: Uma aplicação ao Brasil. *Ensaio: Avaliação e Políticas Públicas Em Educação*, *21*(78), 131–164. https://doi.org/10.1590/S0104-40362013000100008

Ferrão, M. E., & Goldstein, H. (2009). Adjusting for measurement error in the value added model: Evidence from Portugal. *Quality & Quantity*, *43*(6), 951–963. https://doi.org/10.1007/s11135-008-9171-1

Ferrari, R. (2015). Writing narrative style literature reviews. *Medical Writing*, *24*(4), 230–235. https://doi.org/10.1179/2047480615Z.000000000329

Fetzer, J. S. (2011). *Luxembourg as an immigration success story: The Grand Duchy in pan-European perspective*. Lexington Books.

Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 665–694.

Fischbach, A., Colling, J., Levy, J., Pit-ten Cate, I., Rosa, C., Krämer, C., Keller, U., Gamo, S., Hornung, C., Sonnleitner, P., Ugen, S., Esch, P., & Wollschläger, R. (2021). *Befunde aus dem nationalen Bildungsmonitoring ÉpStan vor dem Hintergrund der COVID-19-Pandemie*. https://doi.org/10.48746/BB2021LU-DE-34A

Fischbach, A., Ugen, S., & Martin, R. (2014). *ÉpStan Technical Report*. University of Luxembourg. http://hdl.handle.net/10993/15802

Fisher, Z., Tipton, E., & Hou, Z. (2017). *robumeta: Robust Variance Meta-Regression (R package Version 2.0)* (2.0) [Computer software]. https://CRAN.R-project.org/package=robumeta

Fitzpatrick, C., McKinnon, R. D., Blair, C. B., & Willoughby, M. T. (2014). Do preschool executive function skills explain the school readiness gap between advantaged and disadvantaged children? *Learning and Instruction*, *30*, 25–31. PsycINFO. https://doi.org/10.1016/j.learninstruc.2013.11.003

Fleming, M. L., & Malone, M. R. (1983). The relationship of student characteristics and student performance in science as viewed by meta-analysis research. *Journal of Research in Science Teaching*, *20*(5), 481–495. https://doi.org/10.1002/tea.3660200510

Floden, R. E. (2012). Teacher value added as a measure of program quality: Interpret with caution. *Journal of Teacher Education*, *63*(5), 356–360. https://doi.org/10.1177/0022487112454175

Follmer, D. J. (2018). Executive Function and Reading Comprehension: A Meta-Analytic Review. *Educational Psychologist*, *53*(1), 42–60. https://doi.org/10.1080/00461520.2017.1309295

Friedman, N. P., & Miyake, A. (2017). Unity and diversity of executive functions: Individual differences as a window on cognitive structure. *Cortex*, *86*, 186–204. https://doi.org/10.1016/j.cortex.2016.04.023

Friedman, N. P., Miyake, A., Corley, R. P., Young, S. E., DeFries, J. C., & Hewitt, J. K. (2006). Not All Executive Functions Are Related to Intelligence. *Psychological Science*, *17*(2), 172–179. https://doi.org/10.1111/j.1467-9280.2006.01681.x

Friedman, N. P., Miyake, A., Young, S. E., DeFries, J. C., Corley, R. P., & Hewitt, J. K. (2008). Individual differences in executive functions are almost entirely genetic in origin. *Journal of Experimental Psychology: General*, *137*(2), 201–225. https://doi.org/10.1037/0096-3445.137.2.201

Friso-van den Bos, I., van der Ven, S. H. G., Kroesbergen, E. H., & van Luit, J. E. H. (2013). Working memory and mathematics in primary school children: A meta-analysis. *Educational Research Review*, *10*, 29–44. https://doi.org/10.1016/j.edurev.2013.05.003

Fuhs, M. W., Hornburg, C. B., & McNeil, N. M. (2016). Specific early number skills mediate the association between executive functioning skills and mathematics achievement. *Developmental Psychology*, *52*(8), 1217. https://doi.org/10.1037/dev0000145

Fujisawa, K. K., Todo, N., & Ando, J. (2019). Changes in Genetic and Environmental Influences on Cognitive Ability, Executive Function, and Preacademic Skills in Japanese Preschool Age Twins. *Developmental Psychology*, *55*(1), 38–52. ERIC. http://dx.doi.org/10.1037/dev0000627

Funder, D. C., & Ozer, D. J. (2019). Evaluating Effect Size in Psychological Research: Sense and Nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Gaertner, H., & Brunner, M. (2018). Once good teaching, always good teaching? The differential stability of student perceptions of teaching quality. *Educational Assessment, Evaluation and Accountability*, *30*(2), 159–182.

Ganzeboom, H. B. (2010). *A new International Socio-Economic Index (ISEI) of occupational status for the International Standard Classification of Occupation 2008 (ISCO-08) constructed with data from the ISSP 2002–2007. 1.*

Garon, N., Bryson, S. E., & Smith, I. M. (2008). Executive function in preschoolers: A review using an integrative framework. *Psychological Bulletin*, *134*(1), 31–60. https://doi.org/10.1037/0033-2909.134.1.31

Garvey, J. C., Hart, J., Metcalfe, A. S., & Fellabaum-Toston, J. (2019). Methodological Troubles with Gender and Sex in Higher Education Survey Research. *The Review of Higher Education*, *43*(1), 1–24. https://doi.org/10.1353/rhe.2019.0088

Geary, D. C., & vanMarle, K. (2016). Young Children’s Core Symbolic and Nonsymbolic Quantitative Knowledge in the Prediction of Later Mathematics Achievement. *Developmental Psychology*, *52*(12), 2130–2144. ERIC. https://doi.org/10.1037/dev0000214

Geary, David C., vanMarle, Kristy, Chu, Felicia W., Hoard, Mary K., Nugent, & Lara. (2019). Predicting Age of Becoming a Cardinal Principle Knower. *Journal of Educational Psychology*, *111*(2), 256–267.

Georgiou, S. N., & Tourva, A. (2007). Parental attributions and parental involvement. *Social Psychology of Education*, *10*(4), 473–482. https://doi.org/10.1007/s11218-007-9029-8

Gignac, G. E., & Kretzschmar, A. (2017). Evaluating dimensional distinctness with correlated-factor models: Limitations and suggestions. *Intelligence*, *62*, 138–147. https://doi.org/10.1016/j.intell.2017.04.001

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Gioia, G. A., Isquith, P. K., Guy, S. C., & Kenworthy, L. (2000). TEST REVIEW Behavior Rating Inventory of Executive Function. *Child Neuropsychology*, *6*(3), 235–238. https://doi.org/10.1076/chin.6.3.235.3152

Givens Rolland, R. (2012). Synthesizing the Evidence on Classroom Goal Structures in Middle and Secondary Schools: A Meta-Analysis and Narrative Review. *Review of Educational Research*, *82*(4), 396–435. https://doi.org/10.3102/0034654312464909

Goldacre, B., Morton, C. E., & DeVito, N. J. (2019). Why researchers should share their analytic code. *BMJ*, l6365. https://doi.org/10.1136/bmj.l6365

Goldhaber, D., & Hansen, M. (2010). Using Performance on the Job to Inform Teacher Tenure Decisions. *American Economic Review*, *100*(2), 250–255. https://doi.org/10.1257/aer.100.2.250

Goldhaber, D., & Hansen, M. (2013). Is it just a bad class? Assessing the long-term stability of estimated teacher performance. *Economica*, *80*(319), Article 319. https://doi.org/10.1111/ecca.12002

Gonzalez, O., MacKinnon, D. P., & Muniz, F. B. (2020). Extrinsic Convergent Validity Evidence to Prevent Jingle and Jangle Fallacies. *Multivariate Behavioral Research*, 1–17. https://doi.org/10.1080/00273171.2019.1707061

Gorard, S. (2006). Value-added is of little value. *Journal of Education Policy*, *21*(2), 235–243. https://doi.org/10.1080/02680930500500435

Gorard, S., Hordosy, R., & Siddiqui, N. (2013). How unstable are “school effects” assessed by a value-added technique? *International Education Studies*, *6*(1), 1–9. https://doi.org/10.5539/ies.v6n1p1

Gray, S. A., & Reeve, R. A. (2014). Preschoolers’ Dot Enumeration Abilities Are Markers of Their Arithmetic Competence. *Plos One*, *9*(4), 11. https://doi.org/10.1371/journal.pone.0094428

Graziano, P. A., Garb, L. R., Ros, R., Hart, K., & Garcia, A. (2016). Executive Functioning and School Readiness among Preschoolers with Externalizing Problems: The Moderating Role of the Student-Teacher Relationship. *Early Education and Development*, *27*(5), 573–589. ERIC. https://doi.org/10.1080/10409289.2016.1102019

Greene, J. A. (2022). What Can Educational Psychology Learn From, and Contribute to, Theory Development Scholarship? *Educational Psychology Review*, *34*(4), 3011–3035. https://doi.org/10.1007/s10648-022-09682-5

Grosz, M. P. (2023). Should researchers make causal inferences and recommendations for practice on the basis of nonexperimental studies? *Educational Psychology Review*, *35*(2), 57. https://doi.org/10.1007/s10648-023-09777-7

Grosz, M. P., Ayaita, A., Arslan, R. C., Buecker, S., Ebert, T., Müller, S., Rieger, S., Zapko-Willmes, A., & Rohrer, J. M. (2023). *Natural Experiments: Missed Opportunities for Causal Inference in Psychology* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/dah3q

Grund, S., Robitzsch, Alexander, & Luedtke, Oliver. (2019). *mitml: Tools for Multiple Imputation in Multilevel Modeling* (R package version 0.3-7). https://CRAN.R-project.org/package=mitml

Hadjar, A., & Backes, S. (2021). *Bildungsungleichheiten am Übergang in die Sekundarschule in Luxemburg*. https://doi.org/10.48746/BB2021LU-DE-21A

Hadjar, A., Fischbach, A., & Backes, S. (2018). Bildungsungleichheiten im luxemburgischen Sekundarschulsystem aus zeitlicher Perspektive [Educational inequalities in the Luxembourgish secondary school system from a temporal perspective]. In *A. Hadjar, A. Fischbach & S. Backes (Eds.), Nationaler Bildungsbericht Luxemburg 2018* (pp. 59–83). SCRIPT. https://www.bildungsbericht.lu/media/ul_natbericht_de_web_1.5.pdf

Haertel, G. D., Walberg, H. J., & Weinstein, T. (1983). Psychological models of educational performance: A theoretical synthesis of constructs. *Review of Educational Research*, *53*(1), 75–91. https://doi.org/10.3102/00346543053001075

Hamre, B. K., & Pianta, R. C. (2001). Early Teacher-Child Relationships and the Trajectory of Children's School Outcomes through Eighth Grade. *Child Development*, *72*(2), 625–638. https://doi.org/10.1111/1467-8624.00301

Hanushek, E. A. (1971). Teacher characteristics and gains in student achievement: Estimation using micro data. *The American Economic Review*, *61*(2), 280–288.

Hanushek, E. A. (2019). Testing, Accountability, and the American Economy. *The ANNALS of the American Academy of Political and Social Science*, *683*(1), Article 1. https://doi.org/10.1177/0002716219841299

Hardy, C. L., Bukowski, W. M., & Sippola, L. K. (2002). Stability and Change in Peer Relationships During the Transition to Middle-Level School. *The Journal of Early Adolescence*, *22*(2), 117–142. https://doi.org/10.1177/0272431602022002001

Harrer, M., Cuijpers, P., Furukawa, T., & Ebert, D. (2019). *Dmetar: Companion R Package For The Guide "Doing Meta-Analysis in R". R package version 0.0. 9000.* https://dmetar.protectlab.org

Hartmann-Hirsch, C., & Amétépé, F. S. (2023). The Luxembourg Context: Push-Pull Factors. In C. Hartmann-Hirsch & F. S. Amétépé, *Between Europeanisation and Renationalisation of the Free Movement of Persons* (pp. 35–52). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-40814-5_3

Harvey, H. A., & Miller, G. E. (2017). Executive Function Skills, Early Mathematics, and Vocabulary in Head Start Preschool Children. *Early Education and Development*, *28*(3), 290–307. ERIC. https://doi.org/10.1080/10409289.2016.1218728

Hattie, J. (2008). *Visible Learning: A synthesis of over 800 meta-analyses relating to achievement* (0 ed.). Routledge. https://doi.org/10.4324/9780203887332

Hattie, J. (2012). *Visible learning for teachers: Maximizing impact on learning*. Routledge.

Hattie, J. (2015). The applicability of Visible Learning to higher education. *Scholarship of Teaching and Learning in Psychology*, *1*(1), 79–91. https://doi.org/10.1037/stl0000021

Hattie, J. (2023). *Visible Learning: The Sequel: A Synthesis of Over 2,100 Meta-Analyses Relating to Achievement*. Taylor & Francis.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, *13*(3).

Heckman, J. J. (2008). *Schools, Skills, and Synapses*. *Economic Inquiry*, *46*(3), 289–324. https://doi.org/10.1111/j.1465-7295.2008.00163.x

Helmke, A. (2010). *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts; Franz Emanuel Weinert gewidmet;[Orientierungsband]*. Klett.

Helmke, A., Rindermann, H., & Schrader, F.-W. (2008). Wirkfaktoren akademischer Leistungen in Schule und Hochschule [Determinants of academic achievement in school and university]. In M. Schneider & M. Hasselhorn (Eds.), *Handbuch der pädagogischen Psychologie* (Vol. 10, pp. 145–155). Hogrefe. https://www.google.com/books?hl=de&lr=&id=ujfdQyfaiEgC&oi=fnd&pg=PA38&d

q=lernmotivation+und+interesse+schiefele&ots=1n5fVZRM6L&sig=1Hzwzy8e5W9i3ijzhRKUQYBqQcU

Herbert, B., Fischer, J., & Klieme, E. (2022). How valid are student perceptions of teaching quality across education systems? *Learning and Instruction*, *82*, 101652. https://doi.org/10.1016/j.learninstruc.2022.101652

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *Bmj*, *327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

Hjetland, H. N., Brinchmann, E. I., Scherer, R., Hulme, C., & Melby-Lervåg, M. (2020). Preschool pathways to reading comprehension: A systematic meta-analytic review. *Educational Research Review*, *30*, 100323. https://doi.org/10.1016/j.edurev.2020.100323

Hoffmann, D., Hornung, C., Gamo, S., Esch, P., Keller, U., & Fischbach, A. (2018). *Schulische Kompetenzen von Erstklässlern und ihre Entwicklung nach zwei Jahren.* (Nationaler Bildungsbericht, pp. 84–96). Luxembourg Centre for Educational Testing, Universität Luxemburg; Service de la Coordination de la Recherche et de l'Innovation pédagogiques et technologiques. https://orbilu.uni.lu/bitstream/10993/38687/1/ul_natbericht_de_web_1.6.pdf

Howard, S. J., & Vasseleu, E. (2020). Self-Regulation and Executive Function Longitudinally Predict Advanced Learning in Preschool. *Frontiers in Psychology*, *11*, 49. https://doi.org/10.3389/fpsyg.2020.00049

Hu, Bi Ying, Johnson, Gregory Kirk, Teo, Timothy, Wu, & Zhongling. (2020). Relationship between Screen Time and Chinese Children's Cognitive and Social Development. *Journal of Research in Childhood Education*, *34*(2), 183–207.

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

Hu, T., Zhang, D., & Wang, J. (2015). A meta-analysis of the trait resilience and mental health. *Personality and Individual Differences*, *76*, 18–27. https://doi.org/10.1016/j.paid.2014.11.039

Hughes, J. N., & Cao, Q. (2018). Trajectories of teacher-student warmth and conflict at the transition to middle school: Effects on academic engagement and achievement. *Journal of School Psychology*, *67*, 148–162. https://doi.org/10.1016/j.jsp.2017.10.003

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage.

Hutchison, J., & Phillips, D. (2018, March 27). *Executive Functions: Supporting Foundational Skills for Early Math Learning*. DREME. Development and Research in Early Math Education. https://dreme.stanford.edu/news/executive-functions-supporting-foundational-skills-early-math-learning

Hyde, J. S. (2014). Gender Similarities and Differences. *Annual Review of Psychology*, *65*(1), 373–398. https://doi.org/10.1146/annurev-psych-010213-115057

Hyde, J. S., Bigler, R. S., Joel, D., Tate, C. C., & Van Anders, S. M. (2019). The future of sex and gender in psychology: Five challenges to the gender binary. *American Psychologist*, *74*(2), 171–193. https://doi.org/10.1037/amp0000307

Imbo, I., & LeFevre, J.-A. (2009). Cultural differences in complex addition: Efficient Chinese versus adaptive Belgians and Canadians. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1465–1476. https://doi.org/10.1037/a0017022

Ioannidis, J. P. A. (2016). The Mass Production of Redundant, Misleading, and Conflicted Systematic Reviews and Meta-analyses. *The Milbank Quarterly*, *94*(3), 485–514. https://doi.org/10.1111/1468-0009.12210

Jacob, R., & Parkinson, J. (2015). The Potential for School-Based Interventions That Target Executive Function to Improve Academic Achievement: A Review. *Review of Educational Research*, *85*(4), 512–552. https://doi.org/10.3102/0034654314561338

Jak, S., & Cheung, M. W.-L. (2020). Meta-analytic structural equation modeling with moderating effects on SEM parameters. *Psychological Methods*, *25*(4), 430–455. https://doi.org/10.1037/met0000245

Jansen, T., Meyer, J., Wigfield, A., & Möller, J. (2022). Which student and instructional variables are most strongly related to academic motivation in K-12 education? A systematic review of meta-analyses. *Psychological Bulletin*, *148*(1–2), 1–26. https://doi.org/10.1037/bul0000354

Jewsbury, P. A., Bowden, S. C., & Strauss, M. E. (2016). Integrating the switching, inhibition, and updating model of executive function with the Cattell—Horn—Carroll model. *Journal of Experimental Psychology: General*, *145*(2), 220–245. https://doi.org/10.1037/xge0000119

Johnson, B. T. (2021). Toward a more transparent, rigorous, and generative psychology. *Psychological Bulletin*, *147*(1), 1–15. https://doi.org/10.1037/bul0000317

Johnson, S. M. (2015). Will VAMS reinforce the walls of the egg-crate school? *Educational Researcher*, *44*(2), 117–126. https://doi.org/10.3102/0013189X15573351

Johnston, O., Wildy, H., & Shand, J. (2022). 'That teacher really likes me'—Student-teacher interactions that initiate teacher expectation effects by developing caring relationships. *Learning and Instruction*, *80*, 101580. https://doi.org/10.1016/j.learninstruc.2022.101580

Kane, M. J., Hambrick, D. Z., & Conway, A. R. A. (2005). Working Memory Capacity and Fluid Intelligence Are Strongly Related Constructs: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, *131*(1), 66–71. https://doi.org/10.1037/0033-2909.131.1.66

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment. Research Paper. MET Project*. Bill & Melinda Gates Foundation. https://files.eric.ed.gov/fulltext/ED540959.pdf

Karagiannidis, Y., Barkoukis, V., Gourgoulis, V., Kosta, G., & Antoniou, P. (2015). The role of motivation and metacognition on the development of cognitive and affective responses in physical education les-sons: A self-determination approach. *Motricidade*, 135-150 Pages. https://doi.org/10.6063/MOTRICIDADE.3661

Karr, J. E., Areshenkoff, C. N., Rast, P., Hofer, S. M., Iverson, G. L., & Garcia-Barrera, M. A. (2018). The unity and diversity of executive functions: A systematic review and re-analysis of latent variable studies. *Psychological Bulletin*, *144*(11), 1147–1185. https://doi.org/10.1037/bul0000160

Kassai, R., Futo, J., Demetrovics, Z., & Takacs, Z. K. (2019). A meta-analysis of the experimental evidence on the near- and far-transfer effects among children's executive function skills. *Psychological Bulletin*, *145*(2), 165–188. https://doi.org/10.1037/bul0000180

Keane, K., & Evans, R. R. (2022). The Potential for TEACHER-STUDENT Relationships and the Whole School, Whole Community, Whole Child Model to Mitigate Adverse Childhood Experiences. *Journal of School Health*, *92*(5), 504–513. https://doi.org/10.1111/josh.13154

Kepes, S., McDaniel, M. A., Brannick, M. T., & Banks, G. C. (2013). Meta-analytic Reviews in the Organizational Sciences: Two Meta-analytic Schools on the Way to MARS (the Meta-analytic Reporting Standards). *Journal of Business and Psychology*, *28*(2), 123–143. https://doi.org/10.1007/s10869-013-9300-2

Kesselring, T., & Müller, U. (2011). The concept of egocentrism in the context of Piaget's theory. *New Ideas in Psychology*, *29*(3), 327–345. https://doi.org/10.1016/j.newideapsych.2010.03.008

Kidd, J. K., Lyu, H., Peterson, M. S., Hassan, M. Z., Gallington, D. A., Strauss, L. I., Patterson, A. B., & Pasnak, R. (2019). Patterns, Mathematics, Early Literacy, and Executive Functions. *Creative Education*, *10*(13), 3444–3468. https://doi.org/10.4236/ce.2019.1013266

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., Kennett, C., Slowik, A., Sonnleitner, C., Hess-Holden, C., Errington, T. M., Fiedler, S., & Nosek, B. A. (2016). Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. *PLOS Biology*, *14*(5), e1002456. https://doi.org/10.1371/journal.pbio.1002456

Kincade, L., Cook, C., & Goerdt, A. (2020). Meta-Analysis and Common Practice Elements of Universal Approaches to Improving Student-Teacher Relationships. *Review of Educational Research*, *90*(5), 710–748. https://doi.org/10.3102/0034654320946836

King, R. B., McInerney, D. M., & Watkins, D. A. (2012). How you think about your intelligence determines how you feel in school: The role of theories of intelligence on academic emotions. *Learning and Individual Differences*, *22*(6), 814–819. https://doi.org/10.1016/j.lindif.2012.04.005

Kingdon, D., Cardoso, C., & McGrath, J. J. (2016). Research Review: Executive function deficits in fetal alcohol spectrum disorders and attention-deficit/hyperactivity disorder - a meta-analysis. *Journal of Child Psychology and Psychiatry*, *57*(2), 116–131. https://doi.org/10.1111/jcpp.12451

Kirsch, C., & Seele, C. (2022). Early Language Education in Luxembourg. In M. Schwartz (Ed.), *Handbook of Early Language Education* (pp. 789–812). Springer International Publishing. https://doi.org/10.1007/978-3-030-91662-6_28

Klein, C., & Peltier, F. (2022). *La démographie luxembourgeoise en chiffres*. https://statistiques.public.lu/dam-assets/catalogue-publications/en-chiffres/2022/demographie-en-chiffre-22.pdf

Klieme, E., Lipowsky, F., & Rakoczy, K. (2006). Qualitätsdimensionen und Wirksamkeit von Mathematikunterricht. Theoretische Grundlagen und ausgewählte Ergebnisse des Projekts "Pythagoras". In M. Prenzel & L. Allolio-Näcke (Eds.), *Untersuchungen zur Bildungsqualität von Schule. Abschlussbericht des DFG-Schwerpunktprogramms* (pp. 127–146). Waxmann.

Klieme, E., Schümer, G., & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I:" Aufgabenkultur" und Unterrichtsgestaltung. In *TIMSS-Impulse für Schule und Unterricht* (pp. 43–57). Bundesministerium für Bildung und Forschung.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, *47*, 180–195. https://doi.org/10.1016/j.econedurev.2015.01.006

Kohrs, F. E., Auer, S., Bannach-Brown, A., Fiedler, S., Haven, T. L., Heise, V., Holman, C., Azevedo, F., Bernard, R., Bleier, A., Bössel, N., Cahill, B. P., Castro, L. J., Ehrenhofer, A., Eichel, K., Frank, M., Frick, C., Friese, M., Gärtner, A., … Weissgerber, T. L. (2023). Eleven strategies for making reproducible research and open science training the norm at research institutions. *eLife*, *12*, e89736. https://doi.org/10.7554/eLife.89736

Kolkman, M. E., Hoijtink, H. J. A., Kroesbergen, E. H., & Leseman, P. P. M. (2013). The role of executive functions in numerical magnitude skills. *Learning and Individual Differences*, *24*, 145–151. https://doi.org/10.1016/j.lindif.2013.01.004

Korpershoek, H., Harms, T., de Boer, H., van Kuijk, M., & Doolaard, S. (2016). A Meta-Analysis of the Effects of Classroom Management Strategies and Classroom Management Programs on Students' Academic, Behavioral, Emotional, and Motivational Outcomes. *Review of Educational Research*, *86*(3), 643–680. https://doi.org/10.3102/0034654315626799

Korucu, Irem, Litkowski, Ellen, Schmitt, & Sara A. (2020). Examining Associations between the Home Literacy Environment, Executive Function, and School Readiness. *Early Education and Development*, *31*(3), 455–473.

Kossmeier, M., Tran, U. S., & Voracek, M. (2020a). Power-Enhanced Funnel Plots for Meta-Analysis: The Sunset Funnel Plot. *Zeitschrift Für Psychologie*, *228*(1), 43–49. https://doi.org/10.1027/2151-2604/a000392

Kossmeier, M., Tran, U. S., & Voracek, M. (2020b). Visualizing meta-analytic data with R package metaviz. *R Package Version*, *3*, 1.

Kovacs, K., & Conway, A. R. A. (2016). Process Overlap Theory: A Unified Account of the General Factor of Intelligence. *Psychological Inquiry*, *27*(3), 151–177. https://doi.org/10.1080/1047840X.2016.1153946

Krause, A. (2020). *Peer Aggression and Teacher-Student Relationship Quality: A Meta-Analytic Investigation*. https://doi.org/10.20381/RUOR-25135

Krause, A., & Smith, J. D. (2022). Peer Aggression and Conflictual Teacher-Student Relationships: A Meta-Analysis. *School Mental Health*, *14*(2), 306–327. https://doi.org/10.1007/s12310-021-09483-1

Kroesbergen, E., Van Luit, J., Van Lieshout, E., Van Loosbroek, E., & Van de Rijt, B. (2009). Individual differences in early numeracy: The role of executive functions and subitizing. *Journal of Psychoeducational Assessment*, *27*(3), 226–236. https://doi.org/10.1177/0734282908330586

Kurtz, M. D. (2018). Value-Added and Student Growth Percentile Models: What Drives Differences in Estimated Classroom Effects? *Statistics and Public Policy*, *5*(1), 1–8. https://doi.org/10.1080/2330443X.2018.1438938

Lai, C. L. E., Lau, Z., Lui, S. S. Y., Lok, E., Tam, V., Chan, Q., Cheng, K. M., Lam, S. M., & Cheung, E. F. C. (2017). Meta-analysis of neuropsychological measures of executive functioning in children and adolescents with high-functioning autism spectrum disorder: Meta-analysis of executive functioning in ASD. *Autism Research*, *10*(5), 911–939. https://doi.org/10.1002/aur.1723

Lakens, D., Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC Psychology*, *4*(1), 24. https://doi.org/10.1186/s40359-016-0126-3

Law Insider. (n.d.). *Educational inequality Definition*. Law Insider. Retrieved January 28, 2024, from https://www.lawinsider.com/dictionary/educational-inequality

Leckie, G., & Goldstein, H. (2019). The importance of adjusting for pupil background in school value-added models: A study of Progress 8 and school accountability in England. *British Educational Research Journal*, *45*(3), 518–537. https://doi.org/10.1002/berj.3511

Leff, S. S., Thomas, D. E., Shapiro, E. S., Paskewich, B., Wilson, K., Necowitz-Hoffman, B., & Jawad, A. F. (2011). Developing and Validating a New Classroom Climate Observation Assessment Tool. *Journal of School Violence*, *10*(2), 165–184. https://doi.org/10.1080/15388220.2010.539167

Lehto, J. (1996). Are Executive Function Tests Dependent on Working Memory Capacity? *The Quarterly Journal of Experimental Psychology Section A*, *49*(1), 29–50. https://doi.org/10.1080/713755616

Lehtonen, M., Soveri, A., Laine, A., Järvenpää, J., de Bruin, A., & Antfolk, J. (2018). Is bilingualism associated with enhanced executive functioning in adults? A meta-analytic review. *Psychological Bulletin*, *144*(4), 394–425. https://doi.org/10.1037/bul0000142

Lei, H., Cui, Y., & Chiu, M. M. (2016). Affective Teacher—Student Relationships and Students' Externalizing Behavior Problems: A Meta-Analysis. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01311

Lei, H., Cui, Y., & Chiu, M. M. (2018). The Relationship between Teacher Support and Students' Academic Emotions: A Meta-Analysis. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.02288

Lenz, T., Backes, S., Ugen, S., & Fischbach, A. (2021). *Bereit für die Zukunft? Der dritte Bildungsbericht für Luxemburg*. https://doi.org/10.48746/BB2021LU-DE-1

Lerner, M. D., & Lonigan, C. J. (2014). Executive Function Among Preschool Children: Unitary Versus Distinct Abilities. *Journal of Psychopathology and Behavioral Assessment*, *36*(4), 626–639. https://doi.org/10.1007/s10862-014-9424-3

Levy, J. (2020). *Tertium non datur: Various aspects of value-added (VA) models used as measures of educational effectiveness* [Unpublished doctoral dissertation]. University of Luxembourg.

Levy, J., Brunner, M., Keller, U., & Fischbach, A. (2019). Methodological issues in value-added modeling: An international review from 26 countries. *Educational Assessment, Evaluation and Accountability*, *31*(3), 257–287. https://doi.org/10.1007/s11092-019-09303-w

Levy, J., Brunner, M., Keller, U., & Fischbach, A. (2022). How sensitive are the evaluations of a school's effectiveness to the selection of covariates in the applied value-added model? *Educational Assessment, Evaluation and Accountability*. https://doi.org/10.1007/s11092-022-09386-y

Levy, J., Mussack, D., Brunner, M., Keller, U., Cardoso-Leite, P., & Fischbach, A. (2020). Contrasting Classical and Machine Learning Approaches in the Estimation of Value-Added Scores in Large-Scale Educational Data. *Frontiers in Psychology*, *11*, 2190. https://doi.org/10.3389/fpsyg.2020.02190

Levy, K. N., Ellison, W. D., Scott, L. N., & Bernecker, S. L. (2011). Attachment style. *Journal of Clinical Psychology*, *67*(2), 193–203. https://doi.org/10.1002/jclp.20756

Li, J.-B., Bi, S.-S., Willems, Y. E., & Finkenauer, C. (2021). The Association Between School Discipline and Self-Control From Preschoolers to High School Students: A Three-Level Meta-Analysis. *Review of Educational Research*, *91*(1), 73–111. https://doi.org/10.3102/0034654320979160

Li, X., Bergin, C., & Olsen, A. A. (2022). Positive teacher-student relationships may lead to better teaching. *Learning and Instruction*, *80*, 101581. https://doi.org/10.1016/j.learninstruc.2022.101581

Lin, A. (2013). Classroom code-switching: Three decades of research. *Applied Linguistics Review*, *4*(1), 195–218. https://doi.org/10.1515/applirev-2013-0009

Lin, X., & Powell, S. R. (2022). The Roles of Initial Mathematics, Reading, and Cognitive Skills in Subsequent Mathematics Performance: A Meta-Analytic Structural Equation Modeling Approach. *Review of Educational Research*, *92*(2), 288–325. https://doi.org/10.3102/00346543211054576

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. SAGE publications, Inc.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V.-N., & Martinez, J. F. (2007). The Sensitivity of Value-Added Teacher Effect Estimates to Different Mathematics Achievement Measures. *Journal of Educational Measurement*, *44*(1), 47–67. https://doi.org/10.1111/j.1745-3984.2007.00026.x

Loeb, S., & Candelaria, C. A. (2012). How Stable Are Value-Added Estimates across Years, Subjects and Student Groups? What We Know Series: Value-Added Methods and

Applications. Knowledge Brief 3. *Carnegie Foundation for the Advancement of Teaching*.

LUCET. (2019). *EpStan*. Assessed Competences. Cycle 3.1. Retrieved August 21, 2019, from https://epstan.lu/en/assessed-competences-31/

LUCET. (2021). *Épreuves Standardisées (ÉpStan)*. https://epstan.lu

LUCET. (2023). *Épreuves Standardisées (ÉpStan)*. https://epstan.lu

Luxembourg Centre for Educational Testing (LUCET) & Service de Coordination de la Recherche et de l'Innovation pédagogiques et technologiques (SCRIPT). (2023). *European Public School Report 2023: Preliminary results on student population, educational trajectories, mathematics achievement, and stakeholder perceptions*. Luxembourg Centre for Educational Testing (LUCET) & Service de Coordination de la Recherche et de l'Innovation pédagogiques et technologiques (SCRIPT). https://doi.org/10.48746/EPS2023

Mann, T. D., Hund, A. M., Hesson-McInnis, M. S., & Roman, Z. J. (2017). Pathways to school readiness: Executive functioning predicts academic and social–emotional aspects of school readiness. *Mind, Brain, and Education*, *11*(1), 21–31. https://doi.org/10.1111/mbe.12134

Marsh, H. W., Dowson, M., Pietsch, J., & Walker, R. (2004). Why Multicollinearity Matters: A Reexamination of Relations Between Self-Efficacy, Self-Concept, and Achievement. *Journal of Educational Psychology*, *96*(3), 518–522. https://doi.org/10.1037/0022-0663.96.3.518

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). *Goodness of fit in structural equation models.* Lawrence Erlbaum Associates Publishers.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler's (1999) Findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 320–341. https://doi.org/10.1207/s15328007sem1103_2

Marsh, H. W., Pekrun, R., Parker, P. D., Murayama, K., Guo, J., Dicke, T., & Arens, A. K. (2019). The murky distinction between self-concept and self-efficacy: Beware of lurking jingle-jangle fallacies. *Journal of Educational Psychology*, *111*(2), 331–353. https://doi.org/10.1037/edu0000281

Martens, A., Johns, M., Greenberg, J., & Schimel, J. (2006). Combating stereotype threat: The effect of self-affirmation on women's intellectual performance. *Journal of Experimental Social Psychology*, *42*(2), 236–243. https://doi.org/10.1016/j.jesp.2005.04.010

Martin, R., Ugen, S., & Fischbach, A. (2015). *Épreuves Standardisées: Bildungsmonitoring für Luxemburg*. University of Luxembourg, Luxembourg Centre for Educational Testing (LUCET). https://men.public.lu/dam-assets/catalogue-publications/statistiques-etudes/statistiques-globales/epreuves-standardisees.pdf

Marzano, R. J., & Toth, M. D. (2013). *Teacher evaluation that makes a difference: A new model for teacher growth and student achievement*. ASCD.

Masum, H., Rao, A., Good, B. M., Todd, M. H., Edwards, A. M., Chan, L., Bunin, B. A., Su, A. I., Thomas, Z., & Bourne, P. E. (2013). Ten Simple Rules for Cultivating Open Science and Collaborative R&D. *PLoS Computational Biology*, *9*(9), e1003244. https://doi.org/10.1371/journal.pcbi.1003244

Matt, G. E., & Cook, T. D. (2019). Threats to the validity of generalized inferences from research syntheses. In *H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), The handbook of research synthesis and meta-analysis* (pp. 489–516). Russell Sage Foundation New York, NY. https://doi.org/10.7758/9781610448864.25

Matthews, J. S., Ponitz, C. C., & Morrison, F. J. (2009). Early gender differences in self-regulation and academic achievement. *Journal of Educational Psychology*, *101*(3), 689. https://doi.org/10.1037/a0014240

McCaffrey, D. F., & Lockwood, J. R. (2011). Missing data in value-added modeling of teacher effects. *The Annals of Applied Statistics*, *5*(2A). https://doi.org/10.1214/10-AOAS405

McClelland, M. M., Cameron, C. E., Duncan, R., Bowles, R. P., Acock, A. C., Miao, A., & Pratt, M. E. (2014). Predictors of early growth in academic achievement: The head-toes-knees-shoulders task. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00599

McCoy, Dana C., Gonzalez, Kathryn, Jones, & Stephanie. (2019). Preschool Self-Regulation and Preacademic Skills as Mediators of the Long-Term Impacts of an Early Intervention. *Child Development*, *90*(5), 1544–1558.

McGuire, J. K., Beek, T. F., Catalpa, J. M., & Steensma, T. D. (2019). The Genderqueer Identity (GQI) Scale: Measurement and validation of four distinct subscales with trans and LGBQ clinical and community samples in two countries. *International Journal of Transgenderism*, *20*(2–3), 289–304. https://doi.org/10.1080/15532739.2018.1460735

McKinnon, R. D., Blair, C., & The Family Life Project Investigators. (2018). Does early executive function predict teacher–child relationships from kindergarten to second grade? *Developmental Psychology*, *54*(11), 2053. https://doi.org/10.1037/dev0000584

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for Publication Bias in Meta-Analysis: An Evaluation of Selection Methods and Some Cautionary Notes. *Perspectives on Psychological Science*, *11*(5), 730–749. https://doi.org/10.1177/1745691616662243

Melby-Lervåg, M., & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. *Developmental Psychology*, *49*(2), 270. https://doi.org/10.1037/a0028228

Melby-Lervåg, M., Redick, T. S., & Hulme, C. (2016). Working Memory Training Does Not Improve Performance on Measures of Intelligence or Other Measures of “Far Transfer”: Evidence From a Meta-Analytic Review. *Perspectives on Psychological Science*, *11*(4), 512–534. https://doi.org/10.1177/1745691616635612

MEN-DEP. (1994). *Trois indicateurs de performances des lycées (Les Dossiers d’éducation et Formations)* (Ministère de l’Éducation Nationale et Direction de l’Évaluation et de La Prospective.).

MENJE. (2023). *The Luxembourgish Education System*. https://men.public.lu/dam-assets/catalogue-publications/divers/informations-generales/the-luxembourg-education-system-en.pdf

MENJE, & SCRIPT. (2022). *Education system in Luxembourg Key figures*. https://www.script.lu/sites/default/files/publications/2022-05/2022_chiffres_cles_EN.pdf

Meyer, H. (2004). *Was ist guter Unterricht*. Cornelsen Scriptor.

Meyer, J., Bardach, L., Emslander, V., & Jansen, T. (2023). *Preregistration Evaluating gender similarities and differences using meta-synthesis: An updated summary of meta-analytic findings 10 years after Zell et al. (2015, American Psychologist)*. https://doi.org/10.17605/OSF.IO/T2P67

Milatz, A., Lüftenegger, M., & Schober, B. (2015). Teachers’ Relationship Closeness with Students as a Resource for Teacher Wellbeing: A Response Surface Analytical Approach. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01949

Mills, B., Dyer, N., Pacheco, D., Brinkley, D., Owen, M. T., & Caughy, M. O. (2019). Developmental Transactions Between Self-Regulation and Academic Achievement

Among Low-Income African American and Latino Children. *Child Development*, *90*(5), 1614–1631. https://doi.org/10.1111/cdev.13091

Minaya, V., & Agasisti, T. (2019). Evaluating the Stability of School Performance Estimates over Time. *Fiscal Studies*, *40*(3), 401–425. https://doi.org/10.1111/1475-5890.12201

Ministry of National Education, Children and Youth. (2011). *Elementary School. Cycles 1 -4. The Levels of Competence*. http://www.men.public.lu/catalogue-publications/fondamental/apprentissages/documents-obligatoires/niveaux-competences/en.pdf

Ministry of National Education, Children and Youth. (2018). *L'enseignement luxembourgeois en chiffres: Année scolaire 2016-2017*. MENJE. http://www.men.public.lu/catalogue-publications/themes-transversaux/statistiques-analyses/enseignement-chiffres/2016-2017-depliant/en.pdf

Mischel, W., Ayduk, O., Berman, M. G., Casey, B. J., Gotlib, I. H., Jonides, J., Kross, E., Teslovich, T., Wilson, N. L., Zayas, V., & Shoda, Y. (2011). 'Willpower' over the life span: Decomposing self-regulation. *Social Cognitive and Affective Neuroscience*, *6*(2), 252–256. https://doi.org/10.1093/scan/nsq081

Miyake, A., & Friedman, N. P. (2012). The Nature and Organization of Individual Differences in Executive Functions: Four General Conclusions. *Current Directions in Psychological Science*, *21*(1), 8–14. https://doi.org/10.1177/0963721411429458

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The Unity and Diversity of Executive Functions and Their Contributions to Complex "Frontal Lobe" Tasks: A Latent Variable Analysis. *Cognitive Psychology*, *41*(1), 49–100. https://doi.org/10.1006/cogp.1999.0734

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. (2009). PRISMA Group: Methods of systematic reviews and meta-analysis: Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *J Clin Epidemiol*, *62*, 1006–1012. https://doi.org/10.7326/0003-4819-151-4-200908180-00135

Moksness, L., & Olsen, S. O. (2017). Understanding researchers' intention to publish in open access journals. *Journal of Documentation*, *73*(6), 1149–1166. https://doi.org/10.1108/JD-02-2017-0019

Moore, T. C., Maggin, D. M., Thompson, K. M., Gordon, J. R., Daniels, S., & Lang, L. E. (2019). Evidence Review for Teacher Praise to Improve Students' Classroom Behavior. *Journal of Positive Behavior Interventions*, *21*(1), 3–18. https://doi.org/10.1177/1098300718766657

Moreau, D., & Gamble, B. (2022). Conducting a meta-analysis in the age of open science: Tools, tips, and practical recommendations. *Psychological Methods*, *27*(3), 426–432. https://doi.org/10.1037/met0000351

Morey, R. D., Chambers, C. D., Etchells, P. J., Harris, C. R., Hoekstra, R., Lakens, D., Lewandowsky, S., Morey, C. C., Newman, D. P., Schönbrodt, F. D., Vanpaemel, W., Wagenmakers, E.-J., & Zwaan, R. A. (2016). The Peer Reviewers' Openness Initiative: Incentivizing open research practices through peer review. *Royal Society Open Science*, *3*(1), 150547. https://doi.org/10.1098/rsos.150547

Mou, Y., Berteletti, I., & Hyde, D. C. (2018). What counts in preschool number knowledge? A Bayes factor analytic approach toward theoretical model development. *Journal of Experimental Child Psychology*, *166*, 116–133. https://doi.org/10.1016/j.jecp.2017.07.016

Munafò, M. R., Chambers, C., Collins, A., Fortunato, L., & Macleod, M. (2022). The reproducibility debate is an opportunity, not a crisis. *BMC Research Notes*, *15*(1), 43. https://doi.org/10.1186/s13104-022-05942-3

Murdock, T. B., & Bolch, M. B. (2005). Risk and protective factors for poor school adjustment in lesbian, gay, and bisexual (LGB) high school youth: Variable and person-centered analyses. *Psychology in the Schools*, *42*(2), 159–172. https://doi.org/10.1002/pits.20054

Nagy, G., & Neumann, M. (2010). Psychometrische Aspekte des Tests zu den voruniversitären Mathematikleistungen in TOSCA-2002 und TOSCA-2006: Unterrichtsvalidität, Rasch-Homogenität und Messäquivalenz. In U. Trautwein, M. Neumann, G. Nagy, O. Lüdtke, & K. Maaz (Eds.), *Schulleistungen von Abiturienten. Die neu geordnete gymnasiale Oberstufe auf dem Prüfstand.* (pp. 281–306). VS Verlag für Sozialwissenschaften.

Navarro, J. I., Aguilar, M., Alcalde, C., Ruiz, G., Marchena, E., & Menacho, I. (2011). Inhibitory processes, working memory, phonological awareness, naming speed, and early arithmetic achievement. *The Spanish Journal of Psychology*, *14*(2), 580–588. http://dx.doi.org/10.5209/rev_SJOP.2011.v14.n2.6

Nelson, J. M., James, T. D., Chevalier, N., Clark, C. A. C., & Espy, K. A. (2016). Structure, measurement, and development of preschool executive function. In J. A. Griffin, P. McCardle, & L. S. Freund (Eds.), *Executive function in preschool-age children: Integrating measurement, neurodevelopment, and translational research.* (pp. 65–89). American Psychological Association. https://doi.org/10.1037/14797-004

Newton, X., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Education Policy Analysis Archives*, *18*(23), 1–24. https://doi.org/10.14507/epaa.v18n23.2010

Nguyen, P., McKenzie, J. E., Hamilton, D. G., Moher, D., Tugwell, P., Fidler, F. M., Haddaway, N. R., Higgins, J. P. T., Kanukula, R., Karunananthan, S., Maxwell, L. J., McDonald, S., Nakagawa, S., Nunan, D., Welch, V. A., & Page, M. J. (2023). Systematic reviewers' perspectives on sharing review data, analytic code, and other materials: A survey. *Cochrane Evidence Synthesis and Methods*, *1*(2), e12008. https://doi.org/10.1002/cesm.12008

Nguyen, T., & Duncan, G. J. (2019). Kindergarten components of executive function and third grade achievement: A national study. *Early Childhood Research Quarterly*, *46*, 49–61. https://doi.org/10.1016/j.ecresq.2018.05.006

Nguyen, T., Duncan, R. J., & Bailey, D. H. (2019). Theoretical and methodological implications of associations between executive function and mathematics in early childhood. *Contemporary Educational Psychology*, *58*, 276–287. https://doi.org/10.1016/j.cedpsych.2019.04.002

Niepel, C. (2023). Dynamics in Students' Momentary Perceptions of Instructional Quality in Everyday School Life (DYNAMIQUES) [Grant Proposal]. *Fonds National de La Recherche Luxembourg*.

Niepel, C., Brunner, M., & Preckel, F. (2014). The longitudinal interplay of students' academic self-concepts and achievements within and across domains: Replicating and extending the reciprocal internal/external frame of reference model. *Journal of Educational Psychology*, *106*(4), 1170–1191. https://doi.org/10.1037/a0036307

No Child Left Behind Act, Pub. L. No. 107–110 (2002).

Noble, K. G., McCandliss, B. D., & Farah, M. J. (2007). Socioeconomic gradients predict individual differences in neurocognitive abilities. *Developmental Science*, *10*(4), 464–480. https://doi.org/10.1111/j.1467-7687.2007.00600.x

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Nosek, B. A., & Lindsay, D. S. (2018, February 28). Preregistration Becoming the Norm in Psychological Science. *APS Observer*. https://www.psychologicalscience.org/observer/preregistration-becoming-the-norm-in-psychological-science

Nurmi, J.-E. (2012). Students' characteristics and teacher–child relationships in instruction: A meta-analysis. *Educational Research Review*, *7*(3), 177–197. https://doi.org/10.1016/j.edurev.2012.03.001

Oberauer, K., Schulze, R., Wilhelm, O., & Süß, H.-M. (2005). Working Memory and Intelligence--Their Correlation and Their Relation: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, *131*(1), 61–65. https://doi.org/10.1037/0033-2909.131.1.61

OECD. (2020). *Global Teaching InSights: A Video Study of Teaching*. OECD. https://doi.org/10.1787/20d6f36b-en

OECD & UNESCO Institute for Statistics. (2003). *Literacy Skills for the World of Tomorrow: Further Results from PISA 2000*. OECD. https://doi.org/10.1787/9789264102873-en

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

OpenAI. (2023). *ChatGPT (Mar 14 version) [Large language model]* [Computer software]. https://chat.openai.com

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *International Journal of Surgery*, *88*, 105906. https://doi.org/10.1016/j.ijsu.2021.105906

Paige, M. A., & Amrein-Beardsley, A. (2020). "Houston, We Have a Lawsuit": A Cautionary Tale for the Implementation of Value-Added Models for High-Stakes Employment Decisions. *Educational Researcher*, *49*(5), 350–359. https://doi.org/10.3102/0013189X20923046

Papay, J. P. (2011). Different Tests, Different Answers: The Stability of Teacher Value-Added Estimates Across Outcome Measures. *American Educational Research Journal*, *48*(1), 163–193. https://doi.org/10.3102/0002831210362589

Paré, G., & Kitsiou, S. (2017). Methods for literature reviews. In *Handbook of eHealth evaluation: An evidence-based approach [Internet]*. University of Victoria.

Paré, G., Trudel, M.-C., Jaana, M., & Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management*, *52*(2), 183–199. https://doi.org/10.1016/j.im.2014.08.008

Parkin, J. R., & Beaujean, A. A. (2012). The effects of Wechsler Intelligence Scale for Children—Fourth Edition cognitive abilities on math achievement. *Journal of School Psychology*, *50*(1), 113–128. https://doi.org/10.1016/j.jsp.2011.08.003

Passolunghi, M. C., & Lanfranchi, S. (2012). Domain-specific and domain-general precursors of mathematical achievement: A longitudinal study from kindergarten to first grade: Cognitive precursors of mathematical achievement. *British Journal of Educational Psychology*, *82*(1), 42–63. https://doi.org/10.1111/j.2044-8279.2011.02039.x

Paterson, T. A., Harms, P. D., Steel, P., & Credé, M. (2016). An Assessment of the Magnitude of Effect Sizes: Evidence From 30 Years of Meta-Analysis in Management. *Journal of Leadership & Organizational Studies*, *23*(1), 66–81. https://doi.org/10.1177/1548051815614321

Peixoto, F., Mata, L., Monteiro, V., Sanches, C., & Pekrun, R. (2015). The Achievement Emotions Questionnaire: Validation for Pre-Adolescent Students. *European Journal* *of Developmental Psychology*, *12*(4), 472–481. https://doi.org/10.1080/17405629.2015.1040757

Peláez-Fernández, M. A., Mérida-López, S., Sánchez-Álvarez, N., & Extremera, N. (2021). Managing Teachers' Job Attitudes: The Potential Benefits of Being a Happy and Emotional Intelligent Teacher. *Frontiers in Psychology*, *12*, 661151. https://doi.org/10.3389/fpsyg.2021.661151

Peng, P., Barnes, M., Wang, C., Wang, W., Li, S., Swanson, H. L., Dardick, W., & Tao, S. (2018). A meta-analysis on the relation between reading and working memory. *Psychological Bulletin*, *144*(1), 48–76. https://doi.org/10.1037/bul0000124

Peng, P., & Kievit, R. A. (2020). The Development of Academic Achievement and Cognitive Abilities: A Bidirectional Perspective. *Child Development Perspectives*, *14*(1), 15–20. https://doi.org/10.1111/cdep.12352

Peng, P., Lin, X., Ünal, Z. E., Lee, K., Namkung, J., Chow, J., & Sales, A. (2020). Examining the mutual relations between language and mathematics: A meta-analysis. *Psychological Bulletin*, *146*(7), 595–634. https://doi.org/10.1037/bul0000231

Peng, P., Namkung, J., Barnes, M., & Sun, C. (2016). A meta-analysis of mathematics and working memory: Moderating effects of working memory domain, type of mathematics skill, and sample characteristics. *Journal of Educational Psychology*, *108*(4), 455–473. https://doi.org/10.1037/edu0000079

Peroni, C., Riillo, C. A. F., & Sarracino, F. (2016). Entrepreneurship and immigration: Evidence from GEM Luxembourg. *Small Business Economics*, *46*(4), 639–656. https://doi.org/10.1007/s11187-016-9708-y

Perry, T. (2016). English value-added measures: Examining the limitations of school performance measurement. *British Educational Research Journal*, *42*(6), 1056–1080. https://doi.org/10.1002/berj.3247

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2007). Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Statistics in Medicine*, *26*(25), 4544–4562. https://doi.org/10.1002/sim.2889

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2008). Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of Clinical Epidemiology*, *61*(10), 991–996. https://doi.org/10.1016/j.jclinepi.2007.11.010

Pianta, R. C. (1999). *Enhancing relationships between children and teachers.* American Psychological Association. https://doi.org/10.1037/10314-000

Pianta, R. C. (2001). *STRS Student-Teacher Relationship Scale. Professional manual.* Lutz, FL: Psychological Assessment Resources.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System^{TM}: Manual K-3.* (pp. xi, 112). Paul H. Brookes Publishing Co.

Piccolo, L. R., Merz, E. C., Noble, K. G., & the Pediatric Imaging, Neurocognition, and Genetics Study. (2019). School climate is associated with cortical thickness and executive function in children and adolescents. *Developmental Science*, *22*(1), e12719. https://doi.org/10.1111/desc.12719

Pieper, D., Antoine, S.-L., Mathes, T., Neugebauer, E. A. M., & Eikermann, M. (2014). Systematic review finds overlapping reviews were not mentioned in every other overview. *Journal of Clinical Epidemiology*, *67*(4), 368–375. https://doi.org/10.1016/j.jclinepi.2013.11.007

Polanin, J. R., Maynard, B. R., & Dell, N. A. (2017). Overviews in Education Research: A Systematic Review and Analysis. *Review of Educational Research*, *87*(1), 172–203. https://doi.org/10.3102/0034654316631117

Praetorius, A.-K., & Charalambous, C. Y. (2018). Classroom observation frameworks for studying instructional quality: Looking back and looking forward. *ZDM*, *50*(3), 535–553. https://doi.org/10.1007/s11858-018-0946-0

Praetorius, A.-K., & Charalambous, C. Y. (Eds.). (2023). *Theorizing Teaching: Current Status and Open Issues*. Springer International Publishing. https://doi.org/10.1007/978-3-031-25613-4

Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of Three Basic Dimensions. *ZDM*, *50*(3), 407–426. https://doi.org/10.1007/s11858-018-0918-4

Purpura, D. J., Schmitt, S. A., & Ganley, C. M. (2017). Foundations of mathematics and literacy: The role of executive functioning components. *Journal of Experimental Child Psychology*, *153*, 15–34. https://doi.org/10.1016/j.jecp.2016.08.010

Pustejovsky, J. (2021). *clubSandwich: Cluster-Robust (Sandwich) Variance Estimators with Small-Sample Corrections. R package* (0.5.3) [Computer software]. https://CRAN.R-project.org/package=clubSandwich

Pustejovsky, J. E., & Tipton, E. (2021). Meta-analysis with Robust Variance Estimation: Expanding the Range of Working Models. *Prevention Science*, 1–14. https://doi.org/10.1007/s11121-021-01246-3

Quartagno, M., & Carpenter, J. (2019). *jomo: A package for Multilevel Joint Modelling Multiple Imputation*. https://CRAN.R-project.org/package=jomo

Quin, D. (2017). Longitudinal and Contextual Associations Between Teacher–Student Relationships and Student Engagement: A Systematic Review. *Review of Educational Research*, *87*(2), 345–387. https://doi.org/10.3102/0034654316669434

Quintana, D. S. (2015). From pre-registration to publication: A non-technical primer for conducting a meta-analysis to synthesize correlational data. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01549

Quintana, D. S. (2023). A Guide for Calculating Study-Level Statistical Power for Meta-Analyses. *Advances in Methods and Practices in Psychological Science*, *6*(1), 251524592211472. https://doi.org/10.1177/25152459221147260

Quintana, D. S. (2020, June 24). *Visualising power in meta-analysis effects*. Hormones, Brain, and Behavior. https://www.dsquintana.blog/meta-analysis-power-plot/

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Race to the Top Act, S.844-112th Congress (2011). www.govtrack.us/congress/bills/112/s844

Raghubar, K. P., Barnes, M. A., & Hecht, S. A. (2010). Working memory and mathematics: A review of developmental, individual difference, and cognitive approaches. *Learning and Individual Differences*, *20*(2), 110–122. https://doi.org/10.1016/j.lindif.2009.10.005

Rampton, B. (2017). *Crossing: Language and ethnicity among adolescents*. Routledge.

Rattan, A., Good, C., & Dweck, C. S. (2012). “It’s ok — Not everyone can be good at math”: Instructors with an entity theory comfort (and demotivate) students. *Journal of Experimental Social Psychology*, *48*(3), 731–737. https://doi.org/10.1016/j.jesp.2011.12.012

Redick, T. S., & Lindsey, D. R. B. (2013). Complex span and n-back measures of working memory: A meta-analysis. *Psychonomic Bulletin & Review*, *20*(6), 1102–1113. https://doi.org/10.3758/s13423-013-0453-9

Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*, *25*(1), 30–45. https://doi.org/10.1037/met0000220

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One Hundred Years of Social Psychology Quantitatively Described. *Review of General Psychology*, *7*(4), 331–363. https://doi.org/10.1037/1089-2680.7.4.331

Riggs, N. R., Jahromi, L. B., Razza, R. P., Dillworth-Bart, J. E., & Mueller, U. (2006). Executive function and the promotion of social–emotional competence. *Journal of Applied Developmental Psychology*, *27*(4), 300–309. https://doi.org/10.1016/j.appdev.2006.04.002

Robinson, C. D. (2022). A Framework for Motivating Teacher-Student Relationships. *Educational Psychology Review*, *34*(4), 2061–2094. https://doi.org/10.1007/s10648-022-09706-0

Robitzsch, A., Kiefer, T., & Wu, M. (2019). *TAM: Test Analysis Modules* (3.3-10). https://CRAN.R-project.org/package=TAM

Rogers, A., Castree, N., & Kitchin, R. (2013). Superdiversity. In *A Dictionary of Human Geography*. Oxford University Press. https://www.oxfordreference.com/display/10.1093/acref/9780199599868.001.0001/acref-9780199599868-e-1821

Roorda, D. L., Jak, S., Zee, M., Oort, F. J., & Koomen, H. M. Y. (2017). Affective Teacher–Student Relationships and Students' Engagement and Achievement: A Meta-Analytic Update and Test of the Mediating Role of Engagement. *School Psychology Review*, *46*(3), 239–261. https://doi.org/10.17105/SPR-2017-0035.V46-3

Roorda, D. L., Koomen, H. M. Y., Spilt, J. L., & Oort, F. J. (2011). The Influence of Affective Teacher–Student Relationships on Students' School Engagement and Achievement: A Meta-Analytic Approach. *Review of Educational Research*, *81*(4), 493–529. https://doi.org/10.3102/0034654311421793

Roorda, D. L., Koomen, H. M. Y., Spilt, J. L., & Oort, F. J. (2014). De invloed van affectieve leraar-leerlingrelaties op het schools leren van leerlingen: Verschillen tussen basis- en voortgezet onderwijs. *Pädagogische Studiën*, *91*(2), 97.112.

Roorda, D. L., Zee, M., & Koomen, H. M. Y. (2021). Don't forget student-teacher dependency! A Meta-analysis on associations with students' school adjustment and the moderating role of student and teacher characteristics. *Attachment & Human Development*, *23*(5), 490–503. https://doi.org/10.1080/14616734.2020.1751987

Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence*, *53*, 118–137. https://doi.org/10.1016/j.intell.2015.09.002

Rother, E. T. (2007). Revisão sistemática X revisão narrativa. *Acta Paulista de Enfermagem*, *20*(2), v–vi. https://doi.org/10.1590/S0103-21002007000200001

Rueger, S. Y., Malecki, C. K., & Demaray, M. K. (2008). Gender differences in the relationship between perceived social support and student adjustment during early adolescence. *School Psychology Quarterly*, *23*(4), 496–514. https://doi.org/10.1037/1045-3830.23.4.496

Ryan, R. M., Duineveld, J., Di Domenico, S., Ryan, W. S., Steward, B. A., & Bradshaw, E. L. (2023). *We know this much is (meta-analytically) true: A meta-review of meta-analytic findings evaluating self-determination theory* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/gk5cy

Sabol, T. J., & Pianta, R. C. (2012). Recent trends in research on teacher–child relationships. *Attachment & Human Development*, *14*(3), 213–231. https://doi.org/10.1080/14616734.2012.672262

Sanders, W. L., & Horn, S. P. (1994). The tennessee value-added assessment system (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education*, *8*(3), 299–311. https://doi.org/10.1007/BF00973726

Sanders, W. L., Wright, S. P., & Horn, S. P. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education*, *11*(1), 57–67. https://doi.org/10.1023/A:1007999204543

Sass, T. R. (2008). *The stability of value-added measures of teacher quality and implications for teacher compensation policy. Brief 4* (Brief 4). National Center for Analysis of Longitudinal Data in Education Research. https://eric.ed.gov/?id=ED508273

Sattler, S. (2022). *Curriculum und Mehrsprachigkeit: Planung und Gestaltung sprachlicher Identität in Luxemburg*. Transcript.

Schaffer, H. R., & Emerson, P. E. (1964). The Development of Social Attachments in Infancy. *Monographs of the Society for Research in Child Development*, *29*(3), 1. https://doi.org/10.2307/1165727

Schalken, N., & Rietbergen, C. (2017). The Reporting Quality of Systematic Reviews and Meta-Analyses in Industrial and Organizational Psychology: A Systematic Review. *Frontiers in Psychology*, *8*, 1395. https://doi.org/10.3389/fpsyg.2017.01395

Scheerens, J. (2005). Review of school and instructional effectiveness research. *Paper Commissioned for the EFA Global Monitoring Report*.

Scheffel, C., Korb, F., Dörfel, D., Eder, J., Möschl, M., Schoemann, M., & Scherbaum, S. (2023). Gute wissenschaftliche Praxis und Open Science im Empiriepraktikum: Wissenschaftlicher Kompetenzerwerb durch Replikationsstudien. *Psychologische Rundschau*, *74*(4), 241–243. https://doi.org/10.1026/0033-3042/a000643

Scherer, R., & Emslander, V. (2022a). *Quality Assessment in Meta-Analyses (QuAMA) [Preregistration]*. https://doi.org/10.17605/OSF.IO/3E64F

Scherer, R., & Emslander, V. (2022b). *Quality Assessment in Meta-Analysis (QuAMA) [Data & Materials]*. https://doi.org/10.17605/OSF.IO/NGVCZ

Scherer, R., & Emslander, V. (2023a). *Systematic Review of Second-Order Meta-Analyses (SR-SOMA)*. https://doi.org/10.17605/OSF.IO/RVUSC

Scherer, R., & Emslander, V. (2023b). *Utilizing Primary Study Quality in Meta-Analyses: A Step-by-Step Tutorial* (Quality Assessment in Meta-Analysis (QuAMA)) [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/8emsa

Scherer, R., & Nilsen, T. (2016). The Relations Among School Climate, Instructional Quality, and Achievement Motivation in Mathematics. In T. Nilsen & J.-E. Gustafsson (Eds.), *Teacher Quality, Instructional Quality and Student Outcomes* (Vol. 2, pp. 51–80). Springer International Publishing. https://doi.org/10.1007/978-3-319-41252-8_3

Scherer, R., Nilsen, T., & Jansen, M. (2016). Evaluating Individual Students’ Perceptions of Instructional Quality: An Investigation of their Factor Structure, Measurement Invariance, and Relations to Educational Outcomes. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00110

Scherer, R., Siddiq, F., & Tondeur, J. (2020). All the same or different? Revisiting measures of teachers’ technology acceptance. *Computers & Education*, *143*, 103656. https://doi.org/10.1016/j.compedu.2019.103656

Scherer, R., & Teo, T. (2020). A tutorial on the meta-analytic structural equation modeling of reliability coefficients. *Psychological Methods*, *25*(6), 747–775. https://doi.org/10.1037/met0000261

Scherer, R., Tondeur, J., Siddiq, F., & Baran, E. (2018). The importance of attitudes toward technology for pre-service teachers’ technological, pedagogical, and content knowledge: Comparing structural equation modeling approaches. *Computers in Human Behavior*, *80*, 67–80. https://doi.org/10.1016/j.chb.2017.11.003

Scherrer, J. (2011). Measuring teaching using value-added modeling: The imperfect panacea. *NASSP Bulletin*, *95*(2), Article 2. https://doi.org/10.1177/0192636511410052

Schmidt, F. L., & Oh, I.-S. (2013). Methods for second order meta-analysis and illustrative applications. *Organizational Behavior and Human Decision Processes*, *121*(2), 204–218. https://doi.org/10.1016/j.obhdp.2013.03.002

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*(4), 350.

Schmitt, S. A., Duncan, R. J., Budrevich, A., Korucu, & I. (2020). Benefits of Behavioral Self-Regulation in the Context of High Classroom Quality for Preschoolers' Mathematics. *Early Education and Development*, *31*(3), 323–334.

Schneider, M., & Preckel, F. (2017). Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychological Bulletin*, *143*(6), 565–600. https://doi.org/10.1037/bul0000098

Schneider, W. J., & McGrew, K. S. (2018). The Cattell–Horn–Carroll theory of cognitive abilities. In *In D. P. Flanagan & E. M. McDonough (Eds.), Contemporary intellectual assessment: Theories, tests, and issues, 4th ed.* (pp. 73–163). The Guilford Press.

Schwarzer, G., Chemaitelly, H., Abu-Raddad, L. J., & Rücker, G. (2019). Seriously misleading results using inverse of Freeman-Tukey double arcsine transformation in meta-analysis of single proportions. *Research Synthesis Methods*, *10*(3), 476–483. https://doi.org/10.1002/jrsm.1348

SCRIPT. (2018). *CARAT Ein Schulklima-Modell für Luxemburger Schulen* (p. 40). SCRIPT Service de Coordination de la Recherche et de l'Innovation pédagogiques et technologiques. https://www.script.lu/sites/default/files/publications/2019-12/Carat%20brochure%202018.pdf

SCRIPT. (2023, December 21). *Zesumme wuessen | Alphabetiséierung op franséisch*. https://alpha.script.lu/fr

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, *77*(4), 454–499.

Shea, B. J., Grimshaw, J. M., Wells, G. A., Boers, M., Andersson, N., Hamel, C., Porter, A. C., Tugwell, P., Moher, D., & Bouter, L. M. (2007). Development of AMSTAR: A measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology*, *7*(1), 10. https://doi.org/10.1186/1471-2288-7-10

Shea, B. J., Hamel, C., Wells, G. A., Bouter, L. M., Kristjansson, E., Grimshaw, J., Henry, D. A., & Boers, M. (2009). AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. *Journal of Clinical Epidemiology*, *62*(10), 1013–1020. https://doi.org/10.1016/j.jclinepi.2008.10.009

Shen, J., Wang, Y., Kurpad, N., & Schena, D. A. (2022). A Systematic Review on the Impact of Hot and Cool Executive Functions on Pediatric Injury Risks: A Meta-Analytic Structural Equation Modeling Approach. *Prevention Science*, *23*(3), 366–377. https://doi.org/10.1007/s11121-021-01271-2

Sheng, Z., Kong, W., Cortina, J. M., & Hou, S. (2016). Analyzing matrices of meta-analytic correlations: Current practices and recommendations. *Research Synthesis Methods*, *7*(2), 187–208. https://doi.org/10.1002/jrsm.1206

Siddaway, A. P., Wood, A. M., & Hedges, L. V. (2019). How to Do a Systematic Review: A Best Practice Guide for Conducting and Reporting Narrative Reviews, Meta-Analyses, and Meta-Syntheses. *Annual Review of Psychology*, *70*(1), 747–770. https://doi.org/10.1146/annurev-psych-010418-102803

Skaalvik, E. M., & Skaalvik, S. (2007). Dimensions of teacher self-efficacy and relations with strain factors, perceived collective teacher efficacy, and teacher burnout. *Journal of Educational Psychology*, *99*(3), 611–625. https://doi.org/10.1037/0022-0663.99.3.611

Slavin, R. E. (1995). A model of effective instruction. *The Educational Forum*, *59*(2), 166–176.

Smaldino, P. E., Turner, M. A., & Contreras Kallens, P. A. (2019). Open science and modified funding lotteries can impede the natural selection of bad science. *Royal Society Open Science*, *6*(7), 190194. https://doi.org/10.1098/rsos.190194

Smith, G., & Smith, J. (2005). Regression to the Mean in Average Test Scores. *Educational Assessment*, *10*(4), 377–399. https://doi.org/10.1207/s15326977ea1004_4

Söğüt, M., Göksun, T., & Altan-Atalay, A. (2021). The role of numeracy skills on the Wisconsin card sorting test (WCST) performances of 5- to 8-Year-old turkish children. *British Journal of Developmental Psychology*, *39*(1), 231–246. https://doi.org/10.1111/bjdp.12353

Sonnleitner, P., Krämer, C., Gamo, S., Reichert, M., Keller, U., & Fischbach, A. (2021). *Neue längsschnittliche Befunde aus dem nationalen Bildungsmonitoring ÉpStan in der 3. und 9. Klasse: Schlechtere Ergebnisse und wirkungslose Klassenwiederholungen*. https://doi.org/10.48746/BB2021LU-DE-24A

Soveri, A., Antfolk, J., Karlsson, L., Salo, B., & Laine, M. (2017). Working memory training revisited: A multi-level meta-analysis of n-back training studies. *Psychonomic Bulletin & Review*, *24*(4), 1077–1096. https://doi.org/10.3758/s13423-016-1217-0

Spiegel, J. A., Goodrich, J. M., Morris, B. M., Osborne, C. M., & Lonigan, C. J. (2021). Relations between executive functions and academic outcomes in elementary school children: A meta-analysis. *Psychological Bulletin*, *147*(4), 329–351. https://doi.org/10.1037/bul0000322

Spilt, J. L., Koomen, H. M. Y., & Thijs, J. T. (2011). Teacher Wellbeing: The Importance of Teacher–Student Relationships. *Educational Psychology Review*, *23*(4), 457–477. https://doi.org/10.1007/s10648-011-9170-y

Spruit, A., Goos, L., Weenink, N., Rodenburg, R., Niemeyer, H., Stams, G. J., & Colonnesi, C. (2020). The Relation Between Attachment and Depression in Children and Adolescents: A Multilevel Meta-Analysis. *Clinical Child and Family Psychology Review*, *23*(1), 54–69. https://doi.org/10.1007/s10567-019-00299-9

Starkey, P., Klein, A., & Wakeley, A. (2004). Enhancing young children's mathematical knowledge through a pre-kindergarten mathematics intervention. *Early Childhood Research Quarterly*, *19*(1), 99–120. https://doi.org/10.1016/j.ecresq.2004.01.002

STATEC. (2021, November 8). *Nationalités: Une population de plus en plus cosmopolite*. http://statistiques.public.lu/fr/recensement/nationalites.html

Strauss, L. I., Peterson, M. S., Kidd, J. K., Choe, J., Lauritzen, H. C., Patterson, A. B., Holmberg, C. A., Gallington, D. A., & Pasnak, R. (2020). Evaluation of patterning instruction for kindergartners. *The Journal of Educational Research*, *113*(4), 292–302. https://doi.org/10.1080/00220671.2020.1806195

Stringfield, S. (1994). A model of elementary school effects. In *Advances in school effectiveness research and practice* (pp. 153–187). Elsevier.

Strom, R. E., & Boster, F. J. (2007). Dropping Out of High School: A Meta-Analysis Assessing the Effect of Messages in the Home and in School. *Communication Education*, *56*(4), 433–452. https://doi.org/10.1080/03634520701413804

Suber, P. (2012). *Open access*. MIT Press.

Swanson, H. L., & Jerman, O. (2006). Math Disabilities: A Selective Meta-Analysis of the Literature. *Review of Educational Research*, *76*(2), 249–274. https://doi.org/10.3102/00346543076002249

Takacs, Z. K., & Kassai, R. (2019). The efficacy of different interventions to foster children's executive function skills: A series of meta-analyses. *Psychological Bulletin*, *145*(7), 653–697. https://doi.org/10.1037/bul0000195

Tang, M., Bever, J. D., & Yu, F. (2017). Open access increases citations of papers in ecology. *Ecosphere*, *8*(7), e01887. https://doi.org/10.1002/ecs2.1887

Tao, Y., Meng, Y., Gao, Z., & Yang, X. (2022). Perceived teacher support, student engagement, and academic achievement: A meta-analysis. *Educational Psychology*, *42*(4), 401–420. https://doi.org/10.1080/01443410.2022.2033168

Tekwe, C. D., Carter, R. L., Ma, C.-X., Algina, J., Lucas, M. E., Roth, J., Ariet, M., Fisher, T., & Resnick, M. B. (2004). An Empirical Comparison of Statistical Models for Value-Added Assessment of School Performance. *Journal of Educational and Behavioral Statistics*, *29*(1), 11–36. https://doi.org/10.3102/10769986029001011

Thapa, A., Cohen, J., Guffey, S., & Higgins-D'Alessandro, A. (2013). A Review of School Climate Research. *Review of Educational Research*, *83*(3), 357–385. https://doi.org/10.3102/0034654313483907

Thomas, S., Peng, W. J., & Gray, J. (2007). Modelling patterns of improvement over time: Value added trends in English secondary school performance across ten cohorts. *Oxford Review of Education*, *33*(3), 261–295. https://doi.org/10.1080/03054980701366116

Thompson, S. G., & Pocock, S. J. (1991). Can meta-analyses be trusted? *The Lancet*, *338*(8775), 1127–1130. https://doi.org/10.1016/0140-6736(91)91975-Z

Thorell, L. B., & Nyberg, L. (2008). The Childhood Executive Functioning Inventory (CHEXI): A new rating instrument for parents and teachers. *Developmental Neuropsychology*, *33*(4), 536–552. https://doi.org/10.1080/87565640802101516

Timmermans, A. C., de Wolf, I. F., Bosker, R. J., & Doolaard, S. (2015). Risk-based educational accountability in Dutch primary education. *Educational Assessment, Evaluation and Accountability*, *27*(4), 323–346. https://doi.org/10.1007/s11092-015-9212-y

Tod, D., Booth, A., & Smith, B. (2022). Critical appraisal. *International Review of Sport and Exercise Psychology*, *15*(1), 52–72. https://doi.org/10.1080/1750984X.2021.1952471

Trad, A. (2022). The Societal and Educational Transformation Projects The Middle East's Educational Construct-The Case of the Lebanese Specific Diverse Educational System (LSDES). *The Business and Management Review*, 40. https://doi.org/10.24052/BMR/V13NU03/ART-06

Träff, U., Olsson, L., Östergren, R., & Skagerlund, K. (2020). Development of early domain-specific and domain-general cognitive precursors of high and low math achievers in grade 6. *Child Neuropsychology*, *26*(8), 1065–1090. https://doi.org/10.1080/09297049.2020.1739259

Turnbull, D., Chugh, R., & Luck, J. (2023). Systematic-narrative hybrid literature review: A strategy for integrating a concise methodology into a manuscript. *Social Sciences & Humanities Open*, *7*(1), 100381. https://doi.org/10.1016/j.ssaho.2022.100381

Tymms, P. (1999). Baseline assessment, value-added and the prediction of reading. *Journal of Research in Reading*, *22*(1), 27–36. https://doi.org/10.1111/1467-9817.00066

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., & Newcombe, N. S. (2013). The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin*, *139*(2), 352–402. https://doi.org/10.1037/a0028446

Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How Many Studies Do You Need?: A Primer on Statistical Power for Meta-Analysis. *Journal of Educational and Behavioral Statistics*, *35*(2), 215–247. https://doi.org/10.3102/1076998609346961

van Aken, L., van der Heijden, P. T., Oomens, W., Kessels, R. P., & Egger, J. I. (2019). Predictive value of traditional measures of executive function on broad abilities of the cattell–horn–carroll theory of cognitive abilities. *Assessment*, *26*(7), 1375–1385. https://doi.org/10.1177/1073191117731814

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods*, *45*(2), 576–594. https://doi.org/10.3758/s13428-012-0261-6

Van Der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, *113*(4), 842–861. https://doi.org/10.1037/0033-295X.113.4.842

van der Ven, S. H. G. (2011). *The structure of executive functions and relations with early math learning*. Utrecht University.

Van der Ven, S. H. G., Kroesbergen, E. H., Boom, J., & Leseman, P. P. M. (2012). The development of executive functions and early mathematics: A dynamic relationship: Development executive functions mathematics. *British Journal of Educational Psychology*, *82*(1), 100–119. https://doi.org/10.1111/j.2044-8279.2011.02035.x

Vandenbroucke, L., Spilt, J., Verschueren, K., Piccinin, C., & Baeyens, D. (2018). The Classroom as a Developmental Context for Cognitive Development: A Meta-Analysis on the Importance of Teacher–Student Interactions for Children's Executive Functions. *Review of Educational Research*, *88*(1), 125–164. https://doi.org/10.3102/0034654317743200

Vaughn, S., Wanzek, J., Murray, C. S., & Roberts, G. (2012). Intensive Interventions for Students Struggling in Reading and Mathematics. A Practice Guide. *Center on Instruction. ERIC.*

Verdine, B. N., Golinkoff, R. M., Hirsh-Pasek, K., & Newcombe, N. S. (2017). I. Spatial Skills, Their Development, and Their Links to Mathematics. *Monographs of the Society for Research in Child Development*, *82*(1), 7–30. Embase. https://doi.org/10.1111/mono.12280

Verschueren, K., & Koomen, H. M. Y. (2012). Teacher–child relationships from an attachment perspective. *Attachment & Human Development*, *14*(3), 205–211. https://doi.org/10.1080/14616734.2012.672260

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Viechtbauer, W. (2021). Aggregate Multiple Effect Sizes or Outcomes Within Studies. *Metafor*. https://wviechtb.github.io/metafor/reference/aggregate.escalc.html

Viechtbauer, W., & Cheung, M. W.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, *1*(2), 112–125. https://doi.org/10.1002/jrsm.11

Waack, S. (2018). *Hattie Ranking: 252 Influences And Effect Sizes Related To Student Achievement.* Visible Learning. www.visible-learning.org

Wagner, W., Göllner, R., Werth, S., Voss, T., Schmitz, B., & Trautwein, U. (2016). Student and teacher ratings of instructional quality: Consistency of ratings over time, agreement, and predictive power. *Journal of Educational Psychology*, *108*(5), 705–721. https://doi.org/10.1037/edu0000075

Wang, M. C., Haertel, G. D., & Walberg, H. J. (1993). Toward a Knowledge Base for School Learning. *Review of Educational Research*, *63*(3), 249–294. https://doi.org/10.2307/1170546

Wang, M. C., Haertel, G. D., & Walberg, H. J. (1997). *What Helps Students Learn? Spotlight on Student Success.*

Wang, M.-T., & Degol, J. L. (2016). School Climate: A Review of the Construct, Measurement, and Impact on Student Outcomes. *Educational Psychology Review*, *28*(2), 315–352. https://doi.org/10.1007/s10648-015-9319-1

Wang, M.-T., Degol, J. L., Amemiya, J., Parr, A., & Guo, J. (2020). Classroom climate and children's academic and psychological wellbeing: A systematic review and meta-analysis. *Developmental Review*, *57*, 100912. https://doi.org/10.1016/j.dr.2020.100912

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450. https://doi.org/10.1007/BF02294627

Webb, S. L., Loh, V., Lampit, A., Bateman, J. E., & Birney, D. P. (2018). Meta-Analysis of the Effects of Computerized Cognitive Training on Executive Functions: A Cross-Disciplinary Taxonomy for Classifying Outcome Cognitive Factors. *Neuropsychology Review*, *28*(2), 232–250. https://doi.org/10.1007/s11065-018-9374-8

Weber, M., Ruch, W., & Huebner, E. S. (2013). Adaptation and Initial Validation of the German Version of the Students' Life Satisfaction Scale (German SLSS). *European Journal of Psychological Assessment*, *29*(2), 105–112. https://doi.org/10.1027/1015-5759/a000133

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children—Fourth Edition*. San Antonio, TX: The Psychological Corporation.

Wedderhoff, N., & Bosnjak, M. (2020). Erfassung der Primärstudienqualität in psychologischen Meta-Analysen: Eine systematische Übersichtsarbeit. *Psychologische Rundschau*, *71*(2), 119–126. https://doi.org/10.1026/0033-3042/a000484

Weis, L., Boehm, B., & Krug, A. (2020). *PISA 2018 Luxemburg: Kompetenzen von Schülerinnen und Schülern im internationalen Vergleich [PISA 2018 Luxembourg: Competencies of students in international comparison]*. SCRIPT. https://men.public.lu/dam-assets/catalogue-publications/statistiques-etudes/secondaire/pisa-2018-vergleich.pdf

Weis, M., Mang, J., Baumann, B., & Reiss, K. (2018). Zuwanderung und Erfolg aus Sicht der PISA-Studie: Ein Gesamtüberblick von 2000 bis 2015. In S. Kracht, A. Niedostadek, & P. Sensburg (Eds.), *Praxishandbuch Professionelle Mediation* (pp. 1–14). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-658-18403-2_27-1

Wentzel, K. R. (2022). Does Anybody Care? Conceptualization and Measurement Within the Contexts of Teacher-Student and Peer Relationships. *Educational Psychology Review*, *34*(4), 1919–1954. https://doi.org/10.1007/s10648-022-09702-4

What Works Clearinghouse. (2023, July 19). *WWC | Find What Works!* https://ies.ed.gov/ncee/wwc/FWW

Wiebe, S. A., Espy, K. A., & Charak, D. (2008). Using confirmatory factor analysis to understand executive control in preschool children: I. Latent structure. *Developmental Psychology*, *44*(2), 575–587. https://doi.org/10.1037/0012-1649.44.2.575

Wiliam, D. (2010). Standardized Testing and School Accountability. *Educational Psychologist*, *45*(2), 107–122. https://doi.org/10.1080/00461521003703060

Wilkinson, S. S. (1980). *The Relationship of Teacher Praise and Student Achievement: A Meta-Analysis of Selected Research* [PhD Thesis]. University of Florida.

Willoughby, M., Kupersmidt, J., Voegler-Lee, M., & Bryant, D. (2011). Contributions of Hot and Cool Self-Regulation to Preschool Disruptive Behavior and Academic Achievement. *Developmental Neuropsychology*, *36*(2), 162–180. https://doi.org/10.1080/87565641.2010.549980

Willoughby, M. T., Kupersmidt, J. B., & Voegler-Lee, M. E. (2012). Is preschool executive function causally related to academic achievement? *Child Neuropsychology*, *18*(1), 79–91. https://doi.org/10.1080/09297049.2011.578572

Willoughby, M. T., Magnus, B., Vernon-Feagans, L., Blair, C. B., & Investigators, F. L. P. (2017). Developmental delays in executive function from 3 to 5 years of age predict kindergarten academic readiness. *Journal of Learning Disabilities*, *50*(4), 359–372. https://doi.org/10.1177/0022219415619754

Willoughby, M. T., Wylie, A. C., & Little, M. H. (2019). Testing longitudinal associations between executive function and academic achievement. *Developmental Psychology*, *55*(4), 767–779. https://doi.org/10.1037/dev0000664

Woodcock, R., McGrew, K., & Mather, N. (2001). Test review. *Rehabilitation Counseling Bulletin*, *44*(4), 232–235.

World Bank Group. (2022). *World Bank Open Data*. World Bank Open Data. https://data.worldbank.org

Wu, L. M., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest version 2: Generalised item response modelling software [computer program].* Australian Council for Educational Research.

Yang, X., Peng, P., & Meng, X. (2019). How do metalinguistic awareness, working memory, reasoning, and inhibition contribute to Chinese character reading of kindergarten children? *Infant and Child Development*, *28*(3), e2122. https://doi.org/10.1002/icd.2122

Yang, X., Yan, M., Ruan, Y., Ku, S. Y. Y., Lo, J. C. M., Peng, P., & McBride, C. (2021). Relations among phonological processing skills and mathematics in children: A meta-analysis. *Journal of Educational Psychology*. Advance online publication. https://doi.org/10.1037/edu0000710

Yaniv, B. (1982). *Revealing factors that contribute to teacher dropout* [Unpublished master's thesis]. University of Haifa.

Yeniad, N., Malda, M., Mesman, J., van IJzendoorn, M. H., & Pieper, S. (2013). Shifting ability predicts math and reading performance in children: A meta-analytical study. *Learning and Individual Differences*, *23*, 1–9. https://doi.org/10.1016/j.lindif.2012.10.004

Young, S. E., Friedman, N. P., Miyake, A., Willcutt, E. G., Corley, R. P., Haberstick, B. C., & Hewitt, J. K. (2009). Behavioral disinhibition: Liability for externalizing spectrum disorders and its genetic and environmental relation to response inhibition across adolescence. *Journal of Abnormal Psychology*, *118*(1), 117–130. https://doi.org/10.1037/a0014657

Zelazo, P. D. (2006). The Dimensional Change Card Sort (DCCS): A method of assessing executive function in children. *Nature Protocols*, *1*(1), 297–301. https://doi.org/10.1038/nprot.2006.46

Zelazo, P. D., & Carlson, S. M. (2012). Hot and Cool Executive Function in Childhood and Adolescence: Development and Plasticity. *Child Development Perspectives*, *6*(4), 354–360. https://doi.org/10.1111/j.1750-8606.2012.00246.x

Zell, E., Krizan, Z., & Teeter, S. R. (2015). Evaluating gender similarities and differences using metasynthesis. *American Psychologist*, *70*(1), 10–20. https://doi.org/10.1037/a0038208

12 Appendices

Appendix A. *Exemplary Search Terms as Used in APA PsycINFO*

OR	*AND*	*OR*	*AND*	*OR*
Teacher Student Interaction/ (student* or child* or pupil*) and teacher*) .tw.		(support* or care* or caring or warm* or closeness or compassion* or help* or empath* or trust* or conflict* or attach* or affection* or liking or like* or neglect* or reject* or dislik* or dependency or interact* or connect* or aggress* or violen* or interpersonal or engagement or communicat* or relation*).tw.		systematic review or meta analysis).md. or Meta Analysis/ or systematic review/ or (metaanaly* or (meta adj (analy* or synthes*)) or ((research or evidence) adj3 synthes*)).tw.

Note. The search term was adapted to fit APA PsycInfo. The adaptations to Education Research Complete (EBSCO), ERIC, Scopus, and Web of Science are displayed in Supplemental Material S4. * = the asterisk acts as a wildcard and includes alternate word endings; .tw. = term will be searched in table of contents, Title, Abstract, and key concepts; .md. = term will be searched in the methodology field; adj3 = term will be searched within three words before or after.

***Appendix B**. Percentages of Primary Study Overlap of Included Meta-Analyses*

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
1 Ali et al. 2015	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2 Allen et al. 2018	0	100	0	1	1	0	4	0	0	0	0	3	0	0	0	4	3	2	0	0	1	0	0	0
3 Cherne 2008	0	0	100	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0
4 Chu et al. 2010	0	4	0	100	1	1	2	0	0	0	2	6	0	0	0	3	3	2	0	0	1	0	0	0
5 Cornelius-White 2007	0	2	0	0	100	0	4	0	0	0	0	2	0	0	0	3	3	2	0	0	0	0	0	0
6 Endedijk et al. 2022	0	2	0	1	0	100	4	10	0	28	28	10	6	0	0	10	9	15	10	0	3	7	6	0
7 Givens Rolland 2012	0	4	0	0	2	1	100	0	0	0	0	2	0	0	0	1	1	2	0	0	3	0	3	0
8 Kincade et al. 2020	0	0	0	0	0	1	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0
9 Korpershoek et al. 2016	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10 Krause and Smith 2022	0	0	0	0	0	2	0	0	0	100	12	0	0	0	5	1	1	2	0	0	0	0	0	0
11 Lei et al. 2016	0	0	0	0	0	6	0	0	0	39	100	2	0	0	25	7	6	4	21	0	0	0	0	0
12 Lei et al. 2018	0	4	0	2	1	2	2	0	0	0	2	100	1	0	5	3	3	5	3	0	3	0	0	0
13 Li et al. 2021	0	0	0	0	0	1	0	0	0	0	0	2	100	0	0	1	1	3	0	0	3	54	9	0
14 Moore et al. 2019	0	0	11	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0
15 Nurmi 2012	0	0	0	0	0	0	0	0	0	6	9	2	0	0	100	4	5	3	3	0	0	0	0	0
16 Roorda et al. 2011	0	8	0	1	3	3	2	0	0	6	11	5	1	0	20	100	100	42	17	0	11	0	3	0
17 Roorda et al. 2014	0	6	0	1	3	3	2	0	0	6	9	5	1	0	20	95	100	40	17	0	11	0	3	0
18 Roorda et al. 2017	0	8	0	1	3	9	6	0	0	17	12	14	9	0	25	82	83	100	31	0	30	11	9	0
19 Roorda et al. 2021	0	0	0	0	0	1	0	0	0	0	11	2	0	0	5	5	6	5	100	0	0	0	0	0
20 Strom and Bolster 2007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0
21 Tao et al. 2022	0	2	0	0	0	1	4	0	0	0	0	3	3	0	0	9	9	12	0	0	100	0	6	0
22 Vandenbroucke et al. 2018	0	0	0	0	0	1	0	5	0	0	0	0	22	0	0	0	0	2	0	0	0	100	3	0
23 Wang et al. 2020	0	0	0	0	0	1	2	0	0	0	0	0	4	0	0	1	1	2	0	0	3	4	100	0
24 Wilkinson 1980	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100

Note. The numbers are percentages of overlapping primary studies from the meta-analyses in the columns with the meta-analyses in the rows. For example, Vandenbroucke et al. (2018) had 54% overlap of their 23 studies with Li et al. (2021). Li et al. (2021), however, included 68 studies and, thus, overlapped only 22% with Vandenbroucke et al. (2021).

Appendix C. *Description and Examples of Student Outcome Clusters Coded for the Present Meta-Analyses*

Student Outcome Clusters	Description	Examples
Academic		
Academic Achievement	Measurable accomplishments of students in educational settings, including grades, test scores, and mastery of knowledge, skills, and competencies including some cognitive aspects.	Reading gains (Wilkinson, 1980), grades and achievement (Givens Rolland, 2012)
Behavioral		
Appropriate Behavior	Behavior that aligns with established social norms and are sought-after, encouraged, or beneficial in a school setting.	School conduct (Allen et al., 2018), prosociability (Nurmi, 2012)
Behavior Problems	Maladaptive behavior that interferes with the social, emotional, or academic functioning of the student and is seen as negative or disruptive in school.	Inappropriate social behavior (Cherne, 2008), externalizing behavior (Wang et al., 2020)
Bullying	Intentional and aggressive behavior against another student in the form of physical, verbal, relational, or cyberbullying.	Bullying perpetration and victimization (Krause & Smith, 2022)
Socioemotional		
Academic Emotions	Affective experiences and feelings that students have in relation to their academic tasks, goals, and achievements.	Positive and negative academic emotions (Lei et al., 2018), anger and shyness (Nurmi, 2012)
School Belonging and Engagement	Sense of connectedness, identification, and emotional investment that students experience within the school with engagement being the behavioral expression of belonging.	School belonging (Allen et al., 2018), school engagement (Roorda et al., 2011, 2014, 2017, 2021)
Motivational		
Motivation	Internal processes that activate, direct, and sustain individuals' behavior and efforts toward achieving specific goals.	Personal mastery (Givens Rolland, 2012), motivational outcomes (Korpershoek et al., 2016)
Well-being	A composite of all positive aspects of academic, behavioral, socioemotional, motivational, and cognitive outcomes with the addition of health in students.	Well-being (Chu et al., 2010), cognitive/behavioral/affective outcomes (Cornelius-White, 2007), overall outcome (Korpershoek et al., 2016)
General Cognitive		
Executive functions and Self-Control	Mental processes that regulate human cognition and behavior often defined as response inhibition, mental set shifting, and updating of working memory.	Working memory, inhibition, and cognitive flexibility (Vandenbroucke et al., 2018), self-control (Li et al., 2021),
Positive/Negative Outcomes		
Positive Outcomes	Desirable and beneficial academic, behavioral, socioemotional, motivational, and cognitive student variables.	Appropriate behavior (Moore et al., 2019), personal mastery (Givens Rolland, 2012)
Negative Outcomes	Adverse or undesirable academic, behavioral, socioemotional, motivational, and cognitive student variables.	School dropout (Strom & Boster, 2007), social distress (Wang et al., 2020)

Positive/Negative TSRs		
Positive TSRs	Teacher-student interactions, connections, and emotional bonds characterized by warmth, closeness, support, and friendliness.	Teacher support (Tao et al., 2022), closeness (Kincade et al., 2020)
Negative TSRs	Teacher-student interactions, connections, and emotional bonds characterized by conflict or dependency.	TSR conflict (Krause & Smith, 2022), negative affective TSR (Lei et al., 2016).

Note. For a comprehensive description and examples of positive and negative TSRs, please refer to the Introduction section. TSRs = Teacher Student Relationships.

Appendix D. *Moderators Analyzed in Included Meta-Analyses (Ordered by Meta-Analysis)*

Meta-analysis	Significant moderators	Tested moderators
Ali et al. 2015	student gender	
Allen et al. 2018	school location	publication year, culture/country
Cherne 2008	student age	student disabilities, type of intervention, trigger for praise delivery, study quality, publication status
Chu et al. 2010	No moderator analysis that applied exclusively to the TSR component	
Cornelius-White 2007	TSR informant, teacher gender, teacher ethnic minority	study quality, sample size, student gender, student ethnic minority, student SES, teaching experience, publication year, publication type, school location, and community type
Endedijk et al. 2022	student age, TSR informant, TSR measurement time points	-
Givens Rolland 2012	student age, school location, achievement measurement type	-
Kincade et al. 2020	-	-
Korpershoek et al. 2016	outcome informant, intervention duration	student gender, student age, student SES, student behavior, culture/country
Krause & Smith 2022	student age, TSR informant, study quality	-
Lei et al. 2016	student age, outcome informant, student gender, culture/country	-
Lei et al. 2018	student age, student gender, culture/country	-
Li et al. 2021	No moderator analysis that applied exclusively to the TSR component	
Moore et al. 2019	No moderator analysis	
Nurmi 2012	No moderator analysis	
Roorda et al. 2011	TSR/outcome informant, measurement time points, student age, student gender, student ethnic minority, student SES, student disability, teacher gender, teacher ethnic minority, teaching experience	-
Roorda et al. 2014	student age, student gender, student ethnic minority, student SES, teaching experience, teacher ethnic minority, teacher gender	-
Roorda et al. 2017	-	student age
Roorda et al. 2021	Student age, student ethnic minority, teacher gender, teacher ethnic minority	student gender, teaching experience
Strom and Bolster 2007	-	student age, culture/country
Tao et al. 2022	student age, TSR dimension, outcome measurement type	culture/country
Vandenbroucke et al. 2018	student age, student gender, student SES, level of teacher-student interaction	study design
Wang et al. 2020	No moderator analysis that applied exclusively to the TSR component	
Wilkinson 1980	-	student age, student SES

Note. The list above shows moderators that are frequently tested in TSR meta-analyses. The reported moderators do not necessarily pertain only to TSRs.

Appendix E. *Results of the Categorical Moderators (Including Levels of Nonsignificant Moderators)*

Moderator	k_{MA}	k_{ES}	$\bar{\bar{r}}$ [95% CI]	SE	$R^2_{(2)}$	$R^2_{(3)}$	p
Baseline ($I^2_{(2;3)}$: 49%; 51%)	20	79	.24 [.17, .31]	0.04	–	–	–
Publication type	20	79	–	–	.00	.13	**.10**
Doctoral thesis	2	11	.42 [.20, .64]	0.11	–	–	–
Journal article	18	68	.23 [.15, .30]	0.04	–	–	–
Publication status	20	79	–	–	.00	.13	**.10**
Unpublished	2	11	.42 [.20, .64]	0.11	–	–	–
Published	18	68	.23 [.15, .30]	0.04	–	–	–
Affiliation	20	79	–	–	.03	.00	**.99**
University of Connecticut	1	2	.27 [-.26, .80]	0.26	–	–	–
East China Normal University	3	5	.29 [-.02, .60]	0.16	–	–	–
Harvard University	1	5	.30 [-.21, .81]	0.26	–	–	–
Kansas State University	1	1	.21 [-.35, .77]	0.28	–	–	–
Leiden University	1	2	.27 [-.26, .79]	0.26	–	–	–
Michigan State University	1	1	-.14 [-.72, .44]	0.29	–	–	–
Missouri State University	1	2	.33 [-.20, .86]	0.26	–	–	–
The Education University of Hong Kong	1	2	.21 [-.32, .74]	0.26	–	–	–
University of Amsterdam	2	9	.22 [-.14, .58]	0.18	–	–	–
University of Florida	1	3	.08 [-.44, .60]	0.26	–	–	–
University of Groningen	1	5	.06 [-.47, .59]	0.26	–	–	–
University of Jyväskylä	1	23	.19 [-.31, .69]	0.25	–	–	–
University of Melbourne	1	1	.46 [-.11, 1.03]	0.28	–	–	–
University of Minnesota	2	11	.43 [.06, .79]	0.18	–	–	–
University of Ottawa	1	2	.28 [-.25, .81]	0.27	–	–	–
University of Pittsburgh	1	5	.19 [-.32, .69]	0.25	–	–	–
Country	20	79	–	–	.00	.00	**.95**
USA	9	30	.25 [.11, .38]	0.07	–	–	–
Australia	1	1	.46 [.01, .91]	0.22	–	–	–
Canada	1	2	.28 [-.11, .68]	0.20	–	–	–
China	3	5	.29 [.06, .52]	0.12	–	–	–
Finland	1	23	.19 [-.16, .54]	0.17	–	–	–
Hong Kong	1	2	.21 [-.18, .60]	0.20	–	–	–
Netherlands	4	16	.19 [.01, .38]	0.09	–	–	–
Social minority	20	79	-	-	.02	.13	**.06**
No (baseline)	19	78	.26 [.19, .33]	0.04	–	–	–
Yes	1	1	-.14 [-.54, .26]	0.20	–	–	–

Grade level mode	11	27	-	-	.57	.55	**.28**
Prekindergarten	1	3	.09 [-.14, .32]	0.11	–	–	–
Kindergarten	2	6	.31 [.15, .47]	0.08	–	–	–
Kindergarten, preschool, & elementary school	1	4	.24 [.02, .46]	0.10	–	–	–
Preschool	1	1	.19 [-.07, .44]	0.12	–	–	–
Elementary school	3	6	.20 [.05, .34]	0.07	–	–	–
Elementary school, middle school, & highschool	1	2	.33 [.09, .57]	0.11	–	–	–
Middle school	2	3	.23 [.05, .41]	0.09	–	–	–
Middle school, & highschool	1	1	.26 [.00, .51]	0.12	–	–	–
High school	1	1	-.14 [-.45, .17]	0.15	–	–	–
Prekindergarden	19	71	-	-	.00	1	**.48**
No	16	43	.22 [.17, .27]	0.02	-	-	–
Yes	3	28	.19 [.11, .27]	0.04	-	-	**-**
Kindergarten	18	66	-	-	.00	1	**.89**
No	9	19	.22 [.15, .30]	0.04	-	-	–
Yes	9	47	.22 [.16, .27]	0.03	–	–	–
Preschool	18	66	-	-	.00	1	**.81**
No	10	22	.23 [.16, .29]	0.03	-	-	–
Yes	8	44	.22 [.16, .27]	0.03	-	-	**-**
Elementary school	18	66	-	-	.00	1	**.42**
No	4	9	.26 [.15, .37]	0.05	-	-	–
Yes	14	57	.21 [.17, .26]	0.02	-	-	**-**
Middle school	19	71	-	-	.04	1	**.02***
No	5	37	.15 [.08, .22]	0.03	-	-	–
Yes	14	34	.25 [.20, .30]	0.03	-	-	–
High school	19	71	-	-	.06	1	**.01***
No	5	39	.15 [.09, .22]	0.03	-	-	–
Yes	14	32	.26 [.21, .31]	0.03	-	-	–
TSR level	18	54	-	-	.62	.00	**0.21**
Dyads	10	30	.26 [.14, .37]	0.06	-	-	–
Classroom level	5	19	.16 [.00, .32]	0.08	-	-	**-**
Both	3	5	.40 [.19, .61]	0.10	-	-	**-**
TSR modality	24	79	-	-	.00	.00	**.82**
Negative	6	25	.25 [.15, .36]	0.05	-	-	–
Positive	18	54	.24 [.16, .32]	0.04	-	-	**-**
TSR informant mode	15	65	-	-	.00	.00	**.21**
Student	10	23	.22 [.11, .33]	0.06	-	-	–
Teacher	3	31	.16 [-.01, .34]	0.09	-	-	**-**

Other third party	2	11	.42 [.18, .66]	0.12	-	-	-
Outcome cluster	41	75	-	-	.00	.00	**.94**
Academic achievement	10	15	.22 [.11, .34]	0.06	-	-	–
Academic emotions	3	8	.26 [.09, .42]	0.08	-	-	-
Appropriate behavior	6	13	.24 [.10, .37]	0.07	-	-	-
Behavior problems	6	17	.28 [.16, .40]	0.06	-	-	-
Bullying	1	2	.28 [-.07, .64]	0.18	-	-	-
School belonging and engagement	3	4	.31 [.12, .51]	0.10	-	-	-
Peer relationship quality	1	1	.27 [-.10, .64]	0.19	-	-	-
Motivation	3	5	.30 [.12, .49]	0.09	-	-	-
Well-being	5	6	.30 [.14, .46]	0.08	-	-	-
Executive functions and self-control	1	2	.21 [-.14, .56]	0.17	-	-	-
Outcome modality	24	79	-	-	.00	.00	**.83**
Negative	7	25	.24 [.13, .34]	0.05	-	-	–
Positive	17	54	.25 [.17, .33]	0.04	-	-	-
Outcome informant mode	17	59	-	-	.00	.84	**.90**
Student	10	23	.24 [.15, .32]	0.04	-	-	–
Teacher	7	36	.23 [.14, .32]	0.04	-	-	-
Outcome type mode	37	73	-	-	.00	.00	**.95**
Academic	10	15	.23 [.11, .35]	0.06	-	-	-
Behavioral	11	27	.26 [.16, .36]	0.05	-	-	-
Behavioral & Academic	1	5	.34 [.08, .59]	0.13	-	-	-
Behavioral & Socioemotional	1	2	.35 [.06, .64]	0.15	-	-	-
Behavioral, Socioemotional, & General cognitive outcome	1	2	.36 [.10, .61]	0.13	-	-	-
Socioemotional	7	13	.25 [.12, .38]	0.07	-	-	-
Socioemotional & Behavioral	2	2	.29 [.03, .55]	0.13	-	-	-
General cognitive outcome	1	2	.21 [-.14, .56]	0.18	-	-	-
Motivational	3	5	.31 [.12, .51]	0.10	-	-	-
Moderator analysis (yes/no)	23	61	-	-	.00	.00	**1**
No	8	48	.25 [.13, .36]	.06	-	-	-
Yes	15	31	.25 [.16, .33]	.04	-	-	-

Note. Larger positive effect sizes indicate closer relations between TSRs and student outcomes. A significant *p*-value indicates that there is a statistical difference between the levels of the moderator. $I^2_{(2;3)}$ = Heterogeneity indices at levels two and three, respectively; k_{MA} = Number of included meta-analyses; k_{ES} = Number of effect sizes; $\bar{\bar{r}}$ = Weighted average correlation; $R^2_{(2;3)}$ = Variance explained within (level 2) and between (level 3) meta-analyses.

* $p < .05$.

Appendix F. *Methodological Quality and Reproducibility Indicators Fulfilled by the Included Meta-Analyses*

Meta-analysis	No. of yes	Inclusion criteria described	CI or SE reported	List of included studies	Moderator analysis conducted	Unpublished articles included	Het. index reported	Standardized search string	Het. accounted for	Publication bias assessed	Study description included	Study quality assessed	Double-screening and coding	Data available	List of excluded studies	Study quality considered	Syntax available
Ali et al. 2015	6	yes	yes	yes	no	yes	yes	no	no	no	yes	no	no	no	no	no	no
Allen et al. 2018	9	yes	yes	yes	yes	yes	yes	no	yes	yes	yes	no	no	no	no	no	no
Cherne 2008	9	yes	yes	yes	yes	yes	no	yes	no	no	no	yes	yes	no	no	yes	no
Chu et al. 2010	6	yes	yes	no	yes	yes	yes	yes	no	no	no	no	no	no	no	no	no
Cornelius-White 2007	5	no	yes	yes	yes	yes	no	no	no	no	no	yes	no	no	no	no	no
Endedijk et al. 2022	12	yes	yes	yes	yes	yes	yes	yes	yes	yes	no	yes	yes	yes	no	no	no
Givens Rolland 2012	7	no	yes	yes	yes	yes	yes	no	yes	no	no	yes	no	no	no	no	no
Kincade 2020	7	yes	yes	yes	no	yes	no	yes	no	yes	no	no	no	no	no	no	no
Korpershoek et al. 2016	7	yes	yes	yes	yes	no	yes	no	yes	no	no	no	yes	no	no	no	no
Krause & Smith 2022	12	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	no	no	no	no
Lei et al. 2016	6	yes	yes	yes	yes	no	yes	no	no	yes	no	no	no	no	no	no	no
Lei et al. 2018	8	yes	yes	yes	yes	no	yes	no	no	yes	yes	no	no	no	no	no	yes
Li et al. 2021	10	yes	yes	yes	yes	yes	yes	yes	yes	no	no	no	no	yes	no	no	yes
Moore et al. 2019	9	yes	yes	yes	no	no	no	yes	no	no	yes	yes	yes	no	yes	yes	no
Nurmi 2012	6	yes	yes	yes	no	no	no	yes	no	no	yes	no	no	no	yes	no	no
Roorda et al. 2011	7	yes	yes	yes	yes	yes	no	no	yes	yes	no	no	no	no	no	no	no
Roorda et al. 2014	3	yes	no	no	yes	yes	no	no	no	no	no	no	no	no	no	no	no
Roorda et al. 2017	6	yes	yes	yes	yes	no	yes	no	yes	no	no	no	no	no	no	no	no
Roorda et al. 2021	6	yes	yes	yes	yes	no	no	yes	no	yes	no	no	no	no	no	no	no
Strom & Bolster 2007	5	yes	no	yes	yes	yes	yes	no	no	no	no	no	no	no	no	no	no
Tao et al. 2022	9	yes	yes	yes	yes	no	yes	yes	yes	yes	yes	no	no	yes	no	no	no
Vandenbroucke et al. 2018	8	yes	yes	yes	yes	yes	no	yes	yes	no	yes	no	no	no	no	no	no
Wang et al. 2020	9	yes	yes	no	yes	no	yes	yes	yes	yes	no	yes	yes	no	no	no	no
Wilkinson 1980	3	yes	yes	yes	no	no	no	no	no	no	no	no	no	no	no	no	no
Number of Yes	$\bar{x}$ = 7,3	22	22	21	19	14	14	12	11	10	8	7	6	3	2	2	2
Percentage	45,6	91,7	91,7	87,5	79,2	58,3	58,3	50,0	45,8	41,7	33,3	29,2	25,0	12,5	8,3	8,3	8,3

Note. The methodological quality indicators were adapted from the AMSTAR rating scale by Shea et al (2007). The exact wording can be found in Figure 10 and in the Supplemental Material S18 with additional information on the coding. Number of Yes = Number of methodological quality and reproducibility indicators the meta-analysis met. Het. = Heterogeneity; CI or SE reported= Confidence intervals or standard errors were reported.