Z Erziehungswiss
https://doi.org/10.1007/s11618-022-01108-w

SCHWERPUNKT

Bidirectional effects between reading and mathematics development across secondary school

Timo Gnambs · Kathrin Lockl

Received: 29 November 2021 / Revised: 7 March 2022 / Accepted: 22 March 2022

Abstract Reading and mathematical competencies are important cognitive prerequisites for children's educational achievement and later success in society. An ongoing debate pertains to potential transfer effects between both domains and whether reading and mathematics influence each other over time. Therefore, the present study on $N = 5185$ students from the German *National Educational Panel Study* examined cross-lagged effects between reading and mathematics from Grades 5 to 12. The results revealed, depending on the chosen causal estimand, negligible to small bidirectional effects. Adopting a between-person perspective, students with higher mathematics scores at one point exhibited somewhat higher reading scores at the subsequent measurement. In contrast, when adopting a within-person perspective, both skills predicted longitudinal increases of the other skill in the lower grades but reversed effects in higher grades. Taken together, these findings not only demonstrate that transfer effects between reading and mathematics in secondary education tend to be small but also suggest different patterns of effects depending on the modeling choice.

Keywords Mathematics · Reading · Competencies · Skill transfer · Secondary school

This paper uses data from the National Educational Panel Study (NEPS; Blossfeld and Roßbach 2019). The NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi, Germany) in cooperation with a nationwide network. The study was not preregistered. The data and materials are available at NEPS Network (2020), while the computer code and analysis results are provided at https://osf.io/uejs7/. The authors are employed at the LIfBi. However, the institute had no involvement in the analysis of the data or the writing of the manuscript.

Dr. Timo Gnambs (✉) · Dr. Kathrin Lockl
Leibniz Institute for Educational Trajectories, Wilhelmsplatz 3, 96047 Bamberg, Germany
E-Mail: timo.gnambs@lifbi.de

Dr. Kathrin Lockl
E-Mail: kathrin.lockl@lifbi.de

Published online: 04 July 2022

1 Introduction

Literacy and numeracy skills are the foundation of children's educational careers and later-life economic success (Clements and Sarama 2011; Jordan et al. 2009; Spengler et al. 2018; Watts et al. 2018). Therefore, reading and mathematics represent core subjects in school from primary through secondary education. Numerous longitudinal studies have delineated developmental trajectories for both domains from different perspectives to identify important affordances and hindrances to children's acquisition of sufficient levels of reading and mathematics that allow them to excel in school and beyond (e.g., Little et al. 2021; Scammacca et al. 2020; Skopek and Passaretta 2021; Wicht et al. 2021). However, despite substantial correlations between both domains (Singer and Strasser 2017), an ongoing debate pertains to the co-development of reading and mathematics and their bidirectional influences on each other. While some studies suggested that proficient reading competencies might help children to develop their mathematical competencies (Duncan et al. 2007; Erbeli et al. 2021), results pointing in the opposite direction were less straightforward identifying only modest influences (Bailey et al. 2020; Grimm et al. 2021). To some degree, the heterogeneity in observed effects might be a result of different perspectives adopted in these studies (cf. Hamaker et al. 2015; Mund and Nestler 2019). While recent studies on bidirectional effects between reading and mathematics explicitly examined intraindividual processes to describe causal patterns of effects between both domains, earlier studies primarily focused on between-person mechanisms. Therefore, the present study will make use of recent methodological advancements in longitudinal and causal modeling (Hamaker et al. 2015; Lüdtke and Robitzsch 2021) to rigorously evaluate how one domain longitudinally affects the other domain in a representative sample of German students across a period of eight years. In contrast to previous research that primarily involved children in kindergarten or primary school (e.g., Bailey et al. 2020; Duncan et al. 2007; Erbeli et al. 2021; Hübner et al. 2021), the present study focuses on secondary school from Grades 5 to 12. Thereby, we hope to extend the available body of research to older age groups that have already acquired a basic understanding of reading and mathematics and are about to develop more complex literary and numeracy skills.

1.1 The co-development of reading and mathematics

Reading and mathematical competencies are moderately to highly correlated. A meta-analysis of more than 60 samples estimated a pooled cross-sectional correlation between reading and mathematics of 0.55 (Singer and Strasser 2017). Latent correlations that correct for measurement errors in both domains are often substantially larger and can even reach values exceeding 0.80 (e.g., Lechner et al. 2021b). Similarly, reading and mathematics disorders (i.e., dyslexia and dyscalculia) show pronounced comorbidity. Individuals with one disability have about two times greater chance of also having the other disorder (Joyner and Wagner 2020). Prevalent explanations for these associations typically refer to two major theoretical strains that address either common causes of reading and mathematics or skill transfers between the two domains.

1.1.1 Common causes as explanation

Referring to Cattell's (1987) investment theory, young children supposedly possess largely biologically determined general (fluid) cognitive abilities that set the stage for abstract reasoning and the rate of learning in different tasks. During an individual's life course, individual interests and environmental stimulations at home or in school lead to the acquisition of more complex (crystallized) abilities such as reading or mathematical competencies. These acquired abilities are closely tied to specific contexts and are not expected to affect other crystallized abilities. According to this view, the observed correlation between reading and mathematics is the result of common causes that affect both domains comparably. Indeed, empirical studies corroborated that fluid abilities are a leading predictor of changes in crystallized abilities (Ferrer and McArdle 2004). Consequently, early cognitive abilities such as phonological awareness (Cirino et al. 2018; Vanbinst et al. 2020), rapid automatized naming (Cirino et al. 2018; Korpipää et al. 2017), working memory (Peng et al. 2016, 2018), or reasoning (Peng et al. 2019) were also comparably associated with reading and mathematics achievement.

1.1.2 Skill transfers as explanation

An alternative conjecture that also guides the present research conceptualizes the observed correlation between reading and mathematics as a result of transfer effects between the two domains (see Erbeli et al. 2021, for an overview of specific theories). For example, functional theories of mathematics development (e.g., Dehaene and Cohen 1995; LeFevre et al. 2010) suggest that reading-related skills such as language and phonological processing shape the development of number concepts and support children's learning of mathematics. Thus, reading or, at least, reading-related abilities represent a means to understand mathematical problems (e.g., when mathematical concepts are taught in school settings). Following this line of reasoning, reading competencies should predict changes in mathematical competence during children's cognitive development. In contrast, developmental theories for reading argued for opposite effects with mathematics facilitating reading development. For example, Koponen et al. (2013) argued that fluent counting abilities might be fundamental for the development of visual-verbal associations in long-term memory, which are a precondition for later reading abilities.

Available empirical findings for these bidirectional effects between reading and mathematics are somewhat heterogeneous. Intervention studies that randomly assigned children to a training program for a specific skill (e.g., only reading or only mathematics) or a control group found either substantial (Glenberg et al. 2012), mixed (Sarama et al. 2012), or no pronounced (Fuchs et al. 2013) transfer effects on the other (untrained) domain. However, some of these studies have recently been criticized (Bailey et al. 2020) for exhibiting a rather limited construct validity because it was not clear whether pure reading or mathematical competencies were the focus of the administered trainings (or rather a blend of both, such as mathematical language). Thus, it is difficult to draw definite conclusions on bidirectional effects between reading and mathematics from the available experimental findings. In con-

trast, observational research that studied longitudinal effects of skill development for reading and mathematics found rather consistent support for transfer effects of reading on mathematics (e.g., Bailey et al. 2020; Cameron et al. 2019; Duncan et al. 2007; Erbeli et al. 2021; Grimm 2008; Grimm et al. 2021; Hübner et al., 2021; Purpura et al. 2011). However, the size of the respective effects varied considerably corresponding to correlations of about 0.06 (Bailey et al. 2020; Grimm 2008), 0.20 (Cameron et al. 2019), or even 0.30 (Purpura et al. 2011). Reverse transfer effects with mathematical competencies predicting changes in reading competencies were less consistent with some studies reporting stronger transfer effects (Duncan et al. 2007), comparable transfer effects (Cameron et al. 2019), substantially smaller transfer effects (Bailey et al. 2020), or even no transfer effects at all (Erbeli et al. 2021). Taken together, these findings provide partial support for bidirectional effects between reading and mathematics, albeit with more consistent support for the effects of reading on mathematics than for the other direction.

1.1.3 A developmental perspective

When considering possible bidirectional effects between reading and mathematics, it seems important to have a more differentiated view, depending on which stage in the course of development is being studied. The nature of the developmental association between competence domains can change over time (as it was shown, for example, for the relation between working memory and vocabulary; Gathercole et al. 1992). Moreover, different indicators may be used to capture the constructs in different stages of development (such as vocabulary knowledge in early childhood versus reading comprehension in adolescence). So far, the available research has studied bidirectional effects between reading and mathematics mostly in younger children, that is, in kindergarten and primary school (Bailey et al. 2020; Cameron et al. 2019; Duncan et al. 2007; Koponen et al. 2013; Vanbinst et al. 2020). In these studies, especially when kindergarten children were included, often precursor skills were measured to assess reading or mathematical competencies. For instance, print familiarity, letter recognition, beginning and ending sounds, or vocabulary knowledge were used as indicators of emergent reading competencies, whereas number knowledge, counting, number sequence, or basic addition and subtraction problems were used as indicators of emergent mathematical competencies (Bailey et al. 2020; Duncan et al. 2007; Vanbinst et al. 2020). Studies investigating the relation between reading and mathematical competencies in the school entry phase found a large overlap between both domains (Bailey et al. 2020; Vanbinst et al. 2020) with a correlation above 0.90 between stable latent factors of reading and mathematics (Bailey et al. 2020). An explanation for the high correlation between reading and mathematical competencies at this early stage might be that both rely on phonological codes or an individual's sensitivity to the sound structure of oral language (De Smedt et al. 2010; Koponen et al. 2013; Vanbinst et al. 2020; Vukovic and Lesaux 2013). However, the nature of the relation between reading and mathematics may change when children grow older and the reasons why both domains influence each other might be different ones in secondary school.

So far, only a few studies addressed the relationship between reading and mathematics in secondary school (Björn et al. 2016; Harlaar et al. 2012; Kyttälä and Björn 2014). In comparison to primary school, mathematical tasks become more complex later on. Thus, in secondary education students should develop a more sophisticated mathematical understanding as well as the ability to apply this understanding to solving problems associated with them (Neumann et al. 2013). Often, mathematical word problems are used in the educational context (Riley and Greeno 1988) which require reading the mathematical story or the problem presented before writing down the mathematical operations and then solving the problem. Therefore, it might seem obvious that the ability to solve these tasks relies on reading comprehension and other literacy skills, such as word decoding or vocabulary (Björn et al. 2016; Kyttälä and Björn 2014). In line with this view, Björn et al. (2016) showed that text comprehension at the end of primary school predicted mathematical word problem performance in secondary school. Interestingly, this was true even after controlling for text-reading fluency. In a study with 12 year-olds, Harlaar et al. (2012) also found a genetic overlap between reading comprehension and mathematics, with higher correlations between reading comprehension and mathematics compared to those between word decoding and mathematics. Based on these findings it could be assumed that reading comprehension may be stronger related to mathematics in secondary school than in primary school when children still have to gain experience in decoding or basic reading skills (Gough et al. 1996). Overall, there is some evidence that reading comprehension and mathematics are related in the secondary school years. However, only little is known about the bidirectional relationship during this period and whether the influence of one domain on the other one changes over time.

1.2 Methodological considerations for the analysis of bidirectional effects

Bidirectional effects between reading and mathematics are typically examined using autoregressive models that predict performance in one domain from its initial performance and performance in the other domain (e.g., Cameron et al. 2019; Duncan et al. 2007; Purpura et al. 2011). Analyses that simultaneously examine respective effects for both domains together can be conveniently specified as cross-lagged panel models (CLPM; see the left panel in Fig. 1) in a structural equation modeling (SEM) framework. In these analyses, the lagged effects of, for example, reading on mathematics are interpreted as causal transfer effects. However, this view has been increasingly criticized for representing spurious effects resulting from stable confounds that affect both domains comparably (Bailey et al. 2020; Berry and Willoughby 2017; Lucas 2022). Rather unchanging (or, at least, slowly changing) environmental or personal factors across the observational period such as children's basic cognitive abilities or learning contexts that are unique to each child might simultaneously influence both reading and mathematics. In this case, meaningful causal conclusions, for example, about the impact of reading on mathematics development are infeasible (Berry and Willoughby 2017). Therefore, the random-intercept cross-lagged panel model (RI-CLPM; Hamaker et al. 2015) has been suggested as an intriguing alternative that additionally accounts for stable between-person differences (see the

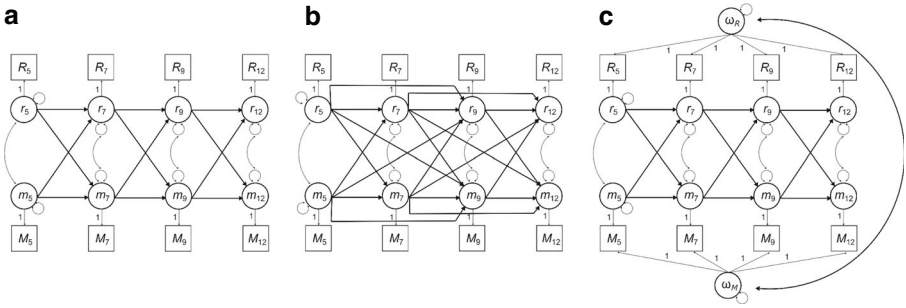

Fig. 1 Path Diagrams for Different Cross-Lagged Panel Models for Reading and Mathematics across Eight Grades. **a** *CLPM* Cross-lagged panel model with lag 1, **b** *CLPM-L2* Cross-lagged panel model with lag 2, **c** *RI-CLPM* Random-intercept cross-lagged panel model, *R* Reading, *M* Math. Squares denote observed variables and circles indicate unobserved variables. The mean structure is not presented

right panel in Fig. 1) and, thus, can overcome the limitations of the CLPM (see also Mund and Nestler 2019 or Grimm et al. 2021). In the RI-CLPM, the competence scores of each person are implicitly ipsatized (i.e., centered within persons). The bidirectional analyses are then based on the residualized scores (net of the stable between-person variances) and not the total scores as in the CLPM. Consequently, in the RI-CLPM the cross-lagged effects indicate whether the deviation of a person's average competence at one point has a prospective effect on the change of the within-person deviation of the other competence. Thus, when a student has a higher reading (or mathematical) competence as compared to his or her usual competence level in one grade, a positive cross-lagged effect would indicate that this student is about to exhibit a subsequent increase in his or her mathematical (or reading) competence in the following. Thus, the RI-CLPM allows for more unambiguous tests of bidirectional hypotheses between different domains.

A major drawback of the RI-CLPM is that it targets a different causal effect as compared to the CLPM. While in the CLPM bidirectional effects try to explain between-person differences (i.e., differential change; Asendorpf 2021), the RI-CLPM captures within-person effects, that is, temporary fluctuations around a person's mean (Lüdtke and Robitzsch 2021). Thus, in contrast to the interpretation of the RI-CLPM outlined above, cross-lagged effects in the CLPM indicate whether students with, for example, a higher reading competence at one point (as compared to other students) are expected to exhibit higher mathematical competence at a subsequent time. To properly examine these between-person effects, Lüdtke and Robitzsch (2021) recently suggested analyzing the CLPM from a causal inference perspective (Pearl et al. 2016) and accounting for potential confounding factors that might impede causal interpretations of bidirectional effects. Importantly, they showed that controlling for the effect from the *two* prior measurement occasions (instead of only *one* as in the CLPM) can help control for many unobserved confounders (see also VanderWeele et al. 2020). Thus, although the RI-CLPM and the CLPM with autoregressive lag 2 effects (CLPM-L2; see the middle panel in Fig. 1) both try to control for stable environmental or personal factors that might affect both domains (either using a latent trait factor or the inclusion of autoregressive effects),

the interpretations of the resulting cross-lagged effects substantially diverge. Which of the two effect interpretations is of greater interest in a given situation depends on the specific research question that one wants to address.

1.3 The present study

Although developmental trajectories of reading and mathematical competencies have been subject to intense longitudinal research in Germany (e.g., Skopek and Passaretta 2021; Wicht et al. 2021) and also internationally (e.g., Little et al. 2021; Scammacca et al. 2020), controversies remain regarding their bidirectional influences over time (e.g., Bailey et al. 2020; Cameron et al. 2019; Duncan et al. 2007; Grimm et al. 2021). Recent methodological research (e.g., Berry and Willoughby 2017; Hamaker et al. 2015; Lüdtke and Robitzsch 2021) suggested that the inconsistent findings in previous research might be partly a consequence of biasing influences from unmodeled confounders that prevented the identification of causal pathways. Therefore, the present study on bidirectional effects between reading and mathematical competencies contrasts two different methodological approaches that overcome this weakness, that is, the RI-CLPM and the CLPM-L2, to study how one domain might predict changes in the other domain. Moreover, prior research on bidirectional effects between reading and mathematics almost exclusively relied on manifest point estimates (e.g., sum scores) although most psychological measures exhibit a less than perfect reliability (e.g., Gnambs 2014, 2015). Because uncorrected measurement errors can seriously distort the validity of path models (Cole and Preacher 2014), we will adopt a latent variable approach to evaluate transfer effects between the two domains on the true score level (see also Mulder and Hamaker 2021).

So far, research on the connection between literacy and numeracy skills mainly concentrated on younger children from kindergarten to primary school (e.g., Bailey et al. 2020; Cameron et al. 2019; Koponen et al. 2013; Vanbinst et al. 2020). Moreover, the focus in these studies has often been placed on the links between text-reading fluency (or word decoding) and mathematical skills. Thus, the existing empirical evidence on bidirectional effects between reading and mathematics is often limited to indicators of early reading skills in primary school. Less is known about the relationship between reading competence and mathematical competence in older age groups (Björn et al. 2016; Harlaar et al. 2012), even though it can be assumed that the role of text comprehension becomes more important for solving complex mathematical tasks (especially involving verbose problem statements) when students are in secondary school. Therefore, we will focus on students in lower and upper secondary education across Germany and study cross-lagged effects of reading and mathematical competencies from Grades 5 to 12.

2 Method

2.1 Sample and procedure

The German *National Educational Panel Study* (NEPS; Blossfeld and Roßbach 2019) follows multiple cohorts of children, adolescents, and adults across their life courses. The present study focuses on Starting Cohort 3 of the NEPS that comprises a representative sample of five graders from different secondary schools across Germany. The sample was drawn using a stratified multistage sampling design as detailed in Steinhauer et al. (2015): First, a random sample of schools at the secondary level offering education in fifth grade was selected that was stratified according to the major school types in Germany. Then, in each school, all students from two randomly drawn classes for whom parental consent could be obtained were eligible to participate. This resulted in a sample of $N = 5185$ students (48% girls) from 234 different secondary schools that were administered the focal competence tests in Grade 5. Their mean age was 10.93 years ($SD = 0.52$) and about 19% of them had a migration background. About 19%, 21%, and 44% of the students attended schools with lower (*Hauptschule*), intermediate (*Realschule*), or upper (*Gymnasium*) secondary education, respectively. The remaining students were enrolled in various specialized school types (e.g., *Gesamtschule*). Follow-up competence assessments in the studied domains were conducted in Grades 7, 9, and 12. The achievement tests were administered in small groups by trained test administrators at the respective schools. In Grade 12, students that left their original school after Grade 9 or chose

Table 1 Sample Characteristics across Measurement Occasions

	Grade 5	Grade 7	Grade 9	Grade 12
Sample size	5185	3830	3266	3241
Percentage non-response	0.0%	26.1%	37.0%	37.5%
Number of schools	234	191	180	78 [c]
Percentage female	48.2%	48.4%	49.1%	49.9%
Mean age (*SD*)	10.9 (0.5)	12.9 (0.5)	14.9 (0.5)	17.9 (0.5)
Percentage migration	19.0%	17.9%	16.6%	15.9%
Mean socio-economic status (*SD*) [a]	55.6 (20.3)	56.5 (20.2)	56.8 (20.0)	57.7 (19.8)
Mean reading competencies (*SD*)	0.0 (1.0)	0.1 (1.0)	0.1 (1.0)	0.2 (1.0)
Mean mathematical competencies (*SD*)	0.0 (1.0)	0.1 (1.0)	0.1 (1.0)	0.2 (1.0)
Mean reasoning abilities (*SD*) [b]	0.0 (1.0)	0.1 (1.0)	0.1 (1.0)	0.1 (1.0)

All reported statistics for reading, mathematics, and reasoning abilities refer to the *z*-standardized scores obtained in Grade 5

[a] Highest parental international socio-economic index of occupational status (Ganzeboom 2010)

[b] Measured with 12 items from Lang et al. (2014)

[c] Only 39.2% of the sample were examined in regular schools (i.e., *Gymnasium*), while most respondents were tested individually outside school

another educational path such as vocational training were tracked and individually tested at their private homes (about 61% of the sample)[1].

Basic information on the sample at the four measurement occasions is summarized in Table 1. We observed pronounced nonresponse rates across grades that ranged between 26% in Grade 7 and 38% in Grade 12[2]. However, the descriptive information in Table 1 did not suggest pronounced selection effects. Although students with migration backgrounds or lower socio-economic status (as measured by the highest parental international socio-economic index of occupational status; Ganzeboom 2010) had a higher propensity for nonresponse in Grade 12, the respective effects were small. Importantly, nonresponse was only weakly associated with reading and mathematical competencies or general cognitive functioning measured in Grade 5. Thus, nonresponse did not introduce a substantial bias in the sample composition across measurement occasions. Detailed attrition analyses are summarized in the supplemental material.

2.2 Instruments

Reading and mathematical competencies were measured with paper-based achievement tests that were specifically constructed for administration in the NEPS. The construction rationales of these tests were linked to established frameworks of other large-scale assessments such as the *Programme for International Student Assessment* (OECD 2017). Thus, they were not curricular (i.e., tied to specific school subjects) but adhered to the literacy concept that focuses on the relevance of competencies for successful participation in society (Weinert et al. 2019). For all tests, a comparable scaling procedure was adopted (see Pohl and Carstensen 2013) that provided unidimensional proficiency scores. To allow for meaningful mean-level comparisons between grades, the tests were linked to a common scale following Fischer et al. (2016, 2019).

Reading competencies were measured with different tests that included 32 items in Grade 5 (Pohl et al. 2012), 40 items in Grade 7 (Krannich et al. 2017), 46 items in Grade 9 (Scharl et al. 2017), and 29 items in Grade 12 (Kutscher and Scharl 2020). To allow for a better test targeting and greater measurement precision, the tests administered in Grades 7, 9, and 12 adhered to a branched testing design (Pohl 2013) that assigned different booklets including either easier or more difficult items to the students depending on their performance in the previous assessment wave. The items used different response formats including multiple-choice or matching tasks. All reading tests followed a common construction framework (see Gehrer

[1] At the last measurement occasion, most participants were assessed individually at home because they left their original school or followed another educational path (e.g., vocational training). Even though only a subsample of respondents attended regular schools (i.e., *Gymnasium*), for convenience, we refer to the last measurement occasion as 'Grade 12'. However, it must be emphasized that the sample was composed of a mixture of respondents from school and non-school settings.

[2] Due to time constraints in the individual assessment at home, a randomly assigned rotation design was applied in Grade 12 which resulted in a reduced number of students with data in the reading and mathematical competence tests. However, because this data was missing completely at random, the nonresponse did not introduce systematic bias in the analyses results.

et al. 2003) and included five different text types (i.e., information, instruction, advertising, commenting, and literary texts). Moreover, the items addressed three different cognitive requirements (i.e., finding information in the text, drawing text-related conclusions, and reflecting and assessing). To prevent memory effects, the tests administered in the different grades did not share common items. Rather, the tests were linked across measurement occasions using an anchor-group design that relied on independent link samples (see Fischer et al. 2016). The proficiency scores resulted in satisfactory marginal reliabilities of 0.77, 0.83, 0.81, and 0.80 in Grades 5, 7, 9, and 12, respectively.

The four tests of *mathematical competencies* included 25 items in Grade 5 (Duchhardt and Gerdes 2012), 23 items in Grade 7 (Schnittjer and Gerken 2017), 34 items in Grade 9 (Van den Ham et al. 2018), and 30 items in Grade 12 (Petersen et al. 2020). Again, the tests administered in Grades 9 and 12 followed the logic of branched testing (Pohl 2013) and included different booklets with either easier or more difficult items. The items used different response formats including multiple-choice and short constructed responses. All mathematics tests followed a common construction rationale (see Neumann et al. 2013) that specified five different content areas (i.e., quantity, space and shape, change and relationship, and data and chance) as well as six cognitive components that were required for a successful task solution. The tests administered in successive grades shared some common items. Therefore, the mathematics tests were linked across grades using an anchor-item design (cf. Fischer et al. 2016). The marginal reliabilities of the proficiency scores in the four grades were good with 0.80, 0.76, 0.81, and 0.77, respectively.

Measured *confounders* included the students' self-reported sex (0 = male, 1 = female) and socio-economic status as reflected by the highest parental international socio-economic index of occupational status (Ganzeboom 2010). Moreover, a Raven-like matrices test (Lang et al. 2014; see also Gnambs and Nusser 2019) administered in Grade 5 was used as an indicator of general cognitive functioning. The 12 items required the identification of a logical rule that completed a figural sequence. Despite its rather short length, the sum score exhibited a good categorical ω reliability of 0.71. Finally, we acknowledged three dummy-coded indicators representing the school type with upper secondary education as the reference category.

2.3 Analysis plan

The bidirectional effects between reading and mathematics were examined using the CLPM and its extensions, the RI-CLPM (Hamaker et al. 2015) and CLPM-L2 (Lüdtke and Robitzsch 2021). The SEMs were estimated in *R* version 4.1.2 (R Core Team 2021) with *lavaan* version 0.6–9 (Rosseel 2012) and *semTools* version 0.5–5 (Jorgensen et al. 2021) using the Yuan and Bentler (2000) test statistic and cluster-robust standard errors (Savalei 2014) that account for the nesting of students within different schools. These analyses modeled latent variables with 30 plausible values estimated with *NEPSscaling* version 2.2.0 (Scharl et al. 2020). For the estimation of plausible values, a large number of variables was specified as background model to increase their precision (cf. Lechner et al. 2021a; Lüdtke and Robitzsch 2017).

Following Weirich et al. (2014), missing values in the background variables were imputed 30 times with classification and regression trees (Burgette and Reiter 2010). Details on the estimation of the plausible values are given in the supplemental material. To account for the nonresponse across measurement occasions, missing values were imputed 30 times. Plausible values were imputed based on the background model in *NEPSscaling* (Scharl et al. 2020) while missing covariates were imputed in *mice* version 3.13.0 (Van Buuren and Groothuis-Oudshoorn 2011). The results of the multiply imputed SEMs were combined following Rubin's (1987) rules.

The competence scores were z-standardized with respect to the means and standard deviations in Grade 5. Thus, all regression parameters refer to the standardized scale (with $M = 0$ and $SD = 1$) of the initial competence assessment. In addition, we also report the percentage of variance explained by these effects in the total reading and mathematics scores[3]. Sensitivity analyses evaluated the stability of the observed results when (not) controlling for four measured confounders.

The goodness of fits of the estimated models were evaluated using three practical fit indices including the robust comparative fit index (CFI; Brosseau-Liard and Savalei 2014), the robust root mean squared error of approximation (RMSEA; Brosseau-Liard et al. 2012), and the standardized root mean squared residual (SRMR). In line with conventional standards (e.g., Schermelleh-Engel et al. 2003), we viewed models with $\text{CFI} \geq 0.95$, $\text{RMSEA} \leq 0.08$, and $\text{SRMR} \leq 0.10$ as "acceptable", while models with $\text{CFI} \geq 0.97$, $\text{RMSEA} \leq 0.05$, or $\text{SRMR} \leq 0.05$ were considered as "good" fitting. Model comparisons were based on differences in the practical fit indices and the Bayesian information criterion (BIC; Schwarz 1978) for which lower values indicate a better fit. Moreover, we also report the results of log-likelihood difference tests between nested models. However, given the large power to identify even negligible effects in the present sample, we give less weight to these results for our interpretations.

2.4 Benchmarks for effect interpretations

Empirical effect size distributions in various psychological fields (e.g., Bosco et al. 2015; Gignac and Szodorai 2016) typically show a median correlation of $r = 0.20$ with an interquartile range of about 0.10 to 0.30. Moreover, a recent meta-analysis on the relationship between fluid intelligence and domain-specific competencies (Peng et al. 2019) showed that intelligence predicted later reading or mathematics performance while partialing out initial performance at $r = 0.17$ to 0.24. Therefore, we will also consider cross-lagged effects that explain about 4% of the variance in the outcome as medium effects, whereas effects explaining less than 1% of variance will be considered small. However, given that competence development is likely to be determined by a multitude of different causes, it has been recently argued that

[3] We refrain from reporting standardized regression coefficients because these are not readily comparable between the CLPM-L2 and the RI-CLPM. Autoregressive and cross-lagged effects in the CLPM-L2 predict the mathematics and reading scores, while in the RI-CLPM they are predicting the residuals (after accounting for stable between-person differences) whose variances are substantially smaller. Therefore, we report the percentage of variance explained in the total scores.

modest effects should be expected to represent the norm in reproducible cumulative science (Götz et al. 2022). Thus, in longitudinal studies, even small bivariate effects controlling for stability (i.e., autoregressive) effects might be viewed as meaningful due to accumulating long-term effects (Adachi and Willoughby 2015).

2.5 Transparency and openness

The study material, detailed information on the assessment procedure, and the analyzed raw data are available to the research community at NEPS Network (2020). Moreover, we provide the computer code including the analysis output that was used to derive the reported findings at https://osf.io/uejs7/.

3 Results

3.1 Descriptive analyses

The density distributions of the latent reading and mathematics scores at the four measurement occasions in Fig. 2 show substantial interindividual differences in both domains and each grade. As expected, the modes of these distributions were sequentially ordered across time indicating that, on average, competencies increased throughout secondary school. However, the increase in mathematical competencies seemed slightly stronger ($d = 2.11$) as compared to the increase in reading competencies ($d = 1.62$; see also Table 2). Reading and mathematics were substantially correlated, cross-sectionally as well as longitudinally (see Table 2). The latent scores correlated within grades between $r = 0.59$ and 0.77. The respective correlations showed a systematic trend across grades and gradually reduced across subsequent measure-

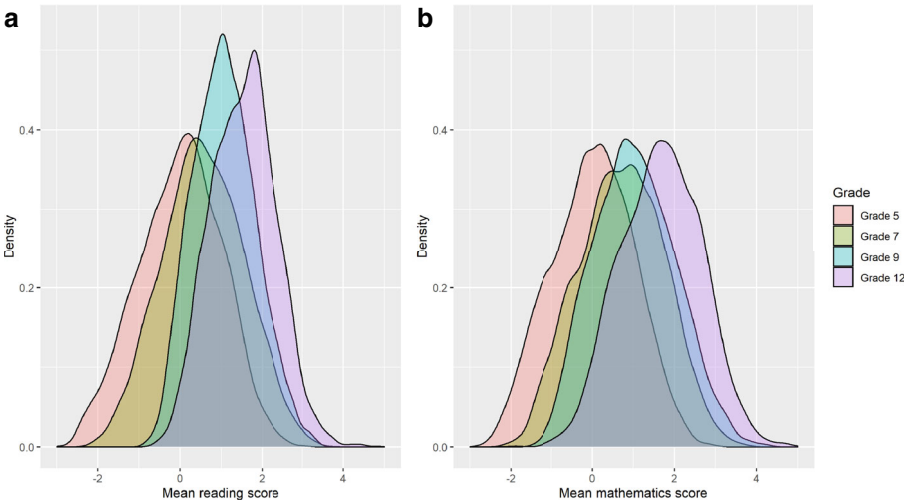

Fig. 2 Average Score Distributions for Reading and Mathematics. Competence scores based on averaged plausible values were z-standardized with respect to the mean and standard deviation in Grade 5

Table 2 Means, Standard Deviations, and Correlations for Reading and Mathematics across Measurement Occasions

				Reading			Mathematics			
	M	*SD*	MV	Grade 7	Grade 9	Grade 12	Grade 5	Grade 7	Grade 9	Grade 12
Reading										
Grade 5	0.00	1.00	0.0%	0.74	0.72	0.68	0.77	0.72	0.70	0.61
Grade 7	0.57	1.08	26.3%	–	0.74	0.71	0.69	0.71	0.68	0.59
Grade 9	1.08	0.84	42.6%	–	–	0.69	0.68	0.67	0.69	0.58
Grade 12	1.64	0.89	51.6%	–	–	–	0.64	0.64	0.63	0.59
Mathematics										
Grade 5	0.00	1.00	0.0%	–	–	–	–	0.89	0.85	0.78
Grade 7	0.70	1.08	26.1%	–	–	–	–	–	0.86	0.79
Grade 9	1.50	1.04	39.9%	–	–	–	–	–	–	0.79
Grade 12	2.11	1.06	49.1%	–	–	–	–	–	–	–

Based upon 30 plausible values. *MV* Percentage of missing values. All correlations are significant at $p < 0.001$

ment waves. The cross-domain correlations across grades were only moderately smaller and fell between 0.58 and 0.72. Thus, descriptively reading and mathematics exhibited substantial bivariate associations within grades and also across grades. However, these correlations do not inform about causal effects between the two domains because unacknowledged confounders might have affected reading and mathematics comparably. Therefore, three bidirectional autoregressive models (see Fig. 1) were estimated to study the longitudinal cross-domain associations between reading and mathematics in greater detail.

3.2 Stable between-Person differences in reading and mathematics

The CLPM (left panel in Fig. 1) showed a rather modest fit with a CFI of 0.92, an RMSEA of 0.13, and an SRMR of 0.06. In contrast, the CLPM-L2 (middle panel in Fig. 2) and the RI-CLPM (right panel in Fig. 1) both exhibited superior fits to the data with CFIs of 0.98, RMSEAs of 0.09 and 0.08, and SRMRs of 0.02 and 0.05, respectively. Model comparisons (see Table 3) emphasized significantly ($p < 0.05$) better fits for the CLPM-L2 (BIC = 77287) and RI-CLPM (BIC = 77287) as compared to the CLPM (BIC = 79634). In contrast, the goodness of fit indicators did not prefer one of the two extensions of the CLPM over the other. Rather, both highlighted the importance of acknowledging stable between-person differences when analyzing bidirectional effects in reading and mathematical competencies.

The CLPM-L2 showed substantial stability between grades for both domains (all $ps < 0.001$). Interestingly, the autoregressive lag 2 effects for reading and mathematics, $Mdn\ (B) = 0.27$ and $Mdn(B) = 0.42$, were only marginally smaller than the respective first-order effects, $Mdn(B) = 0.31$ and $Mdn(B) = 0.44$. Thus, students with higher competencies at a given point also had higher competencies at the two previous time points as compared to other students. Similarly, the random intercepts for reading and mathematics in the RI-CLPM exhibited substantial variances of 0.61, $p < 0.001$, and 0.81, $p < 0.001$, respectively, which indicate pronounced interindivid-

Table 3 Fit Statistics for Different Cross-Lagged Panel Models

	Model	Number of parameters	Degrees of freedom	Chi-squared [a]	*p*	Scaling correction	CFI	RMSEA	SRMR	BIC
Basic models										
1	CLPM	29	15	888.25	<0.001	1.52	0.92	0.13	0.06	79634
2	CLPM-L2	37	7	196.36	<0.001	1.63	0.98	0.09	0.02	77287
	Difference Model 1 – Model 2 [b]	–	8	708.83	<0.001	1.42	–	–	–	–
3	RI-CLPM	32	12	210.11	<0.001	1.79	0.98	0.08	0.06	77287
	Difference Model 1 – Model 3 [b]	–	3	1480.32	<0.001	0.45	–	–	–	–
Controlling for observed confounders										
4	CLPM	77	15	758.19	<0.001	1.47	0.95	0.12	0.03	73661
5	CLPM-L2	85	7	182.97	<0.001	1.63	0.99	0.09	0.01	71788
	Difference Model 4 – Model 5 [b]	–	8	599.50	<0.001	1.33	–	–	–	–
6	RI-CLPM	80	12	196.95	<0.001	1.52	0.99	0.07	0.02	71695
	Difference Model 4 – Model 6 [b]	–	3	452.30	<0.001	1.25	–	–	–	–

CLPM Cross-lagged panels model with lag 1; *CLPM-L2* Cross-lagged panel model with lag 2; *RI-CLPM* Random intercept cross-lagged panel model; *CFI* robust comparative fit index; *RMSEA* robust root mean squared error of approximation; *SRMR* standardized root mean squared residual; *BIC* Bayesian information criterion

[a] Yuan and Bentler (2000) test statistic

[b] Likelihood-ratio test of differences in model fit (Meng and Rubin 1992). All models are based upon 30 plausible values and include stationary constraints for the third and fourth measurement occasion

ual differences between students in both domains. The two random effects correlated at $r = 0.84$, $p < 0.001$, suggesting that students with higher reading competencies had, on average, also higher mathematics competencies.

3.3 Cross-Lagged effects for reading and mathematics

The structural effects of the different autoregressive models are summarized in Table 4. Model comparisons did not support equality constraints for the autoregressive effects or the cross-lagged effects between the different grades. Thus, both dynamic effects changed across the measurement occasions. The CLPM-L2 showed significant ($p < 0.05$) cross-lagged effects for mathematics on reading across grades. However, these cross-lagged effects explained only 1.1% of the variance in reading scores in Grade 12 and even less in Grade 9. The respective effect of mathematics measured in Grade 5 predicting reading in Grade 7 cannot be readily compared to the former effects because they did not include lag effects of the second order. For the other direction, the respective effects were smaller and in Grade 12 not even significant ($p = 0.536$). Thus, reading explained less than 1% in the variance of the observed mathematics scores in Grades 9 and 12. Together, these results suggest that students with higher mathematics scores in Grade 7 or Grade 9 exhibited somewhat higher reading scores at the subsequent measurement point as compared to students with lower mathematics scores.

In contrast to the CLPM-L2, the cross-lagged effects in the RI-CLPM capture only the within-person dynamics of competence development after accounting for the stable traits. The cross-lagged effects showed that when students had higher reading (or mathematics) competencies (as compared to his or her average competence) in Grade 5, they were about to exhibit a subsequent *increase* in the other ability in Grade 7. However, the respective effects were small and explained 2.8% and 1.3% of the variance in the total scores. Across the course of secondary school, these effects gradually reversed. Thus, when individuals had higher reading (or mathematics) competencies as compared to their average competencies in Grade 9, they were expected to have a substantial *decline* in their temporary mathematics (or reading) scores in Grade 12. These effects were substantially larger and explained about 3.2% and 6.6% of the variance in total scores. Together, these results suggest that students with higher than usual competence scores in one domain exhibited substantially lower deviations from their average competence level in the other domain.

3.4 Controlling for observed confounders

Causal inference requires that the analyses controlled for all relevant confounders. Therefore, we repeated the previous analyses using the students' sex, socio-economic status, general cognitive functioning, and the school type as covariates in the analysis models. Thus, we regressed the latent variables for reading and mathematics at each measurement occasion on these covariates. These analyses replicated the previously reported results in large part. The respective results are summarized in the supplemental material. Again, the CPLM exhibited an inferior model fit as compared to the CLPM-L2 and the RI-CLPM (see Table 3). The CLPM-L2 showed small to

Table 4 Structural Coefficients for Different Cross-lagged Panel Models

	CLPM		CLPM-L2		RI-CLPM	
	B (*SE*)	*%Var*	*B* (*SE*)	*%Var*	*B* (*SE*)	*%Var*
Autoregressive effects for reading						
Grades 5 → 7	0.56 (0.02)****	26.8%	0.56 (0.02)****	26.8%	0.12 (0.04)***	0.6%
Grades 7 → 9	0.42 (0.02)****	26.8%	0.31 (0.02)****	14.5%	0.01 (0.01)	0.0%
Grades 9 → 12	0.51 (0.03)****	26.4%	0.30 (0.02)****	9.0%	–0.81 (0.16)****	9.9%
Autoregressive effects for mathematics						
Grades 5 → 7	0.87 (0.02)****	64.9%	0.87 (0.02)****	64.9%	0.34 (0.05)****	2.8%
Grades 7 → 9	0.73 (0.03)****	52.2%	0.44 (0.04)****	19.3%	0.25 (0.05)****	1.5%
Grades 9 → 12	0.77 (0.02)****	62.3%	0.43 (0.03)****	19.7%	–0.07 (0.16)	0.1%
Cross-lagged effects for mathematics on reading						
Grades 5 → 7	0.32 (0.03)****	8.8%	0.32 (0.02)****	8.8%	0.34 (0.07)****	2.8%
Grades 7 → 9	0.22 (0.02)****	7.5%	0.07 (0.03)**	0.7%	–0.12 (0.03)****	0.3%
Grades 9 → 12	0.26 (0.02)****	10.7%	0.08 (0.03)***	1.1%	–0.33 (0.12)***	3.2%
Cross-lagged effects for reading on mathematics						
Grades 5 → 7	0.11 (0.02)****	0.9%	0.11 (0.02)****	0.9%	0.17 (0.03)****	1.3%
Grades 7 → 9	0.14 (0.02)****	1.8%	0.09 (0.02)****	0.7%	0.05 (0.03)*	0.1%
Grades 9 → 12	0.07 (0.02)***	0.3%	0.02 (0.02)	0.0%	–0.70 (0.14)****	6.6%

CLPM Cross-lagged panel model with lag 1; *CLPM-L2* Cross-lagged panel model with lag 2; *RI-CLPM* Random-intercept cross-lagged panel model. *B* Regression coefficient (with standard error in parenthesis); *%Var* Percentage of variance explained in total scores. All models are based upon 30 plausible values and include stationary constraints for the third and fourth measurement occasion. Lag 2 effects of CLPM-L2 are not presented

**** $p < 0.001$, *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

negligible cross-lagged effects of mathematics on reading (0.5% to 1.6% explained variance) and negligible cross-lagged effects of reading on mathematics (explaining less than 1% of variance). In contrast, the RI-CLPM identified positive cross-lagged effects of both domains on the other domain in Grade 7 and a reversed pattern in Grade 12. As compared to the previous analyses, these effects explained a larger amount of variance in Grade 7 (6.6% and 2.0%) and less variance in Grade 12 (0.8% and 4.6%). Thus, it might be speculated that the latter cross-lagged effects are partially spurious results from unacknowledged confounders.

4 Discussion

The main goal of the study was to investigate potential transfer effects between reading and mathematics over the course of lower and upper secondary school. Based on recent discussions (e.g., Asendorpf 2021; Lüdtke and Robitzsch 2021; Orth et al. 2021), different methodological approaches were used to describe how one domain might predict changes in the other domain. These analyses led to several intriguing findings.

First of all, the data of our study confirmed previous findings in showing that reading as well as mathematical competencies—at least on average—steadily increased throughout secondary school (e.g., Rescorla and Rosenthal 2004; Shin et al. 2013). Moreover, the results showed considerable stability of interindividual differences in reading and mathematical competencies over time, albeit the stability was even higher for mathematics than for reading. Consistent with the existent literature (e.g., Adelson et al. 2015; Shin et al. 2013), reading and mathematical competencies were substantially correlated, with the cross-sectional correlations gradually becoming somewhat smaller over time. Combined with the results of the meta-analysis by Singer and Strasser (2017) who showed that the correlations between both domains were independent of age but included only students up to middle school age, the findings might suggest that the association between reading and mathematics tends to become less strong only later, towards the end of secondary school.

To examine the longitudinal cross-domain associations between reading and mathematics three bidirectional autoregressive models were estimated. Whereas the CLPM and the CLPM-L2 adopted a between-person perspective, the RI-CLPM is based on a within-person perspective and indicates whether the deviation of a person's average competence at a given time point influences the change of the subsequent within-person deviation of the other competence. A comparison of the model fits revealed significantly better fits for the CLPM-L2 and RI-CLPM as compared to the CLPM. However, with regard to model fit, none of the two extensions of the CLPM was superior to the other. These findings suggest that both models assuming stable between-person differences provided a more appropriate description of the developmental processes than the CLPM. In this context, it has been argued that differences in model fit do not necessarily indicate which of the models is better from a theoretical perspective (Lüdtke and Robitzsch 2021; Orth et al. 2021).

If we rely on the results of the CLPM which has been most commonly used for decades (Orth et al. 2021), the interpretation would be that students with higher

mathematical competencies (relative to others) will show a subsequent rank-order increase in reading compared to individuals with lower mathematical competencies. Even though corresponding relationships were found for the opposite direction, the size of the bidirectional effects proved to be higher for mathematics on reading than for reading on mathematics. Thus, based on the CLPM, the results confirm findings by Duncan et al. (2007) concerning the predictive power of early mathematical skills in young children and suggest that there might be a transfer effect from mathematics to reading (Koponen et al. 2013). Moreover, they contradict the assumption that reading comprehension becomes more important for solving complex mathematical tasks in secondary school (Björn et al. 2016; Harlaar et al. 2012). However, an alternative interpretation could be that students' mathematical competencies are more reflective than reading competencies of the factors that influence students' learning in general throughout school (Bailey et al. 2020). That is, confounding could have occurred when mathematical competencies were related more strongly to possible common causes such as different facets of children's cognitive abilities or environmental factors than reading competencies. In our study, we controlled for gender, reasoning abilities, school track, and socioeconomic status as possible confounders. Potentially, however, the inclusion of these control variables was not comprehensive enough. For example, we used only a short test to assess students' reasoning abilities and did not include any indicators of working memory (Peng et al. 2016, 2018) or executive functioning (Bull et al. 2008). Thus, the question remains whether the bidirectional effects found could have been partially spurious results from unacknowledged confounders. Apart from that, the fit indices were clearly below the values that would be expected to be considered as a good fit (e.g., Schermelleh-Engel et al. 2003).

In comparison to the CLPM, the advantage of the CLPM-L2 is that the inclusion of prior measures of the variables of interest provides a stronger control for confounding variables (VanderWeele et al. 2020). As the CLPM-L2 did not include measures of reading and mathematics before Grade 5, the respective cross-lagged effects can only be interpreted from Grade 7. Looking at this time frame, the results showed small to negligible effects for mathematics on reading suggesting that students with higher mathematics competencies in Grade 7 or Grade 9 tended to have a somewhat higher increase in reading scores at the subsequent measurement point as compared to students with lower mathematics competencies. Contrary to our expectations, the respective effects for reading on mathematics were even smaller. Taken together, no clear indications were found for any bidirectional effects between reading and mathematics based on the CLPM-L2 and skill transfer does not appear to be a good explanation to describe the developmental changes between the Grades 7 and 12. Again, the assumption that reading comprehension becomes more important for solving complex mathematical tasks in secondary school (Björn et al. 2016; Harlaar et al. 2012) could not be confirmed based on this model. A possible explanation for this unexpected finding could be that most students in the middle of secondary school have already reached a level of reading comprehension that is sufficient to deal with the mathematical tasks that are typical for this age group. Possibly, a further improvement of reading comprehension does not lead to a corresponding improvement in mathematics when the linguistic requirements in the mathematics

tasks do not increase anymore. Rather, the greater difficulty of mathematics tasks may be due to other complexity features more than linguistic characteristics. Compared to the CLPM the bidirectional effects are considerably smaller in the CLPM-L2. As the CLPM-L2 helps to exclude unmeasured confounders (VanderWeele et al. 2020) this finding may be an indication that there are common causes (e.g., reasoning abilities, working memory) that affect both domains and that are not (or not sufficiently) controlled for in the CLPM.

As for the RI-CLPM, the interpretation of the effects is different again because only within-person effects were analyzed that refer to temporary fluctuations around individuals' means. According to this model, the cross-lagged effects showed that, at the beginning of secondary school, students with higher competencies in one domain as compared to their average competence tended to exhibit increased scores in the other competence domain at the subsequent measurement point. Thus, temporary comparatively high competencies in reading or mathematics led not only to increased competencies in the same domain but also in the other domain. Interestingly, the direction of the effects changed across the course of secondary school, more specifically, after Grade 7 for the effect of mathematics on reading and after Grade 9 for the effect of reading on mathematics. Towards the end of secondary school, students with temporary higher reading competencies as compared to their average competence level were likely to show comparatively lower mathematic scores at the next time point. The same was true for the effects of mathematics on reading. A tentative explanation for these effects could be related to the possibilities of domain-specific specialization in upper secondary school. At least in Germany, the school system offers multiple pathways in either further general education or vocational training which might provide more options to specialize in a competence domain (Freund et al. 2021). When students have temporary higher competencies in reading or mathematics than usual this could also be related to affective-motivational factors such as domain-specific self-concepts, motivation, or interest that are known to be associated with the corresponding competencies (Denissen et al. 2007; Gogol et al. 2017; McElvany et al. 2008; Wolter and Hannover 2016). According to the internal/external frame of reference model (Marsh 1986) it could be predicted that domain-specific abilities have positive effects on academic self-concepts in the corresponding domain and negative effects across domains (see also Brunner et al. 2008). Thus, experiencing temporary higher competencies in one domain that are possibly associated with higher self-concept, higher motivation or higher interest in this domain could lead to subsequently decreasing competence scores in the other domain, especially when the school system offers some opportunities for specialization. However, we must acknowledge that this interpretation is speculative. It could also be the case that the reported findings are partially a consequence of using instruments that are not curricular and the fact that the measured competencies do not or only partially correspond to the competencies actually taught in class. Put differently, if more curricular tests were used, possible bidirectional effects could be larger but maybe even smaller or nonexistent, depending on the specific content of the curricular tests (e.g., geometry, algebra, word problems as contents of mathematics or orthography, writing essays, reading literature as contents of language classes).

On a more general level, it is questionable whether the RI-CLPM which captures fluctuations around a person-mean is suited to investigate bidirectional effects in the long-term development of competencies. For example, Andersen (2021) recently emphasized that the RI-CLPM is inherently misspecified if the studied construct is not at equilibrium, that is, stationary with constant mean and variance across the observation period. However, in developmental studies this situation seems to be the norm rather than the exception, for example, when increasing levels of domain-specific competences are expected to occur throughout children's educational careers. As argued before (Asendorpf 2021; Lüdtke and Robitzsch 2021; Orth et al. 2021), the RI-CLPM is a good choice when the goal is to examine oscillations around a constant such as in short-term studies of states. However, from a developmental perspective, the competence level of a student at a given time point can be seen as the result of cumulative learning processes that are based on students' prerequisites (e.g., reasoning ability, working memory) and take place in interaction with their learning environment. Therefore, in this case, differential change can be considered "a continuous drift away from the initial between-person differences" (Asendorpf 2021, p. 829) and the interpretation of the person-mean is not at all clear.

4.1 Limitations and future directions

We want to highlight four aspects of the present study that might limit the generalizability of the reported findings and emphasize the need for follow-up research. First, the administered competence tests operationalized reading and mathematics from a literacy perspective (OECD 2017; Weinert et al. 2019) that describes the ability to understand and use written texts or mathematical concepts to achieve one's goals and to effectively take part in society. Thus, the tests did not emphasize specific content areas of reading or mathematics which might show different bidirectional associations. For example, understanding geometry requires pronounced visual-spatial abilities to apprehend and deconstruct visual forms (Clements 2004). Therefore, reading comprehension might be of less importance for the development of geometric abilities, while it might be more important for learning stochastics for which visual abilities presumably play an ancillary role. Therefore, future research is encouraged to explicate differences in bidirectional effects for relevant subdisciplines in reading and mathematics. Second, to some degree, the reported results could be idiosyncratic to the German language. So far, findings on skill transfer between reading and mathematics are dominated by research on children from English-speaking countries (e.g., Bailey et al. 2020; Cameron et al. 2019; Duncan et al. 2007; Grimm et al. 2021). As compared to English, the German language is characterized by a substantially more complex grammar system that requires the knowledge and application of increasingly elaborate rules to grasp the meaning of single sentences and whole texts. Therefore, understanding written mathematical problems might require higher levels of reading comprehension in the German language than in other languages that incorporate simpler grammar structures. As of yet, the comparability of bidirectional effects between reading and mathematics across different languages is largely uncharted territory. Third, in line with previous research, our analyses modeled linear effects within and between cognitive domains. If some of

these effects, however, are nonlinear, this might have contributed to imprecision in parameter estimates (Voelkle 2008). Moreover, we assumed constant time lags for all participants between adjacent grades. Although this simplification is common in longitudinal research, in practice, measurement periods might vary to some degree because testing times stretched over several weeks to months in each grade. Again, this might have added to parameter uncertainty to some degree. Follow-up studies might, therefore, compare the robustness of the reported findings by examining bidirectional effects using alternative modeling approaches with fewer (or rather, different) assumptions such as continuous time models (Driver and Voelkle 2021) or dynamic measurement models (Dumas et al. 2020). Finally, we want to acknowledge that several alternative autoregressive models such as the latent curve model with structured residuals (Curran et al. 2014) or the bivariate cross-lagged trait–state-error model (Kenny and Zautra 2001) have been developed that also account for unmeasured stable person and environmental effects (for an overview see also Zyphur et al. 2020). However, these are often not very useful in practice because they have substantial data requirements (e.g., regarding the number of measurement points) and more importantly, frequently result in improper solutions or nonconvergence of the SEMs (Orth et al. 2021; Usami et al. 2019).

5 Conclusion

Reading and mathematical competencies showed remarkable stability across lower and upper secondary school. However, in contrast to previous findings from primary school, bidirectional effects between both domains were rather small to negligible in this period. More importantly, the resulting pattern of effects substantially diverged depending on the chosen modeling strategy. Although our analyses emphasized that the traditional CLPM represented an inadequate description of our data, model extensions focusing on within-person effects (RI-CLPM) addressed a different causal effect than is typical of interest in developmental research. Therefore, we concur with Lüdtke and Robitzsch (2021; see also Asendorpf 2021) that differential change is better studied from a causal inference perspective, for example, using the CLPM-L2.

Supplementary Information The online version of this article (https://doi.org/10.1007/s11618-022-01108-w) contains supplementary material, which is available to authorized users.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Adachi, P., & Willoughby, T. (2015). Interpreting effect sizes when controlling for stability effects in longitudinal autoregressive models: Implications for psychological science. *European Journal of Developmental Psychology*, *12*(1), 116–128. https://doi.org/10.1080/17405629.2014.963549.

Adelson, J.L., Dickinson, E.R., & Cunningham, B.C. (2015). Differences in the reading–mathematics relationship: a multi-grade, multi-year statewide examination. *Learning and Individual Differences*, *43*, 118–123. https://doi.org/10.1016/j.lindif.2015.08.006.

Andersen, H.K. (2021). Equivalent approaches to dealing with unobserved heterogeneity in cross-lagged panel models? Investigating the benefits and drawbacks of the latent curve model with structured residuals and the random intercept cross-lagged panel model. *Psychological Methods.*. https://doi.org/10.1037/met0000285.

Asendorpf, J.B. (2021). Modeling developmental processes. In J.F. Rauthmann (Ed.), *The handbook of personality dynamics and processes* (pp. 816–837). Cambridge: Academic Press.

Bailey, D.H., Oh, Y., Farkas, G., Morgan, P., & Hillemeier, M. (2020). Reciprocal effects of reading and mathematics? Beyond the cross-lagged panel model. *Developmental Psychology*, *56*(5), 912–921. https://doi.org/10.1037/dev0000902.

Berry, D., & Willoughby, M.T. (2017). On the practical interpretability of cross-lagged panel models: rethinking a developmental workhorse. *Child Development*, *88*, 1186–1206. https://doi.org/10.1111/cdev.12660.

Björn, P.M., Aunola, K., & Nurmi, J.-E. (2016). Primary school text comprehension predicts mathematical word problem-solving skills in secondary school. *Educational Psychology*, *36*(2), 362–377. https://doi.org/10.1080/01443410.2014.992392.

Blossfeld, H.-P., & Roßbach, H.-G. (Eds.). (2019). *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (2nd edn., Edition ZfE). Wiesbaden: Springer VS.

Bosco, F.A., Aguinis, H., Singh, K., Field, J.G., & Pierce, C.A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, *100*, 431–449. https://doi.org/10.1037/a0038047.

Brosseau-Liard, P.E., & Savalei, V. (2014). Adjusting relative fit indices for nonnormality. *Multivariate Behavioral Research*, *49*, 460–470. https://doi.org/10.1080/00273171.2014.933697.

Brosseau-Liard, P., Savalei, V., & Li, L. (2012). An investigation of the sample performance of two nonnormality corrections for RMSEA. *Multivariate Behavioral Research*, *47*, 904–930. https://doi.org/10.1080/00273171.2012.715252.

Brunner, M., Lüdtke, O., & Trautwein, U. (2008). The internal/external frame of reference model revisited: Incorporating general cognitive ability and general academic self-concept. *Multivariate Behavioral Research*, *43*(1), 137–172. https://doi.org/10.1080/00273170701836737.

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.

Burgette, L.F., & Reiter, J.P. (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology*, *172*(9), 1070–1076. https://doi.org/10.1093/aje/kwq260.

Cameron, C.E., Kim, H., Duncan, R.J., Becker, D.R., & McClelland, M.M. (2019). Bidirectional and co-developing associations of cognitive, mathematics, and literacy skills during kindergarten. *Journal of Applied Developmental Psychology*, *62*, 135–144. https://doi.org/10.1016/j.appdev.2019.02.004.

Cattell, R.B. (1987). *Intelligence: its structure, growth, and action*. North-Holland.

Cirino, P.T., Child, A.E., & Macdonald, K. (2018). Longitudinal predictors of the overlap between reading and math skills. *Contemporary Educational Psychology*, *54*, 99–111. https://doi.org/10.1016/j.cedpsych.2018.06.002.

Clements, D.H. (2004). Geometric and spatial thinking in early childhood education. In D.H. Clements, J. Sarama & A.-M. Di Biase (Eds.), *Engaging young children in mathematics: standards for early childhood mathematics education* (pp. 267–298). Hillsdale: Lawrence Earlbaum Associates.

Clements, D.H., & Sarama, J. (2011). Early childhood mathematics intervention. *Science*, *333*(6045), 968–970. https://doi.org/10.1126/science.1204537.

Cole, D.A., & Preacher, K.J. (2014). Manifest variable path analysis: potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods*, *19*, 300–315. https://doi.org/10.1037/a0033805.

Curran, P.J., Howard, A.L., Bainter, S.A., Lane, S.T., & McGinley, J.S. (2014). The separation of between-person and within-person components of individual change over time: A latent curve model

with structured residuals. *Journal of Consulting and Clinical Psychology*, *82*, 879–894. https://doi.org/10.1037/a0035297.

De Smedt, B. (2018). Language and arithmetic: the potential role of phonological processing. In A. Henik & W. Fias (Eds.), *Heterogeneity of function in numerical cognition* (pp. 51–74). Amsterdam: Elsevier.

Dehaene, S., & Cohen, L. (1995). Towards an anatomical and functional model of number processing. *Mathematical Cognition*, *1*, 83–120.

Denissen, J. J., Zarrett, N. R., & Eccles, J. S. (2007). I like to do it, I'm able, and I know I am: Longitudinal couplings between domain-specific achievement, self-concept, and interest. *Child Development*, *78*(2), 430–447. https://doi.org/10.1111/j.1467-8624.2007.01007.x.

Driver, C. C., & Voelkle, M. C. (2021). Hierarchical continuous time models. In J. F. Rauthmann (Ed.), *The handbook of personality dynamics and processes* (pp. 887–908). Cambridge: Academic Press.

Duchhardt, C., & Gerdes, A. (2012). *NEPS technical report for mathematics—scaling results of starting cohort 3 in fifth grade* (NEPS Working Paper No., Vol. 19). Bamberg: Otto-Friedrich University, National Educational Panel Study.

Dumas, D., McNeish, D., & Greene, J. A. (2020). Dynamic measurement: A theoretical-psychometric paradigm for modern educational psychology. *Educational Psychologist*, *55*(2), 88–105. https://doi.org/10.1080/00461520.2020.1744150.

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.

Erbeli, F., Shi, Q., Campbell, A. R., Hart, S. A., & Woltering, S. (2021). Developmental dynamics between reading and math in elementary school. *Developmental Science*, *24*(1), e13004. https://doi.org/10.1111/desc.13004.

Ferrer, E., & McArdle, J. J. (2004). An experimental analysis of dynamic hypotheses about cognitive abilities and achievement from childhood to early adulthood. *Developmental Psychology*, *40*(6), 935–952. https://doi.org/10.1037/0012-1649.40.6.935.

Fischer, L., Rohm, T., Gnambs, T., & Carstensen, C. H. (2016). *Linking the data of the competence tests* (NEPS Survey Paper No., Vol. 1). Bamberg: Leibniz Institute for Educational Trajectories.

Fischer, L., Gnambs, T., Rohm, T., & Carstensen, C. H. (2019). Longitudinal linking of Rasch-model-scaled competence tests in large-scale assessments: A comparison and evaluation of different linking methods and anchoring designs based on two tests on mathematical competence administered in grades 5 and 7. *Psychological Test and Assessment Modeling*, *61*, 37–64.

Freund, M.-J., Wolter, I., Lockl, K., & Gnambs, T. (2021). Determinants of profiles of competence development in mathematics and reading in upper secondary education in Germany. *PloS ONE*, *16*(10), e258152. https://doi.org/10.1371/journal.pone.0258152.

Fuchs, L. S., Schumacher, R. F., Long, J., Namkung, J., Hamlett, C. L., Cirino, P. T., Jordan, N. C., Siegler, R., Gersten, R., & Changas, P. (2013). Improving at-risk learners' understanding of fractions. *Journal of Educational Psychology*, *105*, 683–700. https://doi.org/10.1037/a0032446.

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*. Presentation at the Annual Conference of the International Social Survey Programme, Lisbon.

Gathercole, S. E., Willis, C. S., Emslie, H., & Baddeley, A. D. (1992). Phonological memory and vocabulary development during the early school years: A longitudinal study. *Developmental Psychology*, *28*(5), 887–898. https://doi.org/10.1037/0012-1649.28.5.887.

Gehrer, K., Zimmermann, S., Artelt, C., & Weinert, S. (2013). NEPS framework for assessing reading competence and results from an adult pilot study. *Journal for Educational Research Online*, *5*, 50–79.

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.

Glenberg, A., Willford, J., Gibson, B., Goldberg, A., & Zhu, X. (2012). Improving reading to improve math. *Scientific Studies of Reading*, *16*(4), 316–340. https://doi.org/10.1080/10888438.2011.564245.

Gnambs, T. (2014). A meta-analysis of dependability coefficients (test-retest reliabilities) for measures of the Big Five. *Journal of Research in Personality*, *52*, 20–28. https://doi.org/10.1016/j.jrp.2014.06.003.

Gnambs, T. (2015). Facets of measurement error for scores of the big five: three reliability generalizations. *Personality and Individual Differences*, *84*, 84–89. https://doi.org/10.1016/j.paid.2014.08.019.

Gnambs, T., & Nusser, L. (2019). The longitudinal measurement of reasoning abilities in students with special educational needs. *Frontiers in Psychology*, *10*(232), 88–92. https://doi.org/10.3389/fpsyg.2019.00232.

Gogol, K., Brunner, M., Martin, R., Preckel, F., & Goetz, T. (2017). Affect and motivation within and between school subjects: development and validation of an integrative structural model of academic self-concept, interest, and anxiety. *Contemporary Educational Psychology*, *49*, 46–65. https://doi.org/10.1016/j.cedpsych.2016.11.003.

Götz, F., Gosling, S., & Rentfrow, J. (2022). Small effects: The indispensable foundation for a cumulative psychological science. *Perspectives in Psychological Science*, *17*(1), 205–215. https://doi.org/10.1177/1745691620984483.

Gough, P. B., Hoover, W. A., & Peterson, C. L. (1996). Some observations on a simple view of reading. In C. Cornoldi & J. Oakhill (Eds.), *Reading comprehension difficulties: processes and intervention* (pp. 1–13). Hillsdale: Erlbaum.

Grimm, K. J. (2008). Longitudinal associations between reading and mathematics achievement. *Developmental Neuropsychology*, *33*(3), 410–426. https://doi.org/10.1080/87565640801982486.

Grimm, K. J., Helm, J., Rodgers, D., & O'Rourke, H. (2021). Analyzing cross-lag effects: a comparison of different cross-lag modeling approaches. *New Directions for Child and Adolescent Development*, *2021*(175), 11–33. https://doi.org/10.1002/cad.20401.

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, *20*, 102–116. https://doi.org/10.1037/a0038889.

Harlaar, N., Kovas, Y., Dale, P., Petrill, S., & Plomin, R. (2012). Mathematics is differentially related to reading comprehension and word decoding: evidence from a genetically sensitive design. *Journal of Educational Psychology*, *104*, 622–635. https://doi.org/10.1037/a0027646.

Hübner, N., Merrell, C., Cramman, H., Little, J., Bolden, D., & Nagengast, B. (2021). *Reading to learn? The co-development of mathematics and reading during primary school*. https://doi.org/10.31219/osf.io/v8h29. OSF Preprints

Jordan, N. C., Kaplan, D., Ramineni, C., & Locuniak, M. N. (2009). Early math matters: Kindergarten number competence and later mathematics outcomes. *Developmental Psychology*, *45*, 850–867. https://doi.org/10.1037/a0014939.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2021). *semTools: Useful tools for structural equation modeling*. R package version 0.5–5. https://CRAN.R-project.org/package=semTools. Accessed 2022-06-24

Joyner, R. E., & Wagner, R. K. (2020). Co-occurrence of reading disabilities and math disabilities: a meta-analysis. *Scientific Studies of Reading*, *24*(1), 14–22. https://doi.org/10.1080/10888438.2019.1593420.

Kenny, D. A., & Zautra, A. (2001). Trait-state models for longitudinal data. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 243–263). American Psychological Association. https://doi.org/10.1037/10409-008.

Koponen, T., Salmi, P., Eklund, K., & Aro, T. (2013). Counting and RAN: predictors of arithmetic calculation and reading fluency. *Journal of Educational Psychology*, *105*(1), 162–175. https://doi.org/10.1037/a0029285.

Korpipää, H., Koponen, T., Aro, M., Tolvanen, A., Aunola, K., Poikkeus, A. M., Lekrkanen, M.-K., & Nurmi, J. E. (2017). Covariation between reading and arithmetic skills from Grade 1 to Grade 7. *Contemporary Educational Psychology*, *51*, 131–140. https://doi.org/10.1016/j.cedpsych.2017.06.005.

Krannich, M., Jost, O., Rohm, T., Koller, I., Pohl, S., Haberkorn, K., Carstensen, C. H., Fischer, L., & Gnambs, T. (2017). *NEPS Technical Report for Reading—Scaling Results of Starting Cohort 3 for Grade 7* (Vol. 14). Bamberg: Leibniz Institute for Educational Trajectories. https://doi.org/10.5157/NEPS:SP14:2.0.

Kutscher, T., & Scharl, A. (2020). *NEPS Technical Report for Reading: Scaling Results of Starting Cohort 3 for Grade 12* (Vol. 67). Bamberg: Leibniz Institute for Educational Trajectories. https://doi.org/10.5157/NEPS:SP67:1.0.

Kyttälä, M., & Björn, P. M. (2014). The role of literacy skills in adolescents' mathematics word problem performance: controlling for visuo-spatial ability and mathematics anxiety. *Learning and Individual Differences*, *29*, 59–66. https://doi.org/10.1016/j.lindif.2013.10.010.

Lang, F. R., Kamin, S., Rohr, M., Stünkel, C., & Williger, B. (2014). *Erfassung der fluiden kognitiven Leistungsfähigkeit über die Lebensspanne im Rahmen des Nationalen Bildungspanels: Abschlussbericht zu einer NEPS-Ergänzungsstudie* (NEPS Working Paper No., Vol. 43). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Lechner, C.M., Bhaktha, N., Groskurth, K., & Bluemke, M. (2021a). Why ability point estimates can be pointless: a primer on using skill measures from large-scale assessments in secondary analyses. *Measurement Instruments for the Social Sciences*, *3*(1), 1–16. https://doi.org/10.1186/s42409-020-00020-5.

Lechner, C.M., Gauly, B., Miyamoto, A., & Wicht, A. (2021b). Stability and change in adults' literacy and numeracy skills: evidence from two large-scale panel studies. *Personality and Individual Differences*, *180*, 110990. https://doi.org/10.1016/j.paid.2021.110990.

LeFevre, J.-A., Fast, L., Skwarchuk, S.L., Smith-Chant, B.L., Bisanz, J., Kamawar, D., & Penner-Wilger, M. (2010). Pathways to mathematics: longitudinal predictors of performance. *Child Development*, *81*, 1753–1767. https://doi.org/10.1111/j.1467-8624.2010.01508.x.

Little, C.W., Lonigan, C.J., & Phillips, B.M. (2021). Differential patterns of growth in reading and math skills during elementary school. *Journal of Educational Psychology*, *113*(3), 462–476. https://doi.org/10.1037/edu0000635.

Lucas, R.E. (2022). *It's time to abandon the cross-lagged panel model*. PsyArXiv Preprints. https://doi.org/10.31234/osf.io/pkec7.

Lüdtke, O., & Robitzsch, A. (2017). Eine Einführung in die Plausible-Values-Technik für die psychologische Forschung. *Diagnostica*, *63*(3), 193–205. https://doi.org/10.1026/0012-1924/a000175.

Lüdtke, O., & Robitzsch, A. (2021). *A critique of the random-intercept cross-lagged panel model*. PsyArXiv Preprints. https://doi.org/10.31234/osf.io/6f85c.

Marsh, H.W. (1986). Verbal and math self-concepts: an internal/external frame of reference model. *American Educational Research Journal*, *23*, 129–149. https://doi.org/10.3102/00028312023001129.

McElvany, N., Kortenbruck, M., & Becker, M. (2008). Lesekompetenz und Lesemotivation. Entwicklung und Mediation des Zusammenhangs durch Leseverhalten [Reading competence and reading motivation. Development and mediation by reading behavior]. *Zeitschrift für Pädagogische Psychologie*, *22*(34), 207–219. https://doi.org/10.1024/1010-0652.22.34.207.

Meng, X. L., & Rubin, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika, 79*(1), 103–111. https://doi.org/10.1093/biomet/79.1.103.

Mulder, J.D., & Hamaker, E.L. (2021). Three extensions of the random intercept cross-lagged panel model. *Structural Equation Modeling*, *28*, 638–648. https://doi.org/10.1080/10705511.2020.1784738.

Mund, M., & Nestler, S. (2019). Beyond the cross-lagged panel model: Next-generation statistical tools for analyzing interdependencies across the life course. *Advances in Life Course Research*, *41*, 100249. https://doi.org/10.1016/j.alcr.2018.10.002.

NEPS Network (2020). *National educational panel study, scientific use file of starting cohort grade 5*. Bamberg: Leibniz Institute for Educational Trajectories (LIfBi). https://doi.org/10.5157/NEPS:SC3:10.0.0.

Neumann, I., Duchhardt, C., Ehmke, T., Grüßing, M., Heinze, A., & Knopp, E. (2013). Modeling and assessing of mathematical competence over the lifespan. *Journal for Educational Research Online*, *5*(2), 80–109.

OECD (2017). *PISA 2015 assessment and analytical framework: science, reading, mathematic, financial literacy and collaborative problem solving*. Paris: OECD Publishing. https://doi.org/10.1787/9789264281820-en.

Orth, U., Clark, D.A., Donnellan, M.B., & Robins, R.W. (2021). Testing prospective effects in longitudinal research: comparing seven competing cross-lagged models. *Journal of Personality and Social Psychology*, *120*(4), 1013–1034. https://doi.org/10.1037/pspp0000358.

Pearl, J., Glymour, M., & Jewell, N.P. (2016). *Causal inference in statistics: a primer*. Hoboken: John Wiley.

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.

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., Wang, T., Wang, C., & Lin, X. (2019). A meta-analysis on the relation between fluid intelligence and reading/mathematics: effects of tasks, age, and social economics status. *Psychological Bulletin*, *145*(2), 189–236. https://doi.org/10.1037/bul0000182.

Petersen, L.A., Litteck, K., & Rohenroth, D. (2020). *NEPS technical report for mathematics: scaling results of starting cohort 3 for grade 12* (Vol. 75). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://doi.org/10.5157/NEPS:SP75:1.0.

Pohl, S. (2013). Longitudinal multistage testing. *Journal of Educational Measurement*, *50*(4), 447–468. https://doi.org/10.1111/jedm.12028.

Pohl, S., & Carstensen, C.H. (2013). Scaling of competence tests in the National Educational Panel Study—many questions, some answers, and further challenges. *Journal for Educational Research Online*, *5*(2), 189–216. https://doi.org/10.25656/01:8430.

Pohl, S., Haberkorn, K., Hardt, K., & Wiegand, E. (2012). *NEPS technical report for reading—scaling results of starting cohort 3 in fifth grade* (NEPS Working Paper No., Vol. 15). Bamberg: Otto-Friedrich University, National Educational Panel Study.

Purpura, D.J., Hume, L.E., Sims, D.M., & Lonigan, C.J. (2011). Early literacy and early numeracy: the value of including early literacy skills in the prediction of numeracy development. *Journal of Experimental Child Psychology*, *110*(4), 647–658. https://doi.org/10.1016/j.jecp.2011.07.004.

R Core Team (2021). *R: A language and environment for statistical computing [Computer software]*. R Foundation for Statistical Computing. https://www.R-project.org

Rescorla, L., & Rosenthal, A.S. (2004). Growth in standardized ability and achievement test scores from 3rd to 10th grade. *Journal of Educational Psychology*, *96*(1), 85–96. https://doi.org/10.1037/0022-0663.96.1.85.

Riley, M.S., & Greeno, J.G. (1988). Developmental analysis of understanding language about quantities and of solving problems. *Cognition and Instruction*, *5*, 49–101. https://doi.org/10.1207/s1532690xci0501_2.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Hoboken: Wiley.

Sarama, J., Lange, A.A., Clements, D.H., & Wolfe, C.B. (2012). The impacts of an early mathematics curriculum on oral language and literacy. *Early Childhood Research Quarterly*, *27*, 489–502. https://doi.org/10.1016/j.ecresq.2011.12.002.

Savalei, V. (2014). Understanding robust corrections in structural equation modeling. *Structural Equation Modeling*, *21*(1), 149–160. https://doi.org/10.1080/10705511.2013.824793.

Scammacca, N., Fall, A.-M., Capin, P., Roberts, G., & Swanson, E. (2020). Examining factors affecting reading and math growth and achievement gaps in grades 1–5: a cohort-sequential longitudinal approach. *Journal of Educational Psychology*, *112*(4), 718–734. https://doi.org/10.1037/edu0000400.

Scharl, A., Fischer, L., Gnambs, T., & Rohm, T. (2017). *NEPS technical report for reading: scaling results of starting cohort 3 for grade 9* (Vol. 20). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://doi.org/10.5157/NEPS:SP20:1.0.

Scharl, A., Carstensen, C.H., & Gnambs, T. (2020). *Estimating plausible values with NEPS data: an example using reading competence in starting cohort 6* (NEPS Survey Paper No., Vol. 71). Bamberg: Leibniz Institute for Educational Trajectories. https://doi.org/10.5157/NEPS:SP71:1.0.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research*, *8*(2), 23–74.

Schnittjer, I., & Gerken, A.-L. (2017). *NEPS technical report for mathematics: scaling results of starting cohort 3 in grade 7* (NEPS Survey Paper No., Vol. 16). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://doi.org/10.5157/NEPS:SP16:1.0.

Schwarz, G.E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136.

Shin, T., Davison, M.L., Long, J.D., Chan, C.-K., & Heistad, D. (2013). Exploring gains in reading and mathematics achievement among regular and exceptional students using growth curve modeling. *Learning and Individual Differences*, *23*, 92–100. https://doi.org/10.1016/j.lindif.2012.10.002.

Singer, V., & Strasser, K. (2017). The association between arithmetic and reading performance in school: a meta-analytic study. *School Psychology Quarterly*, *32*(4), 435–448. https://doi.org/10.1037/spq0000197.

Skopek, J., & Passaretta, G. (2021). Socioeconomic inequality in children's achievement from infancy to adolescence: the case of Germany. *Social Forces*, *100*(1), 86–112. https://doi.org/10.1093/sf/soaa093.

Spengler, M., Damian, R.I., & Roberts, B.W. (2018). How you behave in school predicts life success above and beyond family background, broad traits, and cognitive ability. *Journal of Personality and Social Psychology*, *114*(4), 620–636. https://doi.org/10.1037/pspp0000185.

Steinhauer, H.W., Aßmann, C., Zinn, S., Goßmann, S., & Rässler, S. (2015). Sampling and weighting cohort samples in institutional contexts. *AStA Wirtschafts- und Sozialstatistisches Archiv*, *9*, 131–157. https://doi.org/10.1007/s11943-015-0162-0.

Usami, S., Todo, N., & Murayama, K. (2019). Modeling reciprocal effects in medical research: critical discussion on the current practices and potential alternative models. *PloS one*, *14*(9), e209133. https://doi.org/10.1371/journal.pone.0209133.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03.

Van den Ham, A.-K., Schnittjer, I., & Gerken, A.-L. (2018). *NEPS technical report for mathematics: scaling results of starting cohort 3 for grade 9* (Vol. 38). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study. https://doi.org/10.5157/NEPS:SP38:1.0.

Vanbinst, K., van Bergen, E., Ghesquière, P., & De Smedt, B. (2020). Cross-domain associations of key cognitive correlates of early reading and early arithmetic in 5-year-olds. *Early Childhood Research Quarterly*, *51*, 144–152. https://doi.org/10.1016/j.ecresq.2019.10.009.

VanderWeele, T.J., Mathur, M.B., & Chen, Y. (2020). Outcome-wide longitudinal designs for causal inference: a new template for empirical studies. *Statistical Science*, *35*(3), 437–466. https://doi.org/10.1214/19-STS728.

Voelkle, M.C. (2008). Reconsidering the use of autoregressive latent trajectory (ALT) models. *Multivariate Behavioral Research*, *43*(4), 564–591. https://doi.org/10.1080/00273170802490665.

Vukovic, R.K., & Lesaux, N.K. (2013). The language of mathematics: investigating the ways language counts for children's mathematical development. *Journal of Experimental Child Psychology*, *115*(2), 227–244. https://doi.org/10.1016/j.jecp.2013.02.002.

Watts, T.W., Duncan, G.J., Clements, D.H., & Sarama, J. (2018). What is the long-run impact of learning mathematics during preschool? *Child Development*, *89*(2), 539–555. https://doi.org/10.1111/cdev.12713.

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., Carstensen, C.H., & Lockl, K. (2019). Development of competencies across the life course. In H.-P. Blossfeld & H.-G. Roßbach (Eds.), *Education as a lifelong process: the German National Educational Panel Study (NEPS)* (Edition ZfE, 2nd rev. ed, pp. 57–82). Wiesbaden: Springer. https://doi.org/10.1007/978-3-658-23162-0.

Weirich, S., Haag, N., Hecht, M., Böhme, K., Siegle, T., & Lüdtke, O. (2014). Nested multiple imputation in large-scale assessments. *Large-Scale Assessments in Education, 2*(1), 9. https://doi.org/10.1186/s40536-014-0009-0.

Wicht, A., Rammstedt, B., & Lechner, C.M. (2021). Predictors of literacy development in adulthood: Insights from a large-scale, two-wave study. *Scientific Studies of Reading*, *25*(1), 84–92. https://doi.org/10.1080/10888438.2020.1751635.

Wolter, I., & Hannover, B. (2016). Gender role self-concept at school start and its impact on academic self-concept and performance in mathematics and reading. *European Journal of Developmental Psychology*, *13*(6), 681–703. https://doi.org/10.1080/17405629.2016.1175343.

Yuan, K.H., & Bentler, P.M. (2000). Three likelihood-based methods for mean and covariance structure analysis with non-normal missing data. *Sociological Methodology*, *30*(1), 165–200. https://doi.org/10.1111/0081-1750.00078.

Zyphur, M.J., Allison, P.D., Tay, L., Voelkle, M.C., Preacher, K.J., Zhang, Z., Hamaker, E.L., Shamsollahi, A., Pierides, D.C., Koval, P., & Diener, E. (2020). From data to causes I: building a general cross-lagged panel model (GCLM). *Organizational Research Methods*, *23*(4), 651–687. https://doi.org/10.1177/1094428119847278.