Educational Psychology Review
https://doi.org/10.1007/s10648-022-09662-9

REVIEW

Directional Ordering of Self-Concept, School Grades, and Standardized Tests Over Five Years: New Tripartite Models Juxtaposing Within- and Between-Person Perspectives

Herbert W. Marsh[1,2] · **Reinhard Pekrun**[1,3,4] · **Oliver Lüdtke**[5,6]

Accepted: 1 February 2022

Abstract
Much research shows academic self-concept and achievement are reciprocally related over time, based on traditional longitudinal data cross-lag-panel models (CLPM) supporting a reciprocal effects model (REM). However, recent research has challenged CLPM's appropriateness, arguing that CLPMs with random intercepts (RI-CLPMs) provide a more robust (within-person) perspective and better control for unmeasured covariates. However, there is much confusion in educational-psychology research concerning appropriate research questions and interpretations of RI-CLPMs and CLPMs. To clarify this confusion, we juxtapose CLPMs and RI-CLPMs relating math self-concept (MSCs), school grades, and achievement tests over the five years of compulsory secondary schooling ($N = 3{,}425$). We extend basic models to evaluate: directional ordering among three rather than only two constructs; longitudinal invariance over time (multiple school years) and multiple groups (school tracks); lag-2 paths between non-adjacent waves; and covariates (gender, primary-school math and verbal achievement). Across all basic and extended RI-CLPMs and CLPMs, there was consistent support for the REM bidirectional-ordering hypothesis that self-concept and achievement are each a cause and an effect of the other. Consistent with the logic of these models, extensions of the basic models had more effect on CLPMs, but the direction and statistical significance of cross-lagged paths were largely unaffected for both RI-CLPMs and CLPMs. This substantive-methodological synergy has important implications for theory, methodology, and policy/practice; we support the importance of MSC as a predictor of subsequent achievement and demonstrate a more robust methodological framework for evaluating longitudinal-panel models.

Keyword Directional-ordering · within-person and between-person perspectives · Academic self-concept · Substantive-methodological synergy

✉ Herbert W. Marsh
Herb.Marsh@acu.edu.au

Extended author information available on the last page of the article

Published online: 12 April 2022

Self-concept is a person's perceptions of themselves, formed through their experiences with and interpretations of their environment, and impacted by others' evaluations. It affects how we act, feel, and adjust to a changing environment. In educational settings, the focus of our study, previous research has shown that academic self-concept (ASC) is linked to a variety of educational outcomes, including academic achievement (Marsh & Craven, 2006; Marsh & Martin, 2011; Marsh, et al., 2018a, 2018b; Marsh, Hau, et al., 2005; Marsh, Trautwein, et al., 2005), interest and satisfaction in school, achievement emotions (Marsh, et al., 2018a, 2018b; Pekrun, 2006; Pekrun et al., 2017), course selection (Marsh & Yeung, 1997; Marsh et al., 2019; Parker et al., 2014) persistence, and long-term attainment (Guo et al., 2015; Guo, Marsh, et al., 2015; Guo, Parker, et al., 2015; Guo, Parker, et al., 2015; Marsh & O'Mara, 2008). Particularly good support for the generalizability of the correlation between ASC and achievement comes from the cross-national studies. The positive correlations between ASC and achievement generalize over countries based on studies using Programme for International Student Assessment (PISA) data (Basarkod et al., 2020; Marsh & Hau, 2003; Nagengast & Marsh, 2011; Seaton et al., 2009) and the combined Trends in Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS) database. However, even though ASC and achievement are substantially correlated, a critical question with important theoretical and policy-practice implications is the directional ordering of these constructs. Hence, the critical question is whether this correlation reflects a non-causal association, causal effects of prior ASC on achievement, causal effects of prior achievement on subsequent ASC, or causal effects in both directions.

Here we briefly review the considerable body of research showing that ASC and achievement are reciprocally related over time, supporting a reciprocal effects model (REM). However, nearly all this research is based on a between-person perspective using traditional cross-lag-panel models (CLPMs) of longitudinal data. In contrast to this traditional approach, recent research has challenged CLPMs' appropriateness, arguing that CLPMs fail to uncover the within-person effects linking ASC and achievement (Murayama et al., 2017). CLPMs with random intercepts (Hamaker et al., 2015) have been proposed to provide a more robust (within-person) perspective and better control for unmeasured covariates. However, there is limited research juxtaposing results from these two approaches in educational psychology, and much ambiguity about appropriate use and interpretation of these models and their extensions. Hence, the overarching purpose of the present investigation is to compare these two approaches and demonstrate important extensions to them. In particular, a critical unanswered question is whether the substantial body of REM findings based on the between-person (CLPM) approach holds up for within-person (RI-CLPM) analyses and extensions of the CLPM? Following Marsh and Hau (2006), who originally coined the term, our study is a methodological-substantive synergy that applies a cutting-edge methodology to address substantive issues with implications for theory, methodology, and policy/practice.

Support for the Reciprocal Effects Model

Byrne (1984) proposed three criteria that studies addressing directional ordering must satisfy: (a) a statistical relationship must be established, (b) a clearly established time precedence must be evident, and (c) a causal model must be tested using appropriate statistical techniques such as the use of structural equation models (SEMs). Traditional approaches to this issue (Calsyn & Kenny, 1977) took an "either-or" approach—either prior achievement leads to subsequent ASC (a skill development model) or prior ASC leads to subsequent achievement (a self-enhancement model). However, integrating theoretical and statistical perspectives, Marsh (1990) argued for a dynamic reciprocal effects model (REM) that incorporates both the skill development and the self-enhancement model (see also Pekrun, 1990). This theoretical model predicts that better ASCs lead to better achievement, and that better achievement leads to better ASCs. We emphasize that the theoretical hypothesis is clearly causal. Marsh further noted that it was well established that students base their ASCs at least in part on their prior achievement (the skill development path). Hence the critical issue is whether higher ASCs also lead to higher achievement, regardless of whether this self-enhancement path is larger or smaller than the skill-development path.

The REM hypothesis is clearly causal in nature, a theoretical hypothesis of the causal ordering of variables over time. Indeed, the REM naturally leads to the hypothesis of reciprocal effects that are causal. Thus, Marsh, Trautwein, et al. (2005, p. 397) conclude: Reciprocal effects models of longitudinal data show that ASC is both a cause and an effect of achievement. The REM generated a substantial research literature that treats these reciprocal effects as causal and a number of methodological studies on how to test it (e.g., Usami, Murayama, et al., 2019; Usami, Todo, et al., 2019). For example, from a between-person perspective, a critical empirical question is: For students with the same levels of achievement at wave one, will students with higher ASCs have higher achievement in subsequent waves than those with lower ASCs? Positive evidence answering this question would support REM hypotheses and might reflect a causal effect. However, this interpretation is compromised by a lack of control for a potentially infinite number of covariates (i.e., effects might change if researchers controlled for the right covariates). Our position is that appropriate statistical models capture one key component of causality, namely directionality, by providing empirical tests of directional "causal" hypotheses.

However, there are potential competing interpretations that compromise interpretation—as is always the case with claims of causality. Nevertheless, to deflect concerns about using the broad, potentially ambiguous term of causality (or Granger causality; Granger, 1969), we use the more focused term of "directionality" (i.e., tests of directionality of causal tests effects rather than causal ordering). Indeed, tests of the directionality of effects have a long history concerning this research. Thus, for example, Bandura (1986) noted causal relations between self-efficacy and outcomes are bidirectional (i.e., reciprocal) rather than unidirectional. We also note that the terms directional, unidirectional, and bidirectional are widely used in relation to cross-lagged panel studies (e.g., Bailey et al., 2020;

Wu & Griffin, 2012). Thus, the term directionality (and the reliance on bidirectional models) has a long history concerning our theoretical hypotheses, is more focused than the term causality, and better reflects what is being tested. Nevertheless, we emphasize that the REM hypothizes a bidirectional model that is causal. Hence, the central question is whether empirically demonstrated reciprocal relations between ASC and achievement can be given a causal interpretation that is consistent with the REM hypothesis.

REM's theoretical hypothesis of the directional ordering of causal relations is testable when both ASC and achievement are collected in at least two but preferably three or more waves of data. Following Marsh (1990), there is a substantial research literature supporting REM predictions, including comprehensive meta-analyses and systematic reviews (e.g., Huang, 2011; Valentine et al., 2004; Wu et al., 2021; also see Marsh & Craven, 2006; Marsh & Martin, 2011). Consistent with ASC theory and research, it is not surprising that prior achievement affects ASC. However, the meta-analyses demonstrated that the effect of prior ASC on subsequent achievement, controlling the effects of prior achievement, was also highly significant overall and positive in most of the studies they considered. These authors clearly interpreted support for the REM hypothesis of reciprocal effects as causal effects, noting implications concerning the need for interventions that simultaneously enhance both ASC and achievement. Thus, Marsh and Craven (2006; also see Huang, 2011) argued: "If practitioners enhance self-concepts without improving performance, then the gains in self-concept are likely to be short-lived….If practitioners improve performance without also fostering participants' self-beliefs in their capabilities, then the performance gains are also unlikely to belong-lasting" (p. 159).

In REM studies, achievement is typically assessed by standardized tests or school grades, but the different achievement indicators have different implications. School grades are a particularly salient source of feedback to students and their parents, are easily compared among classmates, and have important implications for academic careers. Hence, school grades tend to be more correlated with ASCs than test scores (Marsh et al., 2014, 2014a, 2014b, 2014c; Marsh et al., 2014; Marsh, Hau, et al., 2005; Marsh, Morin, et al., 2014; Marsh, Trautwein, et al., 2005). However, school grades typically are idiosyncratic to particular teachers, settings, and schools. In particular, teachers typically grade on a curve, allocating the best and worst grades to the relatively better and poorer performing students within a classroom. Hence, teachers use the classroom as a narrow frame of reference in their grading procedure, largely ignoring students' absolute levels of achievement in their class relative to a common metric that generalizes over all students. Although the classic meta-analyses support REM hypotheses for both school grades and test scores, most individual studies have included only one of these achievement indicators. Moreover, few studies have fully juxtaposed the two over an extended developmental period nor evaluated the consistency of effects over time.

In a critique of CLPMs like those used in nearly all REM studies, Hamaker et al. (2015) proposed the RI-CLPM to analyze within-person relations over time. They argue that this approach is more appropriate for evaluating within-person relations between constructs. Specifically, the RI-CLPM shows how within-person deviations

in one construct are related to subsequent within-person deviations in another construct. CLPMs confound the within- and between-person processes.

Based on this within-person perspective, Ehm et al. (2019, 2021) evaluated support for the REM based on CLPMs and RI-CLPMs using the same sample of young primary school students in separate studies of reading and mathematics. For reading constructs, Ehm et al. (2019) found support for REM with CLPMs but not RI-CLPMs. Indeed, their RI-CLPMs provided no support for either skill development or self-enhancement perspectives; there were no statistically significant cross-paths from prior ASC to subsequent achievement or from prior achievement to subsequent ASC. For math outcomes (Ehm et al., 2021), the CLPM and RI-CLPM results were similar and provided partial support for the REM. For both models, the cross-lagged effects were non-significant from Year 1 to Year 2, but there were reciprocal effects for Year 2 to Year 3. In their discussion of limitations and directions for further research, Ehm et al. noted issues specific to the measurement of ASC with young children, developmental processes that are evolving at these young ages, and the need to test the generalizability of the results with older children. Indeed, a variety of theoretical perspectives suggest that children only begin to use social comparison processes as a basis of self-evaluation at the age of 7 or 8 (e.g., Harter, 1998; Piaget & Inhelder, 1969; Ruble, 1983; but also see Marsh et al., 2002). Hence, there might have been a developmental shift in the qualitative nature of ASCs, potentially undermining the rationale of the CLPMs and especially the RI-CLPMs. Emphasizing this issue and noting prior support for the REM was based largely on CLPMs with secondary-school students rather than primary-school students, Ehm et al. (2019, 2021) called for studies comparing CLPMs and RI-CLPMs based on secondary students: *Further research and also open discussions about the appropriateness of different assumptions and corresponding methods for analyzing the longitudinal relations among achievement and self-concept are necessary* (Ehm et al., 2019, p. 33).

In pursuit of this call for further research proposed by Ehm et al. (2019), our study is a methodological-substantive synergy. We juxtapose the application of RI-CLPMs, CLPMs, and extensions of these models to test REM hypotheses for a large longitudinal study of secondary students. Substantively we evaluate the REM for math self-concept (MSC), math achievement test scores, and math school grades over the first five years of secondary school. Methodologically, we critically evaluate and compare results for CLPMs and RI-CLPMs. In addition, we demonstrate extensions of basic CLPMs and RI-CLPMs: focusing on the measurement model, the inclusion of covariates, generalizability of multiple groups, invariance over time, and incorporation of three constructs (MSC, math achievement, and math test scores) into a single model (i.e., tripartite rather than the typical bivariate CLPMs and RI-CLPMs). For these purposes, we provide further analyses and extend the analyses presented by Marsh et al., (2018a, 2018b).

In theory, reciprocal effects linking ASC and achievement can be located at both within- and between-person levels. Specifically, from a cognitive-motivational perspective, effects of achievement on ASC, and effects of ASC on achievement, are built on within-person mechanisms. A typical causal process may involve the following steps. First, achievement (e.g., one's grade in math) is perceived by the individual student and then attributed to ability. Especially with cumulative success

or failure and consistent attributions to ability or lack of ability, respectively, these attributions lead to the formation of self-perceptions of ability that are stored in long-term memory and can be reported as self-concept of ability. Second, when confronted with achievement tasks, the student re-activates task-related ASCs from memory. These ASCs guide motivation to invest effort and make strategic choices, which, in turn, contributes to subsequent achievement (Marsh, Pekrun, Murayama et al., 2016). From this perspective, reciprocal effects of ASC and achievement are located within persons (i.e., within the individual brain) in the first place. However, when repeated over time, the within-person effects can contribute to between-person differences in achievement and ASC and drive the between-person effects that link between-person distributions of the two variables over time, as traditionally analyzed in between-person CLPMs.

Furthermore, it is unclear whether either traditional CLPMs or RI-CLPMs capture these within-person processes. In particular, although the RI-CLPM takes a within-person perspective, it does not actually posit within-person mechanisms to explain the reciprocal effects. Thus, for example, Niepel et al. (2021) argue that these traditional CLPM and RI-CLPM approaches leave the intraindividual dynamics (within-person processes) in a black box. Hence, although the RI-CLPM provides a within-person perspective in terms of the underlying statistical model (a residualized or person-centered statistical analysis), it does not test within-person mechanisms that lead to the reciprocal effects between ASC and achievement. Hence, terms such as person-centered or residualized models might more appropriately describe the RI-CLPM approach than the term within-person perspective. This is an important distinction in evaluating the strengths of the RI-CLPM. Of course, it is possible to extend both the CLPM and RI-CLPMs to test within-person mediating mechanisms. Thus, for example, Marsh, Hau, et al. (2005), Marsh, Trautwein, et al. (2005) evaluated the role of academic interest in tests of the REM based on the CLPM, showing that support for the REM was little affected by the inclusion of math interest measured at each wave of the design. They suggested the need for research that includes a variety of academic choice behaviors to evaluate better mediational processes underpinning the REM.

Methodological Focus: Models of Cross-lagged Panel Data and Reciprocal Effects

Cross-Lagged Panel Data

Cross-lagged panel data are used to test REM hypotheses about relations between MSC and math achievement. In panel designs, the same variables (MSC and math achievement) are measured repeatedly over time. Critical parameters are the stability paths (leading from one variable to the same variable in the next wave) and the cross-lag paths leading from one variable to the other variable in the next wave (e.g., effects of prior MSC on future math achievement, controlling for prior math achievement). The critical results are the directionality of the cross-lag effects—if

there is a directional ordering among the variables and whether it is unidirectional or reciprocal.

Although it is possible to test CLPMs with only two waves, basic RI-CLPMs require at least three waves, and even more waves are desirable. Because most tests of REMs are based on only two waves of data, studies typically considered only lag-1 effects (i.e., paths relating variables in adjacent waves). However, when there are three or more waves, it is possible to consider the invariance of effects over multiple waves and paths between non-adjacent paths (i.e., lag-2 effects). Thus, Marsh and colleagues (Arens et al., 2017; Marsh et al., 2017, 2018a, 2018b) argued that lag-2 effects are typical in CLPMs and should be included in CLPMs. They further noted that the improved fit was achieved primarily by adding lag-2 stability coefficients; lag-2 cross-paths were largely non-significant, whereas lag-1 cross-lag paths were relatively unaffected by the inclusion of additional lag-2 paths. lag-2 paths might have a theoretical basis (e.g., skills mastered in school year t may be directly relevant in year $t+2$ and beyond, in addition to their relevance to year $t+1$). However, including lag-2 paths also provides a more robust control for prior effects and potentially confounding covariates than models based on a single wave (Lüdtke & Robitzsch, 2021; Marsh, et al., 2018a, 2018b; VanderWeele et al., 2020). Hamaker and colleagues (Hamaker et al., 2015; Mulder & Hamaker, 2021) also suggested that CLPMs often have to include lag-2 effects to achieve a goodness-of-fit comparable to the RI-CLPMs. As suggested by Marsh and et al., (2018a, 2018b), for us, the critical issues are whether the addition of lag-2 paths improved fit and particularly whether their inclusion altered the interpretation of the cross-lag paths and support for REMs—a sensitivity test.

Recent research has contrasted a wide variety of complex statistical models that can be applied to cross-lag-panel data (Kenny & Zautra, 1995; McArdle, 2009; Orth et al., 2021; Usami, Murayama, et al., 2019; Usami, Todo, et al., 2019; Zyphur et al., 2020). However, comparisons of the different models based on multiple data sets (Orth et al., 2021) or simulated data (Usami Murayama, & Hamaker, 2019; Usami, Todo, et al., 2019) showed that only RI-CLPMs and particularly CLPMs consistently converged to proper solutions. Although CLPMs consistently converged to proper solutions, RI-CLPMs sometimes did not—even when the RI-CLPM structure was used to generate the simulated data (Usami, et al., 2019a, 2019b). Furthermore, Orth et al. reported that the CLPM produced more consistent cross-lagged effects both within and between samples.

Because CLPMs are nested under RI-CLPMs, RI-CLPMs necessarily result in a better fit for indices that do not correct for CLPMs' greater parsimony. This is sometimes used to argue in favor of RI-CLPMs over CLPMs. However, Orth et al. (2021) noted that the choice of models should also be based on theoretical grounds and appropriate interpretations of the results rather than only goodness-of-fit. Hence, goodness-of-fit should be only one of the considerations in the choice of models and their interpretation. Furthermore, Marsh and et al. (2018a, 2018b, p. 271; also see Lüdtke & Robitzsch, 2021) argued that the addition of lag-2 paths substantially improved the fit of a CLPM and served for "providing stronger controls for preexisting differences." Although they did not consider a RI-CLPM, the fit of their CLPM with covariates and lag-2 effects approached the fit of the corresponding measurement model (in which all constructs

were merely correlated). Because the RI-CLPM and CLPM are both nested under the measurement model, this suggests that their extended CLPM with lag-2 effects would have fit their data as well as a RI-CLPM. If this were the case more generally, goodness-of-fit would no longer be a critical issue in comparing the RI-CLPM and the extended CLPM with lag-2 effects. In the present investigation, we pursue this issue in our juxtaposition of the two models.

Distinguishing Between RI-CLPM (Within-Person) and CLPM (Between-Person) Perspectives

Historically, tests of the REM in ASC research have been based almost entirely on CLPMs, but, following Hamaker and colleagues (Hamaker & Muthén, 2020; Hamaker et al., 2015, 2020; Mulder & Hamaker, 2021), there has been a recent surge in the popularity of RI-CLPMs. Nevertheless, Orth et al. (2021) emphasized that the two models address different questions, result in different interpretations, and are based on different assumptions. Because both models and their juxtaposition are relevant, we argue that it is crucial to understand how the underlying rationales of these two models differ (see Fig. 1).

Structural Characteristics

The primary structural difference between the two models is that RI-CLPMs include a stable trait factor (Tx, Ty, and Tz in Fig. 1), whereas CLPMs do not. In this sense, CLPMs are nested under the RI-CLPM. CLPMs evaluate how the effects of individual differences at each wave are related to those in subsequent waves (an undecomposed between-person perspective). RI-CLPMs evaluate how within-person deviations at each wave differ from a student's stable trait (a decomposed between-person difference), and how these within-person differences from one wave are related to those in the next wave (a within-person perspective).

Importantly, CLPMs and RI-CLPMs differ in the interpretation of the term "between-person." In CLPMs, between-person effects reflect a combination of within-person (i.e., deviations from a global trait) and between-person (e.g., stable trait) effects. This is consistent with the term's use in most individual difference studies of relations among variables and most cross-sectional studies. However, RI-CLPMs decompose these effects into separate components reflecting within- and between-person components. Thus, within the context of each model, the use of the generic term between-person is appropriate. However, to avoid confusion, we refer to these as "decomposed" between-person effects (RI-CLPM) and "undecomposed" between-person effects (CLPM).

For our study based on three latent constructs, the critical parameters for both RI-CLPMs and CLPMs are the auto-regressive stability paths (Bxx, Byy, and Bzz in Fig. 1), and particularly the cross-paths relating achievement and MSC (Bxy, Byx, Bxz, Bzx in Fig. 1), but also cross-paths relating the two indicators (school grades and test scores) of math achievement (Byz and Bzy). If both sets of paths leading from achievement to MSC and from MSC to achievement are statistically significant, the variables are said to be reciprocally related. If only one of the sets of paths

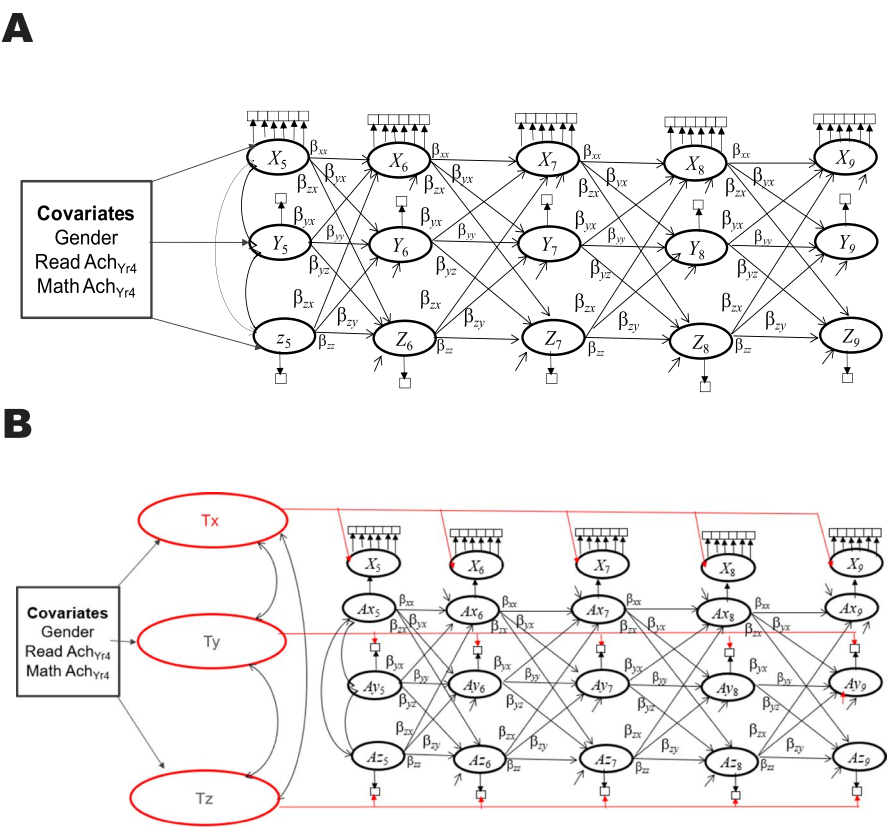

Fig. 1 Diagram of cross-lag-panel-model (A) and random intercept cross-lag-panel-model (B) with covariates. Three constructs were measured in the first five years of secondary school (Years 5 – 9): X = math self-concept; Y = Math grade; Z = Math standardized test. Covariates included verbal and math achievement from the last year of primary school (Year 4). Math self-concept was based on responses to 6 items, but all other constructs were single-item constructs. Excluded in order to avoid clutter are correlated uniquenesses relating responses to the same math self-concept item administered in different years, and correlated residual covariances among the three constructs in Years 6 – 9

is significant (and differs significantly from the other path), the directional ordering is said to be unidirectional rather than reciprocal. Most recent CLPMs tests of REMs are latent, at least for the ASC construct (i.e., there are multiple indicators of the ASC factor—the unlabeled boxes in Fig. 1). However, the latent versions of the RI-CLPM have been developed only recently, so this model has few substantive applications (Mulder & Hamaker, 2020; also Ehm et al., 2019, 2021).

We also note that most applications of latent RI-CLPMs are bivariate models, based on two constructs (e.g., MSC and test scores) or separate analyses of each pair of constructs when more than two are considered (e.g., Ehm et al., 2019, 2021). However, our focus is on relations among three constructs (MSC, test scores, and school grades). Hence we extend the basic bivariate RI-CLPMs and CLPMs to include three constructs: tripartite RI-CLPMs and CLPMs. Our study is one of only

a few applications of the tripartite RI-CLPMs with latent variables, particularly for REM studies of ASC and achievement (but see Van Lissa et al., 2021; Burns et al., 2019; also see Hübner et al., 2022; we elaborate on the importance of this contribution in Discussion section).

RI-CLPMs and CLPMs Address Different Research Questions

For longitudinal panel data, the multiple waves (level 1) are nested under the person (level 2). Here, the level 2 variables in RI-CLPMs are the average levels of ASC and achievement over time for each student (i.e., the random intercepts or latent means; Hamaker et al., 2015). The assumption is that for a given student, these represent stable traits that are consistent over time for the duration of the study but differ from student to student. In contrast, the CLPM is like a single-level model that evaluates relations between ASC and achievement within-waves and over time without controlling for person-level differences in these variables (an undecomposed between-person focus that does not separate within-person and between-person effects). The RI-CLPM estimates relations between these variables after controlling (decomposed) between-person stable trait effects, person-level intercepts; that is, it provides a within-person perspective.

To simplify this distinction, we focus on MSC (X in Fig. 1) and achievement, which could refer to either math school grades (Y in Fig. 1) or math test scores (Z in Fig. 1), or both. However, these same distinctions also apply to other pairs of variables. It is important to note that the undecomposed between-person effects in CLPMs are the effects of individual differences in some construct X (MSC) on change in the individual differences in achievement (Y or Z in Fig. 1). This is based on the covariance of X (wave-1) and achievement (wave-2), controlling wave-1 variables. Thus, for CLPMs, change is based on the residual change in individual differences in achievement from wave-1 to wave-2 and aims to predict change in individual differences (or rank-order change). In contrast, change in the RI-CLPM is based on how within-person deviations for X at wave-1 (i.e., the difference between X at wave-1 from the latent mean of X across all waves) are related to deviations in achievement at wave-2, controlling deviations in wave-1 variables. These two perspectives are easily confounded, but are related to different research questions and often lead to different interpretations. To make this distinction more concrete, we offer the following research questions that are appropriate for each model:

- **CLPMs:** When students have high MSCs (compared to other students), are they more or less likely to experience a subsequent rank-order increase in math school grades (compared to other students)? Likewise, when students have a high achievement (compared to other students), are they more or less likely to experience a subsequent rank-order increase in MSC (compared to other students)? Thus, do individual differences in MSC positively predict rank-order change in relative achievement, and do individual differences in achievement positively predict rank-order change in relative MSC?
- **RI-CLPMs:** When students experience higher than their usual MSC (compared to their long-term average MSC over the duration of the study), are they more or less likely to experience a subsequent higher than their usual

achievement (compared to their long-term average achievement over the duration of the study)? Likewise, when students experience higher than their usual achievement (compared to their long-term average achievement), are they more or less likely to experience a subsequent higher than their usual MSC (compared to their long-term average MSC)?

The Role of Covariates

For both RI-CLPMs and CLPMs, the possible confounding of effects with unmeasured covariates is a potentially serious threat to interpreting results. Advocates of the RI-CLPM argue that the critical advantage of this approach is that it provides greater protection for time-invariant (between-person) covariates that are not measured as part of the study (e.g., Hamaker et al., 2015; Mulder & Hamaker, 2020). The logic is that true time-invariant (between-person) covariates will only affect the global trait factors (Tx, Ty, & Tz in Fig. 1), which are statistically independent of the within-person autoregressive factors (Ax, Ay & Az in Fig. 1). Thus, unmeasured (undecomposed between-person) time-invariant covariates might affect the sizes of the (decomposed between-person) global trait factors. However, they do not affect the within-person autoregressive factors used to test directional ordering (i.e., statistically, the random intercept factor and the within-person components are independent; Hamaker et al., 2015). This has critical interpretational advantages, particularly concerning tests of directional ordering.

RI-CLPMs can incorporate measured covariates as part of the study. For these measured covariates, Mulder and Hamaker (Mulder & Hamaker, 2020) proposed that these should be regressed on the manifest trait scores for variables measured by a single indicator (school grades and test scores in the present investigation). Although they did not consider a fully latent model with covariates, we interpret their approach to mean that covariates are regressed on the undecomposed latent variables (i.e., the X latent factor representing MSC in Fig. 1) rather than the manifest indicators of each trait, the within-person autoregressive factors (the Ax factors in Fig. 1). However, for a fixed covariate with time-invariant effects, all or most of the time-invariant effects will be absorbed by the global trait factors (Tx, Ty & Tz in Fig. 1).

CLPMs also control for measured variables by including them as additional covariates in the statistical models. However, CLPMs assume that all relevant covariates are measured or captured by those included covariates, a selection-on-observables strategy (Little, 2013; Reichardt, 2019). For example, if we are interested in estimating the cross-lagged effect of $X_{t\text{-}1}$ on Y_t, we need to assume that all relevant variables that affect $X_{t\text{-}1}$ as well as Y_t are included. Covariates that are constant across the investigated time period (e.g., demographic variables, achievement in primary school) can be easily included in CLPMs as additional predictors of X_t and Y_t at each wave t. This is also the case for covariates Z_t that vary across time (e.g., grades, math interest), even though the issue of time-varying covariates has received less attention in the application of CLPMs. One challenging aspect of time-varying covariates is that they should not be affected by the treatment. For example, if we

estimate the effect of X_{t-1} (e.g., ASC at wave t–1) on Y_t (e.g., test scores at wave t) and include Z_{t-1} (e.g., grades at wave t–1) as a time-varying covariate in the CLPM, we need to rule out that Z_{t-1} has not been affected by $X_{t-1.}$ Otherwise, Z_{t-1} would not be a confounder but a mediator (i.e., a variable that is on the causal pathway from X_{t-1} to Y_t). In practical applications, it is often difficult to decide whether Z_{t-1} acts as a mediator or confounder, particularly when the time-varying covariate Z_{t-1} is measured at the same time as the treatment X_{t-1} (e.g., see Marsh, Hau, et al., 2005; Marsh, Trautwein, et al., 2005, on the role of time-varying measures of academic interest in REM tests of reciprocal effects between MSC and achievement).

Overall, the CLPMs are based on the assumption that all relevant covariates (time-invariant and time-varying) are measured (or controlled by those that are measured). Thus, compared to RI-CLPMs—which control the effects of time-invariant confounders—CLPMs provide less protection against the confounding effects of unmeasured (time-invariant) confounders. However, both models provide limited protection concerning unmeasured time-varying covariates or fixed-covariates whose effects vary over time.

Neither CLPMs nor the RI-CLPMs provide particularly good controls for unmeasured time-varying covariates (or fixed covariates measured only once with effects that vary from wave to wave, possibly reflecting an unmeasured process). However, support for the consistency of effects over waves (based on invariance tests) suggests that these potentially confounding covariates specific to a particular wave do not substantially affect the results. Furthermore, Marsh et al., (2018a, 2018b) suggest that the extension of the CLPM to include lag-2 effects provides stronger controls for covariates.

The most effective way to control the effects of unmeasured (fixed and time-varying) covariates is to measure them and include them in statistical models. Thus the selection of covariates is crucial for RI-CLPMs and particular CLPMs. Hence, we find it surprising that this issue has been given limited attention in the design, analysis, and interpretation of these models (see VanderWeele, 2019, for discussion of alternative strategies for selecting covariates; also see Hübner et al., 2022).

Cattell's (1966) Data Cube Cattell's (1966) data cube helps distinguish within- and between-person perspectives. It represents data concerning three dimensions: persons, variables, and occasions (Marsh & Grayson, 1994). Voelkle et al. (2014) emphasize that the vast majority of educational and psychological studies focus on relations between variables across persons (undecomposed between-person variation) and, to a much smaller extent, relations on variables across occasions (decomposed within-person variation). In his classic manifesto on an idiographic approach to psychology, Molenaar (2004) noted the need to shift from a focus on the level of inter individual variation in the population to the level of intra individual variation characterizing the life histories of individual subjects. Noting the importance of between-person questions (e.g., MSC predicts achievement), Voelkle et al. (2014) also emphasized the importance of evaluating the consistency of within-person effects over time for a given individual. However, they emphasized that interindividual and intraindividual effects are only likely to be equivalent under very limited conditions (ergodicity): stationarity (invariance over time of means, variances, and covariances) and homogeneity (same relations between variables for all individuals,

such that the same generating model can be applied to all individuals and individuals are not grouped or nested). Because these conditions are unlikely to be met, Voelkle et al. (2014) argued that rather than seeing within- and between-person analyses as competing research paradigms, the focus should be on appropriate research questions and, perhaps, the juxtaposition between interpretations based on the two models.

Juxtaposing CLPMs and RI-CLPMs

CLPMs and RI-CLPMs are very different. They differ in terms of their conceptual and statistical underpinnings and the questions they address. Historically, most studies of unidirectional, bidirectional, and reciprocal effects have used CLPMs (an undecomposed between-person perspective). However, there has been a surge in the popularity of RI-CLPMs (a within-person model based on a decomposed between-person perspective) and much discussion about the relevance of each.

The inability of the traditional CLPM model to disaggregate within (i.e., state-like) and between (i.e., trait-like) effects (Curran & Bauer, 2011) has led to important criticisms of this approach (e.g., Berry & Willoughby, 2017; Hamaker et al., 2015; Mund & Nestler, 2019). These criticisms have led to the emergence of a wide variety of new models specifically designed to address this limitation (e.g., Biaconcini & Bollen, 2018; Curran et al., 2014; Hamaker et al., 2015; Mund & Nestler, 2019; Zyphur et al., 2020). As a result, the recent surge in popularity of the RI-CLPM has created a zeitgeist in which some educational psychology researchers suggest that the decomposed between- and within-person perspective provided by the RI-CLPM is always more appropriate (Nunez-Reueiro et al., 2021; but also see discussion by Asendorpf, 2021; Orth et al., 2021).

However, several researchers recently critiqued considering the RI-CLPM as the default approach to analyzing cross-lagged panel designs (Asendorpf, 2021; Hübner et al., 2022; Lüdtke & Robitzsch, 2021; Orth et al., 2021). More specifically, Lüdtke and Robitzsch (2021) studied the RI-CLPM from a causal inference perspective (e.g., Imbens & Rubin, 2015; Pearl et al., 2016) using mathematical derivation and simulated data. Their overall goal, consistent with our study, was "to provide a more balanced discussion of two main approaches (CLPM and RI-CLPM) for analyzing cross-lagged panel designs, and we would like to emphasize that—despite recent methodological recommendations—there are still good reasons to use the traditional CLPM when estimating cross-lagged effects" (p. 3). Using simulated data, they showed that the RI-CLPM has limited ability to control for unmeasured confounder variables, including fixed confounders (e.g., demographic variables), when their effects vary over time. Drawing in part on early research by Marsh and et al., (2018a, 2018b), Lüdtke and Robitzsch noted that beneficial consequences of including lag-2 effects to provide a stronger control for confounding (also see VanderWeele et al., 2019; 2020). They also noted that the inclusion of lag-2 effects in CLPMs resulted in goodness-of-fit that was as good as RI-CLPMs. Thus, the choice of models is not a question of fit, and this positions CLPMs with lag-2 effects as a viable alternative to the RI-CLPM, even in terms of goodness-of-fit. Noting that there are still issues

with CLPMs, even with the inclusion of lag-2 effects, they argued for a selection-on-observables CLPM approach based on the observed information in the data (previous measures of the treatment and outcome, and additional covariates), instead of stable trait factors (that are based on modeling assumptions in RI-CLPMs). This approach is consistent with VanderWeele et al.'s (2020; see also VanderWeele et al., 2016) perspective on causal inference with longitudinal data and has also been recently emphasized by Hübner et al. (2022).

CLPMs and RI-CLPMs address different questions and often result in different—even contradictory—interpretations. Because the strengths, weaknesses, and appropriate interpretations of each are the basis of ongoing research and extensions of these models, applied researchers need to understand the differences between the two. In the present investigation, we intend merely to present a more balanced view of the different perspectives concerning the issue of the causal ordering of ASC and achievement based on appropriate cross-lagged panel data. We aim this presentation to applied researchers based on what might be the classic application of the cross-lagged panel design in educational psychology, the REM that dates back to at least the 1970s (e.g., Calsyn & Kenny, 1977) and has been the basis of many studies and multiple meta-analyses. Our study is one of the first to juxtapose the theoretical rationale and results based on CLPMs, RI-CLPMs, and extensions of these models in tests of the directional ordering of ASC and achievement. In pursuit of these aims, we operationalize extensions of particularly RI-CLPMs that have mainly been demonstrated with simulated data, consider lag-2 effects, and show their usefulness in our applied study (VanderWeele et al., 2020).

The Present Investigation: Two A Priori Research Hypotheses and Three Research Questions

In the present investigation, we chose for purposes of secondary data analysis what we judged to be the strongest database to juxtapose CLPMs and RI-CLPMs relating ASC, school grades, and achievement tests across secondary school years. The Project for the Analysis of Learning and Achievement in Mathematics (PALMA; Arens et al., 2017; Frenzel et al., 2012; Marsh et al., 2017, 2018a, 2018b; Marsh, Parker, et al., 2016; Marsh, Pekrun, et al., 2016; Pekrun, 2006; Pekrun et al., 2017, 2019) is a large-scale longitudinal study investigating the development of math achievement and its determinants during secondary school years. Although the directional ordering of achievement and math self-concept has been a component of previous PALMA research, this has always been from a between-person perspective. In this sense, PALMA is ideally suited for our purpose of juxtaposing CLPMs and RI-CLPMs in support of the REM. Here we extend these models to test longitudinal invariance over time (multiple school years) and multiple groups (school tracks), lag-2 paths between non-adjacent school years, and covariates (gender; primary school math and verbal achievement). CLPMs and particularly RI-CLPMs are typically based on two variables (bivariate models). However, here we extend the models to include three variables (tripartite models; MSC and the two distinct forms of achievement). The

key issues here involve juxtaposing between- and within-student perspectives on the directional ordering of three variables: MSC and two achievement indicators.

Based on our review of the substantive literature on achievement and MSC, we offer the following two research hypotheses (where there is a clear basis for offering a priori directional hypotheses). In addition, we offer three additional research questions that involve critical issues and extensions of the basic CLPMs and RI-CLPMs for which there is not sufficient basis for offering a priori hypotheses. For the research questions, we also discuss the relevant issues.

Research Hypotheses

Research Hypothesis 1 Directional-Ordering: Model of Reciprocal Effects. We hypothesize a priori that students' MSC will be reciprocally related to both measures of achievement (grades and test scores). The paths from MSC in one wave to both achievement measures in subsequent waves will be significantly positive. Likewise, the paths from both achievement measures in one wave to MSC in the next wave will be significantly positive (see Fig. 1). Our hypotheses are consistent with the REM and extensive research showing that MSC and achievement are reciprocally related (Huang, 2011; Marsh & Craven, 2006; Valentine et al., 2004; Wu et al., 2021). Also consistent with previous research, we view these hypothesized cross-lagged effects as "causal" (but see earlier discussion of the rationale for using the term "directional ordering").

Research Hypothesis 2: Alternative Measures of Achievement. Based on prior research and following from Research Hypothesis 1, we hypothesize a priori that all stability and cross-paths will be positive and statistically significant in models based on test scores, school grades, and the combination test scores and school grades. Also, following previous research (e.g., Marsh, 2007; also see Wu et al's., 2021, meta-analysis), we anticipate that MSC within and across waves will be more highly correlated with school grades than test scores. Compared to models relating MSC to each of these achievement indicators separately, we anticipate that the combined model based on all three will have smaller paths—particularly stability paths for the two achievement measures, and cross-paths relating the two achievement measures and MSC. However, we hypothesize that support for Research Hypothesis 1 will generalize over all models based on alternative measures of achievement.

Research Questions: Juxtaposition and Extensions of CLPMs and RI-CLPMs to Evaluate Sensitivity

Research Question 1: Juxtaposition of CLPM and RI-CLMP Results. The central research question is whether support for the REM (Research Hypothesis 1) differs for CLPMs and RI-CLPMs. Although there is overwhelming support for the REM based on prior research, we note that particularly at the secondary school level, this support is largely based on

CLPMs rather than RI-CLPMs. However, as Hamaker et al. (2015) and others (also see earlier discussion) emphasize, there is no a priori basis for anticipating how estimates from the two models will differ in size or even direction. Hence, we leave this issue as a research question.

Research Question 2: Extended Models: Lag-2 Effects. For both CLPMs and RI-CLPMs, we evaluated extended models with lag-2 effects. Following earlier discussion (e.g., Marsh, et al., 2018a, 2018b; also see Lüdtke & Robitzsch, 2021), this extension should substantially improve the fit of the CLPMs. However, we see this extension of the models as a sensitivity test to determine whether the inclusion of lag-2 effects influences support for Research Hypothesis 1—a substantial issue. Nevertheless, an important methodological contribution is to test the supposition following from Marsh and et al., (2018a, 2018b; also see Lüdtke & Robitzsch, 2021) that adding lag-2 paths to the CLPM will largely eliminate differences in goodness-of-fit and provide stronger controls for preexisting differences.

Research Question 3: Extended Models: Covariates and Multiple Groups. We evaluated extended models for CLPMs and RI-CLPMs that included controls for covariates (gender; prior verbal and math achievement from primary school, before starting secondary school) and multiple groups (school tracks). Although there is substantive interest in how these effects are related to students' achievement and MSC, our primary focus is on how these covariates affect the results concerning directional ordering. Thus we again see this extension of the models as a sensitivity test in relation to support for Research Hypothesis 1, and thus a research question.

Method

Sample

Our study is based on secondary data analysis of data from PALMA, a large-scale longitudinal study investigating the development of math achievement and its determinants during secondary school in Germany. The Data Processing and Research Center (DPC) of the International Association for the Evaluation of Educational Achievement (IEA) conducted sampling and the assessments. Samples were drawn from secondary schools within the state of Bavaria and were representative of the student population of this state in terms of student characteristics such as gender, urban versus rural location, and family background (SES; for details, see Pekrun et al., 2007). The data consisted of five measurement waves spanning Grades 5 to 9 and school grades from the last year of primary school (Year 4). On the basis of the primary school results, students ($N = 3{,}425$; 50% girls; mean age = 11.7 at Year 5, SD = 0.7) were allocated to either the high-achievement (Gymnasium: 37%), middle-achievement (Realschule: 30%), or low-achievement (Hauptschule: 33%) school tracks. Students answered the questionnaire in the first two weeks of July, toward the

end of each successive school year. All instruments were administered in the students' classrooms by trained external test administrators. Participation in the study was voluntary, and parental consent was obtained for all students. The agreement was high (100% for schools and over 90% for students at each data wave), and the final sample closely represented the intended sample and population more generally (Pekrun et al., 2007). Surveys were anonymized to ensure participant confidentiality.

Measures

MSC was measured in each of the five secondary schools Years (5–9) with the same set of six items, using a 5-point Likert scale: not true, hardly true, somewhat true, largely true, or absolutely true. Across the five waves, the alpha estimates of reliability were consistently high (Year 5 $\alpha = 0.88$; Year 6 $\alpha = 0.89$; Year 7 $\alpha = 0.89$; Year 8 $\alpha = 0.91$; Year 9 $\alpha = 0.92$). The items used to measure MSC were: "In math, I am a talented student"; "It is easy for me to understand things in math"; "I can solve math problems well"; "It is easy for me to write tests/exams in math"; "It is easy for me to learn something in math"; "If the math teacher asks a question, I usually know the right answer."

Students' achievement was measured with school grades (math in Years 4–9; German in Year 4) and math standardized achievement test scores (Years 5–9). School grades were end-of-the-year final grades obtained from school documents. The standardized PALMA Math Achievement Test (Murayama et al., 2013; Pekrun et al., 2007) was based on multiple-choice and open-ended items to measure students' modeling and algorithmic competencies in arithmetic, algebra, and geometry. The test was constructed using multi-matrix sampling with a balanced incomplete block design; the number of items increased with each wave, varying between 60 and 90 items across the five waves, with anchor items to allow for the linkage of the two test forms and the five measurement points. The achievement scores were scaled using one-parameter logistic item response theory, confirming the unidimensionality and longitudinal invariance of the test scales (Murayama et al., 2013).

Statistical Analyses

All analyses were done with Mplus (Muthén & Muthén, 2008–20, Version 8). We used the robust maximum likelihood estimator (MLR), which is robust against many violations of normality assumptions.

Missing Data. As is typical in large longitudinal field studies, a substantial portion of the sample had missing data for at least one measurement wave due primarily to absence or students changing schools. Across the five waves, 39% participated in all five measurement waves (i.e., Grades 5 to 9), and 9%, 19%, 15%, and 18% took part in four, three, two, or one of the assessments, respectively. We included all students who responded to at least one wave. Particularly in longitudinal studies, there is increasing awareness of the limitations of traditional approaches to missing data (Enders, 2010). Here, we applied the full-information maximum likelihood (FIML)

method to fully use cases with missing data (Enders, 2010). FIML results in trustworthy, unbiased estimates for missing values even in the case of large numbers of missing values (Enders, 2010) and is an appropriate method to manage missing data in large longitudinal studies (Jeličič et al., 2009). More specifically, as emphasized in classic discussions of missing data (e.g., Newman, 2014), under the missing-at-random (MAR) assumption that is the basis of FIML, missingness is allowed to be conditional on all variables included in the analyses, but does not depend on the values of variables that are missing. This implies that missing values can be conditional on the same variable's values collected in a different wave in a longitudinal panel design. This feature of the data makes it unlikely that MAR assumptions are seriously violated, as the key situation of not-MAR is when missingness is related to the variable itself. Hence, having multiple waves of parallel data provides strong protection against this violation of the MAR assumption. Also, the appropriateness of FIML is further strengthened by support for the invariance of parameter estimates over time (see subsequent discussion of invariance constraints).

Goodness-of-Fit. Applied SEM studies typically focus on fit indices that are relatively sample-size independent (Marsh et al., 2004; Marsh, Hau, & Grayson 2005), such as the root-mean-square error of approximation (RMSEA), the Tucker-Lewis index (TLI), and the comparative fit index (CFI). Population values of TLI and CFI vary along a 0-to-1 continuum, in which values greater than .90 and .95 typically reflect acceptable and excellent fits to the data, respectively. Values smaller than .08 and .06 for the RMSEA support acceptable and good model fits, respectively. For comparing nested models, Cheung and Rensvold (2002) and Chen (2007) suggested that if the decrease in fit for the more parsimonious model is less than .01 for incremental fit indices such as the CFI, there is reasonable support for the more parsimonious model. For indices that incorporate a penalty for lack of parsimony, such as the RMSEA and the TLI, it is also possible for a more restrictive model to result in a better fit than would a less restrictive model. For present purposes, to facilitate communication, we present primarily TLIs in the written summaries of the results. In addition, however, we present the Chi-square, degrees-of-freedom, RMSEA, CFI, and TLI in the corresponding tables. Nevertheless, these cut-off values for these indices constitute only rough descriptive guidelines rather than "golden rules" (Marsh et al., 2004).

Invariance Constraints

A Well-Defined Measurement Model. Particularly as multiple MSC indicators are parallel over the multiple waves, it is relevant to test measurement invariance over time and multiple groups (the three school tracks). For both longitudinal and multiple group data, it is typical to evaluate a set of models that systematically vary the invariance constraints (Marsh et al., 2014; Marsh et al., 2014; Marsh, Morin, et al., 2014; Marsh, Parker, et al., 2016; Meredith, 1993; Millsap, 2012): configural (no invariance constraints), metric (factor loading invariance), and scalar (intercept invariance). For longitudinal data, Marsh et al. (2013) recommended that correlated

uniquenesses relating residual variance terms for the same item in different waves should also be tested (Jöreskog, 1979; Marsh & Hau, 1996). Failure to include them will typically undermine goodness-of-fit and bias parameter estimates. Thus, the measurement model's invariance is relevant to the rationale underlying statistical models of longitudinal data. These invariance constraints also substantially reduce the number of estimated parameters, resulting in a more parsimonious model. Particularly in complex models, this can also improve the convergence behavior of models and increase power. There are two aspects to invariance in the present investigation: invariance over the multiple groups (different academic tracks) and invariance over time.

Importantly, these preliminary tests of the measurement model are not based on any particular model (e.g., CLPM or RI-CLPM) but merely evaluate the extent to which the constructs are well-defined. Nevertheless, unless there is reasonable support for at least configural invariance, the application of subsequent CLPMs and particularly RI-CLPMs is dubious. If there is no support for the invariance of factor loadings, then tests of invariance of other parameters associated with these factors (e.g., stability and cross-paths in CLPMs and RI-CLPMs) are also dubious.

Longitudinal Structural Invariance Constraints For the CLPM studies based on three or more data waves, Marsh and et al. (2018a, 2018b) adopted the term "development equilibrium" for the imposition of invariance for the stability paths and the cross-lagged paths over time. Their focus was mainly on whether the sizes of the stability and cross-paths varied as a function of age for school students. However, this pattern of constraints is also typical in RI-CLPMs (Mulder & Hamaker, 2020), greatly facilitating the interpretation of results. Nevertheless, because this terminology is somewhat idiosyncratic to development studies, we refer to this set of constraints as longitudinal equilibrium. In invoking this constraint, we constrained to be invariant over time the three stability paths (Byy, Bxx, & Bzz; Fig. 1) and the six cross-lag paths (Bxy, Byx, Bxz, Bzx, Byz, Bzy; Fig. 1).

PreliminaryAnalyses

Longitudinal Invariance We tested a series of measurement models based on invariance over time. The models were based on responses to 40 indicators—6 MSC items, one math test score, and one math school grade in each of five waves (i.e., 8 indicators × 5 waves). For present purposes, to facilitate interpretations, all items for MSC were standardized ($Mn = 0$, $SD = 1$) to a common metric, based on wave-1 responses (i.e., Year-5, the first year of secondary school). Marsh et al. (2013) recommended that our a priori model included correlated uniquenesses relating residual variance terms for the same items at different waves (for further discussion, see Marsh et al., 1996; Joreskog, 1979). As expected, the measurement model with no correlated uniqueness (MM0 in Table 1) provided a poorer fit than other measurement models. Measurement model MM1 (configural invariance) model with correlated uniquenesses but no invariance constraints provided a very good fit to the

Table 1 Goodness-of-Fit for Confirmatory Factor Analysis (CFA) Measurement Model: Invariance of the Measurement Factor Structure Over multiple Waves and Multiple Groups

CFA Model	Chi-SQ	Df	RMSEA	CFI	TLI
Longitudinal Invariance					
MM0 No-Correlated Uniquenesses No Invariance	2133	645	.026	.971	.965
MM1 configural M0 with Correlated Uniquenesses	1133	585	.016	.989	.986
MM2 metric M1 with factor loadings invariant	1288	608	.018	.987	.982
MM3 Scalar M2 with intercept invariance	1497	629	.020	.983	.979
Multiple Group Invariance					
MM4 configural with Correlated Uniquenesses	2444	1755	.018	.987	.982
MM5 metric M4 with factor loadings invariant	2496	1805	.018	.987	.983
MM6 Scalar M5 with *i*ntercept invariance	2608	1855	.019	.985	.982
Longitudinal & Multiple Group Scalar Invariance					
MM7 Longitudinal & Multiple Group Scalar Invariance	2900	1895	.021	.980	.976
Covariates					
MM2 + Covariates	1773	754	.020	.983	.978

Summary of Goodness-of-fit statistics for the different factor analyses considered in the present investigation. CFA = confirmatory factor analysis; Chi-SQ = Chi-square; df = degrees of freedom; CFI = comparative fit index; TLI = Tucker-Lewis Index; RMSEA = Root-Mean-Square Error of Approximation. Model; INV = invariance constraints (constraining parameters to be invariant over time); CU = correlated uniqueness (relating residual variances associated with the same item over the multiple waves)

data (RMSEA = .016, CFI = .989, TLI = .986; see results in Table 1). In Model MM2 (metric invariance) with the imposition of factor loading invariance also resulted in a good fit (RMSEA = .018, CFI = .987, TLI = .982). In the final model MM3 (scalar invariance) model, the imposition of intercept invariance resulted in a slightly poorer fit (RMSEA = .020, CFI = .983, TLI = .979), but one that was still excellent based on traditional guidelines. The results demonstrate that the factor structure generalizes well over the multiple waves—the first five years of secondary school.

Multi-Group Invariance Next, we tested a series of measurement models based on invariance over the multiple groups (representing the three academic tracks). We based these models on the same data as the tests of longitudinal invariance. Measurement model MM4 (configural invariance) model with correlated uniquenesses but no invariance constraints over the multiple groups provided a very good fit to the data (RMSEA = .018, CFI = .987, TLI = .982; Table 1). In Model MM5 (metric invariance) with the imposition factor loading invariance also resulted in a good fit (RMSEA = .018, CFI = .987, TLI = .983) as did MM6 (scalar invariance) model (RMSEA = .019, CFI = .985, TLI = .982). The results demonstrate that the factor structure generalizes well over the multiple groups.

Combining Longitudinal Invariance and Multi-Group Invariance When considered separately, our results support scalar invariance over time (longitudinal invariance) and academic track (multigroup invariance). In the final model (MM7; also see Supplemental Materials for Mplus syntax), we simultaneously impose scalar

invariance for both time and group. The fit of this highly constrained model is very good (TLI = .976; Table 1), further demonstrating that the factor structure generalizes over the multiple wave and multiple groups.

We use this final measurement model MM7 for subsequent CLPMs and RI-CLPMs. To facilitate interpretations, we identified all solutions by fixing the factor loading of the first indicator of each MSC factor to a constant value. However, instead of fixing the value to 1.0, we fixed it to the standardized factor loading in the scalar invariance solution. This results in a model in which factor loadings are invariant over time and group. The factor variance is 1.0 in wave-1, but is allowed to vary across waves. In this way, all responses are standardized relative to a common metric (that facilitates comparing parameter estimates in different waves), resulting in an unstandardized solution similar to a standardized solution. This parameterization is particularly useful in the comparison of different cross-lag paths associated with different constructs.

We also note that all subsequent CLPMs and RI-CLPMs considered here are nested under our measurement model MM7. Hence, the measurement model MM7 provides an important basis of comparison for all subsequent CLPMs and RI-CLPMs, the structural invariance constraints imposed on them, and their extensions to include additional lagged parameters. More specifically, because all relations among the 15 factors (MSC, school grades, and test scores in each of the five waves) are freely estimated, this model MM7 is fully saturated in terms of these relations. In contrast, all the various CLPMs and RI-CLPMs place constraints upon these relations. To the extent that the constraints are reasonable, the fit of the constrained CLPMs and RI-CLPMs should approach that of our measurement model MM2. We consider this a fundamentally important contribution to evaluating fit for the CLPMs and RI-CLPMs that is rarely considered—even in studies based on fully latent CLPMs and RI-CLPMs.

Results

Relations Among the Variables

Table 2 is a latent correlation matrix of relations between the 15 factors (MSC, school grades, and test scores in each of the five waves). This is a latent multitrait–multimethod (MTMM) correlation matrix in which time is the "method" factor (for further discussion, see Marsh & Huppert et al., 2020; Marsh, et al., 2010). Thus, the results indicate that all three constructs are highly stable and consistent over the five waves. The average lag-1 correlations (i.e., test–retest correlations in adjacent waves separated by one year) for matching traits are .71 (.68-0.75) for MSC, .82 (.77-.86) for test scores, and .64 (.58-.68) for school grades. Reflecting the typical simplex pattern, lag-2 test–retest correlations are somewhat smaller, but still substantial for all three constructs. Indeed, Year 5 factors are significantly correlated even with Year 9 factors, particularly for test scores ($r = .71$) but also for MSC

Table 2 Latent correlations between Math Self-concept (MSC), Test Scores, and School Grades over five waves (school years 5—9) and background/demographics

Variables	MSC					Test					MGrade					Back				
MSC-Yr5	1																			
MSC-Yr6	.68	1																		
MSC-Yr7	.57	.68	1																	
MSC-Yr8	.53	.60	.73	1																
MSC-Yr9	.50	.55	.65	.75	1															
MTest-Yr5	.38	.33	.29	.28	.24	1														
MTest-Yr6	.33	.33	.31	.30	.25	.77	1													
MTest-Yr7	.29	.31	.33	.32	.28	.74	.80	1												
MTest-Yr8	.29	.31	.31	.34	.30	.73	.81	.83	1											
MTest-Yr9	.28	.28	.27	.31	.31	.71	.77	.79	.86	1										
MGrade-Yr5	.48	.45	.37	.36	.35	.57	.55	.55	.53	.52	1									
MGrade-Yr6	.41	.55	.41	.40	.40	.49	.49	.50	.47	.45	.68	1								
MGrade-Yr7	.30	.38	.56	.47	.45	.39	.43	.45	.44	.42	.53	.58	1							
MGrade-Yr8	.29	.32	.42	.59	.52	.28	.31	.38	.37	.34	.47	.51	.62	1						
MGrade-Yr9	.28	.35	.40	.49	.63	.29	.31	.37	.39	.37	.45	.52	.60	.66	1					
RGrade-Yr4	-.06	-.08	-.06	-.04	-.04	.47	.52	.52	.57	.55	.29	.21	.21	.12	.13	1				
MGrade-Yr4	.25	.21	.20	.21	.18	.65	.67	.66	.70	.69	.46	.41	.35	.26	.26	.65	1			
SEX (M = 2.F = 1)	.28	.23	.24	.23	.20	.13	.09	.04	.07	.07	.03	.03	.04	.01	-.03	-.17	.06	1		
Track-HI	.02	-.02	.04	.06	.00	.43	.51	.48	.57	.53	.17	.09	.10	.04	.04	.62	.58	.03	1	
Track-LO	.02	.06	.09	.05	.08	-.51	-.56	-.56	-.60	-.60	-.18	-.09	-.04	.04	-.01	-.66	-.66	.06	-.56	1

Correlations between math self-concept (MSC), math test scores (MTest), math school grades (MGrade), and Background variables (Gender, final Math and German grades from before the start of secondary school in Year 4, and track). For purposes of this analysis, the three tracks (high, medium, and low) were represented by two dichotomous variables (high track and low track, with the medium track as the left-out level). In the confirmatory factor analyses (CFA), MSCs are latent variables based on responses to six items, whereas grades and test scores are single-variable constructs. All three constructs were highly stable over time (M lag 1 rs = .71 for MSC, .82 for MTest, and .64 for MGrade). Within the same wave, correlations between different constructs are moderate: M r = .34 for MSC and MTest, M r = .56 for MSC and MGrade, and M r = .45 for MGrade and MTest

($r = .50$) and school grades ($r = .45$). These results indicate substantial stability over time for all three constructs, consistent with the rationale for the RI-CLPM.

Compared to the stability (test–retest) correlations, mean correlations among different constructs within the same wave are systematically smaller: .34 for MSC and Test scores; .56 for MSC and school Grades; and .45 for Grades and Test scores. Consistent with expectations (see Research Hypotheses), MSCs are systematically more highly correlated with school grades than test scores. Although lag-1 correlations among the different constructs are lower than those in the same wave, the pattern of correlations remains consistent. These results support the distinctiveness (discriminant validity) of the three constructs.

Also relevant are the relations between covariates (demographic variables) with our measures of MSC, school grades, and test scores (Table 2). Our primary interest is how incorporating these covariates into our model affects support for the REM hypotheses. However, we are also interested in relations of the covariates with MSC and achievement, and their consistency over time. Gender differences consistently favor boys for MSC and, to a lesser extent, math test scores. However, there are almost no gender differences in terms of math school grades. Also consistent with gender stereotypes, achievement at the end of primary school favors girls for verbal achievement and boys for math achievement. Primary school math grades consistently correlated highly with math test scores over the subsequent five years (.65 to .70). However, they were also significantly correlated with math school grades and MSC in subsequent years. Compared to primary school math grades, primary school German grades were less positively correlated with math test scores and grades, and were almost uncorrelated with MSC. Thus, for purposes of the present investigation, primary school grades provide particularly strong covariates to control achievement levels from before the start of secondary school.

The results also demonstrate differences between the tracks. Not surprisingly, the largest differences are for primary school grades that were the main basis for assigning students to secondary school tracks (.62 and .58 for High track, -.66 and -.66 for low track). Differences in test scores are also substantial (.43-.53 for High track, -.60 to -.51 for low track). In contrast, reflecting grading-on-a-curve and well-established frame-of-reference effects, track differences are much smaller for school grades and MSCs (for further discussion, see Marsh, et al., 2018a, 2018b).

Directional Ordering

Directional Ordering: CLPMs. For the basic CLPMs (see Basic Models MB1a-MB3a in Tables 3 and 4), we found support for our a priori (REM) hypotheses for paths from MSC to school grades (MB1a), test scores (MB2a), and the combination school grades and test scores (MB3a). In addition, the model fit was good for all three basic CLPMs. However, the fit for the CLPM MB3a (Table 3, e.g., TLI = .954) was not as good as the corresponding measurement model MM7 (Table 1, e.g., TLI = .976).

The critical parameter estimates of the CLPMs (see Fig. 1) for testing REM's hypotheses are the cross-lag paths relating two achievement measures and MSC (Bxy, Byx, Bxz, Bzx—the values are shaded in Table 4). However, also of interest

Table 3 Relations between Math Self-concept, Standardized Achievement Test Scores, and School Grades over time. Goodness-of-Fit for Basic Cross-lag Panel Models (CLPMs) and Random Intercept Cross-lag Panel Models (CLPMs)

Basic Models (MB)	Chi-SQ	Df	RMSEA	CFI	TLI
Cross-Lag-Panel Model (CLPM)					
MB1a. Grade Only	3117	1636	0.028	0.965	0.962
MB2a. Test Only	3135	1636	0.028	0.965	0.962
MB3a. Grades and Tests	4312	2156	0.029	0.958	0.955
CLPM with Random Intercepts (RI-CLPM)					
MB1b. Grade Only	2732	1629	0.024	0.974	0.972
MB2b. Test Only	2561	1629	0.022	0.978	0.976
MB3b. Grades and Tests	3510	2137	0.023	0.973	0.971
MB3b. Grades and Tests new	3496	2140	0.023	0.974	0.971

RMSEA = root-mean-square error of approximation, CFI = confirmatory fit index, TLI = Tucker-Lewis index.: MSC = math self-concept. For both for CLPMs and RI-CLPMs, we present results separately for models with achievement represented by only school grades (MB1), only test scores (MB2), or both test scores and school grades (MB3, see Fig. 1)

are the stability coefficients (Bxx, Byy, Bzz), and the cross-lag paths relating the two achievement measures (Bzy, Byz). We note that each of these paths was constrained to be equal across the five waves of data. In support of REM hypotheses, all eight of these paths are significantly positive in all three models (MB1a, MB2a, and MB3a in Table 4). Also consistent with Research Hypothesis 2, paths in the MB3a based on all three constructs are somewhat smaller than the corresponding paths in models with MSC and only grades (MB1a) or MSC and only test scores (MB2a).

Nevertheless, even in the more demanding model MB3a, all stability and cross-lag paths are significantly positive and consistent with models MB1a and MB2a. Substantively, the interpretation of the CLPM results is straightforward. There are modest but highly consistent reciprocal effects between MSC, math school grades, and math test scores.

The primary focus for Research Hypotheses 1 and 2 is the autoregressive stability and cross-paths relating constructs from one wave to the next. However, it is also relevant to consider the undecomposed (between-person) within-wave variances and covariances (Table 5). At wave-1, these are substantial, but residual variances and covariances for waves 2–4 (controlling values from the preceding wave) are substantially smaller. Nevertheless, the variance explained in each construct by the same constructs in the immediately preceding wave is substantial (MultR^2 = .55, MSC; .54, tests; .46, grades).

Directional Ordering: RI-CLPMs. For RI-CLPMs (Tables 3, 4 and 5), we evaluated models that parallel the corresponding CLPMs of structural invariance. Again, we found support for the REM (Research Hypothesis 1 and 2) hypotheses for MSC for school grades (MB1b), test scores (MB2b), and the combination school grades and test scores (MB3b). Each of these models had an excellent fit to the data (TLIs:

Table 4 Relations between Math Self-concept, Standardized Achievement Test Scores (Tst), and School Grades Achievement (Grd) Over Time for Basic Cross-lag Panel Models (CLPMs) and Random Intercept Cross-lag Panel Models (RI-CLPMs)

To(Predicted, lag-n) From (Predictor, lag t-1)-t	Test on Test	SE	Test on MSC	SE	Test on Grade	Se	Grade on Grade	SE	Grade on Test	SE	Grade on MSC	SE	MSC on MSC	SE	MSC on Test	SE	MSC on Grade	SE
Cross-Lag Panel Models (CLPM)																		
MB1a. **Grade Only**							.584	.012			.127	.012	.703	.013			.108	.011
MB2a. **Test Only**	.630	.010	.144	.007									.702	.012	.162	.012		
MB3a. **Grades and Tests**	.583	.011	.061	.008	.1189	.008	.450	.013	.278	.014	.070	.012	.675	.014	.136	.013	.067	.012
CLPM with Random Intercept (RI-CLPM)																		
MB1b. **Grade Only**							.283	.025			.137	.027	.435	.04			.12	.022
MB2b. **Test Only**	.184	.02	.102	.018									.501	.038	.129	.029		
MB3b. **Grades and Tests new**	.174	.019	.069	.018	.038	.012	.267	.024	.160	.026	.133	.027	.426	.039	.098	.027	.120	.022
MB3b. **Grades and Tests**	.174	.019	.067	.018	.025	.012	.262	.024	.126	.026	.133	.027	.425	.04	.083	.027	.115	.022

MSC = math self-concept. SE = standard error. Shown are the lag-1 path coefficients that are common to the CLPMs and RI-CLPMs (see Fig. 1). Stability paths link the same construct in adjacent waves: test on test; grade on grade, math self-concept (MSC) on MSC. Cross-paths link one construct to a different construct in adjacent waves. Shaded paths are cross-lag paths involving MSC that are the main focus of the study. For both for CLPMs and RI-CLPMs, we present results separately for models with achievement represented by only school grades (MB1), only test scores (MB2), or both test scores and school grades (MB3, see Fig. 1). All paths are statistically significant ($p < .05$) in relation to standard errors (SEs)

.971-0.976). Furthermore, the fit of MB3b approached the fit of the corresponding measurement model (TLI = .971 for MB3b in Table 3 and .976 for MM7 in Table 1).

The global trait factors representing decomposed between-person differences are the unique feature of RI-CLPMs. Consistent with expectations and in support of the appropriateness of the RI-CLPMs, the global trait factors account for much of the variance in MSC, test scores, and school grades (Table 5). Furthermore, these trait factors are substantially correlated (*r*s = .64 to .80, Table 5). In contrast, the variances and covariances for the within-person components are substantially smaller, although they are all are positive and statistically significant (Table 5). Because the variance components for the global trait factors are substantial, the variance explained in each construct by the same constructs in the immediately preceding wave is modest (MultR^2 = .25, MSC; .08, tests .15 grades).

In support of REM hypotheses, all three models' stability and cross-lag paths are significantly positive (MB1b, MB2b, and MB3b in Table 4). Although the paths in the MB3b based on all three constructs are somewhat smaller than the corresponding paths in MB1b and MB2b, the differences are small. Thus, RI-CLPMs show modest reciprocal effects between MSC, math school grades, and math test scores in support of REM hypotheses.

Directional Ordering: Juxtaposing CLPMs and RI-CLMPs

In Research Question 3, we noted no clear a priori basis for predicting differences in results based on latent CLPMs and RI-CLPMs, and few studies comparing them empirically. In this respect, juxtaposing CLPMs and RI-CLPMs is substantively important for ASC research and better understanding CLPMs and RI-CLPMs. Here we highlight several key findings from comparing results from CLPMs and RI-CLPMs.

Table 5 Variances and Covariances For Math Self-Concept (MSC), Tests, and Grades Based on the RI-CLPM (MB3b Table 4)

Parameters	CLPM			RI-CLPM		
	Estimate	SE	Mult R^2	Estimate	SE	Mult R^2
Global Trait Variance Components						
MSC				.47	.03	
TEST				.38	.01	
Grades				.47	.02	
Global trait UNstandardized Covariances						
MSC & TESTs				.27	.02	
MSC & Grades				.31	.02	
Grades & TESTs				.34	.01	
Global trait standardized correlations						
MSC & TESTs				.64	.02	
MSC & Grades				.65	.02	
Grades & TESTs				.80	.02	
Mean Variance Components Across Waves						
Residual Variance (Wave2-5)						
MSC	.47	.01	.55	.43	.01	.25
TEST	.28	.01	.54	.20	.01	.08
Grades	.56	.01	.46	.49	.01	.15
Variance Component (Wave 1);						
MSC	.79	.03		.38	.03	
TEST	.96	.03		.35	.02	
Grades	.71	.02		.46	.02	
Mean Covariances Across Waves						
UNstandardized Covariances (Wave 1 only);				.09	.01	
MSC & TESTs	.32	.02		.12	.02	
MSC & Grades	.41	.02		.10	.01	
Grades & TESTs	.45	.02				
Standardized correlations (Wave 1 only)				.22	.04	
MSC & TESTs	.43	.02		.32	.05	
MSC & Grades	.48	.02		.25	.03	
Grades & TESTs	.55	.02				
Standardized correlation (Waves 2- 5)				.17	.02	
MSC & TESTs	.22	.01		.48	.02	
MSC & Grades	.23	.01		.09	.02	
Grades & TESTs	.38	.03				
Residual Covariances (Waves 2–5)				.05	.01	
MSC & TESTs	.10	.01		.22	.01	
MSC & Grades	.13	.01		.03	.01	
Grades & TESTs	.11	.01		.47	.03	

For the RI-CLPM, variances and covariances are presented separately for the Global Trait (decomposed between-person) components and the within-wave residual components (within person variances and covariances). For the CLPM, the variance and covariances are the undecomposed between-person estimates. Mult R^2 terms are the amount of variance in MSC, tests scores, and school grades in waves 2 – 5 that can be explained by the same constructs in the preceding wave (i.e., Wave 1 – 4)

First, the fit indices are marginally better for RI-CLPMs than the CLPMs (Table 3). However, given that the CLPMs are nested under RI-CLPMs, this is not surprising. Indeed, it is surprising that the differences are not larger, given the substantial stability of the constructs.

Second, the major difference in the path coefficients (Table 4) is the substantially smaller stability coefficients for the RI-CLPMs. Again, this is consistent with the theoretical rationale and control for random intercepts in RI-CLPMs (see Fig. 1 and earlier discussion on the different conceptualization of between-person differences).

Third, the variance explained in each construct by the same constructs in the immediately preceding wave is substantially larger for CLPMs ($MultR^2 = .46$ to .55) than for RI-CLPMs modest ($MultR^2 = .08$ to .25). This is a natural consequence of the substantial global trait factors in the RI-CLPMs.

Most importantly, both CLPMs and RI-CLPMs support the REM hypotheses, particularly for paths relating the two achievement measures and MSC (Bxy, Byx, Bxz, Bzx—the paths are shaded in Table 4). Each of these cross-lag paths is modest (.061—.136 for CLPM MB3a; .067—0.133 for RI-CLPM MB3b), but highly significant and consistent across CLPMs and RI-CLPMs. In summary, there is good support for REM hypotheses based on both CLPMs and RI-CLPMs.

Extended Models for CLPMs and RI-CLPMs

In the next series of models, we extend the basic CLPMs and RI-CLPMs based on model MB3a (CLPM) and MB3b (RI-CLPM; Tables 3 and 4) in several respects. First (the "ML" models in Tables 5 and 6), we included lag-2 paths (see Research Question 4) between non-adjacent waves (e.g., paths relating factors in wave-1 to wave 3, wave 2 to wave 4, etc.). Second (the "MC" models in Tables 5 and 6), we added covariates to the basic models (see Research Question 5). Finally, we tested models with additional lag-2 paths and covariates (the "MLC" models in Tables 5 and 6).

CLPMs: Lag-2 Paths. We note that the basic model with no lags (MB3a in Table 3) is nested under these ML models (ML1a and ML2a), and all these models are nested under the measurement model (MM7 in Table 1), providing appropriate bases for evaluating goodness-of-fit. Consistent with expectations, additional lag-2 paths noticeably improved the fit of the CLPMs (Table 6). Compared to the fit of the lag-1 CLPM (TLI = 0.955, MB3a in Table 3), the CLPM with lag-2 paths was better (TLI = .972, ML1a in Table 6). In Model ML2a, we showed that this improved fit was primarily a function of the stability paths when we eliminated the lag-2 cross-lagged paths (TLI = .971, ML2a in Table 6). We also note that the fits of ML2a and ML2b (Table 6) were only marginally less than the fit of the corresponding measurement model (TLI = .976, MM7 in Table 1). As noted earlier, this measurement model MM7 is fully saturated in relation to the structural model constraints imposed in the CLPMs. This comparison further supports these constraints and the need to extend to CLPM to include additional lagged effects (also see earlier discussion). Nevertheless, the critical issue is how the inclusion of the lag-2 stability path influences support for the REM (Research Hypotheses 1 and 2).

Compared to the corresponding basic CLPM with no lag-2 paths (MB3a in Table 3), the stability and cross-lagged paths for models with additional lags (ML1a and ML2a, Table 7) are smaller. Not surprisingly, these differences were particularly evident for the stability coefficients (i.e., lag-1 effects were smaller when we added lag-2 paths). Thus, for example, the lag-1 stability paths in ML2a that also included lag-2 stability paths (.458, .401, .531; Table 7) were substantially smaller than the corresponding values in MB3a with only lag-1 paths (.583, .450, .675; Table 3). Although there were also differences in the cross-lagged paths relating achievement and MSC (.045, .060, .099 & .067 for ML1a vs. .061, .070, .136 & .067 for MB3a), these differences were much smaller. Critically, all the stability and cross-lagged paths were significantly positive in all CLPMs with and without lag-2 paths. Importantly, the four critical cross-lag paths that test REM (shaded in Tables 4 and 7) were all significantly positive, even though the values were slightly smaller for the CLPM that included lag-2 paths.

CLPMs: Covariates Models with covariates are not nested under models considered thus far. Hence the fit indices are not directly comparable. To provide a basis of comparison, we fit a model in which we constrained all paths from the three covariates (math and verbal achievement from primary school and gender) to MSC, grades, and test scores to be zero (**MC3a** in Table 6, TLI 0.932). The fit improved when these paths were freely estimated (**MC1a** in Table 6, TLI 0.957). These results indicate that the effects of these covariates were not substantial, even though the covariates correlated substantially with our outcome variables (Table 2). Hence, it is not surprising that the autoregressive stability and cross-lagged path coefficients were not substantially affected by including the covariates (Models MC1a & MC2a in Table 7 compared to model MB3a in Table 4). Indeed, the changes were small and not even consistent in direction for the four critical cross-lagged paths relating achievement and MSC (those shaded in Tables 4 and 6). Nevertheless, there was some evidence that these covariates had effects beyond the first wave of data, consistent with our recommendation that paths from covariates to all data waves should be considered.

CLPM: Additional Lags and Covariates Consistent with earlier discussion, the addition of both lag-2 paths and covariates (MLC models in Tables 6 and 7) led to a better fit (e.g., TLI = 0.957 for MC1a and TLI = 0.971 for MLC1a). Although comparisons with models in which covariate paths were constrained to be zero again showed that there were covariate effects, these effects were modest (e.g., TLI = 0.971 for MLC1a vs. TLI = 0.949 for MLC3a). Unsurprisingly, lag-1 stability paths are smaller for these MLC models than the corresponding MB, ML, and MC models. However, the effects of these lag-1 stability coefficients for the MLC models are similar to the corresponding ML models, suggesting that the effects of lag-2 paths are greater than the effects of the covariates. This pattern of results suggests that the inclusion of lag-2 effects provides some control for unmeasured, time-invariant covariates.

Table 6 Relations between Math Self-concept, Standardized Achievement Test Scores, and School Grades over time. Goodness-of-fit for Extended Cross-lag Panel Models (CLPMs) and Random Intercept Cross-lag Panel Models (CLPMs)

	Chi-SQ	Df	RMSEA	CFI	TLI
Cross-Lag-Panel Models (CLPM)					
Lag-2, No Covariates (ML)					
ML1a. lag-2:Cross- & Stability-paths	3337	2066	.023	.975	.972
ML2a. lag-2: Stability paths only	3491	2120	.023	.973	.971
Covariates No lag-2 (MC)					
MC1a. Covariates: YR5 to YR9	4475	2375	.027	.961	.957
MC2a. Covariates: YR5 only	5083	2477	.03	.951	.947
MC3a. Covariates: Paths fixed to zero	5923	2513	.034	.937	.932
Lag-2 + Covariates (LC)					
MLC1a. lag-2 Covariates-YR5-Yr9	3629	2288	.022	.975	.971
MLC2a. lag-2 Covariates-YR5	4071	2396	.020	.969	.965
MLC3a. lag-2 Covariates-Fixed to 0	4885	2423	.029	.954	.949
Random Intercept CLPM (RI-CLPM)					
Lag-2, No Covariates (ML)					
ML1b. lag-2: Cross- & Stability-paths v2	3261	2050	.022	.976	.973
ML2b. lag-2: Stability paths only	3335	2101	.022	.976	.973
Covariates No lag-2 (MC)					
MC1b. Covariates: Yr5 to YR9	3790	2365	.023	.973	.970
MC2b. Covariates: Yr5 only	4920	2464	.029	.954	.949
MC3b. Covariates: Paths fixed to zero	5074	2500	.03	.952	.948
MC4b. Covariates to Global Trait	4055	2473	.023	.971	.968
Lags + Covariates (LC)					
MLC1b. lag-2 + Covariates-Yr5-Yr9	3548	2275	.022	.976	.972
MLC2b lag-2 + Covariates-Yr5 only	4677	2383	.029	.957	.951
MLC3b. lag-2 + Covariates-Fixed to 0	4828	2410	.029	.955	.949
MLC4b. lag-2 + Covariates to Global Traits	3805	2387	.022	.974	.970

RMSEA = root-mean-square error of approximation, CFI = confirmatory fit index, TLI = Tucker-Lewis index. CLPMs and RI-CLPMs presented here are extensions of the basic models (Tables 3 & 4; also see Fig. 1). lag-2 = models with paths from each wave to the next two waves (lag-2) in addition to lag 1 paths.; covariates = gender and math and verbal achievement measures from year 4 (last year of primary school – see Fig. 1)

The critical parameters for testing the REM hypotheses are the cross-lag paths relating achievement and MSC (shaded in Tables 4 and 7). These four paths in MLC1a (Table 7) are all statistically positive and similar in size to the corresponding paths in the basic CLPM: 0.060, 0.101, 0.086, 0.069 (MLC1a, Table 7) vs. 0.059, 0.091, 0.085, 0.066 (MB3a, Table 4). In summary, the reciprocal effects that were hypothesized by the REM were robust against the inclusion of lag-2 paths and the effects of additional covariates.

Extended Models for RI-CLPMs Although we fit parallel models based on the RI-CLPM structure, the interpretation of these models is fundamentally different. The global trait factors in RI-CLPMs are intended to absorb time-invariant (decomposed between-person) effects. Hence, based on the underlying rationale of the RI-CLPM, we did not expect that extending the models to include additional lags and covariates would have much effect on fit or the (within-person) stability or the cross-lagged paths that are of primary interest. Consistent with these expectations, the cross-lag and stability paths were largely unaffected by the inclusion of covariates and additional lagged effects. This finding demonstrates an important advantage of basic RI-CLPMs compared to basic CLPMs—that do not include lag-2 effects and covariates—concerning the robustness of the interpretations.

The additional lagged paths improved the fit only marginally (TLI = .971 for Model MB3b in Table 3 with no lagged effects vs. .973 for ML1b in Table 6 with lag-2). There were, however, effects of the covariates. This is evident in the difference in fit for MC3b that constrained these effects to be zero (TLI = .948 in Table 3) and MC1b where these effects were freely estimated (TLI = .970, Table 6). However, most of the effects of the covariates were explained in terms of the global trait factors (TLI = .968 in MC4b). This same pattern of results is evident for the corresponding MLC models that include both lag-2 paths and covariates (Table 6). However, adding the lag-2 paths in the MLC models improved the fit only marginally compared to corresponding MC models with no lag-2 path.

The parameter estimates across the RI-CLPMs with additional lag-2 paths and covariates (Table 7) are also similar to those for the basic RI-CLPM MB2 (Table 4) with none of these extensions. Of particular interest, the (within-person)

Table 7 Relations between Math Self-concept, Standardized Achievement Test Scores (Tst) and School Grades Achievement (Grd) over time. Cross-lag Panel Models (CLPMs)

To(Predicted, lag n) From (Predictor, lag n-1)	Test on Test	SE	Test on MSC	SE	Test on Grade	Se	Grade on Grade	SE	Grade on Test	SE	Grade on MSC	SE	MSC on MSC	SE	MSC on Test	SE	MSC on Grade	SE
CLPMs																		
Lag-2 Paths No Covariates (ML)																		
ML1a. lag-2 Cross & Stability paths	.466	.013	.042	.013	.172	.034	.389	.015	.230	.021	.092	.020	.537	.021	.111	.02	.064	.016
ML2a. lag-2 Stability paths only	.458	.012	.045	.009	.075	.009	.401	.015	.207	.017	.060	.014	.531	.019	.099	.016	.067	.013
Covariates No lag (MC)																		
MC1a. Covariates to Yr5-Yr9	.542	.012	.057	.009	.097	.008	.459	.014	.239	.017	.085	.013	.642	.015	.119	.014	.074	.013
MC2a. Covariates to Yr5 only	.612	.012	.058	.009	.109	.009	.483	.015	.291	.021	.071	.014	.697	.015	.14	.016	.046	.013
Lags + Covariates (LC)																		
MLC1a. lag-2+Covariates to Yr5-Yr9	.446	.013	.06	.018	.058	.013	.368	.016	.181	.021	.101	.02	.524	.021	.086	.016	.069	.016
MLC2a lag-2+ Covariates to Yr5 only	.458	.013	.058	.018	.065	.013	.386	.015	.187	.02	.093	.02	.528	.021	.083	.016	.069	.016
RI-CLPM																		
Lags, No Covariates (ML)																		
ML1b. lag-2 Cross & Stability paths v2	**.319**	**.058**	**.064**	**.026**	.012	**.021**	**.264**	**.047**	**.147**	**.043**	**.160**	**.036**	**.447**	**.033**	**.103**	**.034**	**.102**	**.026**
ML2b lag-2 Stability paths only	.27	.057	.077	.018	.021	.016	.257	.037	.176	.033	.154	.031	.465	.031	.118	.029	.099	.022
Covariates No lag (MC)																		
MC1b. Covariates: Yr5 to YR9	.179	.02	.082	.017	.036	.013	.261	.024	.158	.025	.132	.027	.420	.037	.112	.027	.120	.022
MC2b Covariates: Yr5 only	.180	.029	.101	.020	.036	.017	.260	.032	.213	.039	.163	.032	.487	.033	.163	.034	.103	.023
MC4b. Covariates to Global Trait	.181	.02	.077	.018	.035	.013	.265	.023	.160	.025	.131	.026	.428	.037	.106	.027	.118	.022
Lags + Covariates (LC)																		
MLC1. lag-2 + Covariates-Yr5-Yr9	.214	.04	.068	.023	.029	.018	.279	.033	.211	.036	.136	.032	.439	.033	.138	.035	.096	.035
MLC3. lag-2 + Covariates-Yr5 only	.36	.05	.057	.028	047	.021	.313	.034	.238	.036	.12	.032	.46	.034	.133	.037	.0893	.024
MLC5. lag-2+ Covariates to Global Traits	.334	.042	.056	.025	.032	.019	.277	.034	.21	.036	.137	.031	.447	.033	.124	.036	.094	.024

CLPMs and RI-CLPMs presented here are extensions of the basic models (Tables 3 & 4; also see Fig. 1). Cross-paths = autoregressive cross-path; Stability = autoregressivestability paths; covariates = gender and math and verbal achievement measures from year 4 (last year of primary school – see Fig. 1). lag-2 = models with paths fromeach wave to the next two waves (lag-2) in addition to lag 1 paths. Global traits are the (between-person) trait factors for the RI-CLPM (see Fig. 1). Shaded paths are cross-lag paths involving MSC that are the main focus of the study

cross-lagged paths relating achievement to MSC (shaded in Tables 4 and 7) were nearly unaffected. Thus, for RI-CLPM model MLC1b with both lag-2 effects and covariates, these paths (.068-.138, Table 7) are similar in size to the corresponding paths for model MB3b (.067-.133. Table 4). In summary, the additional lagged effects and covariates had relatively little effect on parameter estimates or theoretical interpretations supporting REM hypotheses.

Discussion

In educational psychology, a substantial body of research demonstrates that ASC and achievement are reciprocally related based on CLPMs. However, recent research has challenged CLPMs' appropriateness, arguing that RI-CLPMs provide a stronger (within-person) perspective and better control for unmeasured covariates. Nevertheless, few studies have actually compared results for the directional ordering of ASC and academic achievement based on CLPMs and RI-CLPMs. In this sense, our study is a substantive-methodological synergy; we apply and extend state-of-the-art quantitative research tools to address substantively important issues with implications for theory, methodology, and policy/practice. Whereas our substantive focus is central to educational psychology, we anticipate that methodological issues raised here—and some of the solutions offered to these issues—will have broad cross-disciplinary relevance.

Appropriate Interpretations of Within-Person (RI-CLPM) and Undecomposed Between-Person (CLPM) Effects

The critical difference between CLPMs and RI-CLPMs is the (undecomposed) between-person (single-level) perspective in CLPMs and the within-person (multilevel) perspective in RI-CLPMs. CLPMs are appropriate for comparisons between individuals. However, Hamaker et al., (2015; Mulder & Hamaker, 2021) and others express concerns that CLPMs confound within- and between-person effects in tests of directional ordering. The RI-CLPM evaluates the prospective temporary deviation from the trait level in one construct on change in the temporary deviation from the trait level in a second construct. The auto-regressive factors in the RI-CLPM (Axs, Ays, and Azs in Fig. 1) represent deviations from a student's trait score rather than the individual differences that are the basis of the CLPM. Thus, the stability paths reflect the stability of rank-order differences in CLPMs. However, in RI-CLPMs, they reflect what Hamaker et al. (2015) refer to as within-person carry-over effects (or inertia), and what Kenny and Zautra (2001) refer to as slowly changing autoregressive factors. If these within-person stability paths are positive, elevated scores at one wave are likely to be associated with elevated scores in the next wave (i.e., to have a lasting effect on a later measurement wave beyond the stability captured by the RI component, the global trait factors). Likewise, the cross-lagged paths reflect undecomposed between-person processes in CLPMs, but within-person processes in RI-CLPMs. Hence, these two perspectives address fundamentally different questions

and often result in different conclusions, but are easily confounded. Hence, following our earlier discussion and research questions (Orth et al., 2021), the appropriate interpretations of CLPM and RI-CLPM results in the present investigation are:

- **CLPMs**: When students have higher MSCs (compared to other students), they are likely to experience a subsequent rank-order increase in math achievement (compared to other students). Likewise, when students have higher math achievement (compared to other students), they are likely to experience a subsequent rank-order increase in MSC (compared to other students). Thus, undecomposed individual differences in MSC positively predict rank-order change in the relative position of math achievement, and undecomposed individual differences in math achievement positively predict rank-order change in the relative position of MSC. Moreover, these results are relatively unaffected by the introduction of lag-2 effects or covariates and generalize for achievement based on math school grades and test scores.
- **RI-CLPMs:** When students experience higher than their usual MSC (compared to their long-term average MSC over the duration of the study), they are likely to experience a subsequent increase in their levels of math achievement (compared to their long-term average math achievement over the duration of the study). Likewise, when students experience higher than their usual math achievement (compared to their long-term average math achievement), they are likely to experience a subsequent increase in their levels of MSC (compared to their long-term average MSC). Thus, decomposed within-person differences in MSC positively predict change in achievement, and decomposed within-person differences in math achievement positively change MSC. Moreover, these results are relatively unaffected by the introduction of lag-2 effects or covariates and generalize for achievement based on math school grades and test scores.

The Juxtaposition of Results Based On CLPMs and RI-CLPMs

Recent educational psychology research suggests CLPMs and RI-CLPMs as antagonistic, even suggesting that RI-CLPM's decomposed between- and within-person perspective is always more appropriate (but see discussion by Orth et al., 2021). However, both perspectives provide useful substantive information for understanding ASC and achievement relations over time. In this sense, we see the two approaches as complementary. Each has different strengths and weaknesses, providing different perspectives on longitudinal relations between ASC and achievement. Most educational and psychological research is cross-sectional, focusing on relations among variables across persons (an undecomposed between-person perspective). However, many research questions and interpretations of results are expressed implicitly (or even explicitly) as causal relations, even when the design or statistical analyses are not appropriate for these interpretations. The equivalence in interindividual and intraindividual effects is only likely under limited conditions (e.g., stationarity of parameter estimates over time and homogeneity of relations between

variables across individuals) that are unlikely to be met. Hence, following Voelkle et al. (2014), we argue that rather than seeing within- and between-person analyses as competing research paradigms, the focus should be on commonalities in conjunction with their differences.

In our study, the patterns of cross-lag paths critical to the interpretation of directional-ordering and support for REM hypotheses (Research Hypotheses 1 & 2) are remarkably consistent across variations and extensions of CLPMs and RI-CLPMs. Indeed, even the sizes of the reciprocal effects are similar across both CLPMs and RI-CLPMs. These results suggest that our conclusions are consistent for both undecomposed between-person (CLPM) and within-person (RI-CLPM) perspectives. Importantly, the juxtaposition between results and the different issues faced by both models provides a stronger basis of interpretation of the results than considering either model in isolation.

Methodological Issues

Goodness-of-fit and the Measurement Model. Goodness-of-fit should be an important consideration in model evaluation, as exemplified in our study. Longitudinal studies should routinely begin with a systematic evaluation of the measurement model and its invariance over time. Application of both CLPMs or RI-CLPMs is problematic if there is not good support for at least configural invariance. Metric invariance underpins the routine constraint of critical stability and cross-lag paths over time. Metric invariance also underpins RI-CLPMs' focus on temporal deviation scores (i.e., the difference between the wave and the mean across all waves) for measures at each wave. We also note that manifest models are biased by the presence of measurement errors and correlated measurement errors (i.e., correlated uniquenesses demonstrated in the present investigation in Table 1). A lack of metric invariance does not invalidate the application of CLPMs but compromises the rationale underpinning RI-CLPMs. However, even for CLPMs a lack of invariance complicates the interpretation of results, particularly when there are many waves of data. Hence, all CLPMs and RI-CLPMs should begin with a systematic evaluation of the underlying measurement model.

We also note that the final measurement model is saturated in terms of relations between constructs. Thus, CLPMs and RI-CLPMs used here (and many other theoretical models of longitudinal data) are nested under the corresponding measurement model (MM7 in Tables 1 and 2). Hence, the goodness-of-fit for this measurement model provides a basis of comparison for our subsequent CLPMs and RI-CLPMs independent of assumptions made by these models. If the fit of either the CLPM or the RI-CLPM is meaningfully worse than the measurement model, then researchers may need to explore further the basis of the misfit and whether it has substantively important implications for interpretation of the results. Here we demonstrated that goodness-of-fit indices for the basic CLPMs (Models MB1a-MB3a in Table 3) were good relative to traditional guidelines but not compared to the corresponding

measurement model (MM7 in Table 1). In contrast, the fit of the basic RI-CLPM approached that of the measurement model. We also note that an inspection of estimates based on this preliminary measurement model provides potentially valuable insights into the data. Implicit in this recommendation is the importance of measuring constructs with multiple indicators that allow researchers to test the measurement model.

Goodness-of-fit and Choice of Models Because basic CLPMs are nested under the corresponding RI-CLPMs, RI-CLPMs will necessarily provide a better fit for indices that do not correct for CLPMs' greater parsimony. However, Orth et al., (2021; see also Asendorpf, 2021) argued that because CLPMs and RI-CLPMs address different research questions, the model choice should be based on the study's aims and evaluation of parameter estimates in addition to relying solely on goodness-of-fit. However, we would like to offer a caveat about goodness-of-fit. If RI-CLPMs and the corresponding measurement model fit better than CLPMs, then there is systematic variation or covariation unexplained by CLPMs. The difference in fit between basic CLPMs and RI-CLPMs suggests this is due to global trait factors in RI-CLPMs. Here, the fit of the CLPMs improved (and approximated the fit of the RI-CLPM) with the inclusion of lag-2 stability paths, as will often be the case for CLPMs. This issue is evident even in comparing the CLPM and the corresponding measurement model, even without considering the RI-CLPM. Consistent with our supposition following from Marsh, Pekrun, Murayama, et al. (2018; also see Lüdtke & Robitzsch, 2021), the goodness-of-fit for the extended CLPM that included lag-2 effects fit was similar to that of the RI-CLPM. Because this will generally be the case, goodness-of-fit should no longer be such a critical issue in comparing the RI-CLPM and CLPM with lag-2 effects or, more generally, all lag > 1 effects (e.g., full-forward CLPMs).

Nevertheless, when additional lagged paths are tested (either a priori based on previous research or post hoc), it is important to evaluate support for a priori hypotheses and research questions across these extended models—a sensitivity test. Hence, even when CLPMs are more appropriate for a study's research questions, it is appropriate to evaluate its fit compared to the corresponding measurement model. Of course, if research questions are more appropriate for RI-CLPMs, they would be more appropriate than CLPMs with additional cross-lag paths, even if the two models fit the data equally well. Nevertheless, to the extent that CLPMs with lag-2 paths fit the data as well as RI-CLPMs, then CLPMs become a viable alternative to RI-CLPMs even in relation to goodness-of-fit. More broadly, both perspectives provide potentially useful information for understanding substantive issues. Hence, we include detailed analyses for both models and juxtapose interpretations based on each.

Potential Biases Associated with Covariates Here we considered three covariates: gender and math and German school grades measured in primary school, prior to the start of secondary schooling. We note that these covariates are of interest (e.g., see results in Table 2) in their own right. Thus gender differences were consistent with previous research and gender stereotypes (girls had higher verbal achievement

but lower scores on the math constructs—particularly MSCs). Compared to German primary school grades, math primary school grades are more highly correlated with secondary math grades (.26-.46 vs. .12-.29) and math test scores (.65-.70 vs. .47-.57). However, consistent with previous research, the differences in effects of primary school verbal and math achievement are starker for MSC (+.18 to +.25 vs. -.08 to -.04), where even the direction of relations changes.

Interestingly, test scores collected in the five secondary school years were more highly correlated with primary school grades (0.52 to 0.70) than with school grades in secondary school (0.12 to 0.46). Furthermore, correlations between primary school grades and particularly test scores are reasonably consistent over the first five years of secondary school. Thus, for example, correlations with test scores in Year 9 (0.55 for verbal, 0.69 for math) are only marginally higher than the corresponding test scores in Year 5 (0.47 for verbal, 0.65 for math). These results suggest that school grades in the untracked primary schools are more like test scores, reflecting a more common underlying metric than grades in Years 5–9 in the tracked secondary schools. These findings also have potentially important implications for issues such as grading on a curve and the appropriate interpretations of achievement based on test scores and school grades.

For present purposes, our overriding interest in covariates is how their inclusion or omission affects the results for CLPMs and RI-CLPMs. We can think of these as omitted covariates in the basic models (Tables 3 and 4) that do not include them. Hence, the comparison with models that include covariates demonstrates how the results are affected by their omission. However, the pattern of results, particularly the critical cross-lagged paths used to determine directional-ordering and support for REM hypotheses, was little affected by the inclusion of these covariates in either CLPMs or RI-CLPMs. Thus, even though particularly the primary school grades are powerful covariates substantially related to our outcome variables, the interpretations of the CLPMs and RI-CLPMs were robust concerning their omission.

Potential Biases Associated with Unmeasured Covariates In all studies with non-experimental data, there will always be additional, unmeasured covariates (Reichardt, 2019). These unmeasured covariates can be fixed covariates (measured at the first wave and assumed to be constant across the different waves like our demographic variables) or time-varying covariates (assumed to take on different values at different waves). The effects of fixed covariates can be time-invariant (the same for each wave) or time-varying. Time-varying covariates could be specific to particular waves, or even be auto-regressive covariates that change gradually or systematically over time. However, the nature of such biases and when they are likely to occur has not been given sufficient attention in CLPM and RI-CLPM studies (Schuurman & Hamaker, 2019).

We believe that the main role of covariates is to control for confounding. Therefore, we recommend including covariate effects on all REM variables (e.g., MSC and achievement) at each measurement point. For the CLPMs, fixed covariates that are truly time-invariant (i.e., their effects do not change over the time-frame being considered) are likely to have their greatest direct effect on the first data wave (Lag-1 effects). Although these fixed covariates can continue to have substantial total effects

in subsequent waves, at least some of these total effects are likely to be mediated through time 1 constructs. This reasoning is sometimes used to justify models in which the effects of covariates are only considered for wave 1 outcomes (see discussion by Little, 2013; Marsh, et al., 2018a, 2018b). However, even fixed covariates can have time-varying direct effects on subsequent waves beyond those in the first wave (e.g., sleeper effects). For example, gender might have stronger effects over time, implying that gender differences in math and verbal constructs become more differentiated over time (i.e., the gender differentiation hypothesis). Similarly, it is common for differences among young children to grow larger over time (i.e., Matthew effects). Mund et al. (2021; see also Marsh, et al., 2018a, 2018b) make a similar point. They note that the effects of fixed covariates (e.g., gender, ethnicity) can vary across different waves. From this perspective, fixed covariates can have time-varying or time-invariant effects at all waves. Whether the effects are time-varying or time-invariant is an empirical question that can only be addressed if the data are available and appropriate analyses are pursued. Hence it is always appropriate to test for the effects of fixed—as well as time-varying—covariates across all waves.

RI-CLPMs, due to the addition of global trait factors, provide better control for fixed covariates with time-invariant effects. In particular, the effects of time-invariant covariates on stability and cross-lagged estimates are minimized, because the global trait factors absorb the effects of these covariates. Hence, for RI-CLPMs in the present investigation, neither the within-person stability nor the cross-lagged paths were much affected by the omission of covariates.

Compared to RI-CLPMs, CLPMs are more vulnerable to the effects of unmeasured covariates, even those that are time-invariant. Following VanderWeele et al., (2020; Marsh, et al., 2018a, 2018b), we introduced lag-2 paths as a strategy to control unmeasured covariates. In support of this strategy, we found that the effects of covariates were reduced when lag-2 paths were included in CLPMs (see Table 7). This is important in the comparison of RI-CLPMs and CLPMs with lag-2 effects. However, because the effects of covariates were not substantial (even though the covariates were substantially related to MSC and achievement, Table 2), our tests of this strategy were not strong.

The most severe limitation of both CLPMs and RI-CLPMs is their inability to control the effects of unmeasured time-varying confounders, including fixed covariates whose effects vary over time. Following from Marsh et al. (2018a, 2018b; also see Lüdtke & Robitzsch, 2021), we suggested that extending the CLPM to include lag-2 effects provided a stronger control for covariates. Our rationale is that even if a confounder specific to Wave-T has an effect on REM variables at Wave-T + 1, it is less likely to have an effect on REM variables at Wave-T + 2 after controlling for the effects of REM variables from Wave-T and Wave-T + 1 (see VanderWeele et al., 2020). Certainly, this extended model provides stronger controls for time-varying and time-invariant covariates than CLPM without lag-2 effects. However, further research is needed to determine how and under what circumstances this extended CLPM compares favorably to the RI-CLPM, particularly in relation to controlling for time-varying covariates (see Usami, Murayama, et al., 2019; Usami, Todo, et al., 2019, 2021).

How to Model Time-varying Covariates that Are Measured The best way to control confounders is to measure them and include them in the model. This is relatively straightforward for fixed covariates like demographic variables that causally precede measurement of the autoregressive factors. However, the appropriate modeling of time-varying autoregressive covariates measured in each wave is more complicated. Indeed, we did not specifically consider any time-varying covariates. Similarly, recent extensions of basic RI-CLPMs to include covariates (e.g., Mulder & Hamaker, 2021) only considered fixed covariates (like the demographic variables in our study). However, our tripartite RI-CLPM is appropriate to evaluate time-varying covariates. Thus, for example, it would be possible to reconceptualize our study as a (bivariate) test of the directional ordering of MSC and math test scores, treating math school grades as a time-varying covariate, or a test of MSC and school grades with test scores as a time-varying covariate. Notably, the actual statistical model for this reconceptualization of our study would be the same as presented here (see Mplus syntax in Supplemental Materials)—although the presentation and interpretation of results would differ. Thus, in Table 7, we showed that the introduction of math school grades (MB3 models) reduced the sizes of stability and cross-lag paths in the bivariate models based on MSC and test scores—for both CLPMs and RI-CLPMs. From this perspective, the models considered here demonstrate how to extend basic CLPMs and RI-CLPMs to include time-varying covariates as well as lag-2 effects, time-invariant covariates, and tests of invariance over time and multiple groups. Hence, we recommend considering our tripartite RI-CLPMs and CLPMs to researchers who want to evaluate the effects of time-varying, autoregressive covariates. Nevertheless, we also caution that controlling time-varying covariates with ambiguous directional ordering with the variables under consideration might be problematic—throwing the baby out with the bathwater. Here, for example, it seems inappropriate to consider either test scores or achievement as a time-varying autoregressive control variable—particularly when we also show that these variables are highly correlated and are reciprocally related.

Alternative Approaches to CLPM Data

In providing a balanced view of how best to test REM hypotheses, we note that several extensions and modifications of basic CLPMs, RI-CLPMs, and alternative models have been discussed in the recent methodological literature (see Usami, Murayama, et al., 2019; Usami, Todo, et al., 2019, for an overview). For example, it has been suggested to include additional (linear) change factors at the between-person level in RI-CLPM. In this latent curve model with structured residuals (LCM-SR; Curran et al., 2014), the observations are residualized for interindividual differences in linear change when estimating cross-lagged effects at the within-person level (Nunez-Regueiro et al., 2021). Furthermore, it has also been proposed to directly include stable trait factors or change factors at the level of the undecomposed observations (Bollen & Curran, 2006; Zyphur et al., 2020). Andersen (2021) compares these different modeling approaches and discusses conditions under which they provide similar results (see also Asparouhov & Muthén, 2021). Noting the importance

of considering confounding variables in causal interpretations based on CLPMs and RI-CLPMS, Hübner et al., (2022) proposed using weighting strategies that facilitate controlling for a large number of fixed and time-varying covariates. Niepel et al. (2021) evaluated CLPM data based on intensive experience sampling of momentary state measures of ASC and achievement. Based on their review of research, including CLPM and RI-CLPM studies, they argued (p. 2) that "due to the lack of intensive longitudinal studies on the reciprocal relations between ASC and achievement, the momentary (state) intraindividual (within-person) dynamics between ASC and achievement remain a black box. The existing longitudinal research on their reciprocal relations does not allow inferences to be made about within-person dynamics (see Murayama et al., 2017)." Niepel et al. (2021) applied newly developed dynamic SEM models (Muthén & Muthén, 1998–2019) to show significant reciprocal effects between achievement and ASC on a lesson-to-lesson basis. Relatedly, Asparouhov and Muthén (2021) argued that evidence based on four or fewer time points should be regarded as mostly cross-sectional because there is an insufficient sampling of time points to warrant conclusions about how constructs evolve over time.

Overall, these approaches have in common that they try to account for the potentially biasing effects of confounding variables when estimating cross-lagged effects. It is important for future methodological research to further clarify whether these approaches allow for tests of the REM that are more robust to the presence of unmeasured confounders and the potential dangers of under- or over-adjusting for covariates. However, we also note that alternative approaches to testing REM hypotheses need to be evaluated in relation to testing REM hypotheses rather than additional features that they incorporate that are not specific to REM hypotheses. Indeed, simulation and real data comparisons based on some of these models note that more complex models typically have reduced power concerning specific parameter estimates and frequently have convergence issues (e.g., Orth et al., 2021; Usami, Murayama, et al., 2019; Usami, Todo, et al., 2019).

Tests of the REM hypothesis about the directional ordering are clearly causal in nature and rely on strong assumptions underpinning the model. Although our research extends REM research's scope, threats to the validity of causal interpretations remain. Both the CLPM (with lag-2 effects and covariates) and the RI-CLPM have offsetting strengths and weaknesses. Although new approaches have been posited for CLPM more generally, these have not been widely applied to test the REM. However, an alternative direction for future REM research is to formulate random control trial (RCT) interventions to more formally test implications claimed from non-experimental REM research (e.g., Bailey et al., 2018). Thus, for example, Haney and Durlak's (1998) meta-analysis of self-concept interventions concluded—consistent with REM inferences—that interventions specifically designed to enhance self-concept not only had significant effects on self-concept, but also had positive effects on academic achievement. In this respect, there is experimental evidence that improving academic self-concept will improve subsequent academic performance—the key REM hypothesis. REM research suggests that simultaneously enhancing both ASC and achievement will be more beneficial than enhancing one to the exclusion of the other. Extending Haney and Durlak's (1998) meta-analysis and REM research more generally, this implication can be tested in a 2

(ASC intervention or not)×2 (achievement intervention or not) RCT design. The REM would predict that the group receiving both ASC and achievement interventions would show significant benefits compared to groups receiving only one of the two interventions. The effectiveness of each intervention in isolation could be evaluated in relation to the no-treatment control group that received neither intervention. However, there are likely to be many complications in implementing this design that might compromise the interpretation of the results.

Strengths, Limitations, and Directions for Further Research

Our study is strong in terms of the size and representativeness of the sample of German secondary students and annual waves over all five years of compulsory secondary schooling. However, there is also a need to test the generalizability of our results to other age groups, countries, and school settings.

Perhaps the most significant contribution of our research is the generalizability of support for the REM provided from different modeling approaches (CLPMs and RI-CLPMs). Hamaker et al. (2015) and others have noted that there is no generalizable, a priori empirical basis predicting how results based on CLPMs and RI-CLPMs will differ. Of course, CLPMs and RI-CLPMs will provide similar results in the unlikely situation where the variance of the global trait factors in RI-CLPM is zero. However, this similarity occurs in our study even though the MSC and achievement were highly stable over time. Substantively, the results are important, showing that our support for REM hypotheses generalizes over the alternative interpretations based on CLPMs and RI-CLPMs. Of course, this will not always be the case, and there are examples of where CLPM and RI-CLPM tests of REM hypotheses result in different results (e.g., Ehm et al., 2019; but also see Ehm et al., 2021 and earlier discussion). However, even when there are differences, it is useful to evaluate why there are differences and how these relate to support of the REM hypothesis.

Methodologically, we demonstrate a more robust methodological framework for evaluating directional ordering and extensions of existing research lacking in educational psychology. Nevertheless, our study also provides a challenge to more fully evaluate characteristics that lead to consistent and inconsistent results based on CLPMs and RI-CLPMs. RI-CLPM researchers often note that it is impossible to predict a priori how CLPM and RI-CLPM results will differ (Hamaker et al., 2015; Murayama et al., 2017). Although this might be true without knowing any characteristics of the variables, sample, and study, it is important to establish what constructs, theoretical models of their relations, and study characteristics are associated with consistent and inconsistent results for the CLPMs and RI-CLPMs—particularly in relation to the critical cross-lag paths used to test for directional ordering.

We assessed MSC with self-report measures that might introduce method effects that distort relations. However, this is a complicated issue as students are best suited to judge their own MSCs. For RI-CLPMs, method effects that are stable over time are likely to be absorbed into the global (decomposed between-person) trait effects but have little influence on with-person stability and cross-lag paths (but also see

discussion of measurement error). For CLPMs, such method effects are likely to inflate MSC stability paths. However, we also note that our study is based on relations with MSC and two objective achievement measures. Hence self-report method effects are less worrisome than RI-CLPMs and CLPMs where all the constructs are based on self-reports (or even non-self-report measures likely to be contaminated by shared method effects). It would also be interesting to collect inferred MSC ratings by significant others (teachers, parents, peers) in future research. However, in the literature evaluating self-other agreement, inferred self-concepts are widely recognized to represent a different construct (Marsh, 2006; Marsh & Craven, 2006; Marsh & Martin, 2011; Marsh, Hau, et al., 2005; Marsh, Trautwein, et al., 2005). Thus the addition of inferred self-concept ratings would raise new theoretical questions about the directional ordering of MSCs and inferred MSCs, and how each is related to achievement.

We note insufficient attention is given to the underlying measurement model, particularly RI-CLPMs based mainly on manifest variables. Unless the measurement model is well defined, the application of structural models is dubious. Nevertheless, this is rarely considered, particularly in manifest CLPMs and RI-CLPM. The fit of this measurement model also provides an important basis of comparison for subsequent CLPMs and RI-CLPMs and preliminary insights into the nature of the data (see Table 2). We evaluated the traditional set of factorial invariance constraints (configural, strong, and strict invariance over time) in our longitudinal measurement model. Support for at least metric invariance underpins the rationale for particularly RI-CLPMs, but also longitudinal structural invariance constraints imposed by subsequent CLPMs and RI-CLPMs. For MSC, we had multiple indicators that allowed us to control for method effects idiosyncratic to specific items using the correlated uniqueness approach that are unlikely to be controlled with manifest models.

Nevertheless, like many previous studies, we relied on single measures of each of our achievement indicators. Although it would be possible to treat school grades and test scores as multiple indicators of a latent trait, previous theoretical and empirical research argues that it is important to consider these as separate constructs (Marsh, 2006; Marsh & O'Mara, 2008; Marsh, et al., 2018a, 2018b; Marsh, Hau, et al., 2005; Marsh, Trautwein, et al., 2005). It would also be possible to include estimates of measurement error in the measurement model, but challenging to incorporate the complex error structure typical in longitudinal data (i.e., the contrasting effects of measurement error and correlated uniquenesses) without multiple indicators. There are also conceptual complications in assessing achievement over multiple school years when instructional content changes each year. Although beyond the scope of the present investigation, these are important issues for further research.

We also note that although RI-CLPMs adapt a within-person perspective, they fall short of a fully idiographic approach that models the separate effects of each individual (e.g., Beltz et al., 2016; Molenaar, 2004). Indeed, in RI-CLPMs, the within-person deviations in RI-CLPMs are modeled as typical between-person regressions (i.e., effects are constant across individuals). Thus, for example, RI-CLPMs do not answer the idiographic question of what proportion of the students conforms to REM hypotheses. Hence, both CLPMs and RI-CLPMs fail to articulate within-person processes that underpin the dynamic relations between ASC and

achievement, which remain a black box (Niepel et al., 2021; also see Murayama et al., 2017). A direction for further research is to evaluate the REM from a more idiographic approach such as group iterative multiple model estimation (Beltz et al., 2016) that integrates nomothetic and idiographic approaches. In addition, more idiographic research might better inform policy and practice designed to cater to the distinct needs of individual students.

Implications

Our substantive-methodological synergy has substantive, theoretical, policy/practice, and methodological implications. Substantively, our research shows that MSC and achievement are reciprocally for secondary school students. Furthermore, these findings have important policy implications, demonstrating from a within-person perspective that the enhancement of positive academic self-beliefs and academic achievement is mutually reinforcing.

Our research also has theoretical implications. In contrast to unidirectional (skill development and self-enhancement) models, we found good support for REM hypotheses. MSC and achievement are reciprocally related not only from a between-person (interindividual differences) perspective but also from a within-person perspective. Thus, high MSC is likely to lead to high achievement, and high achievement is likely to lead to higher MSC. This is important for developing interventions that target both achievement and MSC are likely to be more effective than interventions that focus on only one of these constructs.

REM studies of the directional ordering of ASC and achievement in educational psychology are mostly narrowly focused on ASC theory. However, Fredrickson's (2001) broaden-and-build theory posits reciprocal effects between self-beliefs and outcomes that, if sufficiently large, might create positive gain spirals. More broadly, positive reciprocal effects and positive upward spirals are consistent with major psychological theories: social cognitive theory (Bandura, 1986); broaden-and-build theory (Fredrickson, 2001); reciprocal effects models of appraisals, emotions, and achievement (Pekrun, 1992, 2006; Pekrun et al., 2017); job-demand resources model (Bakker & Demerouti, 2014, 2017); and the conservation of resources model (Hobfoll & Shirom, 2001). Thus, in future REM studies, ASC and educational-psychology researchers should draw more broadly on the different theoretical frameworks.

Methodologically, we outline longitudinal design issues, juxtaposing and extending the major statistical models (CLPMs and RI-CLPMs) to test directional ordering. Based on this juxtaposition of the models and our results, we recommend that applied researchers test both models and draw conclusions on comparing results from different models relevant to their research questions. As shown here, CLPMs and RI-CLPMs are not antagonistic; each has counter-balancing strengths and weaknesses. Hence, their juxtaposition is substantively, theoretically, and methodologically informative. More specifically, if researchers want to investigate relations between variables from both undecomposed between-person and within-person perspectives, and theorize that relations exist at both levels, then using both modeling approaches may be helpful. Although relevant to

educational psychology, the theoretical, design, and statistical issues considered here have broad generalizability to other psychological disciplines and applied research more generally.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10648-022-09662-9.

Acknowledgements We would like to acknowledge some of the many colleagues who have contributed to our thinking on issues addressed here, often agreeing with us but sometimes disagreeing in ways that changed our thinking. In no particular order, these include: David Kenny; Ellen Hamaker; Ulrich Orth; Philip D. Parker; Kou Murayama; Jiesi Guo; Geetanjali Basarkod; Theresa Dicke; James Nicholas Donald; Alexandre J.S. Morin; Alexander Robitzsch.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

References

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.* Advance online publication. https://doi.org/10.1037/met0000285

Arens, A. K., Marsh, H. W., Pekrun, R., Lichtenfeld, S., Murayama, K., & vom Hofe, R. (2017). Math self-concept, grades, and achievement test scores: Long-term reciprocal effects across five waves and three achievement tracks. *Journal of Educational Psychology, 109*(5), 621–634. https://doi.org/10.1037/edu0000163

Asendorpf, J. B. (2021). Modeling developmental processes. In J. R. Rauthmann (Ed.), *Handbook of personality dynamics and processes* (pp. 815–835). London, UK. https://doi.org/10.1016/B978-0-12-813995-0.00031-5

Asparouhov, T. & Muthén, B. (2021). *Residual structural equation models.* Technical Report. Version 1. November 1, 2021.

Bailey, D. H., Duncan, G. J., Watts, T., Clements, D. H., & Sarama, J. (2018). Risky business: Correlation and causation in longitudinal studies of skill development. *American Psychologist, 73*(1), 81–94. https://doi.org/10.1037/amp0000146

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*, 912–921. https://doi.org/10.1037/dev0000902

Bakker, A. B., & Demerouti, E. (2014). Job demands-resources theory. In C. L. Cooper (ed.), *Wellbeing: A complete reference guide* (pp. 1–28). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118539415.wbwell019

Bakker, A. B., & Demerouti, E. (2017). Job demands-resources theory: Taking stock and looking forward. *Journal of Occupational Health Psychology, 22*(3), 273–285. https://doi.org/10.1037/ocp0000056

Bandura, Albert. (1986). *Social foundations of thought and action: A social cognitive theory.* Prentice-Hall.

Basarkod, G., Marsh, H., Guo, J., Dicke, T., Xu, K. M., & Parker, P. (2020). *The Big-Fish-Little-Pond Effect for reading self-beliefs: A cross-national exploration with PISA 2018*. https://doi.org/10.35542/osf.io/7wbxj

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

Beltz, A. M., Wright, A. G., Sprague, B. N., & Molenaar, P. C. (2016). Bridging the nomothetic and idiographic approaches to the analysis of clinical data. *Assessment, 23*(4), 447–458. https://doi.org/10.1177/1073191116648209

Biaconcini, S., & Bollen, K. A. (2018). The latent variable-autoregressive latent trajectory model: A general framework for longitudinal data analysis. *Structural Equation Modeling, 25*, 791–808.

Bollen, K. A., & Curran, P. J. (2006). *Latent Curve Models: A Structural Equation Perspective*. Wiley. https://doi.org/10.1002/0471746096

Burns, R. A., Crisp, D. A., & Burns, R. B. (2020). Re-examining the reciprocal effects model of self-concept, self-efficacy, and academic achievement in a comparison of the Cross-Lagged Panel and Random-Intercept Cross-Lagged Panel frameworks. *British Journal of Educational Psychology, 90*(1), 77–91.

Byrne, B. M. (1984). The general/academic self-concept nomological network: A review of construct validation research. *Review of Educational Research, 54*(3), 427–456. https://doi.org/10.3102/00346543054003427

Calsyn, R. J., & Kenny, D. A. (1977). Self-concept of ability and perceived evaluation of others: Cause or effect of academic achievement? *Journal of Educational Psychology, 69*(2), 136–145. https://doi.org/10.1037/0022-0663.69.2.136

Cattell, R. B. (1966). . Patterns of change: Measurement in relation to state dimension, trait change, lability, and process concepts. In R. B. Cattell (ed.), *Handbook of multivariate experimental psychology* (pp. 335–402).

Chen, X., Vallerand, R. J., & Padilla, A. M. (2021). On the role of passion in second language learning and flourishing. *Journal of Happiness Studies,* 1–19.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Curran, P. J., & Bauer, D. J. (2011). The disaggregation of within-person and between-person effects in longitudinal models of change. *Annual Review of Psychology, 62*, 583–619. https://doi.org/10.1146/annurev.psych.093008.100356

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.

Ehm, J.-H., Hasselhorn, M., & Schmiedek, F. (2019). Analyzing the developmental relation of academic self-concept and achievement in elementary school children: Alternative models point to different results. *Developmental Psychology, 55*(11), 2336–2351. https://doi.org/10.1037/dev0000796

Ehm, J.-H., Hasselhorn, M., & Schmiedek, F. (2021). The developmental relation of academic self-concept and achievement in elementary school children in the light of alternative models. *Zeitschrift für Pädagogische Psychologie,* 1–10. https://doi.org/10.1024/1010-0652/a000303

Enders, C. K. (2010). *Applied missing data analysis*. books.google.com.

Fredrickson, B. L. (2001). The role of positive emotions in positive psychology. The broaden-and-build theory of positive emotions. *The American Psychologist*, *56*(3), 218–226. https://doi.org/10.1037/0003-066X.56.3.218

Frenzel, A. C., Pekrun, R., Dicke, A.-L., & Goetz, T. (2012). Beyond quantitative decline: Conceptual shifts in adolescents' development of interest in mathematics. *Developmental Psychology, 48*(4), 1069–1082. https://doi.org/10.1037/a0026895

Granger, C. W. J. (1969). Investigating causal relations by cconometric models and cross-spectral methods. *Econometrica : Journal of the Econometric Society, 37*(3), 424. https://doi.org/10.2307/1912791

Guo, J., Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2015a). Directionality of the associations of high school expectancy-value, aspirations, and attainment: A longitudinal study. *American Educational Research Journal, 52*(2), 371–402. https://doi.org/10.3102/0002831214565786

Guo, J., Parker, P. D., Marsh, H. W., & Morin, A. J. S. (2015b). Achievement, motivation, and educational choices: A longitudinal study of expectancy and value using a multiplicative perspective. *Developmental Psychology, 51*(8), 1163–1176. https://doi.org/10.1037/a0039440

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods, 20*(1), 102–116. https://doi.org/10.1037/a0038889

Hamaker, E. L., Mulder, J. D., & van IJzendoorn, M. H. (2020). Description, prediction and causation: Methodological challenges of studying child and adolescent development. *Developmental Cognitive Neuroscience, 46*, 100867. https://doi.org/10.1016/j.dcn.2020.100867

Hamaker, E. L., & Muthén, B. (2020). The fixed versus random effects debate and how it relates to centering in multilevel modeling. *Psychological Methods, 25*(3), 365–379. https://doi.org/10.1037/met0000239

Haney, P., & Durlak, J. A. (1998). Changing selfesteem in children and adolescents: A metaanalytic review. *Journal of Clinical Child Psychology, 27*, 423–433.

Harter, S. (1998). The development of self-representations. In W. Damon (Ed.), S. Eisenberg (Vol. Ed), Handbook of child psychology (5th ed., pp. 553–617). New York USA: Wiley.

Hobfoll, S., & Shirom, A. (2001). Conservation of resources theory: Applications to stress and management in the workplace. In R. T. Golembiewski (Ed.), *Handbook of organizational behavior* (pp. 57–80). Marcel Dekker.

Huang, C. (2011). Self-concept and academic achievement: A meta-analysis of longitudinal relations. *Journal of School Psychology, 49*(5), 505–528. https://doi.org/10.1016/j.jsp.2011.07.001

Hübner, N., Wagner, W., Zitzmann, S., & Nagengast, B. (2022, January 14). How causal is a reciprocal effect? Contrasting traditional and new methods to investigate the reciprocal effects model of self-concept and achievement. https://doi.org/10.31234/osf.io/f3e8w

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences*. Cambridge University Press. https://doi.org/10.1017/CBO9781139025751

Jelicić, H., Phelps, E., & Lerner, R. M. (2009). Use of missing data methods in longitudinal studies: the persistence of bad practices in developmental psychology. *Developmental Psychology*, *45*(4), 1195–1199. 1.1037/a0015665

Jöreskog, K. G. (1979). *Statistical estimation of structural models in longitudinal investigations.* (J. R. Nesselroade & B. Baltes, Eds.). Academic Press.

Kenny, David 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

Little, T. D. (2013). *Longitudinal structural equation modeling*. Guilford Press.

Littlefield, A. K., King, K. M., Acuff, S. F., Foster, K. T., Murphy, J. G., & Witkiewitz, K. (2021). Limitations of cross-lagged panel models in addiction research and alternative models: An empirical example using project MATCH. *Psychology of Addictive Behaviors*. Advance online publication. https://doi.org/10.1037/adb0000750

Lüdtke, O., & Robitzsch, A. (2021, July 29). A critique of the random intercept cross-lagged panel model. *PsyArXiv*. https://doi.org/10.31234/osf.io/6f85c

Marsh, H. W. (1990). Causal ordering of academic self-concept and academic achievement: A multiwave, longitudinal panel analysis. *Journal of Educational Psychology, 82*(4), 646.

Marsh, H. W. (2006). *Self-concept theory, measurement, and research into practice: The role of self-concept in educational psychology.* (p. 88). British Psychological Society Vernon-Wall Lecture.

Marsh, H. W., Balla, J. R., & Hau, K.-T. (1996). Assessing goodness of fit: Is parsimony always desirable? *The Journal of Experimental Education, 64*(4), 364–390. https://doi.org/10.1080/00220973.1996.10806604

Marsh, H. W., & Craven, R. G. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective: Beyond seductive pleasure and unidimensional perspectives. *Perspectives on Psychological Science, 1*(2), 133–163. https://doi.org/10.1111/j.1745-6916.2006.00010.x

Marsh, H. W., Ellis, L., & Craven, R. G. (2002). How do preschool children feel about themselves? Unravelling measurement and multidimensional self-concept structure. *Developmental Psychology, 38*, 376–393.

Marsh, H. W., & Grayson, D. (1994). Longitudinal stability of latent means and individual differences: A unified approach. *Structural Equation Modeling, 1*, 317–359.

Marsh, H. W., & Hau, K.-T. (1996). Assessing goodness of fit: Is parsimony always desirable? *The Journal of Experimental Education, 64*(4), 364–390.

Marsh, H. W., & Hau, K.-T. (2007). Applications of latent-variable models in educational psychology: The need for methodological-substantive synergies. *Contemporary Educational Psychology, 32*, 151–171. https://doi.org/10.1016/j.cedpsych.2006.10.008

Marsh, H. W., Lüdtke, O., Nagengast, B., Morin, A. J. S., & Von Davier, M. (2013). Why item parcels are (almost) never appropriate: Two wrongs do not make a right—camouflaging misspecification with item parcels in CFA models. *Psychological Methods, 18*, 257–284. https://doi.org/10.1037/a0032773

Marsh, H. W., & Hau, K.-T. (2003). Big-Fish–Little-Pond effect on academic self-concept: A cross-cultural (26-country) test of the negative effects of academically selective schools. *American Psychologist, 58*(5), 364–376. https://doi.org/10.1037/0003-066X.58.5.364

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005a). Goodness of fit evaluation in structural equation modeling. In A. Maydeu-Olivares & J. McArdle (Eds.), Psychometrics: A festschrift to Roderick P. McDonald (pp. 275–340). Hillsdale, NJ: Erlbaum.

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., Huppert, F. A., Donald, J. N., Horwood, M. S., & Sahdra, B. K. (2020). The well-being profile (WB-Pro): Creating a theoretically based multidimensional measure of well-being to advance theory, research, policy, and practice. *Psychological Assessment, 32*(3), 294–313. https://doi.org/10.1037/pas0000787

Marsh, H. W., Kuyper, H., Morin, A. J. S., Parker, P. D., & Seaton, M. (2014a). Big-fish-little-pond social comparison and local dominance effects: Integrating new statistical models, methodology, design, theory and substantive implications. *Learning and Instruction, 33*, 50–66. https://doi.org/10.1016/j.learninstruc.2014.04.002

Marsh, H. W., Kuyper, H., Seaton, M., Parker, P. D., Morin, A. J. S., Möller, J., & Abduljabbar, A. S. (2014b). Dimensional comparison theory: An extension of the internal/external frame of reference effect on academic self-concept formation. *Contemporary Educational Psychology, 39*(4), 326–341. https://doi.org/10.1016/j.cedpsych.2014.08.003

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A new look at the big-five factor structure through exploratory structural equation modeling. *Psychological Assessment, 22*, 471–491. https://doi.org/10.1037/a0019227

Marsh, H. W., & Martin, A. J. (2011). Academic self-concept and academic achievement: Relations and causal ordering. *The British Journal of Educational Psychology, 81*(Pt 1), 59–77. https://doi.org/10.1348/000709910X503501

Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014c). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology, 10*, 85–110. https://doi.org/10.1146/annurev-clinpsy-032813-153700

Marsh, H. W., & O'Mara, A. (2008). Reciprocal effects between academic self-concept, self-esteem, achievement, and attainment over seven adolescent years: Unidimensional and multidimensional perspectives of self-concept. *Personality and Social Psychology Bulletin, 34*(4), 542–552. https://doi.org/10.1177/0146167207312313

Marsh, H. W., Parker, P. D., & Morin, A. J. S. (2016a). Invariance testing across samples and time: cohort-sequence analysis of perceived body composition. In N. Ntoumanis & N. Myers (eds.), *Introduction to Intermediate and AdvancedStatistical Analyses for Sport and Exercise Scientists*. Wiley-Blackwell Publishing, Inc.

Marsh, H. W., Pekrun, R., Lichtenfeld, S., Guo, J., Arens, A. K., & Murayama, K. (2016b). Breaking the double-edged sword of effort/trying hard: Developmental equilibrium and longitudinal relations among effort, achievement, and academic self-concept. *Developmental Psychology, 52*(8), 1273–1290. https://doi.org/10.1037/dev0000146

Marsh, H. W., Pekrun, R., Murayama, K., Arens, A. K., Parker, P. D., Guo, J., & Dicke, T. (2018a). An integrated model of academic self-concept development: Academic self-concept, grades, test scores, and tracking over 6 years. *Developmental Psychology, 54*(2), 263–280. https://doi.org/10.1037/dev0000393

Marsh, H. W., Pekrun, R., Parker, P. D., Murayama, K., Guo, J., Dicke, T., & Arens, A. K. (2018b). 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

Marsh, H. W., Pekrun, R., Parker, P. D., Murayama, K., Guo, J., Dicke, T., & Lichtenfeld, S. (2017). Long-term positive effects of repeating a year in school: Six-year longitudinal study of self-beliefs, anxiety, social relations, school grades, and test scores. *Journal of Educational Psychology, 109*(3), 425–438. https://doi.org/10.1037/edu0000144

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005b). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child Development, 76*(2), 397–416. https://doi.org/10.1111/j.1467-8624.2005.00853.x

Marsh, H. W., Van Zanden, B., Parker, P. D., Guo, J., Conigrave, J., & Seaton, M. (2019). Young women face disadvantage to enrollment in university STEM coursework regardless of prior achievement and attitudes. *American Educational Research Journal, 56*(5), 1629–1680.

Marsh, H. W., & Yeung, A. S. (1997). Coursework selection: Relations to academic self-concept and achievement. *American Educational Research Journal, 34*(4), 691–720. https://doi.org/10.3102/00028312034004691

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology, 60*, 577–605. https://doi.org/10.1146/annurev.psych.60.110707.163612

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543. https://doi.org/10.1007/BF02294825

Millsap, RE 2012 Statistical approaches to measurement invariance. Londn: Routledge. https://doi.org/10.4324/9780203821961

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement, 2*, 201–218.

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

Mund, M., Johnson, M. D., and Nestler, S. (2021). Changes in Size and Interpretation of Parameter Estimates in Within-Person Models in the Presence of Time-Invariant and Time-Varying Covariates. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2021.666928

Mund, M., & Nestler, S. (2019). Beyond the cross-lagged panel model: Next-generation statistical tools for analysing dependencies across the life course. *Advances in Life Course Research, 41*, 100249.

Murayama, K., Goetz, T., Malmberg, L. E., Pekrun, R., Tanaka, A., & Martin, A. J. (2017). Within-person analysis in educational psychology: Importance and illustrations. In P. D. W. & S. K. (eds.), *Psychological Aspects of Education – CurrentTrends: The Role of Competence Beliefs in Teaching and Learning* (pp. 71–87). Wiley.

Murayama, Kou, Pekrun, Reinhard, Lichtenfeld, Stephanie, & vom Hofe, Rudolf. (2013). Predicting Long-Term Growth in Students' Mathematics Achievement: The Unique Contributions of Motivation and Cognitive Strategies. *Child Development, 84*(4), 1475–1490. https://doi.org/10.1111/cdev.12036.

Muthén, L. K., & Muthén, B. O. (2008–19). *Mplus User's Guide*. (Version 8)

Nagengast, B., & Marsh, H. W. (2011). The negative effect of school-average ability on science self-concept in the UK, the UK countries and the world: The Big-Fish-Little-Pond-Effect for PISA 2006. *Educational Psychology, 31*(5), 629–656.

Newman, D. A. (2014). Missing data. *Organizational Research Methods, 17*(4), 372–411. https://doi.org/10.1177/1094428114548590

Niepel, C., Marsh, H. W., Guo, J., Pekrun, R., & Möller, J. (2021). Revealing dynamic relations between mathematics self-concept and perceived achievement from lesson to lesson: An experience-sampling study. *Journal of Educational Psychology*. Advance online publication. https://doi.org/10.1037/edu0000716

Núñez-Regueiro, F., Juhel, J., Bressoux, P., & Nurra, C. (2021). Identifying reciprocities in school motivation research: A review of issues and solutions associated with cross-lagged effects models. *Journal of Educational Psychology.* Advance online publication. https://doi.org/10.1037/edu0000700

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

Parker, P. D., Marsh, H. W., Ciarrochi, J., Marshall, S., & Abduljabbar, A. S. (2014). Juxtaposing math self-efficacy and self-concept as predictors of long-term achievement outcomes. *Educational Psychology, 34*(1), 29–48. https://doi.org/10.1080/01443410.2013.797339

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

Pekrun, R. (1990). Social support, achievement evaluations, and self-concepts in adolescence. In L. Oppenheimer (Ed.), *The self-concept* (pp. 107-119). Springer.

Pekrun, R. (1992). The impact of emotions on learning and achievement: Towards a theory of cognitive/motivational mediators. *Applied Psychology, 41*(4), 359–376. https://doi.org/10.1111/j.1464-0597.1992.tb00712.x

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review, 18*(4), 315–341. https://doi.org/10.1007/s10648-006-9029-9

Pekrun, R., Frenzel, A. C., Goetz, T., & Perry, R. P. (2007). The control-value theory of achievement emotions. In P. Schutz & R. Pekrun (Eds.), *Emotion in Education* (pp. 13–36). Elsevier. https://doi.org/10.1016/B978-012372545-5/50003-4

Pekrun, R., Lichtenfeld, S., Marsh, H. W., Murayama, K., & Goetz, T. (2017). Achievement emotions and academic performance: Longitudinal models of reciprocal effects. *Child Development, 88*(5), 1653–1670. https://doi.org/10.1111/cdev.12704

Pekrun, R., Murayama, K., Marsh, H. W., Goetz, T., & Frenzel, A. C. (2019). Happy fish in little ponds: Testing a reference group model of achievement and emotion. *Journal of Personality and Social Psychology, 117*(1), 166–185. https://doi.org/10.1037/pspp0000230

Piaget, J., & Inhelder, B. (1969). *The psychology of the child*. Basic Books.

Reichardt, C. S. (2019). *Quasi-Experimentation: A guide to design and analysis*. Guilford Press.

Ruble, D. (1983). The development of social comparison processes and their role in achievement-related self-socialization. In E. Higgins, D. Ruble, & W. Hartup (Eds.), *Social cognition and social behavior: Developmental perspectives* (pp. 134–157). Cambridge University Press.

Schuurman, N. K., & Hamaker, E. L. (2019). Measurement error and person-specific reliability in multilevel autoregressive modeling. *Psychological Methods, 24*(1), 70–91. https://doi.org/10.1037/met0000188

Seaton, M., Marsh, H. W., & Craven, R. G. (2009). Earning its place as a pan-human theory: Universality of the big-fish-little-pond effect across 41 culturally and economically diverse countries. *Journal of Educational Psychology, 101*(2), 403.

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selection bias in observational studies. *Psychological Methods, 15*(3), 250–267. https://doi.org/10.1037/a0018719

Usami, S., Murayama, K., & Hamaker, E. L. (2019a). A unified framework of longitudinal models to examine reciprocal relations. *Psychological Methods, 24*(5), 637–657. https://doi.org/10.1037/met0000210

Usami, S., Todo, N., & Murayama, K. (2019b). Modeling reciprocal effects in medical research: Critical discussion on the current practices and potential alternative models. *PLoS ONE, 14*(9), e0209133. https://doi.org/10.1371/journal.pone.0209133

Valentine, J. C., DuBois, D. L., & Cooper, H. (2004). The relation between self-beliefs and academic achievement: A meta-analytic review. *Educational Psychologist, 39*(2), 111–133. https://doi.org/10.1207/s15326985ep3902_3

VanderWeele, T. J., Jackson, J. W., & Li, S. (2016). Causal inference and longitudinal data: A case study of religion and mental health. *Social Psychiatry and Psychiatric Epidemiology, 51*, 457–1466.

VanderWeele, T. J. (2019). Principles of confounder selection. *European Journal of Epidemiology, 34*(3), 211–219. https://doi.org/10.1007/s10654-019-00494-6

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

Van Lissa, C. J., Keizer, R., Van Lier, P. A. C., Meeus, W. H. J., & Branje, S. (2019). The role of fathers' versus mothers' parenting in emotion-regulation development from mid–late adolescence: Disentangling between-family differences from within-family effects. *Developmental Psychology, 55*(2), 377–389. https://doi.org/10.1037/dev0000612

Voelkle, M. C., Brose, A., Schmiedek, F., & Lindenberger, U. (2014). Toward a unified framework for the study of between-person and within-person structures: Building a bridge between two research paradigms. *Multivariate Behavioral Research, 49*(3), 193–213. https://doi.org/10.1080/00273171.2014.889593

Wu, C. H., & Griffin, M. A. (2012). Longitudinal relationships between core self-evaluations and job satisfaction. *Journal of Applied Psychology, 97*(2), 331.

Wu, H., Guo, Y., Yang, Y., Zhao, L., & Guo, C. (2021). A Meta-analysis of the Longitudinal Relationship Between Academic Self-Concept and Academic Achievement. *Educational Psychology Review*, 1–30.

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. *Organizational Research Methods, 23*, 651–687.

Authors and Affiliations

Herbert W. Marsh[1,2] · Reinhard Pekrun[1,3,4] · Oliver Lüdtke[5,6]

Reinhard Pekrun
repekrun@acu.edu.au; pekrun@lmu.de

Oliver Lüdtke
oluedtke@leibniz-ipn.de

1 Institute for Positive Psychology and Education, Australian Catholic University, North Sydney 2060, Australia

2 Oxford University, Oxford, England

3 University of Essex, Colchester, England

4 University of Munich, Munich, Germany

5 Department of Educational Measurement, IPN – Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, 24118 Kiel, Germany

6 Centre for International Student Assessment (ZIB), Kiel, Germany