The Reciprocal Effects Model is Robust to Alternative Modeling Specifications: A Response to Sorjonen et al., 2025

Accepted for publication in Educational Psychology Review

The Reciprocal Effects Model is Robust to Alternative Modeling Specifications: A Response to Sorjonen et al., 2025

Fernando Núñez-Regueiro[1] · Herbert W. Marsh[2,3] · Reinhard Pekrun[2,4,7] · Oliver Lüdtke[5,6] · Jiesi Guo[2]

[1] Contextual Learning Research Laboratory (LaRAC), Université Grenoble Alpes, 38000 Grenoble, France

[2] Institute for Positive Psychology and Education, Australian Catholic University, North Sydney 2060, Australia

[3] Department of Education, Oxford University, Oxford, UK

[4] Department of Psychology, University of Essex, Colchester, UK

[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

[7] Department of Psychology, Ludwig-Maximilians-University of Munich, Munich, Germany

Abstract

In a recent commentary, Sorjonen et al. (2025) reanalyzed simulated data based on Marsh et al. (2024), proposing several alternative models to account for reciprocal effects between academic self-concept (ASC) and achievement (ACH). Their contribution is a welcome addition to the ongoing dialogue on longitudinal modeling, expanding the range of theoretically grounded tests of reciprocal processes. While valuable, the models they proposed relied on specifications that, in our view, raised some concerns about model validity or did not fully capture the temporal structure central to Marsh et al.'s (2024) extended reciprocal effects model (REM), which posits contemporaneous skill-development effects (of ACH on ASC) and lagged self-enhancement effects (of ASC on ACH). In addition to responding to Sorjonen et al., we present a generalizable framework for comparing longitudinal models that differ in their assumptions about change processes, temporal structure, and measurement design. Using this comparative framework, we reanalyzed both the simulated data constructed by Sorjonen et al. and the original PALMA dataset used in Marsh et al. (2024). Although Sorjonen et al. questioned the evidence, our reanalyses provide converging support for the extended REM and, in particular, for an effect of academic self-concept on achievement. The extended REM provided the most probable approximation of the data structure and was robust to several alternative modeling specifications. We provide open-access scripts to replicate results and implement the proposed comparative modeling framework in various research settings (see OSF repository: https://osf.io/84bzp/).

Keywords Reciprocal effects model · Data-generation process · Model testing · Contemporaneous effects · Lag0 and lag1 effects

Introduction

In their commentary, Sorjonen et al. (2025) presented a reanalysis of the data and models reported in Marsh et al. (2024), proposing an alternative set of explanations for the processes underlying reciprocal effects between academic self-concept and achievement, commonly referred to as the reciprocal effects model (REM). Their work is a welcome and valuable endeavor for scientific debate. It makes a strong case for comparing competing models making distinct assumptions about the nature of longitudinal relations between dynamic constructs (Usami et al., 2019a). Their main contribution lies in expanding the range of models explored by Marsh et al., thereby providing a wider range of tests for possible data-generation processes.

In their model comparisons, Sorjonen and colleagues showed that alternative models adjusting for occasion-specific unique variance (stable trait, autoregressive trait and state model—STARTS), positing unidirectional effects of achievement on self-concept (e.g., extended skill development model—ESDM), or assuming intertwined processes of long-term and short-term changes (e.g., latent change score model—LCSM), showed almost equivalent fit to the data as the models presented in Marsh et al. (2024), yet contradicted the claim that academic self-concept had a prospective, positive effect on achievement. They interpreted their analyses as indicating a lack of conclusive evidence for the REM.

Re-analyses such as those conducted by Sorjonen et al. (2025) play a valuable role in testing the robustness of theoretical claims and encouraging refinement of modeling approaches. Our aim in this response is to clarify and empirically test the assumptions underlying Sorjonen et al.'s alternative specifications. We embrace their call for a productive discussion on how best to capture processes occurring at different timescales, disentangle trait and state variance, and incorporate potential sources of confounding or measurement error. At the same time, we emphasize the importance of theoretically and methodologically grounded rationales for model specification—particularly in relation to the comparison of lagged and contemporaneous effects—and propose that model selection should not rely solely on fit indices, but rather on a broader framework combining theory, parsimony, and interpretability. To this aim, we develop a generalizable framework for evaluating competing longitudinal models using real data and theoretically grounded criteria (see open-access scripts at https://osf.io/84bzp/).

The Extended REM Developed in Marsh et al. (2024)

The REM is a theory of motivation positing that academic self-concept contributes to achievement (**the self-enhancement effect**) and, reciprocally, that achievement contributes to academic self-concept (**the skill-development effect**; Marsh & Craven, 2006). The REM has been corroborated by several meta-analyses (Huang, 2011; Valentine et al., 2004; Vu et al., 2024). Studies have nonetheless

questioned the evidence, raising modeling issues such as the confounding of long-term (trait factors) and short-term (state factors) processes of change (Ehm et al., 2019; Hübner et al., 2023; Núñez-Regueiro et al., 2022) and the lack of distinction between "contemporaneous" effects occurring within the same time point (from wave T to wave T) and "lagged" effects occurring across time points (e.g., from wave T to wave $T+1$; Muthén & Asparouhov, 2024).

In light of these critiques, Marsh et al. (2024) refined the analysis of REM processes by disaggregating state and trait factors using the random-intercept cross-lagged panel model (RI-CLPM, Fig. 1A; Hamaker et al., 2015), and by comparing RI-CLPMs contrasting cross-lagged and contemporaneous effects between academic self-concept and academic skills, or random-intercept contemporaneous and cross-lagged panel model (RI-CCLPM, Fig. 1B), following the recommendations by Muthén and Asparouhov (2024). Using data from a large sample of German secondary school students ($N = 3527$) and a 1-year time interval between waves, the study demonstrated that skill development effects were contemporaneous (rather than lagged). In contrast, self-enhancement effects were lagged (rather than contemporaneous). This pattern was interpreted as evidence for mutual influences of varying durations, with performance shaping self-concept in the short term (e.g., via short-term feedback within one single school year), and self-concept influencing achievement over longer periods of time (e.g., through sustained engagement from one school year to the next).

As shown in Fig. 1C, this extended REM (Marsh et al., 2024) is operationalized as a constrained RI-CCLPM—henceforth "RI-CCLPM-C" (corresponding to "RI-CCLPM-M5" in Marsh et al., 2024), which includes contemporaneous (lag0) skill-development effects and lagged (lag1) self-enhancement effects, with their respective cross-lagged and contemporaneous counterparts constrained to zero.

Alternative Models Proposed by Sorjonen et al. (2025)

According to Sorjonen et al. (2025), several alternative models may be considered to account for the observed co-evolution of ASC and ACH. Each of these models rests on specific assumptions about the generative processes underlying reciprocal effects, and thereby provides a different interpretation of what REM effects might capture. In this section, we briefly present the models tested in this study, with a focus on their assumptions regarding temporalities and the nature of causal relations. A unified framework for these models is presented elsewhere (Usami et al., 2019a).

RI-CLPM The first model tested by Sorjonen and colleagues corresponds to the RI-CLPM (Fig. 1A; Hamaker et al., 2015), which assumes that the reciprocal effects between ASC and ACH exist as lag1 effects, i.e., across waves. As the RI-CCLPM-C, this model locates REM effects at the within-person level of dynamic states, by isolating trait factors via random intercepts. However, this RI-CLPM differs from the RI-CCLPM-C of the extended REM (Fig. 1C), because it posits lag1 effects for both skill development and self-enhancement effects, instead of positing alternative

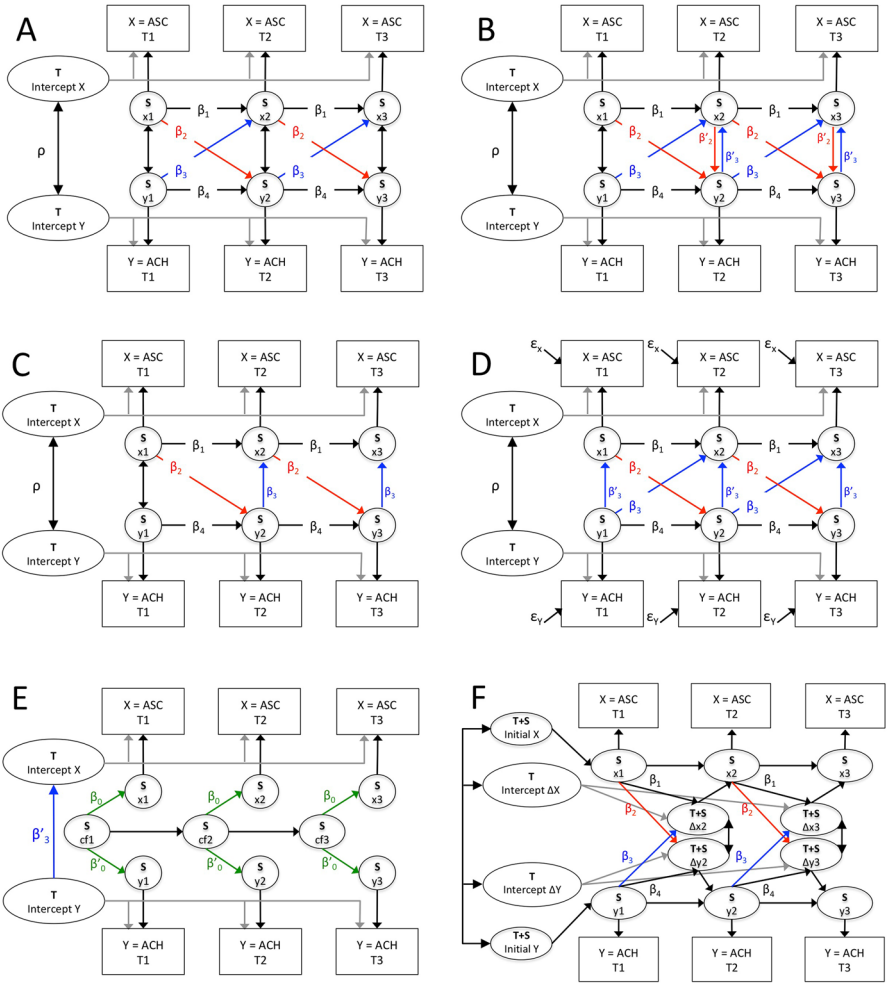

Fig. 1 Alternative Data-Generation Processes Accounting for Covariations Between Academic Self-Concept (ASC) and Achievement (ACH). Note. A) Random-intercept cross-lagged panel model (RI-CLPM; Hamaker et al., 2015); B) RI-CLPM with contemporaneous (lag0) effects (RI-CCLPM; Muthén & Asparouhov, 2024); C) RI-CCLPM with constrained effects (RI-CCLPM-C; Marsh et al., 2024); D) Stable trait, autoregressive trait and state model (STARTS; Kenny & Zautra, 1995) with contemporaneous effects starting on wave 1 (STARTS2 in Sorjonen et al., 2025); E) Extended skill development model (EDSM; Sorjonen et al., 2025); F) Latent change score model (LCSM; Ghisletta & McArdle, 2012; Kievit et al., 2018);. "T" and "S" denote trait- and state-like processes, which are disaggregated (A, B, C, D, E) or accumulated (F) depending on the model (Usami et al., 2019a)

temporal structures (lag0 vs. lag1 effects) as revealed in Marsh et al. (2024). In essence, however, this model does not contradict REM theory, but corresponds to an earlier version of the REM (Marsh et al., 2022).

STARTS A second model corresponds to a RI-CLPM that accounts for error in the measurement of ASC and ACH constructs on each time point, namely the bivariate STARTS (Kenny & Zautra, 1995). Whereas the RI-CLPM assumes that error is ignorable by fixing the residual variances of manifest variables to zero, the STARTS model estimates it to capture additional sources of within-person variation, such as measurement error and residual variance due to occasion-specific (unmeasured) influences. When constructs are modeled using latent variables with multiple indicators—as in Marsh et al. (2024)—measurement error is already accounted for by the measurement model, and the STARTS model primarily captures occasion-specific state variance. In theory, this may clarify whether REM effects are substantive or reflect, instead, heteroscedasticity or covariance between the errors or state variances of the two measures. This could happen, for example, if low- or high-achievers have a greater tendency to overestimate or underestimate their academic competence, introducing self-evaluation bias (a type of error) in the ASC measure (Jamain et al., 2021). In practice, separating residual error from reciprocal effects in the STARTS is challenging. Many waves of data (6–10 or more) are usually recommended to obtain reliable results using maximum likelihood estimates (Kenny & Zautra, 1995; Lüdtke et al., 2018). The error may be overestimated and bias the estimation of reciprocal effects (Usami et al., 2019b).

In Sorjonen's study, two variants of the STARTS were explored with 5 waves of data: a STARTS with lag1 effects (henceforth "STARTS"), and a STARTS with lag1 and lag0 effects (henceforth "STARTS2"; Fig. 1D). As we discuss hereafter, neither timescale structures corresponded to the RI-CCLPM-C in Marsh et al. (2024): The STARTS model included only lag-1 effects, whereas the STARTS2 model combined lag0 and lag1 skill-development effects (as opposed to lag0 only in RI-CCLPM-C) and allowed lag0 effects to begin at wave 1 (rather than wave 2).

ESDM Introduced by Sorjonen et al., the ESDM (Fig. 1E) presents a distinct interpretation. It suggests that REM effects across waves are better explained by a common, latent third-variable confounder (LTVC) that influences both constructs simultaneously (loadings β_0 and $\beta\prime_0$ in Fig. 1E), rather than by true reciprocal influence. This model implies that reciprocal effects observed at the level of wave-specific factors (i.e., REM effects) may be artefacts of unmeasured, wave-specific shocks—such as illness, test anxiety, or mood. Although wave-specific, the LTVC in the ESDM is allowed to be partly stable over time via autoregressive effects; less stable, autocorrelated LTVC structures have also been proposed elsewhere (see LTVC in Kenny & McCoach, 2025). The ESDM also repositions causality at the trait level, assuming that achievement shapes self-concept through stable, enduring differences rather than dynamic states. There is no temporal lag between achievement and self-concept specified in this model. The model therefore contradicts both the bidirectional causality of the REM, but also its claims about the temporal structure of skill development effects.

One potential limitation of the model is its limited causal interpretation. As the skill development effect is posited at the level of trait factors, it can hardly control for endogeneity in ASC and ACH processes (i.e., adjusting for influence of past dynamic states), as opposed to dynamic states in the RI-CLPM and variants that

control for endogeneity via autoregressive effects. Moreover, this model assumes a unidirectional causal effect (ACH trait influencing ASC trait, but not vice-versa), while at the same time positing this effect involves stable traits, that is, unvarying levels in ASC and ACH. In other words, the ESDM posits that unvarying levels in ACH influence (or "modify") unvarying levels in ASC, challenging common perspectives on causal inference (either predictive or experimental; VanderWeele et al., 2016). Finally, reversing the direction of the trait-level effect (i.e., ASC trait influencing ACH trait) to obtain an extended self-enhancement model (ESEM), or changing trait effects to a covariance among traits (random intercept LTVC, or RI-LTVC), would yield exactly the same fit to the data and the same estimates for the confounding factor effects (i.e., β_0 and $\beta\prime_0$ in Fig. 1E; see OSF scripts). In other words, evidence supporting the ESDM would equally support models contradicting the causal directionality of the ESDM (i.e., the ESEM or RI-LTVC), meaning causal directionality cannot easily be assessed. Nonetheless, beyond this issue, the latent confounder posited by the ESDM provides a useful robustness check, notably to test whether estimated reciprocal effects in the RI-CLPM (and lag0 variants) can be accounted for by this latent confounder. In this robustness check, one may verify whether a model including a latent confounder at the state level (e.g., ESDM, RI-LTVC) provides a better approximation of the data structure than a model positing reciprocal effects between states (e.g., RI-CLPM, STARTS or variants).

LCSM In contrast to previous models that separate trait and state factors of change, the LCSM (Fig. 1F) treats traits as accumulating factors that integrate state dynamics over time (Ghisletta & McArdle, 2012; Kievit et al., 2018; Usami et al., 2019a). More specifically, the LCSM decomposes observed scores into latent levels and latent change scores representing the difference in levels from one wave T to the next $T+1$ (for $T \geq 2$). This allows estimating how levels in one construct relate to future changes in another construct (i.e., effects β_2 and β_3 in Fig. 1F). At the same time, the accumulating factors are specified in the form of random intercepts capturing average change scores in the constructs across occasions. Therefore, level effects influence indirectly (via change scores) the value of accumulating factors (Usami et al., 2019a). This means that, per LCSM specification, REM effects reflect how fluctuations in one construct—such as short-term experiences of ACH (e.g., success or failure)—can accumulate into broader developmental trends in ASC (e.g., an emerging trait of self-concept), reflecting trait-state entanglement.

The LCSM remains limited for causal inference, because the trait–state entanglement blurs the distinction between enduring dispositions and momentary fluctuations: What appears to be a short-term level effect may, partly, reflect long-term changes accumulated through traits, and vice versa. This is true even when random intercepts or within-construct coupling effects are included (i.e., β_1 and β_4 in Fig. 1F; Usami et al., 2015). Yet, like the ESDM, the LCSM remains informative as a robustness check, notably to assess the extent to which results obtained under the assumption of trait–state disaggregation (as in the RI-CLPM, STARTS, and variants) hold when this assumption is relaxed. In this approach, one can evaluate whether models that allow trait–state entanglement in structural dynamics (LCSM) yield substantively different results on reciprocal relations and provide a better

approximation of the data structure, compared to models relying on trait–state disaggregation (RI-CLPM, STARTS, and variants).

Sorjonen et al. tested two LCSM variants, one assuming reciprocal effects between ASC and ACH within the same timescales (i.e., lag1 effects, henceforth "LCSM1"), and one assuming a single lag0 effect of ACH on ASC, with no reciprocal effect (henceforth "LCSM2"). Neither variant fully reflects the extended REM (i.e., the RI-CCLPM-C), due to the differences in temporal structure (LCSM1) or reciprocity of effects (LCSM2) and to the trait-state entanglement posited in these models.

Limitations of Simulated Data Used in Sorjonen et al. (2025)

Sorjonen et al. based their simulations on the latent correlation matrix reported in Table 2 of Marsh et al. (2024). This matrix provides a standardized snapshot of construct-level associations derived from a longitudinal CFA with strong invariance properties. All variables were standardized to a common metric ($M = 0$, $SD = 1$), anchored in the first measurement occasion. As such, the matrix omits both scale and variance information and does not provide information on occasion-specific residual variance (e.g., measurement error, state variance) for either ASC or ACH. This has two major implications for the results of the reanalysis proposed in Sorjonen et al. First, as information on measurement error or state variance is absent from the simulated matrix, the STARTS model has little information to locate or estimate the amount of this residual variance in ASC or ACH in the simulated data. Moreover, the manifest STARTS model (i.e., without latent factors) confounds measurement error and state-level variance. Thus, although a STARTS solution may be found, it may not be a good approximation of the measurement error and state-level variance clearly separated in the original PALMA dataset used by Marsh et al. (2024).

Second, simulating data from this matrix implicitly assumes that all ASC and ACH measures have fixed variances of 1. In actual longitudinal data, this assumption rarely holds. Construct variances often evolve over time due to developmental dynamics or environmental differentiation (true-score change or "alpha change"), or changes in measurement precision ("beta change") (Brown, 2015; Marsh & Hau, 1996; Marsh et al., 2010; Widaman et al., 2010). By enforcing equal variances across time, the simulations obscure these evolutions, thereby distorting the scaling of cross-lagged effects, model fit indices and, more generally, eliminating the possibility of testing change processes precisely. Any inferences drawn from such simulations must therefore be interpreted with caution. Although they provide suggestive results, they may not be reliable enough to adjudicate the validity of conclusions in Marsh et al. (2024). The same limitations concern Sorjonen et al.'s reanalysis of their simulated data from the TRAIN dataset (Hübner et al. (2023), which they used to reaffirm their critique of Marsh et al. (2024).

These limitations could have been addressed by using the original PALMA dataset. Although Sorjonen et al. (2025) noted that the dataset was unavailable to them (p. 30), no request for access was made. We would have gladly provided the data

upon request. In light of this, relying on simulated approximations—while perhaps understandable—was not strictly necessary and may have limited the robustness of their conclusions.

In summary, Sorjonen et al. did not simulate the PALMA dataset per se—they simulated an idealized dataset under a correlation matrix they claim is structurally similar to PALMA. However, without access to the raw data or the full covariance matrix, the simulations are approximate and assumption-heavy. Other modeling strategies in Marsh et al. (2024), not detailed here but important for rigorous causal inferences (e.g., inclusion of covariates, lag2 paths, correlated uniquenesses), were also absent from the reanalysis.

Limitations of Model Specifications in Sorjonen et al. (2025)

Nonetheless, the models discussed in Sorjonen et al. (2025) provide a valuable set to obtain a comparative framework in which distinct assumptions about data-generation processes in REM effects can be contrasted. Some of these models (i.e., RI-CLPM with contemporaneous effects, STARTS, LCMS) are highly complex and notoriously difficult to implement (Kenny & Zautra, 1995; Kievit et al., 2018; Muthén & Asparouhov, 2024). Obtaining reliable solutions and valid model comparisons is dependent on key model-building strategies. In this regard, judging from Sorjonen et al.'s (2025) commentary and the analysis script provided by the authors on Open Science Framework (OSF; https://osf.io/9sbvh/), their results were based on models that raise concerns about internal validity and may not have relied on the most informative criteria for model comparisons.

The first issue concerns the LCSM variants "LCSM1" and "LCSM2", noted "m.change" and "m.change2" in the OSF script. Both variants yielded non-positive-definite variance–covariance matrices. An inspection of the matrices shows implausible or impossible values (e.g., negative variances) that make the solutions inadmissible for interpretation. We also note that these models differed from the LCSMs referenced by the authors (Ghisletta & McArdle, 2012; Kievit et al., 2018), which may explain the convergence issues. For example, trait factors representing average change scores (i.e., accumulating traits) were defined across all waves of data (wave 1 to wave 5), instead of starting on wave 2, when change scores begin (Fig. 1F). Furthermore, the predictive effects for within-construct coupling–that is, how prior levels in a construct affect its subsequent change scores, represented by β_1 and β_4 in Fig. 1F–were specified as covariances rather than regression paths. Finally, covariances between initial levels and trait factors, and between change scores, were omitted from the models. Therefore, the negative reciprocal effects between ACH on ASC reported by Sorjonen et al. (i.e., for the LCSM1) are difficult to interpret and, given the model's improper solution, cannot be taken as strong evidence against the REM.

A second issue concerns the integration of "contemporaneous" or lag0 effects into the models. The methodology for lag0 effects was developed as a means to refine the interpretation of "lagged" effects (lag1 effect), notably because such effects assume that the residuals of the effect (e.g., ACH on T3) are uncorrelated

with the predictor (e.g., ASC on T2). When a lag0 effect truly exists but is ignored in the model (e.g., lag0 effect of T2 ACH on T2 ASC fixed to zero), estimates of lag1 effect may be distorted or even biased (Muthén & Asparouhov, 2024; Speyer et al., 2024). However, determining whether lag0 or lag1 effects are most relevant is nontrivial because estimating both simultaneously places strong constraints on the residual structure of state factors. To minimize the risk of misspecification, the recommended strategy (Muthén & Asparouhov, 2024) is to estimate the RI-CCLPM in which both predictors have mutual lag0 and lag1 effects, while applying constraints to avoid negative R^2 values and dual solutions (i.e., $0 < b\prime_2 b\prime_3 < 1$ in Fig. 1B). Provided the solution is admissible, this model enables verifying which lag0 or lag1 effects (or both) are reliable for model interpretation. The RI-CCLPM-C in Marsh et al. (2024) was obtained following this rigorous procedure.

In contrast, Sorjonen et al. (2025) deviated from this methodology when constructing their STARTS2 model. First, as shown in Fig. 1C, the lag0 effect of ACH on ASC was allowed to start on the first wave (i.e., lag0 effect from T1 ACH to T1 ASC), instead of starting on the second wave (T2). As such, without autoregressive effects at the first wave, the model cannot adjust for influences from prior states. As a result, reciprocal effects (lag0 or lag1) absorb unmodeled variance from these prior states, biasing the estimates and inflating their standard errors (potentially rendering them nonsignificant). In other words, such specification introduces endogeneity and compromises the validity of estimated lag0 and lag1 effects (Hamaker et al., 2015; Muthén & Asparouhov, 2024). Second, while the STARTS2 specified a lag1 effect of ASC on ACH and a lag0 effect of ASC on ACH (as in the RI-CCLPM-C), it also specified a lag1 effect of ACH on ASC (Fig. 1C). This means the STARTS2 model specified an asymmetrical structure (2 vs. 1 reciprocal effects). Yet, this structure was not based on results from the RI-CCLPM, making its relevance unknown. For these reasons, the finding reported by Sorjonen et al. (2025) that the STARTS2 contradicted expectations from the REM (i.e., negative and positive lag1 and lag0 effects of achievement on self-concept, nonsignificant lag1 effect of self-concept on achievement), was based on a model that raises concerns about internal validity and should be interpreted with caution.

Finally, a central message from the study in Sorjonen et al. (2025) appears to be that all alternative models to the RI-CCLPM-C (i.e., model for the extended REM) all had almost equivalent fit to the data, leading the authors to question whether evidence for the RI-CCLPM-C was truly conclusive. This claim was based on conventional fit indices such as the Comparative Fit Index (CFI) or the root mean squared error of approximation (RMSEA; see Table 1). These indices are particularly useful for comparing structurally similar models to a hypothesized data structure (e.g., parametric invariance testing; Chen, 2007; Jorgensen et al., 2018). However, their utility diminishes when comparing structurally distinct models based on different assumptions about the data-generating process. As demonstrated in prior work (Kenny & McCoach, 2025; Marsh et al., 2024; Muthén & Asparouhov, 2024), models positing very different processes may yield quasi-identical values on these indices, making them ineffective for model selection—particularly highly complex model that may fit all kinds of data. This is because highly complex models as those tested here (Fig. 1) offer considerable flexibility to reproduce the data structure, which can

Table 1 Fit indices of alternative models of REM effects

Model	χ^2	df	CFI	TLI	RMSEA	SRMR	wAIC (AIC)	wBIC (BIC)
Results from Sorjonen et al. (2025)								
RI-CLPM†	505.5	46	0.981	0.981	0.053	0.049	0 (76,384.1)	0 (76,501.3)
LCSM1††	1366.2	45	0.945	0.945	0.091	0.164	0 (77,246.9)	0 (77,370.2)
ESDM/RI-LTVC	596.7	45	0.977	0.977	0.059	0.039	0 (76,477.4)	0 (76,600.7)
STARTS†	316.5	44	0.989	0.988	0.042	0.039	0 (76,199.2)	0 (76,328.7)
STARTS2††	**265.9**	**44**	**0.991**	**0.991**	**0.038**	**0.043**	**1 (76,148.5)**	**1 (76,278.0)**
LCSM2††	1323.9	46	0.947	0.948	0.089	0.163	0 (77,202.5)	0 (77,319.7)
Reanalysis of simulated data								
RI-CCLPM-C	475.8	47	0.982	0.983	0.051	0.047	0 (76,352.4)	0.007 (76,463.5)
LCSM3	612.0	44	0.977	0.976	0.060	0.063	0 (76,494.6)	0 (76,624.1)
ESDM/RI-LTVC	596.7	45	0.977	0.977	0.059	0.039	0 (76,477.4)	0 (76,600.7)
STARTS3	**457.8**	**46**	**0.983**	**0.983**	**0.050**	**0.045**	**1 (76,336.4)**	**0.993 (76,453.6)**
Reanalysis of PALMA data								
RI-CCLPM-C	1214.8	522	0.983	0.981	0.019	0.038	0.001 (183,812.6)	0.012 (184,694.5)
LCSM3	1276.7	519	0.981	0.979	0.020	0.043	0 (183,890.6)	0 (184,791.0)
ESDM/RI-LTVC	1257.4	520	0.982	0.979	0.020	0.034	0 (183,864.8)	0 (184,759.0)
STARTS3	**1200.9**	**521**	**0.983**	**0.981**	**0.019**	**0.034**	**0.999 (183,797.6)**	**0.988 (184,685.7)**

Note. $N = 3527$ students on 5 waves. odels are presented in Fig. 1. Models in **bold font** correspond to the so-called "most probable" solutions identified based on AIC and BIC weights. Model probabilities based on AIC/BIC weights are conditional on the set of models included; they do not constitute absolute evidence of model superiority

†These models assumed an untested temporal structure (full lag1 effects) not compatible with Marsh et al. (2024)

††These models had inadmissible solutions (e.g., negative variances) or lacked internal validity (e.g., lag0 effects starting on wave 1, spurious lag1 effects)

artificially inflate model performance (e.g., obtaining excellent fit indices), regardless of whether the model is a good approximation of the underlying data-generation process.

Instead, the Akaike (AIC) or Bayesian (BIC) information criteria offer a more robust basis for model comparison in such contexts (Burnham & Anderson, 2002; Preacher & Yaremych, 2022; Wagenmakers & Farrell, 2004). As other information criteria, they are based on a combination of the model's chi-square statistic measuring data fit (low bias), plus a penalty term to account for the model's degree of parsimony (low uncertainty), thereby balancing data fit with model simplicity. Across low- to high-dimensional solutions, the solutions with lower AIC will generalize better across samples. The AIC is nonetheless sensitive to sample size and will select more complex models as N grows, increasingly capturing idiosyncrasies

(overfitting; Marsh & Balla, 1994). By contrast, the BIC will tend to favor low-dimensional solutions more robust to sample idiosyncrasies, but possibly overlooking nontrivial structural complexity (underfitting; Burnham & Anderson, 2002; Preacher & Yaremych, 2022). The AIC and BIC thus provide balanced information on the fit-to-parsimony tradeoff.

Furthermore, using AIC and BIC weights, one can estimate the relative probability ($0 < weight < 1$) that a given model is the most suitable available. For instance, although Sorjonen et al. (2025) suggest equivalent fit across models, judging from AIC and BIC weights (Table 1) the STARTS2 model appears to have the highest probability of being the most suitable among the models that were compared. However, this should not be taken to imply that STARTS2 offers a fully valid representation of the underlying data structure. As discussed earlier, the model's internal validity could be improved by first testing key structural assumptions (e.g., contrasting lag-0 vs. lag-1 effects), and by letting lag0 effects start on wave 1 (instead of wave 2). Moreover, the comparison is made against models with inadmissible solutions (e.g., LCSM1, LCSM2), raising further concerns. Using this case as an example, a broader cautionary note can be derived: Model selection based on AIC or BIC may not always offset limitations in model specification or theoretical grounding.

Consistent with Marsh et al.'s (2004) caution against relying on "golden rules", model evaluation requires a careful assessment of the strategies used to estimate parameters (e.g., obtaining proper solutions, testing components rigorously), as well as a substantive interpretation of the structural parameters in light of the research hypotheses (e.g., contemporaneous vs. lagged effects linking ASC and ACH). Neither component of such an evaluation was fully addressed in the critique by Sorjonen et al. (2025). A reanalysis of alternative models is warranted.

Reanalysis of Alternative Models Using the Simulated Data in Sorjonen et al. (2025)

We propose a reanalysis that addresses the modeling issues in Sorjonen et al. (2025), while reintegrating the substantive issue of lag0 versus lag1 effects proposed in Marsh et al. (2024), which received limited attention in the critique. For comparison, we apply this reanalysis on the same simulated data as Sorjonen et al. (2025), with reproducible scripts accessible on the OSF repository (https://osf.io/27vb4). A reanalysis with the original PALMA data is provided next (see next section).

The first step of the reanalysis consists of applying the methodology for contrasting lag0 and lag1 effects (Muthén & Asparouhov, 2024). To do so, we estimate the RI-CCLPM (Fig. 1B) on the simulated data, using random starts on Mplus. This model yielded a proper solution with interpretable results (see Fig. 2A.1). More precisely, the effect of ACH on ASC was significant as a lag0 effect ($\beta\prime_3 = [.358, .370], p < .001$) but was nonsignificant as a lag1 effect ($\beta_3 = [.012, .013], p = .491$). Conversely, the effect of ASC was significant as a lag1 effect ($\beta_2 = [.105, .113], p < .001$), but non-significant as a lag0 effect ($\beta\prime_2 = [.128, .132], p = .059$). In other words, contrasting the two kinds of effects establishes the predominance of a lagged self-enhancement effect (lag1 effect) and a

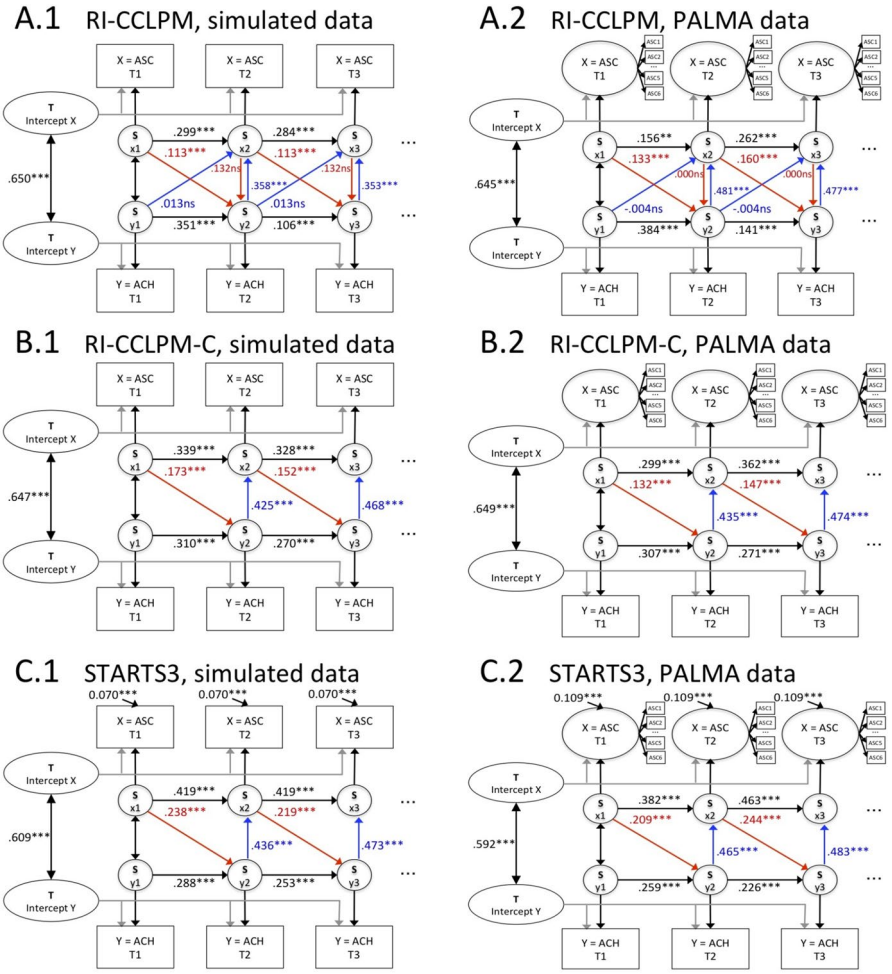

Fig. 2 Full Reciprocal Model (RI-CCLPM), Lag0-Lag1 REM Model (RI-CCLPM-C), and State Variance Adjustment (STARTS3) Across Simulated and PALMA Datasets. Note. The simulated (Sojonen et al., 2025) and PALMA (Marsh et al., 2024) datasets contained five waves of measurement but for simplicity the figure presents only the first three waves. In both datasets, the STARTS3 cannot estimate the residual variance of ASC and ACH simultaneously (negative variances or no convergence). Only ASC residual variances were estimated. ns = not significant; $*p < 0.05$; $**p < 0.01$; $***p < 0.001$

contemporaneous skills development effect (lag0 effect), consistent with the findings reported in Marsh et al. (2024).

The second step consists of correcting the alternative models while accounting for the valid temporal structure of REM effects (lag0 vs. lag1) and ensuring model convergence. This mainly concerned the RI-CLPM and the STARTS2 models, resulting in the RI-CCLPM-C revealed by Marsh et al. (Fig. 2B) and in a modified "STARTS3" model (Fig. 2C). We also re-specified the LCSM (i.e., traits loading on change scores, initial levels covarying with traits, including within-construct

coupling effects and covariances between change scores) to obtain a proper solution that aligns with more conventional LCSMs (as proposed by Ghisletta & McArdle, 2012; Kievit et al., 2018), yielding the "LCSM3". Comparing these re-specified models with the additional ESDM (or, equivalently, the RI-LTVC) revealed three findings. First, unlike the STARTS models in Sorjonen et al. (2025), the STARTS3 could not estimate construct errors (i.e., state variance and measurement error) for ASC and ACH simultaneously, as this yielded improper solutions (non-convergence, negative variances). This finding suggests that the construct errors in the STARTS or STARTS2 (in Sorjonen et al., 2025) may have been overestimated, potentially reflecting artifacts introduced by the use of lag0 and lag1 effects that do not align with the underlying data structure (Usami et al., 2019b). Second, unlike the LCSM variants in Sorjonen (LCSM1, LCSM2), the LCSM3 yielded an admissible solution and provided support for reciprocal effects between ASC and ACH ($\beta_2 = [.239, .290], p < .001$, $\beta_3 = [.267, .270], p < .001$, see details on OSF, https://osf.io/27vb4.).

Third, although all models showed satisfactory fit according to conventional fit indices, the STARTS3 stood out as the most plausible model for the underlying process that generated the data ($wAIC = 1$, $wBIC = 0.993$), even if these probabilities are sensitive to sample size and model complexity, and should not be interpreted as definitive evidence in isolation. Fourth, the STARTS3 (Fig. 2, B3) showed REM effects similar to those of the RI-CCLPM-C (Fig. 2, B2), differing only by slightly stronger effects of ASC on ACH (e.g., $\beta_{2,T1-T2} = .238 vs.\ \beta_{2,T1-T2} = .173$), and by slightly lower correlation among traits (i.e., $r = .609 vs. r = .647$). In other words, the STARTS3 was fully concordant with the lag0-lag1 REM and gave even stronger support for an effect of ASC on ACH, both larger in size and robust to residual variance specific to each time point.

Reanalysis of Alternative Models Using the Original PALMA Data

Results from the reanalysis based on the original PALMA datasets mirrored those of the simulated data, indicating that, in this instance, the simulation matrix happened to align well — though such alignment should not be taken for granted. As for the simulated data, and as reported before by Marsh et al. (2024), the full reciprocal model RI-CCLPM (Fig. C.1) showed that the skill development effects of ACH on ASC were significant as lag0 effects ($\beta\prime_3 = [.436, .481], p < .001$) but not as lag1 effects ($\beta_3 = -.004, p = .832$). Conversely, the self-enhancement effects of ASC on ACH were significant as lag1 effects ($\beta_2 = [.133, .171], p < .001$), but not as lag0 effects ($\beta\prime_2 = 0, p = .999$, boundary condition). As for the simulated data, the STARTS3 provided the best-fitting solution per AIC and BIC weights (Table 1), and showed model parameters concordant with the RI-CCLPM-C (Fig. 2, B2 and B3), only differing by a slightly larger effect of ASC on ACH (e.g., $\beta_{2,T1-T2} = .209 vs.\ \beta_{2,T1-T2} = .132$), and by slightly lower correlation among traits (i.e., $r = .592 vs. r = .649$). In essence, the reanalysis of the PALMA confirm the validity of the findings reported in Marsh et al. (2024), showing their robustness to alternative data-generation processes including state variance confounding

(STARTS3), state-trait entanglement (LCSM3), and time-varying third-variable confounding (ESDM, RI-LTVC).

Discussion

We view the re-analysis by Sorjonen et al. (2025) as a valuable contribution to the ongoing effort to interrogate and strengthen the empirical foundations of the REM. Critical comparisons, such as those reported by Sorjonen et al., help clarify the scope and boundaries of theoretical claims by encouraging closer scrutiny of the mechanisms underlying different model specifications. By examining how reciprocal effects are represented—whether through lagged or contemporaneous paths, disaggregated or accumulated trait-state structures, or alternative sources of covariation such as latent confounding—such work advances our understanding of what drives observed patterns in longitudinal data. Our reanalysis of the data and models proposed by Sorjonen et al. (2025) offers further support for the extended REM with lag0–lag1 effects first identified by Marsh et al. (2024), strengthening the evidence for this temporal structure.

While the commentary by Sorjonen et al. (2025) broadened the scope of models tested, some of their interpretations may reflect the challenges associated with applying complex model specifications (e.g., for the LCSM, the RI-CLPM, and the STARTS model). In this response, we aimed to clarify some of these complexities and to outline the substantive rationales and modeling strategies that can help produce interpretable results. Using both the data from Sorjonen et al. (2025) and the original PALMA dataset, our analyses suggest that reciprocal effects—particularly the effect of academic self-concept on achievement—provide a plausible explanation for the observed covariations between these dynamic constructs. Although Sorjonen et al. raised questions about the strength of the evidence, our analyses suggest that the effect of academic self-concept on achievement is both probable given the data and consistent across the modeling strategies examined in this response. Furthermore, assuming that ASC influences ACH remains theoretically well-supported, in light of prior findings that self-concept is associated with processes influencing achievement, such as emotions (e.g., Pekrun et al., 2019), engagement, and effort (e.g., Marsh et al., 2016).

Moving forward, we encourage researchers to adopt stringent modeling standards when testing dynamic motivational theories across time. In particular, we showed that the substantive-methodology synergy for integrating contemporaneous (lag0) and lagged (lag1) reciprocal effects between dynamic constructs was underexplored in Sorjonen et al.'s commentary, either by not fully engaging with recommended guidelines (Muthén & Asparouhov, 2024) or by not elaborating on the rationales for contrasting temporalities of effects in the extended lag0-lag1 REM model (Marsh et al., 2024). This resulted in models positing effects that—as our reanalysis showed—were not clearly supported by the data (e.g., lag1 effects of ACH on ASC when controlling for lag0 effects). Other modeling choices with uncertain justification in their commentary were also discussed in relation to the competing data-generation processes considered.

While the comparison of these models is essential, we underscore the importance of theory-driven modeling strategies that align analytic tools with the underlying research questions. A concern for causal inference, in particular, requires precise conceptualization about the processes of change (e.g., disaggregation or entanglement of state and trait factors) and of the likely sources of confounding (e.g., covariate effects, endogenous dynamics, measurement error; Marsh et al., 2024; Usami et al., 2019a). These theoretical considerations should inform the development of empirical models, ensuring the interpretability of the results. At the same time, it is important to acknowledge the potential heuristic value of comparing models even in the absence of strong theoretical priors. For example, models with imperfect causal assumptions—such as the LCSM or ESDM—may still serve a diagnostic function: By testing whether traits are entangled with states (LCSM), whether third-variable confounders drive reciprocal effects (ESDM), and whether these models provide a better account of the data than a theorized model (e.g., the RI-CCLPM). Ultimately, these perspectives converge on a common recognition that model evaluation is not a mechanical exercise in fit statistics, but a judgment-based process requiring integration of empirical performance with conceptual clarity (Marsh et al., 2004). In this exercise, overreliance on any single index, including information criteria as used in this study (e.g., AIC, BIC), risks privileging formally parsimonious models over those that better reflect the substantive constructs under investigation. Model selection instead requires a comprehensive evaluation integrating statistical fit, theoretical coherence, and the interpretability of model parameters, as illustrated in the development of the extended REM (Marsh et al., 2024). We hope this exchange fosters continued dialogue about the use of longitudinal models and illustrates how re-analyses—such as those of Sorjonen et al. (2025)—enrich our practices by stimulating deeper reflection on modeling strategies..

Funding This work was supported by a grant from the Australian Research Council awarded to Jiesi Guo (DE230100300).

Data Availability Data and code to reproduce the analyses are openly available at the following OSF repository: https://osf.io/84bzp/.

References

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford Publications.

Burnham, K. P., & Anderson, D. R. (Eds.). (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. Springer. https://doi.org/10.1007/b97636

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

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

Ghisletta, P., & McArdle, J. J. (2012). Latent curve models and latent change score models estimated in R. *Structural Equation Modeling: A Multidisciplinary Journal, 19*(4), 651–682. https://doi.org/10.1080/10705511.2012.713275

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

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. (2023). How strong is the evidence for a causal reciprocal effect? Contrasting traditional and new methods to investigate the reciprocal effects model of self-concept and achievement. *Educational Psychology Review, 35*(1), 6. https://doi.org/10.1007/s10648-023-09724-6

Jamain, L., de Place, A.-L., Bouffard, T., & Pansu, P. (2021). Students' self-evaluation bias of school competence and its link to teacher judgment in elementary school. *Social Psychology of Education, 24*(3), 919–938. https://doi.org/10.1007/s11218-021-09638-7

Jorgensen, T. D., Kite, B. A., Chen, P.-Y., & Short, S. D. (2018). Permutation randomization methods for testing measurement equivalence and detecting differential item functioning in multiple-group confirmatory factor analysis. *Psychological Methods, 23*(4), 708–728. https://doi.org/10.1037/met0000152

Kenny, D. A., & McCoach, D. B. (2025). Ruling out latent time-varying confounders in two-variable multi-wave studies. *Multivariate Behavioral Research*, 1–17. https://doi.org/10.1080/00273171.2025.2503829

Kenny, D. A., & Zautra, A. (1995). The trait-state-error model for multiwave data. *Journal of Consulting and Clinical Psychology, 63*(1), 52.

Kievit, R. A., Brandmaier, A. M., Ziegler, G., van Harmelen, A.-L., de Mooij, S. M. M., Moutoussis, M., Goodyer, I. M., Bullmore, E., Jones, P. B., Fonagy, P., Lindenberger, U., & Dolan, R. J. (2018). Developmental cognitive neuroscience using latent change score models: A tutorial and applications. *Developmental Cognitive Neuroscience, 33*, 99–117. https://doi.org/10.1016/j.dcn.2017.11.007

Lüdtke, O., Robitzsch, A., & Wagner, J. (2018). More stable estimation of the STARTS model: A Bayesian approach using Markov chain Monte Carlo techniques. *Psychological Methods, 23*(3), 570–593. https://doi.org/10.1037/met0000155

Marsh, H. W., & Balla, J. (1994). Goodness of fit in confirmatory factor analysis: The effects of sample size and model parsimony. *Quality and Quantity, 28*(2), 185–217. https://doi.org/10.1007/BF01102761

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

Marsh, H. W., Guo, J., Pekrun, R., Lüdtke, O., & Núñez-Regueiro, F. (2024). Cracking chicken-egg conundrums: Juxtaposing contemporaneous and lagged reciprocal effects models of academic self-concept and achievement's directional ordering. *Educational Psychology Review, 36*(2), 53. https://doi.org/10.1007/s10648-024-09887-w

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal, 11*(3), 320–341. https://doi.org/10.1207/s15328007sem1103_2

Marsh, H. W., Pekrun, R., & Lüdtke, O. (2022). Directional ordering of self-concept, school grades, and standardized tests over five years: New tripartite models juxtaposing within- and between-person perspectives. *Educational Psychology Review, 34*(4), 2697–2744. https://doi.org/10.1007/s10648-022-09662-9

Marsh, H. W., Pekrun, R., Lichtenfeld, S., Guo, J., Arens, A. K., & Murayama, K. (2016). 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. https://doi.org/10.1037/dev0000146

Marsh, H. W., Scalas, L. F., & Nagengast, B. (2010). Longitudinal tests of competing factor structures for the Rosenberg Self-Esteem Scale: Traits, ephemeral artifacts, and stable response styles. *Psychological Assessment, 22*(2), 366. https://doi.org/10.1037/a0019225

Muthén, B., & Asparouhov, T. (2024). Can cross-lagged panel modeling be relied on to establish cross-lagged effects? The case of contemporaneous and reciprocal effects. *Psychological Methods*, No Pagination Specified-No Pagination Specified. https://doi.org/10.1037/met0000661

Núñez-Regueiro, F., Juhel, J., Bressoux, P., & Nurra, C. (2022). Identifying reciprocities in school motivation research: A review of issues and solutions associated with cross-lagged effects models. *Journal of Educational Psychology, 114*(5), 945–965. https://doi.org/10.1037/edu0000700

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

Preacher, K. J., & Yaremych, H. E. (2022). Model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling*. Guilford Publications.

Sorjonen, K., Melin, B., & Nilsonne, G. (2025). Inconclusive evidence for a prospective effect of academic self-concept on achievement: A simulated reanalysis and comment on Marsh et al. (2024). *Educational Psychology Review, 37*(2), Article 30. https://doi.org/10.1007/s10648-025-10008-4

Speyer, L. G., Zhu, X., Yang, Y., Ribeaud, D., & Eisner, M. (2024). On the importance of considering concurrent effects in random-intercept cross-lagged panel modelling: Example analysis of bullying and internalising problems. *Multivariate Behavioral Research*. https://doi.org/10.1080/00273171.2024.2428222

Usami, S., Hayes, T., & McArdle, J. J. (2015). On the mathematical relationship between latent change score and autoregressive cross-lagged factor approaches: Cautions for inferring causal relationship between variables. *Multivariate Behavioral Research, 50*(6), 676–687. https://doi.org/10.1080/00273171.2015.1079696

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), Article 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*(11), 1457–1466. https://doi.org/10.1007/s00127-016-1281-9

Vu, T., Scharmer, A. L., van Triest, E., van Atteveldt, N., & Meeter, M. (2024). The reciprocity between various motivation constructs and academic achievement: A systematic review and multilevel meta-analysis of longitudinal studies. *Educational Psychology, 44*(2), 136–170. https://doi.org/10.1080/01443410.2024.2307960

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review, 11*(1), 192–196. https://doi.org/10.3758/BF03206482

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives, 4*(1), 10–18. https://doi.org/10.1111/j.1750-8606.2009.00110.x