Studies in Educational Evaluation 83 (2024) 101377

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/locate/stueduc

Estimating the dimensionality of learning: The model for decomposed change

Denis Federiakin [a,*], Olga Zlatkin-Troitschanskaia [a], William B. Walstad [b]

[a] Department of Business and Economics Education, Johannes Gutenberg University of Mainz, Mainz, Germany
[b] Department of Economics, University of Nebraska-Lincoln, USA

ARTICLE INFO

Keywords:
Psychometrics
Item response theory
Learning score analysis
Learning
Forgetting
Growth and change

ABSTRACT

Psychometrics traditionally has assumed that the changes occurring during learning have the same dimensionality as the targeted construct. This approach is akin to understanding learning as climbing a mountain: for each ability a student can increase their ability, decrease it, or not change it. However, an alternative approach has recently been suggested – learning score analysis. This analysis is grounded in the assumption that students "slide through" the curriculum rather than "climb it": they forget something to learn something. This approach suggests that positive and negative dynamics can occur at the same time. This suggestion is also supported by the evidence from cognitive psychology that learning and forgetting are separate cognitive processes. Hence, we contrast the traditional approach to conceptualizing growth and change to the alternative approach. We use IRT-based modeling to compare the results from both approaches, and show that the alternative approach provides more insight into learning.

1. Introduction

One of the key issues of contemporary psychometrics is the analysis of growth and change in student learning (Cai & Houts, 2021). Educational studies as well as formative and summative assessments for test-takers often utilize a repeated measures design to track the dynamics of changes (Chang & Wimmers, 2017). The advantage of the repeated measures designs is that they can provide more information for institutional or individual decision-making than is possible with single measure designs (Kristensen & Hansen, 2004).[1]

In educational data, growth and change are almost universally synonymous with learning (Muthén & Khoo, 1998). Although learning encompasses a range of definitions within educational science and is undeniably complex (De Houwer et al., 2013), psychometrics operationalizes it as a change over time between two measurement occasions (Panik, 2014). Traditionally, indicators of growth and change are conceptualized as variations in response profiles between these occasions. If an individual's responses to items change from the first to the second measurement, such changes are attributed to growth and change, or "learning" in educational terms (e.g., Talbot, 2013). This approach to measuring learning is fundamental to evidence-based education (Mislevy, 2018). Consequently, it raises a critical question: "Are the changes in item responses random or systematic?" Psychometric modeling serves as a tool to address this question, distinguishing between random noise and meaningful signal (Ruspini, 2003).

The modeling of latent growth is a flourishing and rapidly developing area of psychometric research in the last decades. Numerous psychometric models have been developed for the analysis of (longitudinal) data from repeated measures. In Item Response Theory (IRT), they include such fundamental models as Andersen's model for repeated measures (Andersen, 1985) and models for growth and change (Embretson, 1991) adopted from factor analysis (Jöreskog, 1970). In structural equation modeling, they include latent growth curve models (Duncan & Duncan, 2004), random intercept cross-lagged panel models (Mulder & Hamaker, 2021), and other cross-lagged models (Lüdtke & Robitzsch, 2021).

However, all these approaches to the analysis of growth and change are based on a similar conceptualization of latent growth and change (for details, see Section 2). They assume that latent growth and change have the same *dimensionality* as the measured construct. If a construct is unidimensional, then an individual's ability can increase, decrease, or stay constant over time. Thus, positive and negative change are assumed

* Correspondence to: Jakob-Welder-Weg 9, room 01/249, 55128 Mainz, Germany.
E-mail address: denis.federiakin@uni-mainz.de (D. Federiakin).

[1] Single measure design studies can be useful and they require less data collections than repeat measure design. For example, PISA assessments, based on changing samples, analyze changes in educational system dynamics (Grek, 2009), although they avoid reporting results for individual decision-making (Wu, 2005).

https://doi.org/10.1016/j.stueduc.2024.101377
Received 20 December 2023; Received in revised form 15 March 2024; Accepted 16 June 2024
Available online 1 July 2024

to be *opposite poles of the same continuum* for the ability construct. This widely established conceptualization of latent growth and change is intuitive and straightforward, which may explain why researchers have applied it in many psychometric studies (e.g., Byrne et al., 2008).

However, this dichotomous assumption can be less useful for the valid measurement of the actual development of increasingly complex, multidimensional, and multifaceted knowledge and skill constructs, which can develop very dynamically and differently over time (Piacentini et al., 2023). Therefore, this study presents an alternative conceptualization of the latent growth and change of student learning to the traditional analysis. To illustrate our approach, we use student learning data obtained from a study with repeated measures. Based on prior research (Walstad & Wagner, 2016), we assume that there are three separate dimensions of latent change: positive learning, negative learning, and "no-change". Using educational measurement methodology, we provide the psychometric validity evidence on this conceptualization of the latent change by contrasting our approach to one of the more established approaches.

The structure of this paper is as follows: First, we illustrate the traditional approach to latent growth with one of the main models for the analysis of repeated measures – Embretson's model for growth and change (1991). Second, we describe our alternative approach to the conceptualization of growth and change in terms of cognitive and educational interpretations. Then, we propose an IRT model based on our conceptualization – the Model for Decomposed Change (MDC). After that, we provide an analysis of real data applying our model, demonstrating how much more information it provides than the traditional approach. Finally, we discuss our approach and show how it contributes to contemporary psychometrics.

2. Traditional conceptualization of growth and change

Thirdly, the traditional conceptualization of growth and change assumes that positive and negative change are mutually exclusive. That is, the more a test taker's ability grows, the less it decays. One of the clearest examples of this logic is Embretson's model for growth and change. This model can be seen as a predecessor to many models for repeated measures in the IRT paradigm (Wilson et al., 2012), much like its prototype from factor analysis – Jörsekog's simplex model (1970).

Embretson's model follows Eq. (1) in the case of two measurement occasions and dichotomous items. On the path diagrams, this model can be represented as in Fig. 1.

$$\begin{cases} P\left(U_{ip}^{t=1}=1\right)=\dfrac{1}{1+\exp\left(\beta_i-\alpha_i\theta_p^{t=1}\right)} \\ P\left(U_{ip}^{t=2}=1\right)=\dfrac{1}{1+\exp\left(\beta_i-\alpha_i\left(\theta_p^{t=1}+\theta_p^{t=2}\right)\right)} \end{cases}, \tag{1}$$

Where $P\left(U_{ip}^{t=m}=1\right)$ is the probability that the response of student $p(p=1,\ldots,P)$ to item $i(i=1,\ldots,I)$ at measurement occasion $m(m=1,\ 2)$ is correct,

α_i is the discrimination parameter of item i,

β_i is the item threshold of item i,

$\theta_p^{t=1}$ is the ability of student p at the measurement occasion 1 (baseline),

$\theta_p^{t=2}$ is the change in the ability of student p from the measurement occasion 1 to the measurement occasion 2.

The item parameters (α_i and β_i) are time-invariant, which is necessary for establishing a comparable scale across the latent dimensions of

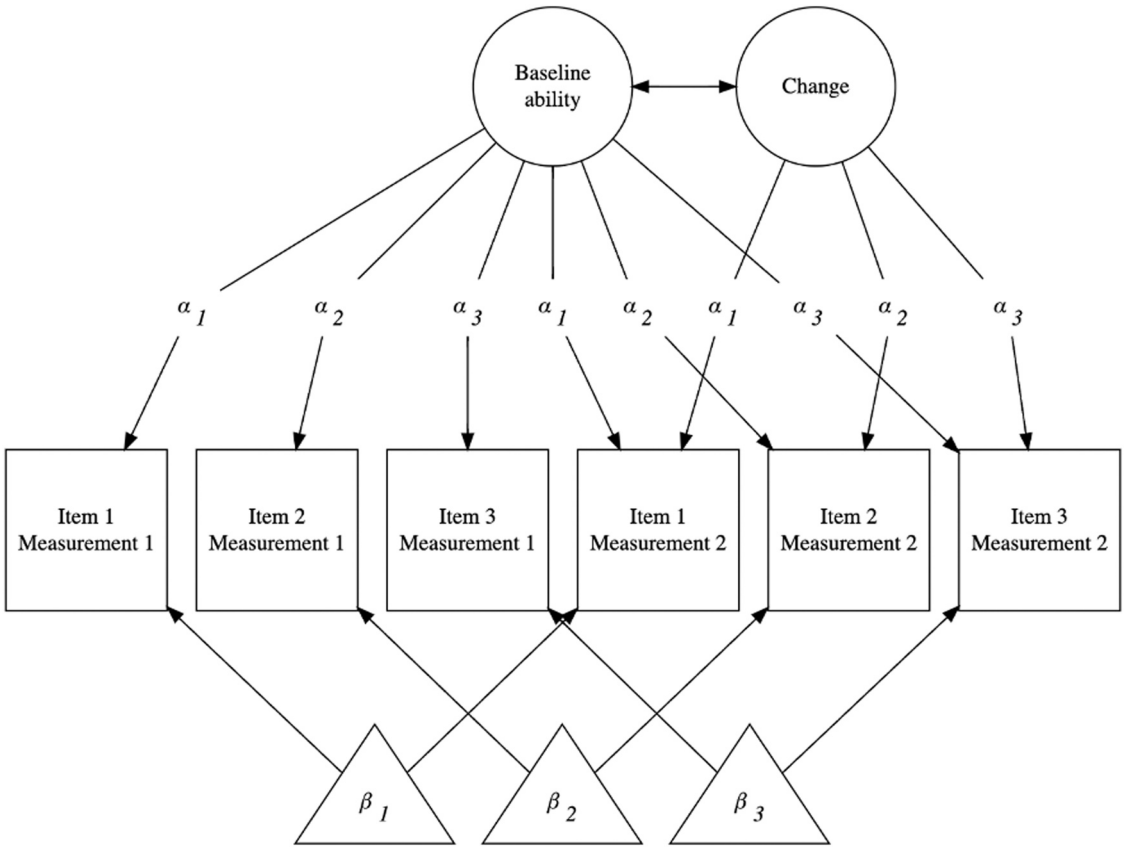

Fig. 1. An exemplary path diagram for Embretson's model for growth and change for 3 items and 2 measurement occasions. *Note.* According to the notations of path diagrams, squares are observed variables, circles are unobserved variables, triangles are intercepts. One-headed arrows are regression dependencies, and two-headed arrows are correlations.

student parameters. In this way, it is possible to estimate the variance of change ($\theta_p^{t=2}$) and its correlation with the initial ability ($\theta_p^{t=1}$). The item parameters constrained across measurement occasions allow for the full identification of the scale of change and, consequently, for estimating the average of the change dimension. Parameters of sample distribution of the change dimensions ($\theta^{t=2} \sim N(\mu, \sigma)$) should be interpreted in comparison to the initial ability dimension, where they are constrained ($\theta^{t=1} \sim N(0, 1)$). This brings up the issue of longitudinal measurement invariance (Liu et al., 2017) – if the item parameters have changed from one measurement occasion to another, the item cannot be used as a common item, because its psychological interpretation has changed.

Embretson's model controls for the baseline level of ability in the second measurement by loading the variables from the second measurement on the baseline factor. Thus, all changes that occurred from the measurement at $t = 1$ to $t = 2$ are reflected in the second latent factor. This logic has a straightforward extension to cases where there are more than 2 measurements. Then, the baseline factor loads all items, the second factor loads all items starting from the second measurement, the third factor loads all items starting from the third measurement, and so on. Thus, each later factor is nested within all earlier factors. Embretson's model does not make any assumptions about functional relationships between the time elapsed between measurement occasions and the amount (and direction) of growth and change; rather, it simply describes the change that occurred by the time of the measurement. This distinguishes this model from latent growth curve models, which assume that the amount of growth and change is a parametric (typically, polynomial) function of the time elapsed between measurements (Wilson et al., 2012).

Embretson's model clearly exemplifies the traditional approach to conceptualizing growth and change. The fact that the second factor consolidates information on the change into a single latent variable implies that each student can either exhibit positive change or negative change (by sliding up or down on this latent continuum), or maintain the same level of ability by the second measurement. In the next section, we propose an alternative approach to understanding educational growth and change.

3. An alternative conceptualization of growth and change

Our conceptualization of growth and change is based on the theory of different types of learning proposed by Walstad and Wagner (2016) who proposed studying learning patterns in item responses instead of an aggregated change in ability. They compared dichotomous item scores from the same pre-test ($U_{pi}^{t=1}$) and post-test ($U_{pi}^{t=2}$), and inferred four types of learning indicators: positive learning (L_{ip}^{+}), negative learning (L_{ip}^{-}), zero learning (L_{ip}^{0}), and retained learning (L_{ip}^{R}). The derivation of the item-wise learning indicators is described in Table 1. According to this data recoding procedure, each item repeated across two measurement occasions produces four dichotomous mutually exclusive variables, reflecting the learning type that occurred on an item between the measurements. These learning indicators were analyzed for the purposes of individual feedback and content improvement (for details, Walstad & Zlatkin-Troitschanskaia, 2024).

The idea of learning indicators has mainly been applied in classical test theory on raw test scores, with only limited interest from other psychometric paradigms. For example, Schmidt et al. (2019) used a two-parameter logistic IRT model to describe the distribution of these learning indicators conditional on the latent variables. This approach assumes a common cause behind the distributions of learning indicators of the same type (Epskamp et al., 2018) in exactly the same manner as Embretson's model, which assumes that there is a common cause directly behind item scores. However, Schmidt et al. (2019) used separate unidimensional IRT models and studied only positive and negative learning dimensions. Using separate unidimensional models in a consecutive approach is not ideal, as their likelihood functions sacrifice information on the multivariate distribution of latent variables. Thus, we aim to utilize a multidimensional IRT model in this study to examine the details of relationships between the estimates of different types of learning as well as to increase their reliability (De La Torre & Patz, 2005).

Given the unorthodox nature of this conceptualization of change, we discuss the cognitive interpretation of such an approach and contrast it with the traditional approach to conceptualizing growth and change. In the traditional approach, learning growth is assumed to be a fairly linear progression through the curriculum (although, there is a variation in such a conceptualization too; e.g., Wilson, 2009). A student "climbs" the content in terms of its complexity: from the simplest topics to the most difficult ones (given appropriate curriculum design), accumulating knowledge. Thus, the ability either grows, decays, or stays the same, reflecting the aggregate probability of a correct response on the same items.[2]

In the understanding of the latent change proposed in this study, a student does not necessarily accumulate knowledge over time. Particularly, due to the multidimensionality of change, a student might exhibit to some extent "opposite" types of learning simultaneously – for example, negative and positive. This may happen in the case of having acquired knowledge in some content areas while simultaneously losing it in others. In some cases, this might even result in situations where the traditional growth and change estimate has not changed, but its content and meaning have. Such a conceptualization assumes that students rather "slide" through the curriculum content, instead of "climbing" it – to learn something, they forget something else and develop misconceptions. Thus, breaking the traditional estimate of change down into such components allows us to see a more detailed picture of the dynamic of student learning.

Moreover, such a conceptualization of change might be supported by contemporary cognitive psychology. For example, cognitive science still struggles with the definitive understanding of the exact relationships between learning and forgetting, especially in the context of 'conceptual change' (Vosniadou, 1994). Some cognitive studies argue that student knowledge has a significant impact on their forgetting rates (Kopelman, 1985; Mary et al., 2013; Walsh et al., 2014; Yang et al., 2016; Loftus, 1985). At the same time, some other research found that forgetting rates are independent from the initial degree of information familiarity (Slamecka & McElree, 1983; Giambra & Arenberg, 1993; Rivera-Lares et al., 2022) and age (Rivera-Lares et al., 2023). This means that the pattern forgetting rates remain the same for different degrees of initial retention. This result aligns with the suggestion that remembering and forgetting are two different cognitive processes (Sasin, et al., 2017). Correspondingly, modeling different directions of learning progress in the same manner – as independent but correlated continuums reflecting different causes – appears to be a promising topic of further investigation for both diagnostic and instructional purposes.

In such a case, however, the interpretation of zero and retained

Table 1
The original map of data recoding from Walstad and Wagner (2016).

Response pattern		Learning indicator			
$U_{ip}^{t=1}$	$U_{ip}^{t=2}$	L_{ip}^{+}	L_{ip}^{-}	L_{ip}^{0}	L_{ip}^{R}
0	1	1	0	0	0
1	0	0	1	0	0
0	0	0	0	1	0
1	1	0	0	0	1

[2] Operationally speaking, items may not be exactly the same across the repeated measures, but generally at least some item parameters are constrained across the measurements to establish the comparability of scales (Chapman & Johnson, 1994). This defines the interpretation of ability estimate as a tool to track the changes in the probability of the correct response on the same item.

learning as separate dimensions poses problems. Particularly, both dimensions reflect the absence of change, rather than any process behind it. The only difference between them is that, in the case of zero learning, the item was too difficult for a student in both measurement occasions, while in the case of retained learning it was too easy. Based on this interpretational amendment to the original Walstad and Wagner (2016) methodology, we merge the two learning indicator types (zero and retained learning) into a single learning type called "no-change". Despite the fact that, from the educational practitioner's point of view, zero and retained learning might be two different phenomena, for the sake of measurement illustration, we unite them in a single dimension. This makes the map of item scores recoding from Table 1 into Table 2 – the map that we use in this study.

Recoding data into such learning indicators allows us to derive three types of learning indicators for each item, as shown in Table 2. This results in $I \times 3$ indicators, where I is the number of repeating items. Thus, we use a between-item (Adams et al., 1997) 3-dimensional 2PL model, where each dimension describes a single learning type (the Model for Decomposed Change, MDC, Fig. 2):

$$P\left(L_{ip}^{X}=1\right)=\frac{1}{1+\exp\left(\beta_{i}^{X}-\alpha_{i}^{X}\theta_{p}^{X}\right)},$$

where $P\left(L_{ip}^{X}=1\right)$ is the probability of a learning indicator of type X (+ for positive, - for negative, *NC* for no-change) for student p on item i to take the value of "1",

β_{i}^{X} is the item threshold of a learning indicator of type X for item i,

α_{i}^{X} is the item discrimination of a learning indicator of type X for item i,

θ_{p}^{X} is the estimate of learning of type X for student p ($\boldsymbol{\theta} \sim MVN(\mathbf{0},\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is the correlation matrix (with the diagonal elements constrained to unity) for the purposes of model identification).

The recoding of data demonstrated in Table 2, however, makes the data incompatible with Embretson's model. The number of observed variables increases by a factor of 1.5 (from 2 variables per item to 3), subsequently inflating the likelihood of our multidimensional model, thus rendering it incomparable to Embretson's model. Moreover, this data recoding collapses some information from the original dataset, as, in the case of a "no change" indicator taking the value of "1", it is impossible to know if the original item responses were both correct or both incorrect in the original data.

It is critical to recognize that such decomposition renders the data "ipsative" by nature. Specifically, when one indicator – representing a type of learning for an item (a "pseudo-item") – is assigned a value of "1", the indicators for other learning types for that item are necessarily set to "0". As a result, this recoding process generates a data structure consisting of I blocks, each containing 3 pseudo-items. These pseudo-items signify the dominance of a particular learning type over others within each block, akin to the response format found in forced-choice questionnaires. In such questionnaires, respondents are required to (for example) select the statement that best represents them from a set, with each statement loading different dimensions. This structure leads to a misspecification of the normative 2PL model because the model's likelihood function is designed to describe the probability of all responses being "1" simultaneously – an impossibility in ipsative data.

Table 2
The map of data recoding used in this study.

Response pattern		Learning indicator		
$U_{ip}^{t=1}$	$U_{ip}^{t=2}$	L_{ip}^{+}	L_{ip}^{-}	L_{ip}^{NC}
0	1	1	0	0
1	0	0	1	0
0	0	0	0	1
1	1	0	0	1

Nevertheless, a correspondence between parameter estimates derived from models of ipsative data and those from normative models suggests that while ipsative data introduces limitations on the precision of parameter estimation, it does not invalidate the utility of our approach substantially (e.g., Morillo et al., 2019).

4. Sample, instrument and data

For a real data example, we use a sample of 25 items from the *Test 1* (Walstad et al., 2013) and the *Test 2* (Walstad et al., 2007). Items were sampled to represent the most important content areas according to the German higher education Economics curriculum (*Authors*, 2021). Based on a series of validations studies according to The Standards for Educational and Psychological Testing (AERA, APA, NCME, 2014; e.g., Walstad & Wagner, 2016; Schmidt et al., 2019), the items were purposefully chosen to represent two levels of cognitive taxonomy (according to Anderson & Krathwohl, 2001): knowledge (9 items) and (implicit and explicit) application (16 items), across 13 selected fundamental economics concepts. These include:

- Basic Concepts (1 item per concept): Decision-Making and Marginal Analysis, Economic Systems and Allocation Mechanisms, Economic Incentives (including Prices, Wages, Profits, etc.), and Economic Institutions.
- Microeconomics Concepts: Supply and Demand (3 items), Role of Government (3 items), Competition (2 items), Labor Markets and Income (2 items), and International Microeconomics (1 item).
- Macroeconomics Concepts: Money and Inflation (3 items), Aggregate Supply and Demand (4 items), Fiscal and Monetary Policy (2 items), and International Macroeconomics (1 item).

Despite the apparent heterogeneity in item content, the selection process was aimed at reflecting the general economic competency of students in German higher education. These items collectively form a unidimensional scale, aligning with the conceptualization of economic competency as an integrated composite construct (Authors, 2022). While alternative approaches to the operationalization of economic competencies exist (e.g., Macha & Schuhen, 2011), our previous research supports the validity of a unidimensional representation of this multifaceted construct (*Authors*, 2021).

The first measurement occasion was conducted in the winter term 2016/17 at 32 universities Germany-wide and the second in the winter term 2017/18 at 27 universities. The total sample size used in this study consisted of 929 students of economics from 27 universities who took part in both measurements (for more study details, Schmidt et al., 2019) without any missing responses. The test was conducted as part of a broader representative data collection project across Germany (for details, Schmidt et al., 2019). Administering procedures for the test conducted by trained test administrators on sites were standardized across all participating universities and at each time point to ensure consistency.

5. Results

All models were estimated using the TAM package, version 4.1–4 (Robitzsch et al., 2022), within R software, version 4.3.0. The results of the initial model comparison are presented in Table 3. For the model fit evaluation, we used absolute global model fit indices: SRMR (Hu & Bentler, 1999) and SGDDM (Levy et al., 2015). These indices quantify the discrepancy between the model-expected variance-covariance matrix of item residuals to the observed in the data. To estimate individual student abilities across all models, we employed the Expected-a-Posteriori (EAP) method, which additionally generates an estimate of the posterior standard deviation. This estimate can be interpreted as the standard error of measurement for an individual's ability (Bock & Mislevy, 1982). Consequently, this approach was

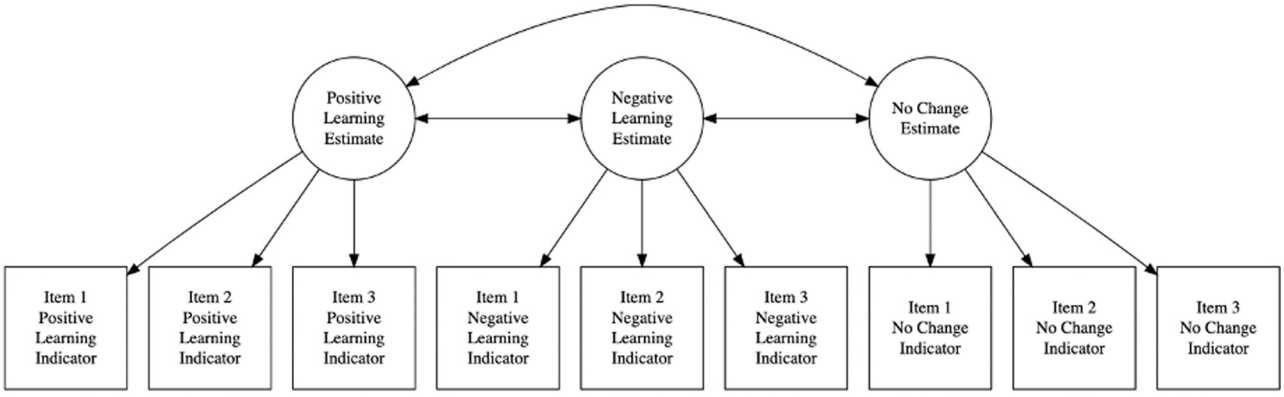

Fig. 2. An exemplary path diagram for the proposed Model for Decomposed Change for 3 items and 2 measurement occasions (complementary to the Fig. 1).

Table 3
Model comparison.

Statistics	Embretson's model	MDC
Sample	929	
Number of estimated parameters	53	153
Deviance	55899.47	70980.88
AIC	56005	71287
SRMR	0.031	0.039
SGDDM	0.036	0.044

utilized to calculate the empirical EAP-reliability of our measures (Adams, 2005).

Table 3 provides a clear indication that both models demonstrate a globally adequate fit to the data. Particularly, the good fit of Embretson's model demonstrates that the test itself is almost perfectly unidimensional. The MDS has slightly increased fit indices, but they are still below the critical values. Given this observation, we proceeded to conduct a detailed analysis of the parameter estimates across the models.

The variance-covariance matrix and the reliability of the student parameters derived from Embretson's model and the MDC are depicted in Tables 4 and 5, respectively. We evaluate these parameters, as they provide insight into the underlying patterns of variance and covariance in the data, as well as the degree to which these models can reliably estimate individual differences in learning outcomes. While both models show a generally good fit, there may be nuanced differences in how they account for variability in the data. A careful inspection of these parameters can shed light on the differential effectiveness of the two models in capturing the latent dynamics of learning.

The findings outlined in Table 4 are instructive. Embretson's model presents a relatively low variance for the change dimension, suggesting that the shifts in ability over time among students aren't particularly wide-ranging. However, the mean of the change dimension is slightly higher than the initial ability dimension, indicating that, on average, students have improved their ability over the course of the study. Notably, the primary source of variability in the data stems from the differences in students' initial abilities rather than their learning trajectories over the course of the study. This highlights the important role

Table 4
The variance-correlation matrix, averages, and reliability from Embretson's model.

	The baseline	The change
The baseline	1	0.232
The change	0.232	0.271
Mean	0	0.181
Reliability	0.828	0.370

Table 5
The correlation matrix and reliability from the 3-dimensional IRT Model for Decomposed Change.

	Positive Learning	Negative Learning	No-change
Positive Learning	1	0.369	-0.761
Negative Learning	0.369	1	-0.825
No-change	-0.761	-0.825	1
Reliability	0.571	0.653	0.724

of baseline abilities in determining subsequent performance outcomes. The relatively low reliability of the change scores could be attributed to the inherent nature of Embretson's model as a within-item multidimensional (Adams et al., 1997) model and the paucity of information on the change scale, coupled with the low variance within the sample on this scale. This finding supports the assumption that the reliability estimates of difference scores are typically lower than their true reliability (Trafimow, 2015). Furthermore, the correlation between the baseline and the change is modest but positive, hinting at a moderate 'Matthew effect' (Walberg & Tsai, 1983). This effect posits that students who start out stronger tend to benefit more from learning, further widening the ability gap between stronger and weaker students.

Firstly, we analyze the reliability estimates presented in Tables 4 and 5. Reliability can generally be understood as a measure of the consistency or non-randomness in item responses (Adams, 2005). The reliabilities shown in Table 4 indicate that many item response changes from the first to the second measurement occasion appear to be random. However, after categorizing these changes into different types of dynamics, Table 5 reveals a significant presence of meaningful information within these scores. Furthermore, while these reliability estimates may not be exceptionally high, their comparison with the change dimension reliabilities from Embretson's model (Table 4) suggests that delineating different types of dynamics provides more accurate insights into learning processes. Most importantly, the reliability levels are not adequate for high-stakes decision-making at the individual level but are suitable for low-stakes decision-making at the group level (e.g., Wells & Wollack, 2003).

In Table 5, the means are not presented as they were constrained by zeros according to the specification of the standard normal distribution for the sake of model identification. Interestingly, there is a positive correlation between negative and positive learning latent variables, which implies that students may not be retaining information as their studies progress. In other words, the more they forget, the more they learn, suggesting a stable effect of simultaneous knowledge gain and loss. The correlations between no-change and both positive and negative learning are strong and negative. This could be attributed to the nature of these indicators: the higher the dominance of absence of learning, the less likely it is for other types of learning (designating the presence of change) to attain higher values.

To verify these interpretations, we examined the *between-model* correlations of all dimensions, as shown in Fig. 3. However, these correlations are not population-based estimates but sample-based, as they are obtained by simply correlating student parameter estimates from the models. As a result, these estimates are somewhat biased. Therefore, estimates of correlations presented in Tables 4 and 5 differ from these sample-based estimates in strength.

Fig. 3 reveals that the abilities from the MDC are highly heteroscedastic. Although, the individual student parameters were marginalized to a multivariate normal distribution, they have a changing variance when conditioned on each other. This is meaningful, since the individual item indicators are mutually exclusive. For example, the only way to have a high estimate of no-change is to have low estimates of both positive and negative learning. However, when the room for positive and negative learning increases (the amount of no-change decreases), variability of possible estimates for both positive and negative learning increases too, resulting in the cone-shaped scatterplots. This logic goes beyond interdependencies in the MDC. For example, the same logic can be applied within Embretson's model when the same test was administrated in both measurements. If the estimate of the baseline ability is very high and the baseline and change dimensions are positively correlated, the variance of change is going to decrease with the increase in baseline due to ceiling effects – there is less possibility of a high change estimate when the student's baseline ability is already very high.

The most important insights of Fig. 3 concern the comparisons across the models. The red lines in these scatterplots show the absent change in terms of Embretson's model. When comparing Embretson's change with any type of learning, the dispersion of the parameter estimates along these red lines shows that, even if Embretson's model claims that a student did not experience any change, the MDC shows that there is significant variability across people in terms of different types of learning. In other words, students can get better in some content areas while getting worse in others, keeping their overall ability the same as far as Embretson's model is concerned.

Still, almost all scatterplots comparing parameter estimates across the models are highly heteroscedastic. The comparisons of positive and negative learning with baseline ability are similar graphically, but for different reasons. The expectation (and variation) of positive learning decreases with the increase in the baseline ability because of the ceiling effect. At the same time, the same picture for the negative learning is due to the fact that Embretson's change and baseline dimensions are positively correlated, revealing 'Matthew's effect': The higher the baseline ability was, the less negative learning students tended to exhibit. The comparison of positive learning with Embretson's change dimension shows that even students with the negative Embretson's change exhibit an approximately average amount of positive learning. However, as the amount of Embretson's change grows, the variability in positive learning increases. This shows that even generally negative Embretson's change can correspond to positive learning on some test items, even if the test is unidimensional. We interpret it as a sign that our model is more sensitive to capturing the complexity of learning progress. The comparison of negative learning with Embretson's change dimension shows a generally expected picture of negative relations. This is due to the positive orientation of scale of Embretson's change dimension. The comparison of Embretson's baseline with no-change in MDC reveals that they are positively correlated. This means that the higher Embretson's baseline ability was, the more students preserved it. This, again, is coherent with both the 'Matthew's effect' and the 'ceiling effect'. Relations similar in the correlation direction but reversed in the direction of the cone-spreading are observed between Embretson's change and no-change in MDC. We can explain these relations via the positive correlations of both variables with Embretson's baseline dimension, hence their positive correlation with each other. While the negative values of Embretson's change correspond to the lower values of no-change in MDC, there is a significant variation in no-change on the positive side of

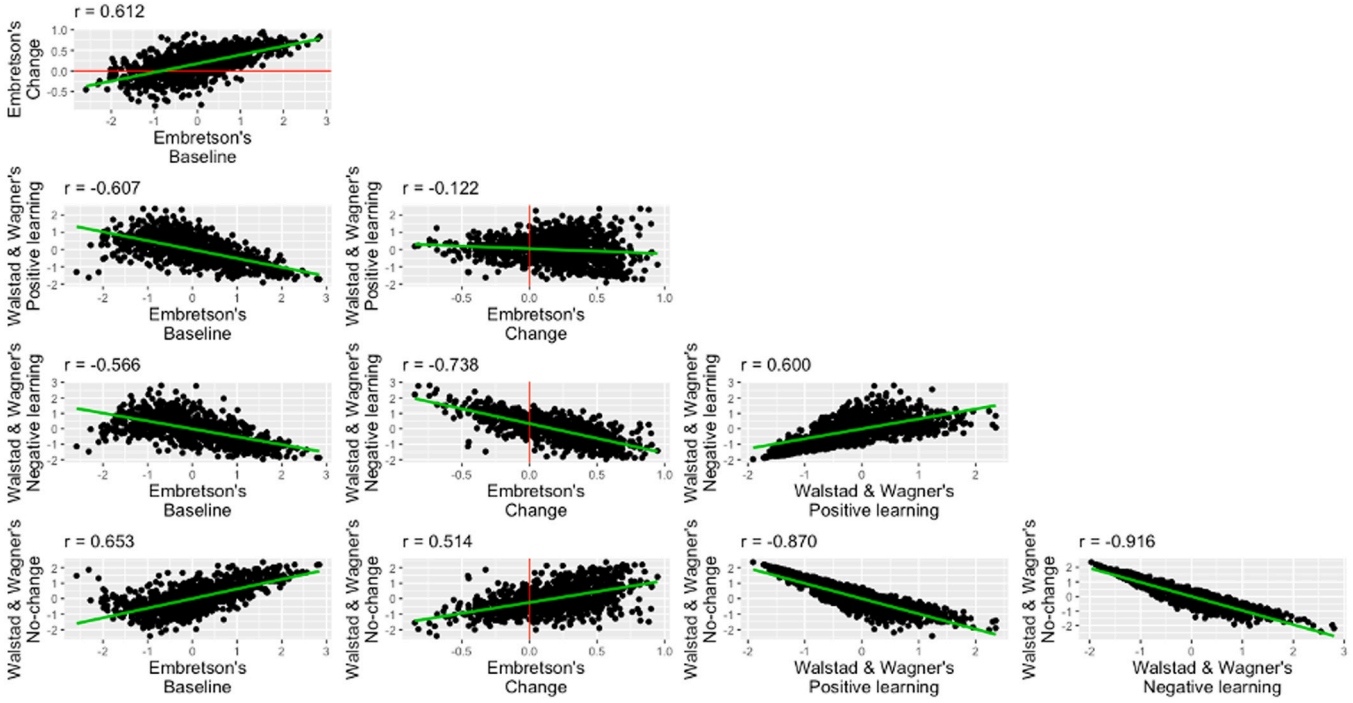

Fig. 3. Sample-based Pearson correlations and scatterplots of student parameters within and between models. *Note.* The visualization consists of a grid of scatterplots that compare person parameters across all models. In this grid, the axes of the scatterplots are uniformly matched; i.e., scatterplots within a single column share the same variable on the x-axis, while those within a single row share the same variable on the y-axis. This arrangement would, in essence, mirror a symmetric variance-covariance matrix, complete with redundant elements. Therefore, to avoid repetition and enhance clarity, we display only the lower triangle of this grid. Each point is a student from the sample. The red lines designate the absent change according to Embretson's model (numerical value of "0"). The green lines designate the linear trend on each scatterplot based on the Pearson correlation.

Embretson's change. In other words, highly positive values of Embretson's change overlook the preservation of the initial baseline ability to a high degree.

Next, we analyzed the item parameters from different models. In Table 6, we present item statistics. Since in Embretson's model, items have the same parameters relative to both dimensions, we describe only one set of item parameters for this model, even though the items have been administered twice. Item parameters from the 3-dimensional model for learning scores are estimated relative to a numerically identical sample distribution on each dimension ($N(0,1)$). Comparing these parameters between dimensions can provide an understanding of the comparative difficulty of demonstrating learning types.

In Embretson's model, the average item intercept of -0.328 indicates that the test was easier than the average ability in the baseline measurement. One item (13) has a non-significant item intercept, meaning that this item is just on par with the average level of sample ability in the baseline, as the model identification was fixed to zero. Also, all items have statistically significant discrimination estimates.

The 3-dimensional model for learning scores, however, shows that in MDC it is much easier to demonstrate no-change than other types of learning. The average item intercept there is -0.694, while the average estimate of negative learning is zero. The average item intercepts on other dimensions are 1.596 for positive learning and 1.792 for negative learning. This means that it is marginally easier to demonstrate positive learning than negative learning.

Importantly, in the MDC, a set of items (1, 6, 17, 23) systematically exhibit non-significant discrimination estimates, suggesting these items diverge in measuring learning dynamics compared to the remainder of the test. Discrimination estimates gauge an item's effectiveness in assessing the same construct as the other items. Therefore, the learning

Table 6
Item parameter estimates.

Item	Embretson's model		MDC					
	β_i	α_i	β_i^+	α_i^+	β_i^-	α_i^-	β_i^{NC}	α_i^{NC}
1	0.161 *	0.382 *	1.141 *	-0.048	1.314 *	0.088	-0.184 *	0.001
	(0.028)	(0.044)	(0.077)	(0.100)	(0.080)	(0.098)	(0.066)	(0.076)
2	-0.960 *	1.065 *	1.816 *	1.150 *	2.079 *	0.930 *	-0.871 *	1.023 *
	(0.021)	(0.061)	(0.095)	(0.123)	(0.100)	(0.115)	(0.078)	(0.109)
3	-0.063 *	0.816 *	1.141 *	0.226 *	1.868 *	0.420 *	-0.472 *	0.218 *
	(0.025)	(0.053)	(0.077)	(0.101)	(0.096)	(0.115)	(0.068)	(0.08)
4	-1.250 *	1.231 *	1.891 *	0.569 *	2.497 *	0.946 *	-1.254 *	0.758 *
	(0.020)	(0.065)	(0.096)	(0.121)	(0.114)	(0.122)	(0.081)	(0.102)
5	-0.388 *	0.644 *	1.296 *	0.388 *	1.893 *	0.667 *	-0.568 *	0.538 *
	(0.023)	(0.050)	(0.080)	(0.106)	(0.096)	(0.113)	(0.071)	(0.089)
6	0.837 *	0.377 *	1.579 *	-0.056	1.294 *	0.045	-0.463 *	-0.074
	(0.036)	(0.045)	(0.087)	(0.113)	(0.080)	(0.098)	(0.067)	(0.078)
7	-1.692 *	1.359 *	2.488 *	1.454 *	2.971 *	1.696 *	-1.593 *	1.510 *
	(0.019)	(0.067)	(0.113)	(0.127)	(0.125)	(0.126)	(0.092)	(0.125)
8	-0.495 *	0.683 *	1.388 *	0.187	1.561 *	0.511 *	-0.471 *	0.287 *
	(0.023)	(0.051)	(0.082)	(0.108)	(0.087)	(0.106)	(0.068)	(0.082)
9	-0.121 *	0.569 *	1.696 *	0.270 *	1.217 *	0.603 *	-0.410 *	0.406 *
	(0.025)	(0.048)	(0.091)	(0.118)	(0.080)	(0.101)	(0.069)	(0.084)
10	-0.247 *	0.588 *	1.605 *	0.250 *	1.300 *	0.584 *	-0.424 *	0.407 *
	(0.024)	(0.048)	(0.088)	(0.115)	(0.081)	(0.102)	(0.069)	(0.084)
11	-0.756 *	0.929 *	1.762 *	0.639 *	2.022 *	0.766 *	-0.933 *	0.719 *
	(0.022)	(0.058)	(0.092)	(0.117)	(0.099)	(0.115)	(0.076)	(0.098)
12	0.224 *	0.262 *	1.050 *	0.077	1.738 *	0.229 *	-0.359 *	0.085
	(0.029)	(0.042)	(0.075)	(0.098)	(0.092)	(0.112)	(0.067)	(0.078)
13	0.023	0.801 *	1.645 *	0.403 *	1.795 *	0.572 *	-0.768 *	0.470 *
	(0.026)	(0.053)	(0.089)	(0.115)	(0.093)	(0.111)	(0.072)	(0.089)
14	-1.733 *	1.022 *	2.269 *	1.083 *	2.790 *	1.191 *	-1.588 *	1.212 *
	(0.019)	(0.063)	(0.106)	(0.123)	(0.123)	(0.124)	(0.089)	(0.115)
15	-0.651 *	0.770 *	1.765 *	0.362 *	1.692 *	0.839 *	-0.730 *	0.549 *
	(0.022)	(0.054)	(0.093)	(0.120)	(0.090)	(0.110)	(0.072)	(0.091)
16	-0.795 *	0.677 *	1.765 *	0.324 *	1.464 *	0.716 *	-0.614 *	0.548 *
	(0.021)	(0.053)	(0.093)	(0.120)	(0.085)	(0.106)	(0.071)	(0.090)
17	0.462 *	0.693 *	1.313 *	0.031	1.549 *	0.081	-0.458 *	0.019
	(0.030)	(0.049)	(0.080)	(0.105)	(0.086)	(0.106)	(0.067)	(0.078)
18	-0.135 *	0.551 *	1.520 *	0.235 *	1.556 *	0.348 *	-0.577 *	0.267 *
	(0.025)	(0.047)	(0.086)	(0.112)	(0.086)	(0.106)	(0.069)	(0.082)
19	0.565 *	0.557 *	1.120 *	-0.175	1.968 *	0.239 *	-0.524 *	-0.100
	(0.031)	(0.047)	(0.076)	(0.099)	(0.100)	(0.121)	(0.068)	(0.078)
20	-1.012 *	0.762 *	1.922 *	0.668 *	1.665 *	0.778 *	-0.807 *	0.767 *
	(0.021)	(0.055)	(0.097)	(0.120)	(0.09)	(0.109)	(0.075)	(0.099)
21	0.327 *	0.726 *	1.562 *	0.022	1.259 *	0.287 *	-0.416 *	0.191 *
	(0.028)	(0.050)	(0.087)	(0.113)	(0.079)	(0.098)	(0.067)	(0.079)
22	-1.028 *	0.843 *	1.965 *	0.592 *	1.681 *	0.848 *	-0.852 *	0.813 *
	(0.021)	(0.057)	(0.098)	(0.122)	(0.090)	(0.110)	(0.076)	(0.101)
23	1.028 *	0.865 *	1.513 *	-0.080	1.950 *	0.083	-0.821 *	-0.055
	(0.035)	(0.052)	(0.085)	(0.111)	(0.099)	(0.122)	(0.071)	(0.082)
24	-0.700 *	0.774 *	1.626 *	0.358 *	1.690 *	0.479 *	-0.718 *	0.478 *
	(0.022)	(0.054)	(0.088)	(0.115)	(0.090)	(0.109)	(0.072)	(0.089)
25	0.205 *	0.960 *	1.068 *	0.240 *	1.996 *	0.436 *	-0.473 *	0.241 *
	(0.027)	(0.056)	(0.076)	(0.100)	(0.100)	(0.119)	(0.068)	(0.081)
Mean	-0.328	0.756	1.596	0.367	1.792	0.575	-0.694	0.451
	(0.025)	(0.053)	(0.088)	(0.113)	(0.094)	(0.111)	(0.072)	(0.090)
SD	0.732 (0.005)	0.255	0.356	0.392	0.438	0.378	0.346	0.402
		(0.006)	(0.009)	(0.008)	(0.012)	(0.008)	(0.007)	(0.013)

type indicators on items 1, 6, 17, and 23 seem to capture different dynamics than those on the rest, indicating that the observed changes in scores for these items stem from distinct causes. Furthermore, a few items (8, 12, 19, 21) show non-significant discrimination parameters sporadically, implying that while some learning types on these items (e.g., negative learning) align with the general test dynamics, other learning types (e.g., positive learning) do not. A closer content analysis reveals that items 6 and 12 are related to students' understanding of the Role of Government, and items 17, 19, and 23 pertain to Aggregate Supply and Demand. This pattern suggests that mastering these concepts might involve distinct skills and cognitive structures from those required for other concepts. However, forgetting these concepts happens by the same mechanisms, as in the rest of the test. The remaining items exhibiting non-significant discrimination parameters encompass a range of topics, including Decision-Making and Marginal Analysis (item 1, the only item on this topic), Competition (item 8, 1 of 2 items on this topic), and International Macroeconomics (item 21, the only item on this topic). The variety of these topics, combined with the limited number of items assessing each, complicates any definitive conclusions regarding their distinctiveness from the broader set of topics. Yet, remarkably, most of these items have positive item intercepts in Embretson's model, meaning that they are generally more difficult for the sample than the majority of items.

Next, we studied the item fit for both of these models, to investigate if the models used are appropriate for studying the test items and learning indicators used. Table 7 presents an analysis of item fit, which is crucial to validate that all items and derived learning indicators are compliant with the model assumptions. These assumptions are integral to the model's performance, and a good item fit ensures that the variables used can be adequately explained by the model. To estimate the item fit, we use the bias-corrected RMSD fit statistics (Köhler et al., 2020), which, due to the utilization of parametric bootstrapping, are better estimates the population mean of RMSD fit statistic for the fitting items. For the cutoff values of this statistic, we use the values accepted for the model RMSEA statistics in Structural Equation Modeling due to the similar nature of item-specific RMSD and the model-specific RMSEA (Yamamoto et al., 2013): $0 \leq$ good fit $< 0.05 \leq$ marginal fit $< 0.08 \leq$ unacceptable fit.

Table 7
Item-dimension bias-corrected RMSD fit statistics.

Item	Embretson's model		The Model for Decomposed Change		
	The baseline	The 2nd measurement	Positive Learning	Negative Learning	No-change
1	0.027	0.026	0.018	0.005	0.007
2	0.033	0.047	0.042	0.038	0.052
3	0.049	0.039	0.040	0.021	0.039
4	0.023	0.024	0.025	0.025	0.033
5	0.037	0.029	0.030	0.026	0.017
6	0.046	0.033	0.020	0.032	0.022
7	0.020	0.025	0.041	0.033	0.060
8	0.022	0.023	0.032	0.015	0.017
9	0.068	0.058	0.038	0.027	0.035
10	0.046	0.051	0.020	0.030	0.040
11	0.012	0.021	0.031	0.021	0.033
12	0.052	0.057	0.034	0.015	0.018
13	0.014	0.027	0.017	0.027	0.021
14	0.031	0.026	0.033	0.020	0.033
15	0.028	0.033	0.033	0.035	0.040
16	0.038	0.046	0.025	0.023	0.046
17	0.019	0.016	0.025	0.025	0.028
18	0.032	0.022	0.019	0.018	0.018
19	0.066	0.060	0.034	0.017	0.040
20	0.035	0.043	0.025	0.029	0.026
21	0.051	0.053	0.018	0.020	0.036
22	0.051	0.050	0.025	0.032	0.031
23	0.041	0.029	0.029	0.023	0.040
24	0.039	0.045	0.024	0.030	0.037
25	0.053	0.052	0.042	0.016	0.034

The results presented in Table 7 show that, although some items (9, 12, 19, 21, 22, 25, and to a lesser extent 10) exhibit marginal fit to Embretson's model, almost all learning indications fit MDC (except the marginal fit of no-change indicators for items 3 and 7). This allows us to conclude that learning indicators are compliant with the assumption of their relation to the learning estimates. Therefore, we conclude that the results in general are reliable.

6. Conclusion

6.1. Discussion

Measuring learning progress has historically been one of the main driving forces behind the development of psychometrics. For example, the development of the Rasch models started as a response to the demand for learning progress assessment (Rasch, 1993). Over the next decades, it has become one of the most important areas of psychometric research. Numerous psychometric models exist primarily for this purpose: Andersen's model for repeated measures (Andersen, 1985), models for growth and change (Embretson, 1991) and simplex models (Jöreskog, 1970), latent growth curve models (Duncan & Duncan, 2004, Wilson et al., 2012), and various cross-lagged panel models (Mulder & Hamaker, 2021, Lüdtke & Robitzsch, 2021) to name a few. However, they are all based on the same conceptual assumption, namely that the dimensionality of learning progress on a construct is the same as the dimensionality of the construct itself. Even other psychometric techniques, not tied to a certain longitudinal model, such as vertical equating (Loyd & Hoover, 1980) or vertically moderated standard setting (Cizek & Agger, 2012), tend to incorporate this assumption in one form or another. This is a reflection of the rather intuitive logic that the ability of a student can increase, decrease, or stay the same. This intuition boils down to the basic assumption that positive and negative change are the opposite poles of the same continuum.

This paper aims to introduce the idea that positive and negative change in the targeted construct (e.g., content knowledge) might be conceptualized as a multidimensional characteristic. This approach tries to complement the traditional approach to modeling growth and change by suggesting that a student can exhibit to some extent both positive and negative learning at the same time. This assumption is also rooted in the relatively recent proposition from cognitive psychology, which states that learning and forgetting are separate cognitive processes (Sasin et al., 2017). Thus, it makes sense to model them as separate causes for the observed scores – separate latent variables. In terms of educational interpretation, this means that students might not just accumulate knowledge as a course progresses, but rather they "slide" through the course content, forgetting some parts of it while learning others. The traditional conceptualization of learning progress, however, hides these realities in education, collapsing this complexity into a simple logic of increasing or decreasing (or unchanging) ability.

We showcase the analysis of learning progress via the analysis of item learning scores. Particularly, we recode the data on the same items (see Table 2) to receive separate learning indicators for positive, negative, and no-change learning. We provide the results of IRT-based analysis of these learning scores, demonstrating that they are compliant with the model assumptions, as well as showing the complexity of learning dynamics. We show that the traditional approaches to studying learning dynamics do not allow the conception of this complexity. Particularly, we showed that students tend to simultaneously exhibit positive and negative learning, which is not possible in the traditional approach, as these tendencies cancel each other out. We also show that students who show the absent change in ability according to the traditional approach actually vary significantly in terms of their decomposed learning types.

6.2. Limitations of this study and further research directions

This study has several limitations that present opportunities for further research and exploration in the future. First, the recoded learning scores are mutually exclusive (see Tables 1 and 2): If a single learning indicator for an item takes the value of "1", all others are automatically forced to take the values of 0. This creates the "ipsativity" of the data, making IRT models for the forced-choice items more appropriate for modeling such learning scores (e.g., Brown & Maydeu-Olivares, 2011). However, IRT models for the forced-choice data are considerably more complex than the models used in this study, and correspondingly they have significantly more limitations and difficulties in their application, which make its use for such data almost impossible (Bürkner et al., 2019).

Moreover, the choice of the exact model for the forced-choice scores (e.g., Stark et al., 2005; Brown & Maydeu-Olivares, 2011; Morillo et al., 2016; Joo et al., 2021, see Nie et al., 2024) needs to be grounded in its psychological interpretation. To date, however, the rigorous and comprehensive comparison of these models in terms of their assumptions about the response process has not been conducted. Instead, there are only their partial comparisons in terms of their statistical properties (Hontangas et al., 2016). Therefore, we cannot select the IRT model for forced-choice data reflecting the cognitive processes occurring during learning, despite the attractiveness of such models in this setup. Alternatively, Item Response Tree models (De Boeck & Partchev, 2012; Jeon & De Boeck, 2016) can be suggested to model the decomposed learning scores. However, the appropriate development of such models for learning scores is a topic for further research.

Second, the plots of bivariate latent variables distributions (Fig. 3) show that the results of our modeling approach are highly heteroscedastic. This effect manifests in the cone-shaped patterns in the scatterplot. This is to be expected, because the learning indicators of different types for the same item are mutually exclusive. In other words, the more a student exhibits one type of learning, the less room they have to exhibit others. This means that the simple traditional assumption of multivariate normality of latent variables results in the mis-specified models. This leads to the inappropriateness of judging the relations between the latent variables within out modeling approach by the simple Pearson correlations. Applying IRT models for forced-choice data or appropriate Item Response Tree models could possibly resolve this issue. However, copula-based IRT and factor analysis models (Nikoloulopoulos & Joe, 2015) can also serve this purpose, which requires further investigation.

Third, the analyses outlined in the paper also require the longitudinal measurement invariance to hold (Liu et al., 2017). Investigating longitudinal measurement invariance is necessary, because it checks if items have changed their psychometric properties. If it holds, researchers interpret it as a fact that items have not changed their cognitive meaning, and the construct itself has preserved its structure and interpretation across measurement occasions. However, all such procedures are developed for the traditional conceptualization of learning progress, and no procedures exist for the learning scores to date. In this paper, we assume that, since the test content has not changed, the construct also has not changed, which is a weak assumption. How to prove that the construct has not changed its meaning in the analyses similar to those described in this paper is a topic for further research.

Fourth, in this paper, we resort to 2PL models of learning scores. In the case of one resorting to the Rasch models (and if the data fits such models), one might build a Wright Map relating learning indicators to the learning estimates (e.g., Bond et al., 2020). Additionally, if the test used is a composite in the sense that content areas are distributed non-uniformly across items, one might build a Q-matrix for learning indicators. Then, building a Wright Map for content areas in relation to different learning-type estimates is also possible (given that a Linear Logistic Test Model (Fischer, 1973) fits the data). Such visualization tools might be extremely powerful for applied educators, instructional designers, curriculum developers, and policy-makers, showing which topics are the most difficult to exhibit certain types of learning. However, due to model fit issues, we resort only to 2PL modeling, which does not allow for building Wright Maps.

Fifth, further exploration into the content-related factors influencing positive, negative, and absent learning dynamics holds considerable interest. Our framework facilitates a detailed examination of the causes and correlates associated with distinct learning types. Investigating variables such as curriculum differences and teaching practices at the institutional level, alongside study time and effort at the individual level, could provide deeper insights into the unique characteristics of each learning type. Such research has the potential to significantly enrich educational theory and practice by identifying key factors that encourage diverse types of learning. Additionally, exploring the relationship between differences in student ability and item difficulty represents another promising avenue. While these aspects were beyond the scope of our current study, we recognize that a thorough investigation into why certain items are prone to specific learning types, especially in conjunction with item content analysis, could greatly enhance our understanding of learning processes.

Despite these limitations, we show in this paper that positive, negative learning, and no-change can be conceptualized as independent but correlated continuums of learning. This approach suggests that they can occur simultaneously, implying that learning is a much more complex process than the simple accumulation of knowledge. Moreover, this approach can be further supported by the notions from cognitive psychology that learning and forgetting are different cognitive processes that should be investigated separately. The "no-change" dimension, however, is a composite dimension, combining the absent learning and the retained learning. Therefore, a detailed investigation of the cognitive process behind this dimension is necessary and further study is required. Investigations of the correlates of those types of learning and the specifics of their manifestations in the educational environment may reveal valuable practical implications for educators.

CRediT authorship contribution statement

Olga Zlatkin-Troitschanskaia: Writing – review & editing, Project administration, Methodology, Funding acquisition. **William B. Walstad:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Conceptualization. **Denis Federiakin:** Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT to enhance the readability of the text. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Acknowledgements

This study was funded by the German Federal Ministry of Education and Research with the funding number 01PK15001A.

We also would like to thank the two anonymous reviewers who provided constructive feedback and helpful guidance in the revision of this manuscript.

References

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31*(2-3), 162–172. https://doi.org/10.1016/j.stueduc.2005.05.008

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1–23. https://doi.org/10.1177/0146621697211001

AERA, A. P. A., & NCME. (2014). *Standards for Educational and Psychological Testing*. Washington DC: American Educational Research Association.
Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika, 50*, 3–16. https://doi.org/10.1007/BF02294143
Anderson, L. W., & Krathwohl, D. R. (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. *Addison Wesleyan Longman, Inc*.
Authors, 2022.
Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*(4), 431–444. https://doi.org/10.1177/014662168200600405
Bond, T., Yan, Z., & Heene, M. (2020). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Routledge. https://doi.org/10.4324/9780429030499
Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement, 71*(3), 460–502. https://doi.org/10.1177/0013164410375112
Bürkner, P. C., Schulte, N., & Holling, H. (2019). On the statistical and practical limitations of Thurstonian IRT models. *Educational and Psychological Measurement, 79*(5), 827–854. https://doi.org/10.1177/0013164419832063
Byrne, B. M., Lam, W. W., & Fielding, R. (2008). Measuring patterns of change in personality assessments: An annotated application of latent growth curve modeling. *Journal of Personality Assessment, 90*(6), 536–546. https://doi.org/10.1080/00223890802388350
Cai, L., & Houts, C. R. (2021). Longitudinal analysis of patient-reported outcomes in clinical trials: Applications of multilevel and multidimensional item response theory. *Psychometrika, 86*(3), 754–777. https://doi.org/10.1007/s11336-021-09777-y
Chang, E. K., & Wimmers, P. F. (2017). Effect of repeated/spaced formative assessments on medical school final exam performance. *Health Professions Education, 3*(1), 32–37. https://doi.org/10.5455/njppp.2023.13.01035202323012023
Chapman, G. B., & Johnson, E. J. (1994). The limits of anchoring. *Journal of Behavioral Decision Making, 7*(4), 223–242. https://doi.org/10.1002/bdm.3960070402
Cizek, G. J., & Agger, C. A. (2012). Vertically moderated standard setting. In G. J. Cizek (Ed.), *Setting Performance Standards: Foundations, Methods, and Innovations* (2nd ed..,, pp. 467–484). Routledge. https://doi.org/10.4324/9780203848203.
De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software, 48*, 1–28. https://doi.org/10.18637/jss.v048.c01
De Houwer, J., Barnes-Holmes, D., & Moors, A. (2013). What is learning? On the nature and merits of a functional definition of learning. *Psychonomic Bulletin & Review, 20*, 631–642. https://doi.org/10.3758/s13423-013-0386-3
De La Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics, 30*(3), 295–311. https://doi.org/10.3102/10769986030003295
Duncan, T. E., & Duncan, S. C. (2004). An introduction to latent growth curve modeling. *Behavior Therapy, 35*(2), 333–363. https://doi.org/10.1016/S0005-7894(04)80042-X
Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56*, 495–515. https://doi.org/10.1007/BF02294487
Epskamp, S., Maris, G., Waldorp, L. J., & Borsboom, D. (2018). Network psychometrics. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley Handbook of Psychometric Testing: a Multidisciplinary Reference on Survey, Scale and Test Development* (pp. 953–986). Wiley Blackwell. https://doi.org/10.1002/9781118489772.ch30.
Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*(6), 359–374. https://doi.org/10.1016/0001-6918(73)90003-6
Giambra, L. M., & Arenberg, D. (1993). Adult age differences in forgetting sentences. *Psychology and Aging, 8*(3), 451–462. https://doi.org/10.1037/0882-7974.8.3.451
Grek, S. (2009). Governing by numbers: The PISA 'effect' in Europe. *Journal of Education Policy, 24*(1), 23–37. https://doi.org/10.1080/02680930802412669
Hontangas, P. M., Leene, I., Torre, J. D. L., Ponsoda Gil, V., Morillo Cuadrado, D., & Abad García, F. J. (2016). Traditional scores versus IRT estimates on forced-choice tests based on a dominance model. *Psicothema*. https://doi.org/10.7334/psicothema2015.204
Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: a Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118
Jeon, M., & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior Research Methods, 48*, 1070–1085. https://doi.org/10.3758/s13428-015-0631-y
Joo, S. H., Lee, P., & Stark, S. (2021). Modeling multidimensional forced choice measures with the Zinnes and Griggs pairwise preference item response theory model. *Multivariate Behavioral Research, 58*(2), 241–261. https://doi.org/10.1080/00273171.2021.1960142
Jöreskog, K. G. (1970). Estimation and testing of simplex models. *British Journal of Mathematical and Statistical Psychology, 23*(2), 121–145. https://doi.org/10.1111/j.2044-8317.1970.tb00439.x
Köhler, C., Robitzsch, A., & Hartig, J. (2020). A bias-corrected RMSD item fit statistic: An evaluation and comparison to alternatives. *Journal of Educational and Behavioral Statistics, 45*(3), 251–273. https://doi.org/10.3102/1076998619890566
Kopelman, M. D. (1985). Rates of forgetting in Alzheimer-type dementia and Korsakoff's syndrome. *Neuropsychologia, 23*(5), 623–638. https://doi.org/10.1016/0028-3932(85)90064-8
Kristensen, M., & Hansen, T. (2004). Statistical analyses of repeated measures in physiological research: a tutorial. *Advances in Physiology Education, 28*(1), 2–14. https://doi.org/10.1152/advan.00042.2003
Levy, R., Xu, Y., Yel, N., & Svetina, D. (2015). A standardized generalized dimensionality discrepancy measure and a standardized model-based covariance for dimensionality assessment for multidimensional models. *Journal of Educational Measurement, 52*(2), 144–158. https://doi.org/10.1111/jedm.12070
Liu, Y., Millsap, R. E., West, S. G., Tein, J. Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological Methods, 22*(3), 486–506. https://doi.org/10.1037/met0000075
Loftus, G. R. (1985). Evaluating forgetting curves. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*(2), 397–406. https://doi.org/10.1037/0278-7393.11.2.397
Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, 179–193. https://doi.org/10.1111/j.1745-3984.1980.tb00825.x
Lüdtke, O., & Robitzsch, A. (2021). A Critique of the Random Intercept Cross-Lagged Panel Model. *PsyArXiv Preprint*. https://doi.org/10.31234/osf.io/6f85c
Macha, K., & Schuhen, M. (2011). Framework of measuring economic competencies. *Journal of Social Science Education*. https://doi.org/10.4119/jsse-570
Mary, A., Schreiner, S., & Peigneux, P. (2013). Accelerated long-term forgetting in aging and intra-sleep awakenings. *Frontiers in Psychology, 4*, 750. https://doi.org/10.3389/fpsyg.2013.00750
Mislevy, R. J. (2018). *Sociocognitive Foundations of Educational Measurement*. New York: Routledge. https://doi.org/10.4324/9781315871691
Morillo, D., Abad, F. J., Kreitchmann, R. S., Leenen, I., Hontangas, P., & Ponsoda, V. (2019). The journey from Likert to forced-choice questionnaires: Evidence of the invariance of item parameters. *Revista Dëlelőtt Psicología del Trabajo York Dëlelőtt las Organizaciones, 35*(2), 75–83. https://doi.org/10.5093/jwop2019a11
Morillo, D., Leenen, I., Abad, F. J., Hontangas, P., de la Torre, J., & Ponsoda, V. (2016). A dominance variant under the multi-unidimensional pairwise-preference framework: Model formulation and Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 40*(7), 500–516. https://doi.org/10.1177/0146621616662226
Mulder, J. D., & Hamaker, E. L. (2021). Three extensions of the random intercept cross-lagged panel model. *Structural Equation Modeling: a Multidisciplinary Journal, 28*(4), 638–648. https://doi.org/10.1080/10705511.2020.1784738
Muthén, B. O., & Khoo, S. T. (1998). Longitudinal studies of achievement growth using latent variable modeling. *Learning and Individual differences, 10*(2), 73–101. https://doi.org/10.1016/S1041-6080(99)80135-6
Nie, L., Xu, P., & Hu, D. (2024). Multidimensional IRT for forced choice tests: A literature review. *Heliyon*. https://doi.org/10.1016/j.heliyon.2024.e26884
Nikoloulopoulos, A. K., & Joe, H. (2015). Factor copula models for item response data. *Psychometrika, 80*, 126–150. https://doi.org/10.1007/s11336-013-9387-4
Panik, M. J. (2014). *Growth curve modeling: theory and applications*. John Wiley & Sons. https://doi.org/10.1002/9781118763971
Piacentini, M., Foster, N., & Nunes, C. (2023). Next-generation assessments of 21st Century competencies: Insights from the learning sciences. In N. Foster, & M. Piacentini (Eds.), *Innovating Assessments to Measure and Support Complex Skills*. OECD Publishing. https://doi.org/10.1787/fe01601f-en.
Rasch, G. (1993). Probabilistic models for some intelligence and attainment tests. MESA Press. web-address: ⟨www.rasch.org⟩.
Rivera-Lares, K., Logie, R., Baddeley, A., & Della Sala, S. (2022). Rate of forgetting is independent of initial degree of learning. *Memory & Cognition, 50*, 1706–1718. https://doi.org/10.3758/s13421-021-01271-1
Rivera-Lares, K., Sala, S. D., Baddeley, A., & Logie, R. (2023). Rate of forgetting is independent from initial degree of learning across different age groups. *Quarterly Journal of Experimental Psychology, 76*(7), 1672–1682. https://doi.org/10.1177/17470218221128780
Robitzsch, A., Kiefer, T., & Wu, M. (2022). *Package 'TAM' (Test Analysis Modules), v. 4.1-4*. ⟨https://cran.r-project.org/web/packages/TAM/TAM.pdf⟩.
Ruspini, E. (2003). *An introduction to longitudinal research*. Routledge. https://doi.org/10.4324/9780203167229
Sasin, E., Morey, C. C., & Nieuwenstein, M. (2017). Forget me if you can: Attentional capture by to-be-remembered and to-be-forgotten visual stimuli. *Psychonomic Bulletin & Review, 24*(5), 1643–1650. https://doi.org/10.3758/s13423-016-1225-0
Schmidt, S., Zlatkin-Troitschanskaia, O., & Walstad, W. B. (2019). IRT modeling of decomposed student learning patterns in higher education economics. *Frontiers and advances in Positive Learning in the Age of Information (PLATO)*, 237–251. https://doi.org/10.1007/978-3-030-26578-6_17
Slamecka, N. J., & McElree, B. (1983). Normal forgetting of verbal lists as a function of their degree of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9*(3), 384–397. https://doi.org/10.1037/0278-7393.9.3.384
Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement, 29*(3), 184–203. https://doi.org/10.1177/0146621604273988
Talbot, R. M., III (2013). Taking an item-level approach to measuring change with the force and motion conceptual evaluation: An application of item response theory. *School Science and Mathematics, 113*(7), 356–365. https://doi.org/10.1111/ssm.12033
Trafimow, D. (2015). A defense against the alleged unreliability of difference scores. *Cogent Mathematics, 2*(1), 1064626. https://doi.org/10.1080/23311835.2015.1064626

Vosniadou, S. (1994). Capturing and modeling the process of conceptual change. *Learning and Instruction, 4*(1), 45–69. https://doi.org/10.1016/0959-4752(94)90018-3

Walberg, H. J., & Tsai, S. L. (1983). Matthew effects in education. *American Educational Research Journal, 20*(3), 359–373. https://doi.org/10.3102/00028312020003359

Walsh, C. M., Wilkins, S., Bettcher, B. M., Butler, C. R., Miller, B. L., & Kramer, J. H. (2014). Memory consolidation in aging and MCI after 1 week. *Neuropsychology, 28*(2), 273–280. https://doi.org/10.1037/neu0000013

Walstad, W. B., Rebeck, K., & Butters, R. B. (2013). *Test of Economic Literacy: Examiner's Manual*. National Council on Economic Education.

Walstad, W. B., & Wagner, J. (2016). The disaggregation of value-added test scores to assess learning outcomes in economics courses. *The Journal of Economic Education, 47*(2), 121–131. https://doi.org/10.1080/00220485.2016.1146104

Walstad, W. B., Watts, M., & Rebeck, K. (2007). *Test of Understanding in College Economics: Examiner's Manual*. Council on Economic Education.

Walstad, W. B., & Zlatkin-Troitschanskaia, O. (2024). Learning Scores and Economics Instruction. *The American Economist, 69*(1), 6–20. https://doi.org/10.1177/05694345231207868

Wells, C.S., & Wollack, J.A. (2003). *An instructor's guide to understanding test reliability*. Testing & Evaluation Services University of Wisconsin. ⟨https://testing.wisc.edu/Reliability.pdf⟩.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching, 46*(6), 716–730. https://doi.org/10.1002/tea.20318

Wilson, M., Zheng, X., & McGuire, L. (2012). Formulating latent growth using an explanatory item response model approach. *Journal of Applied Measurement, 13*(1), 1–22.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31*(2-3), 114–128. https://doi.org/10.1016/j.stueduc.2005.05.005

Yamamoto, K., Khorramdel, L., & Von Davier, M. (2013). Scaling PIAAC cognitive data. *In Technical Report of the Survey of Adult Skills (PIAAC)* (pp. 408–440). OECD Publishing,.

Yang, J., Zhan, L., Wang, Y., Du, X., Zhou, W., Ning, X., Sun, Q., & Moscovitch, M. (2016). Effects of learning experience on forgetting rates of item and associative memories. *Learning & Memory, 23*(7), 365–378. https://doi.org/10.1101/lm.041210.115