Investigation of adolescents' developmental stages in deductive reasoning: An application of a specialized confirmatory mixture IRT approach

Minjeong Jeon¹ · Karen Draney² · Mark Wilson² · Yinghao Sun³

Published online: 20 March 2019
© The Psychonomic Society, Inc. 2019

Abstract

In this paper, we propose a specialized confirmatory mixture IRT model to analyze complex cognitive assessment data that is designed to evaluate adolescents' developmental stages in deductive reasoning. The model is specified for the following purposes: (1) to measure multiple deductive reasoning traits, (2) to identify adolescents' differential developmental stages based on their ability levels in the multiple dimensions, (3) to quantify the differences in dimension-specific performance between developmental stages, and (4) to examine the difficulty levels of test design factors. A Bayesian estimation of the model is described. The overall goodness-of-fit of the model is assessed as well as its parameter recovery to validate the application of the model to the data.

Keywords Deductive reasoning · Developmental stages · Piaget's theory · Mixture IRT · Multi-dimensionality · Item properties · Saltus model

Introduction

In this study, we present an analysis of a complex cognitive assessment that is designed to examine adolescents' developmental stages in deductive reasoning. We propose a specialized confirmatory mixture IRT model where its specification is motivated by the need to analyze data with a convoluted data structure.

The model is designed to serve the following purposes: (1) to measure multiple deductive reasoning traits, (2) to identify the adolescents' different developmental stages based on their multiple ability levels, (3) to quantify the differences in the adolescents' dimension-specific

performance between developmental stages, and (4) to examine the cognitive complexity of the test design factors.

In the following “[Specialized confirmatory mixture IRT model](#)”, we first describe our motivating data example. In the subsequent “[Data and analysis](#)”, we describe the rationale, formulation, and features of the proposed model for analyzing the data. We also discuss relations with other existing models and the benefits of a confirmatory approach in mixture IRT modeling. In “[Results](#)”, we then describe the Bayesian estimation of the described model and present the results from the data analysis in “[Model validation](#)”. In “[Discussion](#)”, we provide the evaluation of the goodness-of-fit of the model as well as its parameter recovery. We conclude the paper with a summary and a discussion on contributions and other applications.

Electronic supplementary material The online version of this article (<https://doi.org/10.3758/s13428-019-01221-5>) contains supplementary material, which is available to authorized users.

✉ Minjeong Jeon
mjjeon@ucla.edu

¹ Department of Education, University of California, Los Angeles, 3141 Moore Hall, 457 Portola Avenue, Los Angeles CA, 90024, USA

² University of California, Berkeley, CA, USA

³ Ohio State University, Columbus, OH 43210, USA

Background: deductive reasoning assessment

The Competence Profile Test of Deductive Reasoning - Verbal test (DRV; Spiel, Gluck, & Gossler, 2001; Spiel & Gluck 2008) was designed to assess the competence profile and level of participants' deductive reasoning (Draney, Wilson, Gluck, & Spiel, 2007). Deductive reasoning is a logical process where a conclusion is drawn based on the concordance of multiple premises that are generally presumed to be true. As stated in Piaget (1971)'s

cognitive-developmental theory, which is one of the most influential theories of deductive reasoning, children transition through four developmental stages: the sensorimotor, the pre-operational, the concrete operational, and the formal operational stages. These stages qualitatively differ in the cognitive processes (Spiel et al., 2001).

The DRV test focuses on assessing school-aged participants who are either in the concrete operational or formal operational stages. A transition between these stages occurs during the beginning of adolescence (roughly at age 12), while the transition from one stage to another is characterized by complex growth in reasoning ability which involves a vital restructuring of the thinking process (Draney et al., 2007).

According to Piaget, concrete operational participants interpret all inferences as bi-conditional and hence, are able to only perform logical operations on concrete objects. In comparison, formal operational participants can perform logical operations on abstractions as well as concrete objects and thus, can draw correct inferences on all problems (Spiel et al. 2001; Draney et al. 2007). However, empirical studies suggest that these are theoretical classifications and do not fully reflect reality. For instance, researchers observed that older adolescents were not able to apply formal operational strategies to all problems (e.g., Neimark 1975), while others reported that even adults could not distinguish deductive arguments from non-deductive problems (e.g., Chazan 1993; Spiel et al. 1997; Staudenmeyer & Bourne 1972; Wildman & Fletcher 1977).

For these reasons, most neo-Piagetians argue that cognitive development is more gradual than what Piaget theorized (e.g., Rips 1994). It should be noted, however, that there is still a strong interest in applying stage-like development analysis in various areas, including adult development (e.g., Commons, Trudeau, Stein, Richards, & Krause, 1998; Fischer, Hand, & Russel, 1984) and cognitive development (e.g., Bond 1995a, b; Bond & Bunting 1995; Demetriou & Efklides 1994, 1989; van Hiele 1986).

To examine adolescents' development in deductive reasoning in more detail, the DRV test was developed by utilizing three design factors: type of inference, content of the conditional, and presentation of the antecedent (Spiel et al., 2001). First, type of inference differentiates four inference types based on premises and conclusions: (a) Modus Ponens (MP; A, therefore B), (b) Negation of Antecedent (NA; Not A, therefore B or not B), (c) Affirmation of Consequent (AC; B, therefore A or not A), and (d) Modus Tollens (MT; Not B, therefore not A). The MP and MT types come with a "yes" or "no" binary response (that is, bi-conditional conclusions). The NA and AC types that have an additional "perhaps" response option, are also known as a 'logical fallacy' type in the sense that those provoke the choice of a seemingly straightforward

but logically incorrect solution. An example of an AC item is "Susan is lying in her bed. Is Susan sick?" (in this case, the correct answer is "perhaps"). Second, content of the conditional differentiates three content types: (a) Concrete (CO), (b) Abstract (AB), and (c) Counterfactual (CF). An example CF item is given as "If an object is put into boiling water, it becomes cold". Third, presentation of the antecedent differentiates two types based on the presence of negation: (a) with negation (NE) or without negation (NN). An example of a negation item is "If the sun does not shine, Tom wears blue pants". Investigators showed that logical fallacy (NA and AC) items are generally more difficult than bi-conditional (MT and MP) items, and the probability of correctly solving the NA and AC items usually increases as cognitive development progresses. Other researchers found that Concrete (CO) items are easier to solve than Abstract (AB) and Counterfactual (CF) items; however, there seems to be no clear performance difference between abstract (AB) and counterfactual (CF) items (e.g., Overton 1985). Items become more difficult to solve when they are presented with negation (NE) than without negation (NN) (e.g., Roberge & Mason 1978). Table 1 summarizes the DRV test design factors based on their complexity levels.

Previous studies

Two prior studies have been conducted with the DRV data for the purpose of identifying the developmental stages of adolescents. Spiel et al. (2001) employed a mixture Rasch model and reported that three latent classes were identifiable, where Class 1 could be described as the concrete-operational stage (with those who correctly solved MP and MT items only), Class 2 as the formal-operational stage (with those who showed better performance with the NA and AC items than Class 1 participants, and Class 3 as the transition stage (with those who showed mixed performance compared to Class 1 and Class 2 participants).

To validate the results from Spiel et al. (2001), Draney et al. (2007) applied a confirmatory mixture Rasch model called the Saltus model (Wilson 1989; Mislevy and Wilson 1996, Draney et al. 2007). The Saltus model is a confirmatory model in that the number and nature of the latent classes are pre-specified (i.e., developmental stages) and the analysis goal is to verify the hypothesized latent class structure. The model imposes a set of constraints on item parameters using prior knowledge on items and item-class relationship, so that the model can differentiate participants in different latent classes. Draney et al. (2007) reported that logical fallacy items operated satisfactorily to differentiate adolescents in the formal-operational stage from the concrete-operational one, and concluded that two

Table 1 Elements of the three design factors of the DRV assessment (listed based on relative complexity)

Design factor	Less complex	More complex
Type of inference	Bi-conditional (MP, MT)	Logical fallacy (NA, AC)
Content of conditional	Concrete (CO)	Abstract (AB), Counterfactual (CF)
Presentation of antecedent	No negation (NN)	Negation (NE)

latent classes, not three, were clearly differentiable with their confirmatory modeling approach.

These previous studies showed agreement on the presence of heterogeneous participant clusters in the DRV data, which correspond to different developmental stages of deductive reasoning. Both studies, however, seem to have minimally utilized in their analyses the information that the DRV test could offer. For instance, this complex design assessment allows us to measure more than one deductive reasoning trait and to check whether/how adolescents in different developmental stages may present different (or similar) reasoning levels in each deductive reasoning dimension. Hence, there is a room for further analysis and investigation.

Specialized confirmatory mixture IRT model

Rationale

We propose a customized confirmatory mixture IRT model for the DRV assessment data based on the following reasoning: First, we wish to measure two deductive reasoning traits based on the two types of test items that differ in the presentation of the antecedent (NE/NN design factor). The two dimensions (NE and NN) are defined with non-overlapping 12 items, each consisting of six MT/MP and six NA/AC items and four AB, four CF, and four CO items. Among the three design factors, we choose the presentation of an antecedent design factor because research has shown that solving negation (NE) or no negation (NN) items require distinctive cognitive ability levels. The other two design factors appear inappropriate for defining dimensions of traits; the content of conditional factor includes empirically indistinguishable types (e.g., abstract (AC) and counterfactual (CF) items). The type of inference (logical fallacy vs. bi-conditional items) is used to define the item group and for latent class differentiation (this will be explained in detail later). Second, based on the prior research (Spiel et al., 2001; Draney et al., 2007), we intend to classify participants into two developmental stages (concrete-operational vs. formal-operational stages), but across two deductive reasoning dimensions. Third, as in the prior studies, we utilize logical fallacy (NA/AC) items (vs. bi-conditional (MT/MP) items) to differentiate

formal-operational participants from concrete-operational participants. This item group is utilized both in NE and NN dimensions for class differentiation. Fourth, we would like to evaluate the cognitive complexity levels of the DRV test design factors to validate or invalidate previous study results on those factors. Specifically, for the type of inference factor, the NA, AC, and MT types are used with MP as the reference type and for the content of the conditional factor, the AB and CF factors are utilized with CO as the reference content.¹

Model formulation

For dichotomous response y_{ij} (1 if correct, 0 if not) to item i ($i = 1, \dots, I$) for subject j ($j = 1, \dots, N$) who belong to latent class $z_j = g$ ($g = 1, \dots, G$), we construct the model as follows:

$$\text{logit}(\Pr(y_{ij} = 1 | \theta_{jg}, z_j = g)) = \sum_{k=1}^K r_{ik} \theta_{jkg} - \sum_{q=0}^Q b_q X_{iq} + \epsilon_i + \sum_{k=1}^K r_{ik} \tau_{ghk}. \quad (1)$$

First, the regression coefficient b_q represents the average difficulty of the q th design factor X_{iq} ($q = 1, \dots, Q$), rather than the difficulty of individual test items (note that b_q has subscript q rather than i). The global intercept b_0 represents the average logit probability of a correct response when $X_{iq} = 0$ for all q for subjects with average ability levels. The idea of modeling average difficulties of item features, rather than individual items, was utilized in the linear logistic test model (LLTM; Fischer 1973). We additionally include the error term ϵ_i in the model, where $\epsilon_i \sim N(0, \sigma_\epsilon^2)$, in order to allow for the items with the same design features to have different levels of difficulty, as proposed by e.g., Janssen, Schepers, and Peres (2004).

Next, τ_{ghk} is the structural parameter that is introduced to differentiate subjects in different latent classes. Technically, this structural parameter indicates the effect of item group h ($h = 1, \dots, G$) on the logit probability of subjects in class g ($g = 1, \dots, G$) correctly solving item group h in dimension k . Note that introducing this parameter is equivalent to

¹The presentation of an antecedent factor was used to define two dimensions of the DRV test rather than used as item covariates.

imposing a set of planned constraints to a standard mixture IRT model, which are tailored to the multidimensional assessment structure under investigation. The number of item groups is set equivalent to the number of latent classes because each item group represents the performance of a corresponding latent class (Wilson, 1989; Mislevy & Wilson, 1996; Draney, 2007). Conceptually, the parameter τ_{ghk} represents an average advantage (or disadvantage) that subjects in class g have in solving item group h in dimension k , compared with the subjects in the reference latent class (e.g., $g = 1$). If the structural parameter is positive, it means that the items in group h are easier for the subjects in class g . However, if this structural parameter is negative, this indicates that the group h items are more for the subjects in class g . This type of parameter was devised in the Saltus model for unidimensional tests (Wilson, 1989; Mislevy & Wilson, 1996). In the current paper, we extend the Saltus modeling idea for a multidimensional assessment by allowing for the amount of advantage (or disadvantage) that subjects in class g have (in solving items in group h) to differ across dimensions of the test. Note that this kind of information is unobtainable with existing mixture IRT models or the original Saltus model.

For model identification and ease of interpretation, we set $\tau_{1,k} = \tau_{1k} = 0$ for each dimension k . Such an identification constraint was proposed by Wilson (1989) and Mislevy and Wilson (1996) for a unidimensional case and it means that the first latent class ($g = 1$) and the first item group ($h = 1$) are used as the reference groups per dimension. When applying this rule, two structural parameters are estimated (τ_{221} and τ_{222} , for each of the two dimensions) for the DRV assessment data.

Next, in order to let the model know which test items tap into which dimensions (or latent traits that are measured in the corresponding dimensions), the $I \times K$ score matrix R is introduced in Eq. 1, where r_{ik} indicates the (i, k) th element (that takes value 0 or 1) of the score matrix, denoting whether the i -th item belongs to the k -th dimension (or latent trait). This score matrix has a simple structure, meaning that each item is associated with only one dimension. For the DRV data, the score matrix is a 24×2 matrix, where its first column corresponds to the dimension of negation (NE) and the second column corresponds to the dimension of no negation (NN). For each latent class g , subject j has a K -dimensional latent trait vector, $\theta_{jg} = (\theta_{j1g}, \dots, \theta_{jKg})'$ that is assumed to follow a multivariate normal distribution, $\theta_{jg} \sim N(\mu_g, \Sigma_g)$. The means and covariance of the K -dimensional latent traits are freely estimated for each latent class g , except for the means of the reference latent class (e.g., $\mu_g = 0$ when $g = 1$ is specified as the reference group). That is, the specified model allows for latent classes to have a different correlational structure for the latent traits of

Fig. 1 A hypothetical joint mixture distribution of the two latent traits (θ_1 and θ_2) for Class 1 (C1) and Class 2 (C2)

interest. Such an approach is not only more realistic but also more informative compared with the approaches that assume the same correlational structure across latent classes and the approaches that assume zero correlation between latent traits (or apply a separate unidimensional model to each test dimension).

Figure 1 illustrates a hypothetical distribution of the two latent classes in the two-dimensional coordinate system. The figure shows that the two latent traits (θ_1 and θ_2) marginally follow a mixture of normal distributions (on the x -axis and the y -axis, respectively). The two-dimensional latent traits jointly follow a mixture of two bivariate normal distributions with a class-specific mean vector and variance-covariance matrix as in De Boeck and Wilson (2004).

Lastly, Eq. 1 is a conditional model given a particular latent class. The marginal probability model can be specified with respect to all latent classes:

$$\Pr(y_{ij} = 1 | \theta_{jg}) = \sum_{g=1}^G \pi_g \Pr(y_{ij} = 1 | \theta_{jg}, z_j = g), \quad (2)$$

where π_g is the probability of classifying into class g ($g = 1, \dots, G$). The parameter π_g is also referred to as a mixing proportion (with $\sum_{g=1}^G \pi_g = 1$).

Further discussion

Related models

The model described in “[Model formulation](#)” is a special type of finite mixture IRT model that accommodates the

difficulties of item features and multidimensionality of latent traits. Similar extensions have been presented in the literature. For instance, Mislevy and Verhelst (1990) presented a mixture IRT model that estimates the difficulties of a few item properties, rather than individual test items. Multidimensional mixture IRT models have been applied in several contexts, for instance, for large-scale cross-country data analysis (e.g., De Jong & Steenkamp 2010), longitudinal data analysis (e.g., Cho, Cohen, Kim, & Bottge, 2010; Cho, Cohen, & Bottge, 2013; von Davier, Xu, & Carstensen, 2011), and differential item functioning investigations (e.g., De Boeck, Cho, & Wilson, 2011). In addition, Cho, Cohen, and Kim (2014) and Huang (2016) presented extended mixture IRT models for bifactor and higher-order test structures, respectively. Finch and Finch (2013) applied a multilevel and multidimensional extension of a mixture IRT model and Choi and Wilson (2015) presented a mixture random weights linear logistic test model, which can be seen as a mixture extension of the Rasch model with item features and multidimensionality.

A distinctive feature that our model offers, compared with these existing models, is that our model includes structural parameters that quantify latent class differences. Such a construction has been presented for unidimensional binary responses (Wilson, 1989; Mislevy & Wilson, 1996) and extended for polytomous responses (Draney, 2007) and with item covariates (Draney & Wilson, 2008). In addition, Jeon (2018) presented several modifications of the Saltus model with item discrimination parameters, person predictors, and ordinal item responses. To our knowledge, however, the original Saltus model has not been adapted to accommodate multiple dimensions and/or to permit heterogeneity in the item feature difficulties (with a random error term) as we did in the current application.

Exploratory vs. confirmatory mixture modeling

Another important feature of our model is that it is a confirmatory approach, whereas most other mixture IRT models are an exploratory approach. Some researchers who may be more accustomed with a more traditional, exploratory use of mixture IRT modeling may feel that such a confirmatory use of mixture modeling is somewhat unusual. In an exploratory approach, mixture IRT models are utilized assuming (1) the number and nature of latent classes are unknown and (2) no structure is imposed in the class-specific item parameters prior to data analysis. In a confirmatory approach, researchers utilize prior knowledge on the number/character of latent classes as well as test items for model construction. Although relatively less common compared with an exploratory approach, confirmatory mixture IRT modeling have been adopted in various applications.

For example, Mislevy and Verhelst (1990) investigated examinee's two types of item solution strategies (guessing and ability-based strategy) during tests (also see e.g., Schnipke & Scrams 1997; Yamamoto & Everson 1997; Boughton & Yamamoto 2007). Molenaar, Oberski, Vermunt, and De Boeck (2016) hypothesized two types of intelligence (slow and fast) as latent classes and investigated whether and how examinees adopt different types of intelligence during tests. Tijmstra, Bolsinova, and Jeon (in press) assumed two response styles as latent classes and studied examinee's differential item solution behavior. Jin, Chen, and Wang (in press) presumed inattentive response behavior that is differentiated from normal response behavior and applied a confirmatory mixture IRT approach to rating-scale data with two latent classes. We would like to point out that our contribution is to present a special application of customized confirmatory mixture IRT modeling to measurement and applied researchers.

It is important to note that research using confirmatory approaches differs from research based on exploratory approaches. Exploratory mixture modeling aims to identify the number and characteristics of latent classes which are unknown prior to data analysis, whereas confirmatory modeling intends to validate a researcher's hypothesis on the nature of postulated latent classes and/or latent class differentiation. Note that in factor analysis contexts, confirmatory and exploratory approaches are utilized in a similar way. Exploratory factor analysis is adopted to identify the number of factors and item-factor relationship, while confirmatory factor analysis is applied to verify a factor structure hypothesized prior to data analysis. Both confirmatory and exploratory factor analyses are extensively utilized in applied research.

Data and analysis

Data

The DRV test data collected by Spiel et al. (2001) and described in "[Background: deductive reasoning assessment](#)" were used for data analysis. Specifically, the DRV test was administered to 418 secondary school students (162 females and 256 males) in Graz, Austria. There were approximately the same number of students in grades 7 through 12 (age 11 through 18).

The students' responses were coded dichotomously, with 1 for correct, and 0 for incorrect responses. The DRV questionnaire was administered in classrooms during regular class hours with no set time limits. To control for order effects, two task versions (A and B) were implemented with different random orders of the items. Half of the participants were presented with version A and the

other half with version B. To see whether the test version plays a role in the results, we conducted preliminary data analysis with the regular Saltus model and our proposed model by including the main effect of the test version. We found from both analyses that the version effect was not significantly different from zero at the 5% level, assuring us that the two test versions did not make a difference in the results.

Although Piaget theorized that the transition from concrete-operational stage to formal-operational stage occur around age 12, numerous empirical studies reported that the transition might not happen even in late adolescence and early adulthood. For that reason, Spiel et al. (2001) targeted school-aged-participants (in the age range of 11 to 18) for their investigation.

Analysis

To fit the model constructed in “[Model formulation](#)” we applied a Bayesian approach with Gibbs sampling. To obtain full conditional distributions, prior distributions are specified for the model parameters as follows: First, for the item regression parameters b_q ($q = 0, \dots, Q$) as well as the structural parameter τ_{ghk} ($g = 1, \dots, G, h = 1, \dots, H, k = 1, \dots, K$), a normal prior is specified with a relatively large variance:

$$b_q \sim N(0, 10),$$

$$\tau_{ghk} \sim N(0, 10).$$

For the inverse of the variance parameter σ_ϵ^2 , a slightly informative gamma prior (with the shape and rate parameters in parenthesis) is assigned:

$$\{\sigma_\epsilon^2\}^{-1} \sim \text{gamma}(1, 1).$$

For the ability distributions $\theta_{jg} \sim N(\mu_g, \Sigma_g)$, a normal prior is assigned to the mean parameter μ_{kg} and for the inverse of the variance-covariance matrices Σ_g^{-1} , a Wishart distribution with the scale matrix I is specified:

$$\mu_{kg} \sim N(0, 10),$$

$$\Sigma_g^{-1} \sim \text{Wishart}(I, \nu),$$

where I is the $K \times K$ identity matrix and ν is the degree of freedom that is set to K . A Dirichlet distribution is specified for the mixing proportion π_g of the latent class g :

$$\pi_g = (\pi_1, \dots, \pi_G)' \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_G),$$

where the hyperparameter α_g is 1.

Posterior samples were obtained using the JAGS software (Plummer 2011, 2003), based on 100,000 iterations (with a 90,000 burn-in), while fifth sampled value was retained. Three chains were used with three different starting values.

For convergence checking, the Gelman and Rubin (1992) method and the Geweke (1992) method were utilized along with graphical checks. In this application, no label switching problem was observed between and within chains for the fitted model. The JAGS code is provided as the supplementary material (Appendix A).

In addition, we also fit the original Saltus model for a comparison with the proposed model. The Saltus model was set up for the DRV data with two latent classes (concrete- and formal-operational stages) and two item groups (NA/AC vs. MT/MP items). The Saltus model is a uni-dimensional model; hence, two test dimensions are indistinguishable in the Saltus model. The basic Saltus model includes the following parameters: (1) the difficulty parameters for the 24 individual DRV test items, (2) the means and standard deviations of the two latent classes (except for the mean of the first latent class (as the reference class) that was fixed at 0), and (3) the structural parameter that differentiates Class 2 from Class 1 based on the NA/AC items. The Saltus model was estimated by using Gibbs sampling similar to the extended model utilized in the current study. Details of the formulation of the Saltus model and the specification of the priors are provided in the supplementary material (Appendix B).

Results

Table 2 lists the parameter estimates (posterior means) and standard errors (posterior standard deviations) of the proposed model utilized in the current study. We also discuss the parameter estimates of the basic Saltus model when such comparison is appropriate. The estimates of all model parameters of the Saltus model are provided in the supplementary material (Appendix B).

First, the structural parameter estimates (standard errors) are $\hat{\tau}_{221} = -4.51$ (0.22) in the NE dimension and $\hat{\tau}_{222} = -4.21$ (0.22) in the NN dimension. These two values are significant at the 5% level based on the Wald test. From the basic Saltus model, the structural parameter is estimated as -4.31 (0.29), which is also significant at the 5% level. The structural parameters are the effect of the NA/AC items on the probability of success for adolescents in Class 2, and represent an extra degree of ‘easiness’ on the NA/AC item group for the adolescents in Class 2. The negative estimated values of the structural parameters indicate that the NA/AC items are more difficult for adolescents in Class 2 than in Class 1.

Thus, our result suggests that Class 1 students have advantages in solving the NA and AC items compared with Class 2 students, while the amount of the advantage is similar in both dimensions of the test (that is, regardless of whether or not the items are presented with negation).

Table 2 Parameter estimates (posterior means; Est) and standard errors (posterior standard deviations; SE)

	Par	Est	SE
Structural parameters			
Dim1: NE	τ_{221}	−4.51	0.22
Dim2: NN	τ_{222}	−4.21	0.22
Item parameters			
	b_0	−1.04	0.50
	$b_1(\text{NA})$	−0.31	0.58
	$b_2(\text{AC})$	0.23	0.57
	$b_3(\text{MT})$	0.85	0.57
	$b_4(\text{AB})$	0.55	0.49
	$b_5(\text{CF})$	0.67	0.50
	σ_ϵ	0.98	0.17
Class proportion			
Class 1	π_1	0.51	0.03
Class 2	π_2	0.49	0.03
Mean (Class 2)			
Dim1: NE	μ_{12}	1.56	0.17
Dim2: NN	μ_{22}	1.51	0.15
Covariance (Class 1)			
SD (NE)	σ_{111}	1.16	0.10
SD (NN)	σ_{221}	0.82	0.07
Cov	σ_{121}	0.88	0.14
Corr	ρ_1	0.92	0.03
Covariance (Class 2)			
SD (NE)	σ_{112}	0.59	0.10
SD (NN)	σ_{222}	0.53	0.08
Cov	σ_{122}	0.22	0.08
Corr	ρ_2	0.70	0.11

Item regression parameters: b_0 to b_5 ; Standard deviation of the item regression: σ_ϵ ; Structural parameters: τ_{ghk} , $k = 1, 2$; Mixing proportion for class g : π_g , $g = 1, 2$; Means of latent traits: μ_{kg} ; Standard deviations and covariances of latent traits: σ_{kkg} ; Correlation between latent traits: ρ_g

Given that solving the NA/AC items requires a higher cognitive level typical of the formal-operational developmental stage, we conclude that Class 1 describes the higher, formal-operational developmental stage, while Class 2 describes the lower, concrete-operational stage.

Second, the analysis showed that approximately 51% and 49% of the adolescents are classified into Class 1 and Class 2, respectively. These proportions were equivalent to those from the basic Saltus model. Although the proportion of adolescents in each of the two classes is similar, there are clear differences in the latent trait distributions between the two classes.

Specifically, for Class 2 we have $\hat{\mu}_{12} = 1.56$ and $\hat{\mu}_{22} = 1.51$ in the NE and NN dimensions, respectively (for Class 1, $\hat{\mu}_{11} = \hat{\mu}_{21} = 0$). From the Saltus model, the Class 2 mean is estimated as 2.41 (0.40).

To compute the actual class means of the two dimensions, we must take into account the item intercept parameter ($\hat{b}_0$). That is, for the NE dimension, the mean for Class 2 is

$\hat{\mu}_{12} + \hat{b}_0 = 1.56 - 1.04 = 0.5$, while the mean for Class 1 is $\mu_{11} + \hat{b}_0 = 0 - 1.04 = -1.04$. For the NN dimension, the mean for Class 2 is $\hat{\mu}_{22} + \hat{b}_0 = 1.51 - 1.04 = 0.47$, while the mean for Class 1 is $\mu_{21} + \hat{b}_0 = 0 - 1.04 = -1.04$ (while μ_{11} and μ_{21} are fixed at 0 for model identification).

This result tells us that the overall proficiency levels for Class 1 adolescents (formal-operational stage) are lower than Class 2 adolescents (concrete-operational stage) in both dimensions.² This result is somewhat counterintuitive because one would expect that adolescents at the higher developmental stage (Class 1) would show higher proficiency levels. This phenomenon may indicate that adolescents in the formal-operational stage (or Class

²Note that the means of the two dimensions (or latent-traits) are fixed at 0 for the reference latent class. In addition, three item groups (MT, AB, CF) serve as anchor items as their effects (on the probability of a correct response) are set to be equal across latent classes (in both dimensions). These settings enable us to compare latent trait means across latent classes.

1) overgeneralize MP/MT items (which are easier than NA/AC items) and tend to get those items wrong. Similar phenomena have been reported in the literature (e.g., Markovits, Fleury, Quinn, & Venet, 1998; Draney 2007).

Third, the estimated standard deviations for the latent traits are somewhat greater for Class 1 adolescents than Class 2 adolescents. Specifically, in the NE dimension, $\hat{\sigma}_{111} = 1.16$ for Class 1 and $\hat{\sigma}_{112} = 0.59$ for Class 2, while in the NN dimension, $\hat{\sigma}_{221} = 0.88$ for Class 1 and $\hat{\sigma}_{222} = 0.53$ for Class 2. From the Saltus model, the Class 1 standard deviation is 0.96 (0.34) and the Class 2 standard deviation is 0.29 (0.08). This result implies that the proficiency level of Class 1 adolescents (formal-operational stage) are more variable than Class 2 adolescents (concrete-operational stage). Furthermore, the estimated correlation between the two latent traits ($\hat{\rho}_g = \frac{\hat{\sigma}_{12g}}{\hat{\sigma}_{11g}\hat{\sigma}_{22g}}$, for Class g) is somewhat higher in Class 1 ($\hat{\rho}_1 = 0.92$) than in Class 2 ($\hat{\rho}_2 = 0.70$). This suggests that the performance is more consistent across the two dimensions for Class 1 adolescents than Class 2 adolescents. In other words, adolescents' performance in the formal-operational stage is less influenced by whether or not items are presented with negation than adolescents in the concrete-operational stage. To our knowledge, this is a new finding regarding the formal-operational stage adolescents' performance in deductive reasoning.

For a visual illustration, we compare the distributions of the estimated proficiency scores (posterior means) of the two latent traits for adolescents in Class 1 and Class 2. See Fig. 2. This figure confirms our earlier findings: (1) Class 2 participants tend to show higher proficiency levels than

Fig. 2 A bivariate mixture distribution of the proficiency levels in the NE (Dim1) and NN (Dim2) dimensions for Class 1 and Class 2

Class 1 participants in both dimensions, (2) participants' proficiency levels in Class 1 show a larger range (or is more variable) than Class 2 in both dimensions (Class 1 distribution ranges from -2 to 3, while Class 2 ranges from 1 to 3), and (3) the distribution for Class 1 is centered around 0, while the distribution for Class 2 is centered around 1.5 (class means).

Fourth, the MT items ($\hat{b}_3 = 0.85$) are on average more difficult than the NA items ($\hat{b}_1 = -0.31$) and the AC items ($\hat{b}_2 = 0.23$). In addition, the CF items ($\hat{b}_5 = 0.67$) are on average slightly more difficult than the AB items ($\hat{b}_4 = 0.55$).

Here, $\hat{b}_1$ and $\hat{b}_2$ represent the average difficulty levels of the NA and AC items only for Class 1 adolescents. Hence, the result suggests that for Class 1 adolescents the MT items are more difficult than the NA, AC, AB, and CF items. The average difficulty levels of the NA and AC items are modified for Class 2 adolescents, however, which can be computed by adding $-\hat{\tau}_{hgk}$ for the NE dimension and the NN dimension to the $\hat{b}_1$ and $\hat{b}_2$ estimates. That is, in the NE dimension, the NA items have the average difficulty level $\hat{b}_1 + (-\hat{\tau}_{221}) = 4.20$ and the AC items have the average difficulty level $\hat{b}_2 + (-\hat{\tau}_{221}) = 4.74$. In the NN dimension, the NA items have the average difficulty level $\hat{b}_1 + (-\hat{\tau}_{222}) = 3.90$ and the AC items have the average difficulty level $\hat{b}_2 + (-\hat{\tau}_{222}) = 4.44$.

These results allude to interesting relationships between DRV test design factors as well as between latent classes. To see this more clearly, Fig. 3 is provided.

Figure 3 shows that (1) for the students in the formal-operational stage (Class 1), the MT items are more difficult than the NA/AC items, whereas for the students in the

Fig. 3 Difficulty levels of MT, NA, and AC item groups in the NE and NN dimensions for Class 1 and Class 2

concrete-operational stage (Class 2), the NA/AC items are more difficult than the MT items; and (2) for the students in the concrete-operational stage (Class 2), the NA/AC items are more difficult when they are presented with negation (NE) than without negation (NN), while the presentation with negation does not make a difference in the difficulty level of the NA/AC items for the students in the formal-operational stage (Class 1). These are intriguing findings that suggest potential across-latent-class and between-design-factor interactions for the DRV assessment.

Fig. 4 Frequency distributions of the sum scores for the observed and replicated data for the NE (Dim1) and NN (Dim2) dimensions

Model validation

Model goodness-of-fit

To further validate the model utilized in this study, we assessed the absolute goodness-of-fit of the model using posterior predicted checking (e.g., Gelman, Carlin, Stern, & Rubin, 2004; Sinharay, Johnson, & Stern, 2006).

To apply the method, we first generated $D = 100$ replicated datasets from the posterior, and calculated the posterior predictive score distribution as the median of the generated frequency distributions over replicates. We then compared the posterior predictive score distribution with a frequency distribution of the observed sum scores. If the model fits the data, the two distributions should be similar. In addition, we computed a posterior predictive p value using the sum of the squared Bayesian residual (De Jong & Steenkamp, 2010). Specifically, the p value was calculated as the proportion of draws in which the sum of the squared Bayesian residual based on replicated data exceeded the realized value based on the observed data. A p value near 0.5 would indicate a good fit, whereas more extreme values (close to 0 or 1) indicate a poor fit (Li, Cohen, Kim, & Cho, 2009).

Figure 4 shows the posterior predictive sum score distributions compared with the observed sum score distribution for each dimension. The predictive sum score patterns from the model appear to emulate the observed sum score patterns quite well in both dimensions, implying that the model fits the data well. In addition, the calculated p values are close to 0.5 for all items, indicating that the replicated data generated under the model are not significantly different from the observed data, suggesting an adequate model fit.

Simulation study

We also conducted a simulation study to verify that the estimated parameter values were recoverable. Assuming the test setting same as in the empirical study ($N = 418$, $I = 24$) and using the estimated parameter values (reported in Table 2) as the data generating values, we generated 100 datasets. We then fit the presented model to each simulated dataset by using the same estimation setting as in the empirical study. Table 3 summarizes the results.

The results suggest that for all model parameters, the average values of the posterior mean estimates are quite close to the data generating values. In addition, we find that for most parameters the bias is not significantly different from zero at the 5% level except the three variance and covariance parameters, σ_{221} ($t = 6.39$, $p < 0.01$), σ_{112} ($t = 4.09$, $p < 0.01$), and σ_{122} ($t = 4.47$, $p < 0.01$).

Table 3 Parameter recovery of the proposed model over 100 replications

	Par	True	$\bar{Est}$	$\overline{SD(Est)}$
Structural parameters				
Dim1: NE	τ_{221}	−4.51	−4.50	0.20
Dim2: NN	τ_{222}	−4.21	−4.25	0.20
Item parameters				
	b_0	−1.04	−1.10	0.49
	$b_1(NA)$	−0.31	−0.36	0.57
	$b_2(AC)$	0.23	0.24	0.57
	$b_3(MT)$	0.85	0.82	0.56
	$b_4(AB)$	0.55	0.60	0.50
	$b_5(CF)$	0.67	0.76	0.49
	σ_ϵ	0.98	1.00	0.10
Class proportion				
Class 1	π_1	0.51	0.50	0.02
Class 2	π_2	0.49	0.49	0.02
Means (Class 2)				
Dim1: NE	μ_{12}	1.56	1.56	0.15
Dim2: NN	μ_{22}	1.51	1.53	0.14
Covariance (Class 1)				
SD (NE)	σ_{111}	1.16	1.18	0.08
SD (NN)	σ_{221}	0.82	0.86	0.07
Cov	σ_{121}	0.88	0.88	0.12
Covariance (Class 2)				
SD (NE)	σ_{112}	0.59	0.62	0.08
SD (NN)	σ_{222}	0.53	0.59	0.08
Cov	σ_{122}	0.22	0.19	0.07

True is data-generating true values; *Est* is the posterior mean estimates; $\bar{Est}$ is the mean of the posterior mean estimates; $\overline{SD(Est)}$ is the mean of the posterior standard deviation estimates

Even for those three parameters, the size of the bias appears minor, being less than 0.1.

In addition, we evaluated classification accuracy for the two latent traits (dimensions 1 and 2). Overall, subjects were accurately classified into the expected classes. Specifically, a total of 418 subjects were classified with an average 81% accuracy into class 1 (ranged from 0.72 to 0.94) and into class 2 (ranged from 0.74 to 0.94).

These simulation study results assure that the parameter estimates and subjects' class membership from the empirical study could be well recovered.

Discussion

In this study, we analyzed a complex cognitive assessment that was designed to examine adolescents' developmental stages in deductive reasoning. We proposed a specialized confirmatory mixture IRT model for the following purposes: (1) to measure multiple latent traits, (2) to identify adolescents in different developmental stages, (3) to investigate multivariate latent trait distributions of different developmental stages, (4) to quantify dimension-specific

performance differences between adolescents in different developmental stages, and (5) to examine the difficulty levels of the test design factors per dimension.

The constructed model was successfully estimated with a Bayesian approach. We showed that the overall goodness-of-fit of the model was adequate and the parameter estimates were recoverable. From the analysis, we found that the formal-operational adolescents showed more variable performance levels in both deductive reasoning dimensions (NE, NN) compared with the concrete-operational adolescents, while their levels of performance were more consistent across the NE and NN dimensions than the concrete-operational adolescents, which is a new and interesting finding. The overall proficiency level for the formal-operational adolescents was somewhat lower in both dimensions than the concrete-operational adolescents. This result implies that adolescents in the formal-operational stage might have incorrectly answered easier items (e.g., MP/MT items) by extrapolating too far. This somewhat counterintuitive finding has been consistently reported by other researchers in the literature (e.g., Markovits et al. 1998; Draney 2007).

In addition, we found that the formal-operational and concrete-operational developmental stages were clearly distinguished in both dimensions by adolescents' performance on the logical fallacy (NA/AC) items. The performance differences on the logical fallacy items between the adolescents in the formal-and concrete-operational stages were similar across the NA and NE dimensions. We also found that for the adolescents in the concrete-operational stage the logical fallacy items were more difficult in the NE dimension than in the NN dimensions, meaning that logical fallacy items tended to be more difficult when the items were presented with negation. In addition, the concrete items were easier than the logical fallacy items for the concrete-operational stage adolescents, but the opposite is true for the formal-operational stage adolescents. These interesting findings suggest potential between-stages and between-design-factor interactions for the DRV assessment, which could be a topic of future investigations.

From a modeling perspective, our model can be seen as a multidimensional extension of the Saltus model with item covariates and a random error. Such an extension has not been made for the Saltus model. Our larger contribution is to present a special application of confirmatory mixture IRT modeling to the measurement field. Further, we provided interesting substantive results as summarized above, which could be of interest to deductive reasoning researchers.

Lastly, although the construction of our model was motivated by a particular cognitive test, the proposed model is applicable to a variety of situations. Generally, our model can be applied to scenarios where a researcher would like to test a hypothesis on structural differences between latent classes with multidimensional assessments. For instance, one may consider utilizing the proposed model to analyze data from a large-scale assessment, such as the Trends in International Mathematics and Science Study (TIMSS). The TIMSS mathematics items tap into three cognitive domains and three content domains. Cognitive domains (knowing, applying, and reasoning) specify the types of thinking processes required to solve the items. Content domains (number, geometric shapes and measures, and data display) specify the subject matters that are assessed. These content and cognitive domains can be used as test design factors for the item regression, while the content domains are used to define test dimensionality and cognitive areas to define item groups. In this case, the proposed model can be utilized to differentiate students who may need extra support (from 'proficient' students) as well as to evaluate whether the designed test items (e.g., applying and reasoning) function effectively as expected in terms of differentiating a potentially disadvantaged group of students. In future studies, we will further investigate these and other possible applications of the proposed model.

Author Note We would like to acknowledge that the initial work of the current study has been published in the International Meeting of Psychometric Society (IMPS) conference proceedings (Jeon, Draney, & Wilson, 2015).

References

- Bond, T. G. (1995a). Piaget and measurement II: Empirical validation of the Piagetian model. *Archives de Psychologie*, 63, 155–185.
- Bond, T. G. (1995b). Piaget and measurement I: The Twain really do meet. *Archives de Psychologie*, 63, 71–87.
- Bond, T. G., & Bunting, E. M. (1995). Piaget and measurement III: Reassessing the methode clinique. *Archives de Psychologie*, 63, 231–255.
- Boughton, K. A., & Yamamoto, K. (2007). A HYBRID model for test speededness. In D. von, & C. Carstensen (Eds.) *Multivariate and mixture distribution Rasch models*, (pp. 147–156). New York: Springer.
- Chazan, D. (1993). High school geometry students' justification for their views of empirical evidence and mathematical proof. *Educational Studies in Mathematics*, 24, 359–387.
- Cho, S.-J., Cohen, A., Kim, S.-H., & Bottge, B. (2010). Latent transition analysis with a mixture IRT measurement model. *Applied Psychological Measurement*, 34, 583–604.
- Cho, S.-J., Cohen, A. S., & Bottge, B. A. (2013). Detecting intervention effects using a multilevel latent transition analysis with a mixture IRT model. *Psychometrika*, 78, 576–600.
- Cho, S.-J., Cohen, A. S., & Kim, S.-H. (2014). A mixture group bifactor model for binary responses. *Structural Equation Modeling: A Multidisciplinary Journal*, 21, 375–39.
- Choi, I.-H., & Wilson, M. (2015). Multidimensional classification of examinees using the mixture random weights linear logistic test model. *Educational and Psychological Measurement*, 75, 78–101.
- Commons, M. L., Trudeau, E. J., Stein, S. A., Richards, F. A., & Krause, S. R. (1998). Hierarchical complexity of tasks shows the existence of developmental stages. *Developmental Review*, 18, 237–278.
- De Boeck, P., Cho, S.-J., & Wilson, M. (2011). Explanatory secondary dimension modelling of latent DIF. *Applied Psychological Measurement*, 35, 583–603.
- De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.
- De Jong, M. G., & Steenkamp, J.-B. E. M. (2010). Finite mixture multilevel multidimensional ordinal IRT models for large-scale cross-cultural research. *Psychometrika*, 75, 3–32. <https://doi.org/10.1007/s11336-009-9134-z>
- Demetriou, A., & Efklides, A. (1989). The persons' conception of the structures of developing intellect: Early adolescence to middle age. *Genetic, Social, and General Psychology Monographs*, 115(371), 423.
- Demetriou, A., & Efklides, A. (1994). Structure, development, and dynamics of mind: A meta-Piagetian theory. In A. Demetriou, & A. Efklides (Eds.) *Intelligence, mind, and reasoning: Structure and development*. Amsterdam, The Netherlands: North-Holland, (pp. 75–109).
- Draney, K. (2007). The Saltus model applied to proportional reasoning data. *Journal of Applied Measurement*, 8, 438–455.
- Draney, K., & Wilson, M. (2008). A LLTM approach to the examination of teachers' ratings of classroom assessment tasks. *Psychology Science*, 50, 417–432.
- Draney, K., Wilson, M., Gluck, J., & Spiel, C. (2007). Mixture models in a developmental context. In R. Hancock, & K. M. Samuelson (Eds.) *Latent variable mixture models* (p. 199–216) Charlotte, NC: Information Age.

- Finch, W. H., & Finch, M. E. H. (2013). Investigation of specific learning disability and testing accommodations based differential item functioning using a multilevel multidimensional mixture item response theory model. *Educational and Psychological Measurement*, 73, 973–993.
- Fischer, G. (1973). Linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37, 359–374.
- Fischer, K. W., Hand, H. H., & Russel, S. (1984). The development of abstractions in adolescence and adulthood. In M. L. Commons, F. A. Richards, & C. Armon (Eds.) *Beyond formal operations: Late adolescent and adult cognitive development*, (pp. 43–73). New York: Praeger.
- Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). *Bayesian data analysis*. New York: Chapman and Hall.
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457–472. <https://doi.org/10.1214/ss/1177011136>
- Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.) *Bayesian statistics*, (pp. 169–193). Oxford: Clarendon Press.
- Huang, H.-Y. (2016). Mixture IRT model with a higher-order structure for latent traits. *Educational and Psychological Measurement*, 77, 275–304.
- Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. D. Boeck, & M. Wilson (Eds.) *Explanatory item response models*, (pp. 198–212). New York: Springer.
- Jeon, M. (2018). A constrained confirmatory mixture IRT model: Extensions and estimation of the Saltus model using mplus. *The Quantitative Methods for Psychology*, 14, 120–136.
- Jeon, M., Draney, K., & Wilson, M. (2015). Multidimensional Saltus linear logistic test model for modeling children's cognitive development. In R. Millsap, D. Bolt, L. van der Ark, & W.-C. Wang (Eds.) *Quantitative psychology research: Proceedings of the 78th annual meeting of the Psychometric Society*, (p. 73). New York: Springer.
- Jin, K.-Y., Chen, H.-F., & Wang, W.-C. (in press). Mixture item response models for inattentive responding behavior. *Organizational Research Methods*.
- Li, F., Cohen, A., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement*, 33, 353–373. <https://doi.org/10.1177/0146621608326422>.
- Markovits, H., Fleury, M.-L., Quinn, S., & Venet, M. (1998). The development of conditional reasoning and the structure of semantic memory. *Child Development*, 69, 742–755. <https://doi.org/10.2307/1132201>
- Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, 55, 195–215. <https://doi.org/10.1007/BF02295283>
- Mislevy, R. J., & Wilson, M. (1996). Marginal maximum likelihood estimation for a psychometric model of discontinuous development. *Psychometrika*, 61, 41–71. <https://doi.org/10.1007/BF02296958>.
- Molenaar, D., Oberski, D., Vermunt, J., & De Boeck, P. (2016). Hidden Markov IRT models for responses and response times. *Multivariate Behavioral Research*, 51, 606–626.
- Neimark, E. D. (1975). Longitudinal development of formal operations thought. *Genetic Psychology Monographs*, 91, 171–225.
- Overton, W. F. (1985). Scientific methodologies and the competence-moderator performance issue. In E. D. Neimark, R. de Lisi, & J. L. Newman (Eds.) *Moderators of competence*, (pp. 15–41). Hillsdale: Erlbaum.
- Piaget, J. (1971). *Biology and knowledge*. Chicago: University of Chicago Press.
- Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing (DSC 2003)*, Vienna, Austria.
- Plummer, M. (2011). JAGS version 3.1.0 user manual. (<http://sourceforge.net/projects/mcmc-jags/files/>)
- Rips, L. J. (1994). *The psychology of proof*, MIT Press, Cambridge.
- Roberge, J. J., & Mason, E. J. (1978). Effects of negation on adolescents' class and conditional reasoning abilities. *The Journal of General Psychology*, 98, 187–195. <https://doi.org/10.1080/00221309.1978.9920872>.
- Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, 34, 213–232.
- Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, 30, 198–321. <https://doi.org/10.1177/0146621605285517>.
- Spiel, C., Gittler, G., Sirsch, U., & Gluck, J. (1997). Application of the Rasch model for testing Piaget's theory of cognitive development. In R. Langeheine, & J. Rost (Eds.) *Applications of latent trait and latent class models in the social sciences*, (pp. 111–117). Germany: Waxmann.
- Spiel, C., & Gluck, J. (2008). A model based test of competence profile and competence level in deductive reasoning. In J. Hartig, E. Klieme, & D. Leutner (Eds.) *Assessment of competencies in educational contexts: State of the art and future prospects*, (pp. 41–60). Gottingen: Hogrefe.
- Spiel, C., Gluck, J., & Gossler, H. (2001). Stability and change of unidimensionality: The sample case of deductive reasoning. *Journal of Adolescent Research*, 16, 150–168. <https://doi.org/10.1177/0743558401162003>.
- Staudenmeyer, H., & Bourne, L. E. (1972). Learning to interpret conditional sentences: A developmental study. *Psychology*, 13, 616–623.
- Tijmstra, J., Bolsinova, M. A., & Jeon, M. (in press). Generalized mixture IRT models with different item-response structures: A case study using Likert-scale data. *Behavior Research Methods*.
- van Hiele, P. M. (1986). *Structure and insight: A theory of mathematics education*. Orlando: Academic Press.
- von Davier, M., Xu, X., & Carstensen, C. H. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika*, 76, 318–336.
- Wildman, T. M., & Fletcher, H. J. (1977). Developmental increases and decreases in solutions of conditional syllogism problems. *Developmental Psychology*, 13, 630–636.
- Wilson, M. R. (1989). Saltus: a psychometric model of discontinuity in cognitive development. *Psychological Bulletin*, 105, 276–289. <https://doi.org/10.1037/0033-2909.105.2.276>
- Yamamoto, K., & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost, & R. Langeheine (Eds.) *Applications of latent trait and latent class models in the social sciences*, (pp. 89–98). New York: Waxmann Verlag GmbH.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.