Behaviormetrika (2018) 45:457–474
https://doi.org/10.1007/s41237-018-0070-z

ORIGINAL PAPER

Learning meets assessment

On the relation between item response theory and Bayesian knowledge tracing

Benjamin Deonovic[1] · Michael Yudelson[1] · Maria Bolsinova[1] · Meirav Attali[1] · Gunter Maris[1,2]

Received: 28 February 2018 / Accepted: 11 October 2018 / Published online: 20 October 2018

Abstract
Few models have been more ubiquitous in their respective fields than Bayesian knowledge tracing and item response theory. Both these models were developed to analyze data on learners. However, the study designs that these models are designed for differ; Bayesian knowledge tracing is designed to analyze longitudinal data while item response theory is built for cross-sectional data. This paper illustrates a fundamental connection between these two models. Specifically, the stationary distribution of the latent variable and the observed response variable in Bayesian knowledge Tracing are related to an item response theory model. This connection between these two models highlights a key missing component: the role of education in these models. A research agenda is outlined which answers how to move forward with modeling learner data.

Keywords Item response theory · Bayesian knowledge tracing · Network psychometrics

Ronny Scherer and Marie Wiberg.

✉ Benjamin Deonovic
benjamin.deonovic@act.org

Michael Yudelson
michael.yudelson@act.org

Maria Bolsinova
maria.bolsinova@act.org

Meirav Attali
meirav.attali@act.org

Gunter Maris
gunter.maris@act.org

[1] ACTNext, 500 ACT dr, Iowa City, IA 52243, USA

[2] University of Amsterdam, Amsterdam, The Netherlands

1 Introduction

Learning and assessment deal with related, yet distinct concepts. Learning can be defined as the acquisition of knowledge, skills, values, beliefs, and habits through experience, study, or instruction. Assessments are instruments designed to observe behavior in a learner and produce data that can be used to draw inference about the knowledge, skills, values, beliefs, and habits that the learner has. Although learning and assessment are both key to education, the statistical models used to describe learning data and assessment data have significantly diverged and grown to leverage the salient features and distinct assumptions that are embodied in their respective data sets. The fields of educational data mining and learning analytics harness the dynamic, temporal, and large-scale nature of learning data to construct models which can be used to predict learner performance, personalize and adapt instructional content, recommend intervention and curriculum changes, and provide information visualization to track progress. On the other hand, the same objectives are targeted by the field of psychometrics, using cross-sectional assessment data rather than longitudinal data. Specifically, this paper will explore the connections between Bayesian knowledge Tracing (BKT) and item response theory (IRT). BKT, a statistical model in educational data mining, is the most ubiquitous model used for data obtained from intelligent tutoring systems, which are systems constructed to provide immediate and customized instruction to learners. IRT, a modeling framework developed in the field of psychometrics, was designed for constructing and analyzing assessments.

Historically, the research in BKT and IRT models has had little overlap, as on the surface these models seem to be completely different and incompatible. Both these models fit their respective data sets well, but each has flaws. Due to the relationship between the longitudinal learning data and the cross-sectional assessment data, we posit that there exists a relationship between BKT and IRT models. Indeed we will show that there is an intimate connection between these two models that places BKT and IRT under an umbrella of general models of learning and assessment data. First in Sect. 2, the BKT model is explained in detail and extensions to the standard model that have been described in the literature are also listed. Section 3 describes the IRT model and its extensions. Section 4 discusses the shortcomings of the respective models in the context of learning. Section 5 describes the connection between the BKT models and IRT models and how the shortcomings of each model can be addressed by incorporating concepts from the other.

It should be noted that this paper does not describe a novel statistical model nor an algorithm to fit a statistical model to either the longitudinal learning data or cross-sectional assessment data. Rather this paper identifies a key theoretical connection between two existing and popular models. However, this result is not inconsequential. The connection between BKT and IRT highlights that there is a crucial ingredient missing from both. That crucial ingredient is education. Only when learning, assessment, and education go hand in hand can there be hope to make progress. Hence in Sect. 7, we end this paper with sketching a research agenda for achieving this.

2 BKT

Bayesian Knowledge Tracing or BKT (Corbett and Anderson 1995) is a modeling paradigm frequently used in the field of Intelligent Tutoring Systems (ITS) where it is tasked with continuously tracking the process of student knowledge acquisition and serves as the basis for selecting the next problem set or skill that a student should work on, once mastery has been attained on the current problem set or skill. In BKT, skills are modeled as (latent) binary variables (mastered/not-mastered) and learning is characterized as a transition between these states. Let Z_{pkt} denote the dichotomous state (i.e., mastery/not-mastery) of the kth skill for the pth person at attempt t for $k = 1, \ldots, K$, $p = 1, \ldots, n$ and $t = 1, \ldots, T_p$. Originally, BKT was developed with cognitive theory of learning in mind. Each model addresses one skill and assumes that it is relatively fine grained (e.g., addition, subtraction, division) (Corbett and Anderson 1995). Granularity of the skills is a subject of experimental research and, for example, addition could be split to single-digit addition and multi-digit addition if the data indicate that the split is warranted (Koedinger et al. 2013).

The standard BKT model utilizes five global (i.e., shared among all individuals in the model) parameters per skill $\boldsymbol{\pi}_k = (\pi_{0k}, \pi_{\ell k}, \pi_{\phi k}, \pi_{sk}, \pi_{gk})^{\mathbf{T}}$. The probability of learning skill k, after completing an item utilizing that skill, is denoted by $\pi_{\ell k}$, the probability of forgetting skill k once learned is denoted by $\pi_{\phi k}$, the probability of an incorrect answer on an item when skill k is mastered (a slip) is denoted by π_{sk}, the probability of a correct answer when skill k is unmastered (a guess) is denoted by π_{gk}, and π_{0k} is the probability that a learner is in the mastery state for that skill before beginning the assessment at hand. Note for the standard model these parameters are not individual specific, each skill is assumed to be independent of the other skills, and once a skill is learned it cannot be forgotten (i.e., $\pi_{\phi k} = 0$). The BKT model is equivalent to a two-state hidden Markov model (HMM) (see Fig. 1), where Z_{pkt} represents the hidden or latent state, with a dichotomous emission X_{pkt}. The notation and description used in the original Corbett and Anderson (1995) paper are listed in Table 1.

The standard BKT model only utilizes longitudinal performance data and does not permit features such as student-specific or item-specific parameters. A plethora of extensions to BKT has been developed, including variants that measure the effect of students' individual characteristics (Pardos and Heffernan 2011; Lee and Brunskill 2012; Yudelson et al. 2013; Khajah et al. 2014a, b), assessing the effect of help in a tutor system (Beck et al. 2008; Sao Pedro et al. 2013), controlling for item difficulty (Gowda et al. 2011; Pardos and Heffernan 2011; Schultz and Tabor 2013), measuring impact of time between attempts (Qiu et al. 2010), incorporating forgetting (Nedungadi et al. 2015; Khajah et al. 2016), and measuring the effect of subskills (Xu and Mostow 2010).

Table 1 The learning and performance parameters as described in the original Corbett and Anderson (1995) paper along with the original notation

π_{0k}	$p(L_0)$	Initial learning	The probability a skill is in the learned state prior to the first opportunity to apply the skill
$\pi_{\ell k}$	$p(T)$	Acquisition	The probability a skill will make the transition from the unlearned to the learned state following an opportunity to apply the skill
π_{gk}	$p(G)$	Guess	The probability a student will guess correctly if a skill is in the unlearned state
π_{sk}	$p(S)$	Slip	The probability a student will slip (make a mistake) if a skill is in the learned state
$\pi_{\phi k}$		Forgetting	The probability a skill will make the transition from the learned to the unlearned state following an opportunity to apply the skill

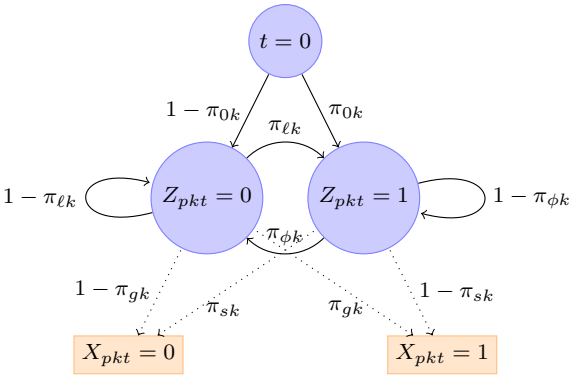

Fig. 1 The transition and emission probabilities of the standard BKT model with forgetting. It is equivalent to a two-state hidden Markov model, where Z_{pkt} represents the hidden or latent state for the kth skill of person p at time t, with a dichotomous emission X_{pkt} representing the observed response (correct/incorrect) for the pth person of the kth skill at attempt t

3 IRT

Statistical models utilizing IRT have played an extensive role in assessment and educational measurement. The history of IRT can be traced back to pioneering work by Louis Thurstone in the 1920s, and seminal work by Bert Green, Alan Birnbaum, Frederic Lord, and Georg Rasch in the 1950s and 60s (Green 1950; Lord 1951; Birnbaum 1967; Rasch 1960). The IRT model consists of three basic assumptions: (1) the probability a person correctly answers an item follows a specific parameteric functional form called the item characteristic curve (ICC) or item response function (IRF), which depends on parameter(s) for that person and parameter(s) for the item; (2) this IRF is monotonically increasing function with respect to a person's ability; and (3) given the person's ability, the items are considered conditionally independent. One specific IRT model is the four-parameter logistic (4PL) model which models the probability that individual p answers item i correctly by the logistic function (Fig. 2)

$$P(Y_{pi} = 1|\theta_p, a_i, b_i, c_i, d_i) = c_i + (d_i - c_i)\frac{\exp\left[a_i(\theta_p - b_i)\right]}{1 + \exp\left[a_i(\theta_p - b_i)\right]}, \tag{1}$$

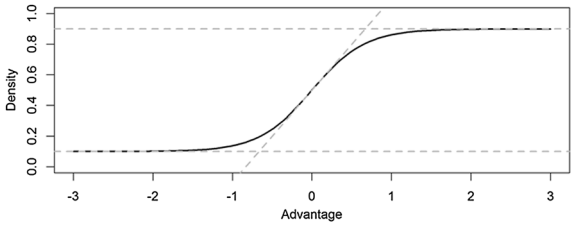

Fig. 2 Item response function (IRF) of the 4PL model. The horizontal axis is the advantage, $\theta_p - b_i$. The IRF is the black solid line, the upper gray dashed line is the upper asymptote d_i, the lower gray dashed line is the lower asymptote c_i, and the last gray dashed line is the maximum slope of the IRF which is $a_i(d_i - c_i)/4$

where Y_{pi} is the response of person p on item i, θ_p represents the person's ability, b_i represents an item's difficulty, a_i represents an item's discrimination, c_i represents guessing, and d_i represents inattention (slips). Subsets of the 4PL include the 3PL ($d_i = 1$), 2PL ($d_i = 1, c_i = 0$), and 1PL ($d_i = 1, c_i = 0, a_i = 1$) which is often refereed to as the Rasch model.

Originally, the first IRT models that were developed treated ability as a static, unidimensional parameter. This framework made IRT especially well suited for ranking a set of individuals by their ability, hence its use in assessment, particularly summative assessments. However, these assumptions make IRT inappropriate for analyzing data that are not cross-sectional in nature, such as data collected by continuous assessments, or sometimes formative assessments. Historically the field of psychometrics has been mostly concerned with the analysis of cross-sectional data, such as summative assessments, while the fields of educational data mining and learning analytics have focused on longitudinal data such as data produced by continuous and formative assessments.

The research on IRT is extensive and thorough. Countless extensions and generalizations of the 4PL and other IRT models have been developed, too many to list and cite here. For a review of IRT and its extensions see van der Linden (2016–2018). Notably, for the context of this paper, three extensions of IRT which make it more amenable to learning data are briefly described: adaptive item administration, multidimensionality of ability, and time-varying ability.

First, traditional assessments consist of fixed forms comprising a set of items. This may not be appropriate in a learning context in which the strengths and weakness of different learners can be harnessed to present the learner with more appropriate items for their skill level. One way to administer an adaptive set of items, rather than a fixed form, proposed by Lord (1977), is to administer an item which would maximize the Fisher information given the current estimate of a learner's ability. Selecting the item which maximizes the Fisher information is an efficient way to select an item as maximizing the Fisher information is equivalent to minimizing the lower bound of the variance of the ability estimate. This method serves the goal of testing, which is to efficiently estimate an individuals ability and, however, requires a large pool of items.

Although adaptive item administration is a key feature of a learning system, models which utilize a unidimensional representation of ability, such as the standard IRT model, are incapable of inferring what aspects of the material the individual has mastered or not mastered. Multidimensional IRT (MIRT) extends IRT to allow for a multidimensional ability parameter (Reckase 1972). In the multidimensional 4PL (MD-4PL), the ability parameter and the discrimination parameter are expanded to be k dimensional $\boldsymbol{\theta}_p = (\theta_{p1}, \ldots, \theta_{pk})^{\mathbf{T}}$ and $\mathbf{a}_i = (a_{i1}, \ldots, a_{ik})^{\mathbf{T}}$ with the following IRF:

$$P(Y_{pi} = 1 | \boldsymbol{\theta}_p, \mathbf{a}_i, \beta_i, c_i, d_i) = c_i + (d_i - c_i) \frac{\exp(\mathbf{a}_i^{\mathbf{T}} \boldsymbol{\theta}_p + \beta_i)}{1 + \exp(\mathbf{a}_i^{\mathbf{T}} \boldsymbol{\theta}_p + \beta_i)}. \tag{2}$$

In Eq. (2), β_i is labeled an item intercept and is related to the item difficulty. This model is able to estimate a more complex construct rather than just a unidimensional

estimate of a learners ability. It is important to note that this formulation of MIRT is compensatory. This means that a high ability in one dimension of the multidimensional $\boldsymbol{\theta}_p$ can compensate for low ability in the other dimensions. Non-compensatory MIRT has also been defined (Sympson 1978).

Finally, several extensions to traditional IRT have been proposed to allow for ability to be time dependent (to be able to fit longitudinal learning data). Extensions that have been published include extensions of MIRT that allow for longitudinal data (Embretson 1991); state space modeling approaches that have been used to model attitudinal changes (Martin and Quinn 2002) and growth in reading ability (Wang et al. 2013); a deterministic moment-matching method to estimate dynamic ability with real-time continuously streaming data (Weng 2017); and a multidimensional state-space approach (Ekanadham and Karklin 2017). Additionally, two comprehensive theses have been written which introduce several dynamic IRT models (van Rijn 2008; Studer 2012). A simple example of a dynamic IRT model (in the state-space modeling approach) is as follows:

$$\text{System equation:} \quad \theta_{p,t} = \theta_{p,t-1} + \epsilon_{p,t} \tag{3}$$

$$\text{Observation equation:} \quad P(Y_{pit} = 1|\theta_{pt}, b_i) = \frac{\exp(\theta_{pt} - b_i)}{1 + \exp(\theta_{pt} - b_i)}. \tag{4}$$

This is a dynamic extension of the 1PL model, where θ_{pt} represents person p's ability at time t. The various models mentioned in the literature above vary on the specific form of these equations and how the parameters are estimated.

Another extension of IRT that accounts for longitudinal data comes from the learning data modeling literature. These extensions include the Additive Factor Analysis (AFM) (Cen et al. 2006) and the Performance Factor Analysis (PFA) (Pavlik et al. 2009) models. These models extend IRT to longitudional data by dropping the requirement of conditional independence for the same items. Instead the dependence is modeled by linear factors involving the number of attempts on the item along with other factors. These models and more complex versions have been shown (Maclellan et al. 2015) to consistently outperforming BKT across 5 data sets in better prediction via cross-validation.

4 Criticisms of BKT and IRT for learning systems

The core issue with both BKT and IRT is their lack of a placeholder for education in the model. Although the BKT model can estimate the rate at which learning occurs through the parameter $\pi_{\ell k}$ and the IRT model is capable of estimating the learning that has occurred (i.e., the student's faculties) through the ability parameter θ_p, there is no component in either model to denote teaching or education that is occurring to the learners, nor how differences in teaching lead to differences in learning outcomes (IRT), or the learning process (BKT).

The BKT model is basically a ballistic model, where the learning process is closer to firing a cannon, with the path being almost entirely determined by the initial conditions (i.e., parameters), than it is to flying a plane, with a pilot (i.e., education) steering and changing the course of the plane as needed. This is one way in which education interacts with assessment and learning. Education can be seen as setup up the canon (e.g., a system powered by BKT), firing it (i.e., having learners go through the system answering questions and fitting the BKT parameters), seeing where the cannon ball lands (i.e., interpreting and analyzing the resulting BKT parameters), and then reconfiguring the system and doing it all over to optimize some criterion. We believe that this process should be made more holistic, with the effects of education incorporated into the model.

IRT models on the other hand are inherently cross-sectional, and the aim to explain observed differences in what has been learned. Such models, however, have little to offer in explaining how these observed differences came into existence, or what measures could reduce or alter them.

Furthermore, there is a requirement that the skills in the BKT model (and assignment/tagging of particular items to skills) need to be done *apriori*. One could argue that this process serves as a sort of placeholder for education to some extent. However, this aspect is quite removed from the model itself and may be arbitrary. It is possible to forego this issue by considering each unique set of problems to be its own "skill". This can then be used as a proxy mapping for knowledge components and the BKT model can be fit to the data. This allows one to identify the BKT learning rate parameters without skill tags at the expense of external validity (you cannot describe what the skills are without consulting content experts) and some degree of overfitting (it is unlikely that you have as many unique skills as unique problem sets in most learning environments).

A similar issue arises in MIRT analysis with the $\mathbf{a}_i$ parameters: should it be specified which $\mathbf{a}_i$ are nonzero beforehand (confirmatory MIRT) or should they all be freely estimated with some identifiability constraint (exploratory MIRT)? These skills are also considered to be independent in BKT, which may not be appropriate, and similarly in MIRT the components are often estimated to be orthogonal to each other but then rotated to obtain some kind of interpretable result (Fig. 3).

5 Connecting BKT and IRT

In this section, we develop a unified framework encompassing both the standard BKT model and the IRT family of models. To motivate the construction of the unified framework, we demonstrate for the BKT model that at equilibrium the distribution of the latent mastery variable and the distribution for a correct response both follow an IRT model.

Let the transition and emission matrices of the BKT model be denoted by $\mathbf{A}$ and $\mathbf{B}$, respectively

$$\mathbf{A} = \begin{pmatrix} 1-\pi_{\phi k} & \pi_{\phi k} \\ \pi_{\ell k} & 1-\pi_{\ell k} \end{pmatrix} \qquad \mathbf{B} = \begin{pmatrix} 1-\pi_{gk} & \pi_{gk} \\ \pi_{sk} & 1-\pi_{sk} \end{pmatrix}. \tag{5}$$

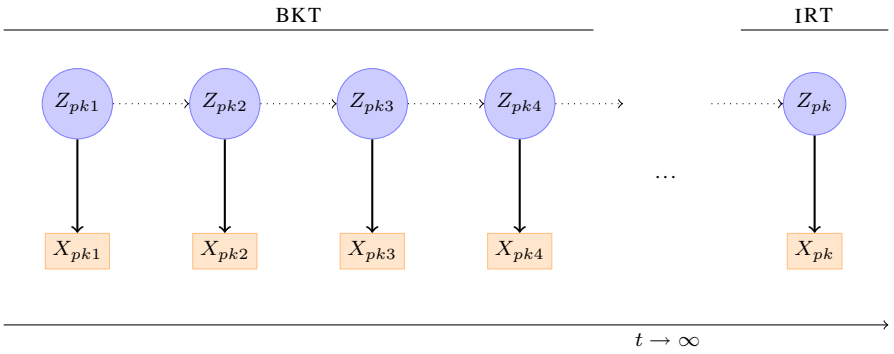

Fig. 3 The left hand side illustrates the HMM for the BKT data for a specific skill mastery of a single individual. The right hand side illustrates the equilibrium distribution of the latent mastery variable which can be shown to follow an IRT distribution. The fact that IRT is placed at the end of the BKT's state-transition chain serves to indicate that in the limit the distribution of the latent skill in BKT converges to IRT. Also, traditionally, summative tests are often given after learning has occurred. However, note that there is no requirement that precludes administering such a test before or during the targeted period of learning

Let $\{Z_{pkt}\}_{t=1,\ldots,T_p}$ denote the Markov chain formed by the transitions of the latent state variable in the BKT model. This Markov chain has a finite state space and is irreducible meaning that it is possible to get to any state from any state (as long as $\pi_{\phi k}$ and $\pi_{\ell k}$ are not equal to zero). Furthermore, this Markov chain is aperiodic, meaning every state can be returned to at every time point. Therefore, the stationary distribution for this Markov chain, $\boldsymbol{\lambda}^{\mathbf{T}} = (\lambda_0, \lambda_1)$, exists and is the unique solution which satisfies $\boldsymbol{\lambda} = \mathbf{A}^{\mathbf{T}}\boldsymbol{\lambda}$ (Ross 2014) where $\lambda_0 = P(Z_{pk} = 0)$, $\lambda_1 = P(Z_{pk} = 1)$, and Z_{pk} is the latent mastery variable at equilibrium, i.e., a binary random variable for which $Z_{pk} = 1$ indicates person p has mastered skill k. After some simple algebra, this stationary distribution can be shown to be

$$\boldsymbol{\lambda}^{\mathbf{T}} = \left(\frac{\pi_{\phi k}}{\pi_{\ell k} + \pi_{\phi k}}, \frac{\pi_{\ell k}}{\pi_{\ell k} + \pi_{\phi k}} \right). \tag{6}$$

This is similar to a Rasch model, which after reparameterizing $\theta_k = \log \pi_{\ell k}$ and $b_k = \log \pi_{\phi k}$ we get

$$\frac{\pi_{\ell k}}{\pi_{\ell k} + \pi_{\phi k}} = \frac{\exp(\theta_k - b_k)}{1 + \exp(\theta_k - b_k)} \tag{7}$$

$$\frac{\pi_{\phi k}}{\pi_{\ell k} + \pi_{\phi k}} = 1 - \frac{\exp(\theta_k - b_k)}{1 + \exp(\theta_k - b_k)}. \tag{8}$$

This is something similar to the 1PL IRT model for the hidden mastery variable. However, note that instead of a person-specific ability parameter and an item-specific difficulty there is a skill-specific ability and skill-specific difficulty. It must be noted that the identifiability of BKT parameters has been discussed at length (Beck and Chang 2007; Sande 2013; Gweon et al. 2015; Doroudi and Brunskill 2017) and without constraints the parameters are not identifiable, especially if the forgetting parameter is included. This becomes quite evident in Eqs. (7) and (8) in which the parameters θ_k and b_k, which are indexed by the same skill, are not identifiable. For now let us put aside this issue and see this connection through to the end.

Let X_{pk} be the random variable corresponding to a response of person p to an item utilizing skill k when their learning state has reached equilibrium. By adding in the emissions probabilities, we obtain

$$\begin{aligned} P(X_{pk}=1) &= P(X_{pk}=1|Z_{pk}=1)P(Z_{pk}=1) \\ &\quad + P(X_{pk}=1|Z_{pk}=0)P(Z_{pk}=0) \end{aligned} \tag{9}$$

$$= (1-\pi_{sk})\frac{\pi_{\ell k}}{\pi_{\ell k}+\pi_{\phi k}} + \pi_{gk}\frac{\pi_{\phi k}}{\pi_{\ell k}+\pi_{\phi k}} \tag{10}$$

$$= (1-\pi_{sk})\frac{\exp(\theta_k - b_k)}{1+\exp(\theta_k - b_k)} + \pi_{gk}\left(1 - \frac{\exp(\theta_k - b_k)}{1+\exp(\theta_k - b_k)}\right) \tag{11}$$

$$= \pi_{gk} + ((1-\pi_{sk}) - \pi_{gk})\frac{\exp(\theta_k - b_k)}{1+\exp(\theta_k - b_k)}. \tag{12}$$

Thus, at equilibrium the BKT model corresponds to a (4 - 1)PL (four minus one parameter logistic) skill-centric IRT model, i.e., a 4PL model with discrimination parameter set to 1; item-specific difficulty, guessing, and slipping replaced with skill-specific quantities; and individual ability replaced with skill-specific ability. Note that if we further restricted the BKT guess and slip parameters to be 0 and 1, respectively, the resulting equilibrium distribution would be the 1PL IRT model. Any restrictions imposed on the BKT model parameters, for purposes of identifiability for example, that leave the Markov chain formed by the latent variable distributions irreducible and aperiodic will have a corresponding IRT-like model for the equilibrium distribution of the response with analogous restrictions.

So far we have shown how a limiting distribution associated with BKT is related to a type of IRT model. This model, however, is quite strange. The parameters in the model are not all identifiable and the model does not separate between person parameters and item parameters. The strength of IRT models lies in separating the effects of individuals and specific items. What sort of hidden Markov structure would then lead to the 4PL IRT model described above in Sect. 3? We must construct separate HMMs for every learner-item pair. These HMMs are the

same as the ones described for BKT, except that the learner-skill latent mastery variable Z_{pkt} is replaced by a learner-item latent mastery variable W_{pit} and at equilibrium the latent mastery variable Z_{pk} is replaced by W_{pi}. The emission parameter X_{pkt} is replaced with Y_{pit} and at equilibrium X_{pk} is replaced by Y_{pi}. Additionally, the standard BKT parameters $\boldsymbol{\pi}_k = (\pi_{0k}, \pi_{\ell k}, \pi_{\phi k}, \pi_{sk}, \pi_{gk})^{\mathbf{T}}$ must be replaced by $(\pi_{0i}, \pi_{\ell p}, \pi_{\phi i}, \pi_{si}, \pi_{gi})^{\mathbf{T}}$, where π_{0k} is replaced by an item specific initial probability π_{0i}, $\pi_{\ell k}$ is replaced by a learner-specific value $\pi_{\ell p}$, $\pi_{\phi k}$ is replaced by an item-specific forgetting rate $\pi_{\phi i}$, and guessing and slipping values are item rather than skill specific. Thus, at equilibrium of this new HMM we get the following:

$$\begin{aligned} P(Y_{pi} = 1) &= P(Y_{pi} = 1 | W_{pi} = 1) P(W_{pi} = 1) \\ &\quad + P(Y_{pi} = 1 | W_{pi} = 0) P(W_{pi} = 0) \end{aligned} \tag{13}$$

$$= (1 - \pi_{si}) \frac{\pi_{\ell p}}{\pi_{\ell p} + \pi_{\phi i}} + \pi_{gi} \frac{\pi_{\phi i}}{\pi_{\ell p} + \pi_{\phi i}} \tag{14}$$

$$= (1 - \pi_{si}) \frac{\exp(\theta_p - b_i)}{1 + \exp(\theta_p - b_i)} + \pi_{gi} \left(1 - \frac{\exp(\theta_p - b_i)}{1 + \exp(\theta_p - b_i)} \right) \tag{15}$$

$$= \pi_{gi} + ((1 - \pi_{si}) - \pi_{gi}) \frac{\exp(\theta_p - b_i)}{1 + \exp(\theta_p - b_i)}. \tag{16}$$

Equation (16) can be recognized to be almost the 4PL IRT model described in Sect. 3 with discrimination parameter set to 1. Note that since this derivation deals with the equilibrium distribution the initial probability parameter π_{0i} was not involved and thus the actual specification of this parameter (and whether it is skill specific or item specific) is inconsequential.

It should be noted that the above derivations correspond to a "fixed-effects" version of the IRT model. However, the derivations are still valid if instead of a standard HMM we consider a mixed HMM (Altman 2007), where the transition parameters are allowed to be drawn from some distribution. This is turn will result in a "random-effects" IRT model. Furthermore, it should be noted that his stationary distribution is different from that obtained from standard BKT model in which the forgetting parameter is constrained to be zero. Because of this constraint the traditional BKT model will have a stationary distribution that is $P(X_{pk} = 1) = 1 - \pi_{sk}$ since in the long term the learner will always converge to the mastered state. The IRT form of the stationary distribution arises from maintaining a non-zero forget rate and also from characterizing some parameters as person specific and some to be item specific.

This embedding of the two models into a unifying framework highlights their relationship. One of the reasons the BKT model has been so successful for modeling learning data is its requirement of fine-grained skills (Corbett and Anderson 1995). The finer the skills modeled in BKT are, the closer the model comes to approximating an IRT model at equilibrium. Furthermore, this connection explains the success of extensions to BKT which allow for person-specific learning parameters (Pardos and Heffernan 2011; Lee and Brunskill 2012; Yudelson et al. 2013; Khajah et al. 2014b, a).

By formally connecting BKT and IRT models, we obtain a new interpretation of some of the key IRT parameters. Ability is seen to be nothing other than (a function of) the probability to move from the unlearned to the learned state. That is, a person is characterized by his or her *ability to learn*. Item difficulty is seen to be nothing other than (a function of) the (item specific) probability to move from the learned state back to the unlearned one and, hence, modulates how long people stay in the learned state, once they have entered it.

6 Simulation

A small simulation illustrates the concordance between the BKT model and IRT. A total of $n = 1000$ people are simulated with learning rate $\pi_{\ell p}$ and a total of $m = 100$ items are simulated with forgetting rate $\pi_{\phi i}$. The learning and forgetting parameters are drawn from a uniform distribution between 0 and 1 for p in $1, \ldots, n$ and i in $1, \ldots, m$. Ability and difficult are then calculated as $\theta_p = \log \pi_{\ell p}$ and $b_i = \log \pi_{\phi i}$ for p in $1, \ldots, n$ and i in $1, \ldots, m$.

Each person maintains a latent mastery variable for each item, Z_{pit} that starts off as unmastered $Z_{pit} = 0$. The states of this latent variable are then sampled according to its corresponding transition matrix determined by the learning and forgetting rates for a specific number of iterations. This number of iterations is varied and takes on values 2, 5, and 50. Each individual then submits an answer to each item. The probability a correct response is given is $1 - \pi_s$ if they have mastered the item and π_g is they have not mastered it, where $\pi_s = \pi_g = 0.1$ This process is then repeated over 1000 random replications.

Figure 4 shows the results of the simulation. The horizontal axis is the difference between ability and difficulty $\theta_p - b_i$. The vertical axis is the proportion of the 1000 simulations in which the correct answer was submitted. The leftmost plot corresponds to number of HMM iterations equal to 2, the center plot is 5 iterations, and the rightmost plot is after 50 iterations. Each plot has superimposed on the 4PL IRF (the dashed white curve). From the figure, we can see that as the number of iterations increases, the distribution of the latent states approaches that of their equilibrium distribution and thus correspond to the 4PL IRF.

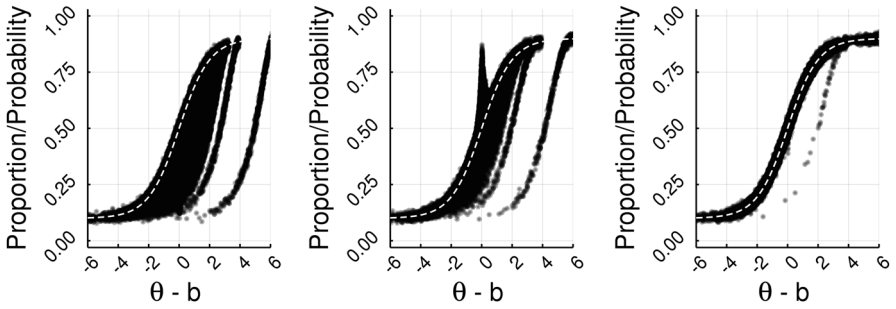

Fig. 4 Simulation results. The horizontal axis is the difference between ability and difficulty $\theta_p - b_i$. The vertical axis is the proportion of the 1000 simulations in which the correct answer was submitted. The leftmost plot corresponds to number of HMM iterations equal to 2, the center plot is 5 iterations, and the rightmost plot is after 50 iterations. The 4PL IRF is the superimposed dashed white line

7 Future directions

So where do we stand, after having explicated the formal relation between BKT and IRT models, with respect to our criticism of both for Learning Systems? Assessment has met learning, but neither has met education thus far. One could argue that no progress has been made. After all, integrating two models, neither one suitable for thinking about education, does not by itself lead to a model suitable for thinking about education.

We argue that the outlook is not so bleak. The integration of BKT and IRT is a point of departure for a new research agenda for the learning sciences and psychometrics *together*; an agenda aimed at factoring the role of education into the learning equation. Both the outcome of learning and the process of learning crucially depend on it. In this section, we outline some key features that we believe models for learning and assessment should have.

Let us start off with a question: Why is it that no educational system starts off in primary education by teaching long division, and then slowly working towards counting? Every systems starts off with counting, followed by addition, and slowly works its way to long division. Counting is a clear prerequisite for addition. That is, even though both learning processes may be adequately described by a BKT-IRT model, the learning processes are not independent. Leveraging such inherent dependencies is what education is all about. In Doignon and Falmagne (1985), a set theoretic approach is taken for describing these dependencies which the authors called knowledge spaces. Although the authors in this paper provide The structures and language to describe these sorts of dependency structures, they do not provide the algorithmic procedures for the assessment of knowledge. Such a modeling framework which is capable of taking these hierarchical dependencies into account is the models that come from the literature on network psychometrics.

Network psychometrics is a new conceptual model that formalizes such (mutual) dependencies. Starting from the mutualism process model of intelligence (Maas et al. 2006), network psychometrics has rapidly expanded and matured over the past decade. Whereas the mutualism model focussed on interactions and dependencies

between unobservable quantities, the recent literature has focused on interactions between observable variables (Borsboom 2008; Borsboom and Cramer 2013; Cramer et al. 2010). The primary innovation in this new conception is the construction of a graphical network in which the nodes are the observable features that are mutually reinforcing based on their connections by causal relations. Recent advances in network psychometrics have highlighted the connection between several graphical models to those of psychometric models (Marsman et al. 2018). Specifically, Marsman et al. (2018) shows a statistical equivalence between the Ising model (a graphical model in statistical physics) and the multidimensional 2PL IRT model.

The necessity of incorporating the dependency structure of skills can be easily seen (we would not want to teach children long division before counting); however, there are other less tangible and more nuanced psychological phenomena and findings that have been replicated over many studies and data sets that our models should explain. These phenomena include positive manifold (i.e., positive correlation between scores of different tests) (Spearman 1904), Matthew effect (Merton 1968) (sometimes summarized by the adage that "the rich get richer and the poor get poorer"), and high predictive validity, the extent to which a score on a scale or test predicts scores on some criterion measure (Cronbach and Meehl 1955). However, neither BKT nor IRT explains why these phenomena occur. It is not necessary in either model, for example, to have high math scores be positively correlated with high English scores. Fortunately, research in network psychometrics has identified how these well replicated phenomena are actually intrinsic to the models proposed by network psychometrics, for example, see Kan et al. (2016); Kovacs and Conway (2016); Savi et al. (2018).

Finally, our model should address how this network of dependencies between skills grows and changes. Learners will learn new skills or the connections between existing skills will change, typically by adding in more connections and dependencies, but this is also where forgetting can be incorporated. The paper by Savi et al. (2018) introduces a network model for intelligence based on the Fortuin–Kasteleyn model. In this paper, a method for studying growth dynamics is also described.

We have outlined some major features that we would wish a future model to exhibit, many stemming from innovations and research done recently in network psychometrics. In light of this, it should be noted that the BKT and IRT model framework is also a subset of these network psychometrics models. It has already been noted above that IRT models such as MIRT have been shown to be related to the Ising model. It is not much of an extension from there to include BKT as well. Figure 5 illustrates this connection. On the right, we have an illustration of an Ising model which is equivalent to an MIRT model. The Ising model portrayed is fully connected. The $\mathbf{a}_i$ parameters of the corresponding MIRT model are related to eigenvalue decomposition of the connectivity matrix Σ of the Ising model, where σ_{ij} is the interaction strength between nodes i and j. The Ising model is difficult to evaluate numerically if there are many nodes in the network and thus it is often simulated using Monte Carlo methods such as the Metropolis algorithm or Glauber dynamics (Newman and Barkema 1999). The structure of the Ising model along with the Monte Carlo sampling scheme used for simulating the Ising model constitutes an interacting particle system. An interacting particle system

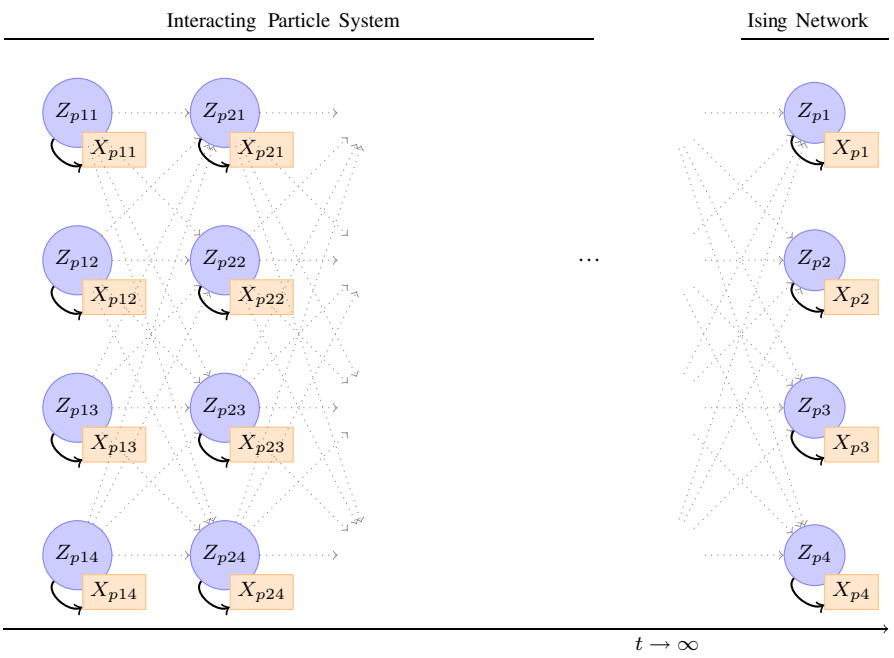

Fig. 5 An interacting particle system in which Z_{pkt} and X_{pkt} are the same as in Fig. 1 but are allowed to interact with each other rather than be separate, independent models, thus producing a hidden Markov field model

is defined by a graphical network, along with the transition probabilities for each node (Liggett 2012). Figure 5 shows an extension of the ideas presented in Fig. 3. Whereas Fig. 3 illustrates a single latent variable evolving over time to its stationary distribution, Fig. 5 depicts an interconnected network of multiple variables that evolve over time to a stationary distribution that is an MIRT model (or equivalently, an Ising model).

This generalization resolves two issues that were described with BKT and IRT as they relate to learning systems. First, there is now a placeholder for education, namely the connections between the various observable nodes and the associated strength of interaction parameters in the Ising model. Learning/education is depicted as the addition of connections between these nodes. Second, the independence of skills is not required as in BKT, but rather the dependencies between latent skills are implied by the causal relationships of the observable nodes. Whereas in IRT, the items are assumed to be conditionally independent given the latent trait, the Ising model does not require this. Indeed, the items in the Ising model exhibit dependence and this dependence is not necessarily due to a latent trait.

We have shown how the standard models for learning and assessment are related and how they in fact fall under a larger umbrella of models in network psychometrics. The success and popularity of these models may be attributable to the powerful phenomenon that fall out of these network models. Further implications of network psychometrics on the connection between BKT and IRT need to be explored.

Compliance with ethical standards

Conflicts of interest The authors declare that they have no conflict of interest.

References

Altman RM (2007) Mixed hidden markov models. J Am Stat Assoc 102(477):201–210. https://doi.org/10.1198/016214506000001086

Beck JE, Chang K (2007) Identifiability: A fundamental problem of student modeling. In: Conati C, McCoy K, Paliouras G (eds) User modeling 2007 UM 2007 Lecture notes in computer science, vol 4511. Springer, Berlin, Heidelberg, pp 137–146. https://doi.org/10.1007/978-3-540-73078-1_17

Beck JE, Chang K, Mostow J, Corbett A (2008) Does help help? introducing the Bayesian evaluation and assessment methodology. In: Woolf BP, Aïmeur E, Nkambou R, Lajoie S (eds) Intelligent tutoring systems, ITS 2008. Lecture notes in computer science, vol 5091. Springer, Berlin, Heidelberg, pp 383–394. https://doi.org/10.1007/978-3-540-69132-7_42

Birnbaum A (1967) Statistical theory for logistic mental test models with a prior distribution of ability. ETS Res Bull Ser. https://doi.org/10.1002/j.2333-8504.1967.tb00363.x

Borsboom D (2008) Psychometric perspectives on diagnostic systems. J Clin Psychol 64(9):1089–1108. https://doi.org/10.1002/jclp.20503

Borsboom D, Cramer AO (2013) Network analysis: an integrative approach to the structure of psychopathology. Annu Rev Clin Psychol 9:91–121

Cen H, Koedinger K, Junker B (2006) Learning factors analysis – A general method for cognitive model evaluation and improvement. In: Ikeda M, Ashley KD, Chan TW (eds) Intelligent tutoring systems, ITS 2006. Lecture notes in computer science, vol 4053. Springer, Berlin, Heidelbergr, pp 164–175. https://doi.org/10.1007/11774303_17

Corbett AT, Anderson JR (1995) Knowledge tracing: modeling the acquisition of procedural knowledge. User modeling and user-adapted interaction 4(4):253–278

Cramer AOJ, Waldorp LJ, van der Maas HLJ, Borsboom D (2010) Comorbidity: a network perspective. Behav Brain Sci 33(2–3):137150. https://doi.org/10.1017/S0140525X09991567

Cronbach LJ, Meehl PE (1955) Construct validity in psychological tests. Psychol Bull 52(4):281

Doignon J, Falmagne J (1985) Spaces for the assessment of knowledge. Int J Man Mach Stud 23:175–196

Doroudi S, Brunskill E (2017) The misidentified identifiability problem of bayesian knowledge tracing. In: Hu X, Barnes T, Hershkovitz A, Paquette L (eds) Proceedings of the 10th international conference on educational data mining, Wuhan, Hubei, China, June 25–28, 2017, pp 143–149

Ekanadham C, Karklin Y (2017) T-skirt: online estimation of student proficiency in an adaptive learning system. arXiv preprint arXiv:170204282

Embretson SE (1991) A multidimensional latent trait model for measuring learning and change. Psychometrika 56(3):495–515

Gowda SM, Rowe JP, Baker RS, Chi M, Koedinger KR (2011) Improving models of slipping, guessing, and moment-by-moment learning with estimates of skill difficulty. In: Pechenizkiy M, Calders T, Conati C, Ventura S, Romero C, Stamper J (eds) Proceedings of the 4th international conference on educational data mining, EIndhoven, the Netherlands, July 6–8, 2011, pp 179–188

Green B (1950) A general solution for the latent class model of latent structure analysis. Research Bulletin No. RB-50-38, Educational Testing Service, Princeton, NJ

Gweon GH, Lee HS, Dorsey C, Tinker R, Finzer W, Damelin D (2015) Tracking student progress in a game-like learning environment with a monte carlo bayesian knowledge tracing model. In: Proceedings of the fifth international conference on learning analytics and knowledge, ACM, pp 166–170. https://doi.org/10.1145/2723576.2723608

Kan KJ, van der Maas HLJ, Kievit RA (2016) Process overlap theory: strengths, limitations, and challenges. Psychol Inquiry 27(3):220–228. https://doi.org/10.1080/1047840X.2016.1182000

Khajah M, Lindsey RV, Mozer MC (2016) How deep is knowledge tracing? arXiv preprint arXiv:1604.02416

Khajah M, Wing R, Lindsey R, Mozer M (2014a) Integrating latent-factor and knowledge-tracing models to predict individual differences in learning. In: Stamper J, Pardos Z, Mavrikis M, McLaren BM (eds) Proceedings of the 7th international conference on educational data mining, London, United Kingdom, July 4–7, 2014, pp 99–106

Khajah MM, Huang Y, González-Brenes JP, Mozer MC, Brusilovsky P (2014b) Integrating knowledge tracing and item response theory: A tale of two frameworks. In: Proceedings of workshop on personalization approaches in learning environments (PALE 2014) at the 22th international conference on user modeling, adaptation, and personalization, University of Pittsburgh, pp 7–12

Koedinger KR, Stamper JC, McLaughlin EA, Nixon T (2013) Using data-driven discovery of better student models to improve student learning. In: Lane HC, Yacef K, Mostow J, Pavlik P (eds) Artif Intell Education. Springer, Berlin Heidelberg, Berlin, pp 421–430

Kovacs K, Conway ARA (2016) Process overlap theory: a unified account of the general factor of intelligence. Psychol Inquiry 27(3):151–177. https://doi.org/10.1080/1047840X.2016.1153946

Lee JI, Brunskill E (2012) The impact on individualizing student models on necessary practice opportunities. In: Yacef K, Zaïane O, Hershkovitz H, Yudelson M, Stamper J (eds) Proceedings of the 5th international conference on educational data mining. Chania, Greece, June 19–21, 2012, pp 118–125

Liggett TM (2012) Interacting particle systems, vol 276. Springer, Berlin

Lord F (1977) Practical applications of item characteristic curve theory. J Educ Meas 14(2):117–138. https://doi.org/10.1111/j.1745-3984.1977.tb00032.x

Lord FM (1951) A theory of test scores and their relation to the trait measured. Research Bulletin No. RB-51-13, Educational Testing Service, Princeton, NJ

Maclellan C, Liu R, Koedinger K (2015) Accounting for slipping and other false negatives in logistic models of student learning. In: Proceedings for the 8th international conference on educational data mining

Marsman M, Borsboom D, Kruis J, Epskamp S, van Bork R, Waldorp LJ, van der Maas HLJ, Maris G (2018) An introduction to network psychometrics: relating ising network models to item response theory models. Multivar Behav Res 53(1):15–35. https://doi.org/10.1080/00273171.2017.1379379

Martin AD, Quinn KM (2002) Dynamic ideal point estimation via markov chain monte carlo for the us supreme court, 1953–1999. Polit Anal 10(2):134–153

Merton RK (1968) The matthew effect in science: the reward and communication systems of science are considered. Science 159(3810):56–63

Nedungadi P, Remya M (2015) Incorporating forgetting in the personalized, clustered, bayesian knowledge tracing (pc-bkt) model. In: 2015 International Conference on cognitive computing and information processing (CCIP) IEEE, pp 1–5

Newman M, Barkema G (1999) Monte Carlo methods in statistical physics chapter 1–4. Oxford University Press, New York, USA

Pardos ZA, Heffernan NT (2011) KT-IDEM: Introducing item difficulty to the knowledge tracing model. In: Konstan JA, Conejo R, Marzo JL, Oliver N (eds) User modeling, adaption and personalization UMAP 2011 Lecture notes in computer science, vol 6787. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-22362-4_21

Pavlik PI, Cen H, Koedinger KR (2009) Performance factors analysis –a new alternative to knowledge tracing. In: Proceedings of the 2009 conference on artificial intelligence in education: building learning systems that care: from knowledge representation to affective modelling, IOS Press, Amsterdam, The Netherlands, The Netherlands, pp 531–538. http://dl.acm.org/citation.cfm?id=1659450.1659529

Qiu Y, Qi Y, Lu H, Pardos Z, Heffernan N (2010) Does time matter? Modeling the effect of time with Bayesian knowledge tracing. In: Pechenizkiy M, Calders T, Conati C, Ventura S, Romero C, Stamper J (eds) Proceedings of the 4th international conference on educational data mining, Eindhoven, the Netherlands, July 6–8, 2011, pp 139–148

Rasch G (1960) Probabilistic Models for Some Intelligence and Attainment Tests. Studies in mathematical psychology, Danmarks Paedagogiske Institut, https://books.google.com/books?id=aB9qLgEACAAJ

Reckase MD (1972) Development and application of a multivariate logistic latent trait model. PhD thesis, Syracuse University

Ross S (2014) Introduction to probability models. Elsevier. https://books.google.com/books?id=cehTngEACAAJ

Sao Pedro M, Baker R, Gobert J (2013) Incorporating scaffolding and tutor context into bayesian knowledge tracing to predict inquiry skill acquisition. In: D'Mello SK, Calvo RA, Olney A (eds) Proceedings of the 6th international conference on educational data mining, Mepmphis, TN, USA, July 6–9, 2013

Savi AO, Marsman M, van der Maas HLJ, Maris G (2018) The wiring of intelligence. PsyArXiv preprint. https://doi.org/10.31234/osf.io/32wr8

Schultz S, Tabor T (2013) Revisiting and extending the item difficulty effect model. In: In Proceedings of the 1st workshop on massive open online courses at the 16th annual conference on artificial intelligence in education, pp 33–40

Spearman C (1904) "general intelligence," objectively determined and measured. Am J Psychol 15(2):201–292. http://www.jstor.org/stable/1412107

Studer C (2012) Incorporating learning into the cognitive assessment framework. PhD thesis, Carnegie Mellon University

Sympson JB (1978) A model for testing with multidimensional items. In: Proceedings of the 1977 computerized adaptive testing conference, University of Minneapolis, Department of Psychology, Psychometric Methods Program Minneapolis, MN, 00014

van de Sande B (2013) Properties of the bayesian knowledge tracing model. J Educ Data Min 5(2):1

van der Linden WJ (2018) Handbook of item response theory, three volume set. Boca Raton, CRC Press

van der Maas HL, Dolan CV, Grasman RP, Wicherts JM, Huizenga HM, Raijmakers ME (2006) A dynamical model of general intelligence: the positive manifold of intelligence by mutualism. Psychol Rev 113(4):842

van Rijn PW (2008) Categorical time series in psychological measurement. PhD thesis, Universiteit van Amsterdam

Wang X, Berger JO, Burdick DS (2013) Bayesian analysis of dynamic item response models in educational testing. Ann Appl Stat 7(1):126–153

Weng RCH, Coad DS (2017) Real-time bayesian parameter estimation for item response models. Bayesian Analysis 13(1):115–137

Xu Y, Mostow J (2010) Using logistic regression to trace multiple sub-skills in a dynamic bayes net. In: Pechenizkiy M, Calders T, Conati C, Ventura S, Romero C, Stamper J (eds) Proceedings of the 4th international conference on educational data mining, EIndhoven, the Netherlands, July 6–8, 2011

Yudelson MV, Koedinger KR, Gordon GJ (2013) Individualized Bayesian knowledge tracing models. In: Lane HC, Yacef K, Mostow J, Pavlik P (eds) Artificial intelligence in education, AIED 2013. Lecture notes in computer science, vol 7926. Springer, Berlin, Heidelberg, pp 171–180. https://doi.org/10.1007/978-3-642-39112-5_18