Applying Prerequisite Structure Inference to Adaptive Testing

Sam Saarinen
Brown University
Providence, Rhode Island
sam_saarinen@brown.edu

Evan Cater
Brown University
Providence, Rhode Island
Evan.ecater@gmail.com

Michael L. Littman
Brown University
Providence, Rhode Island
mlittman@cs.brown.edu

ABSTRACT

Modeling student knowledge is important for assessment design, adaptive testing, curriculum design, and pedagogical intervention. The assessment design community has primarily focused on continuous latent-skill models with strong conditional independence assumptions among knowledge items, while the prerequisite discovery community has developed many models that aim to exploit the interdependence of discrete knowledge items. This paper attempts to bridge the gap by asking, "When does modeling assessment item interdependence improve predictive accuracy?" A novel adaptive testing evaluation framework is introduced that is amenable to techniques from both communities, and an efficient algorithm, Directed Item-Dependence And Confidence Thresholds (DIDACT), is introduced and compared with an Item-Response-Theory based model on several real and synthetic datasets. Experiments suggest that assessments with closely related questions benefit significantly from modeling item interdependence.

CCS CONCEPTS

• **Applied computing** → **Education**; • **Human-centered computing** → *Empirical studies in collaborative and social computing*; • **Computing methodologies** → *Logical and relational learning*.

KEYWORDS

prerequisite inference, adaptive testing, knowledge models, assessment design

ACM Reference Format:
Sam Saarinen, Evan Cater, and Michael L. Littman. 2020. Applying Prerequisite Structure Inference to Adaptive Testing. In *Proceedings of the 10th International Conference on Learning Analytics and Knowledge (LAK '20), March 23–27, 2020, Frankfurt, Germany*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3375462.3375541

1 INTRODUCTION

This paper attempts to bridge the gap between two communities of knowledge-modeling research. The paper is specifically built around the question, "When does modeling assessment item interdependence improve predictive accuracy?" This introduction will provide context for the paper and distinguish this work from related work in the literature.

LAK '20, March 23–27, 2020, Frankfurt, Germany

ACM ISBN 978-1-4503-7712-6/20/03…$15.00
https://doi.org/10.1145/3375462.3375541

1.1 We are Discovering Prerequisite Structures

Although this paper uses an adaptive testing evaluation framework, the techniques are most closely related to the Prerequisite Inference literature. There are many educational uses for identifying dependencies between topics, concepts, or questions. These uses include defining constraints on curricular order (what order should topics be taught in to maximize student learning) [4], providing course recommendations [2] , designing adaptive testing systems (and inferring student knowledge) [12], and efficiently validating new test questions. Although the exact form of such relational structures has varied across the literature, this paper will call all such devices **dependency maps**. Prior work has attempted to deduce such dependency maps from a variety of data sources using a variety of techniques and evaluation methods. See Table 1 for a summary.

This work is motivated by the problem of detecting student knowledge efficiently using student-sourced questions, a promising approach to scalable assessment generation and adaptation [17]. Due to minimal expert oversight, there is no ground-truth source of skill-labelings for questions assessing the same skill or knowledge, nor is there a ground-truth dependency map to validate against (so it cannot be used to measure performance of algorithms designed for this problem). Furthermore, because the student-contributed questions are often written without global awareness of the other questions available, many questions are related or equivalent. This motivates an adaptive testing system that attempts to minimize the number of questions needed to accurately predict student performance. (Note that even with expert-authored questions, experts may wish to validate their own dependency maps empirically, or to save themselves the effort of creating one manually.)

This paper aims to learn a dependency map on the basis of explaining (or predicting) the observed data, so the works closest to this paper are the attempts to use Bayesian inference to infer prerequisite relationships among latent skills, given the mapping from assessment questions to required skills [4, 7, 10]. Although those approaches are promising and able to reproduce small artificially-generated or expert-defined structures, they suffer from two primary limitations. First, the ground-truth mapping from questions to measured latent skills is not available in the problem domain considered here. Second, Bayesian inference methods are generally both approximate and slow, limiting their scalability. This paper considers structures with an order of magnitude more nodes than those studied in prior work.

The algorithm explored here, DIDACT, also bears resemblance to the prior Probabilistic Association Rules Mining work [8]. The work presented here differs primarily in that this paper explicitly considers the problem of predicting or filling in values in the dataset, and the algorithm has been generalized to allow item equivalence.

Table 1: Approaches to prerequisite map inference are grouped broadly by approach to validation, then by exact validation method, then by source of data. This paper introduces a new evaluation framework for dependency maps and evaluates a novel technique inspired by several existing ones.

Data Source	Validation Method	Technique	Reference
Expert (or Simulated) Dependency Map Recovery			
Student Answers to Test Questions	Plausible Structure Recovery	Expectation Maximization on Pairwise Relationships	[4]
Pairwise Interaction Features	Human Evaluation	Various Regression Algorithms	[3, 6]
Course Enrollment and Grades	Reproducing Existing Course Prerequisites	Ranking by Conditional Success Ratios	[2]
Probabilistic Student Knowledge States from Test Questions	Rediscovery of Simulated and Expert Structure	Probabilistic Association Rules Mining	[8]
Student Answers to Test Questions	Rediscovery of Simulated and Expert Structure	Bayesian Model Selection	[7, 10]
Data Self-Supervision			
Student Answers to Test Questions	Leave-One-Out Cross Validation	Structural EM for Bayesian Model Selection	[7]
Student Answers to Test Questions	Data Reconstruction Error	Restricted Bayesian Inference (DIDACT)	**this paper**

There is also a fascinating body of work into Dependency Map learning from natural language sources (Adorni et al. [1], for example), but those techniques require a large text corpus (such as a textbook), are not designed for relating assessment items, and the evaluation method presented here is fundamentally different.

There is also work on predicting student responses using supervised learning [11], but that work only applies to predicting responses to a fixed set of questions given responses to a different fixed set of questions, making it inapplicable for either detecting prerequisite relationships or facilitating adaptive testing.

Finally, we also note that the methods presented here exploit algorithms on directed acyclic graphs (DAGs) to explicitly simplify the output and enforce global constraints in the dependency map, a technique that has not appeared in the prior literature.

1.2 We do *NOT* use a Q-Matrix

Many approaches to inferring dependency maps aim to simplify the problem through use of a **Q-matrix**, which maps a number of assessment items to a smaller number of latent knowledge variables. Q-matrix Q has $Q_{ij} = 1$ if question i uses skill j, and 0 otherwise. If an exam is built by experts, a Q-Matrix may be hand-coded. In our setting however, we do not use a Q-matrix. Instead, we design an inference algorithm that scales well to large numbers of assessment items.

1.3 We are Doing Adaptive Testing

Computer Adaptive Testing (CAT), or simply Adaptive Testing, has a rich history in the literature, dating back to 1985 (Weiss [20]). In recent years, many innovations in Knowledge Modeling have been carried over to an Adaptive Testing setting [16]. We continue this tradition, but with a novel evaluation framework for adaptive testing that provides rich information around the tradeoff between data-efficiency and accuracy.

1.4 We compare to Item Response Theory

Item Response Theory (IRT) assumes that students have skills which influence their question answers. In Item Response Theory, the simplest model is known as the 1-parameter logistic model, or the 1PL model. In a 1PL model, the ith learner is modeled by a single parameter θ_i called ability or proficiency, and the question/item is modeled by a difficulty parameter d_j. If we add a parameter a_j that specifies the discrimination ability of the question, the model is known as a 2PL model. If we incorporate a parameter c_j that specifies the likelihood of a guess, we have a 3PL model. Each question in an IRT theory has an associated item response function, often a logistic function. The difficulty, discrimination, and guess parameters reshape the logistic function as follows:

$$p_j(\theta) = c_j + \frac{1 - c_j}{1 + e^{-a_j(\theta - d_j)}}$$

If the ability and difficulty parameters are allowed to be multi-dimensional, the framework is called Multi-Dimensional Item Reponse Theory (MIRT). At the time of writing, no general framework for MIRT model learning operates directly from student response data with no expert input [5].

Plajner [15] introduces a straightforward method for building CAT models with IRT. They use empirical bayesian estimates of the latent parameters based on answers, and compute the information provided by asking a given question, consequently picking the question that maximizes the information at each timestep. The use of IRT in adaptive testing is well-established [18].

Although it should now be apparent how IRT and Dependency Map inference can be both used within an adaptive testing framework, it may be beneficial to clarify their differences. Fundamentally, IRT is rooted in the assumption that there are (at most) a small number of latent continuous skills that independently predict correctness on each item. This assumption of conditional independence among the items given the skills is very elegant, allowing efficient model inference and preserving the simplicity of the model. In contrast, Dependency Map inference is fundamentally premised on the idea that assessment items exhibit interdependence.

1.5 We are *NOT* Doing Knowledge Tracing

Knowledge tracing is the task of modelling student knowledge *over time* to accurately predict future student performance [14]. When systems can accurately model student knowledge, content can be suggested to students based on individual needs. In the literature it is common to use a Bayesian model of the knowledge of a student, updating learner's latent knowledge using a hidden Markov model as learners interact with exercises [19]. Recent models propose using recurrent neural networks to predict student responses based on their past activity [13, 14]. The fundamental difference between knowledge tracing (KT) and CAT is that while in KT system designers are trying to maximize the student's knowledge through exercises that teach concepts, CAT is focused on *testing* a student's knowledge, as accurately and efficiently as possible. This is not to say that the two tasks are unrelated—both KT and CAT use models of student knowledge. For example, the use of IRT and MIRT models for knowledge representation, Bayesian networks, and Q-Matrices are used throughout the both the KT and CAT literatures.

1.6 Contributions

This paper has three primary contributions. First, a quantitative evaluation framework for adaptive testing is introduced that allows control of the tradeoff between data efficiency and accuracy through a settable parameter γ. Second, a fast algorithm for mining dependency relationships and doing adaptive testing is presented. This algorithm does a restricted form of Bayesian reasoning that achieves high accuracy, brief runtime, and high data-efficiency. Third, experiments on real and simulated data suggest that modeling of item interdependencies has a significant impact on predictive power when the assessment is narrow in scope.

2 VALIDATION METHOD

The value of a model should ultimately be measured by how well it predicts unseen/new data. This perspective is inherently captured by the adaptive testing problem, where the goal is to ask questions until the student's responses to the remaining items can be predicted with high accuracy. There are two primary objectives involved in adaptive testing systems. The first is efficiency—to minimize the number of questions asked. The second is robustness. Adaptive testing suffers from asymmetrical error conditions whereby asking unnecessary questions is much less expensive than mislabeling student knowledge of an item. These two kinds of error are difficult to compare directly in terms of, for example, total cost in student time, so we use a proxy condition: All inferred student responses should be provided with at least some minimum **accuracy threshold** denoted γ. For example, $\gamma = .95$ indicates a model should only predict the student's response to a question if it is at least 95% likely to get it right. This requires the model to both have high accuracy and to *know* that it has high accuracy. This setup motivates the following active-learning-style problem:

(1) Train on a dataset of previous student correctness scores on a variety of assessment items, possibly with missing values.
(2) For each (test set) student, repeatedly select a question to ask and then receive a response, or issue a stop command.
(3) After the stop command, predict the student's responses to any remaining questions.
(4) For every predicted response that is correct, give score 1. For every predicted response that is incorrect, give score $-\frac{\gamma}{1-\gamma}$. This penalty gives expected score 0 when the algorithm has exactly confidence γ. Note that questions that were asked (not predicted) receive score 0.

This scoring scheme is simple and allows traditional train/test splits, cross validation, or online learning evaluations. It is in the best interest of the tested algorithm to only predict responses that it believes it will get correct with probability greater than γ and to ask the question if its confidence is less than γ. If its confidence is exactly γ, guessing or asking yield the same score in expectation.

This metric allows us to explore the tradeoff between data efficiency and accuracy by adjusting γ. With γ equal to 0, the score is the number of questions that were inferred correctly without being asked. If the score is normalized by the total number of questions, this is a fairly direct measure of the "efficiency" of the adaptive testing system — how many questions (on average) the system is able to predict responses to without asking them. Here, the baseline to compare to is an algorithm that just guesses that each student will do what the majority do on each item (get it correct or incorrect). This baseline asks no questions of new students, so the only way to improve over its score is by using responses to some questions to improve the accuracy of predictions made on the rest (by modeling student ability or inter-question relationships, for example). Note that it is difficult to achieve scores near 1 when there are only a small number of assessment items, due to the proportional cost of gathering information. However, as the number of assessment items grows, the opportunities for modeling to accurately predict responses to a large fraction of the items increases.

At high γ, the model is primarily concerned with accuracy; with such a steep penalty for wrong predictions, the model will be willing to ask many questions in order to ensure that each remaining inference is correct. Here, the baseline is an algorithm which asks all questions, achieving score 0 every time. While this baseline is not very efficient, it is perfectly accurate, and so suffers no penalties. The danger for algorithms based on models is that the models must not over-estimate their confidence of a student's response - otherwise they stand to suffer large negative penalties for incorrect guesses. At $\gamma = 1$, there is an infinite penalty for even a single incorrect inference, so any score above 0 is highly impressive. Note that if questions can be guessed (or mistakes made) with some probability ϵ (the maximum noise in the observation), models should simply ask most questions when $\gamma > 1 - \epsilon$. Along this line of thinking, the most practically relevant range on these plots is the range from $\frac{1}{2} \leq \gamma \leq 1 - \epsilon$. In this range, there will be questions for which majority rule is no longer a safe guessing strategy, but careful modeling still has a chance of inferring responses accurately. In terms of interpretation, $\gamma = \frac{1}{2}$ is the point at which asking a single question has about the same cost as simply teaching that content.

3 A FAST DISCRETE MODEL

Bayes Nets are quite general and very data efficient, but can be computationally slow to learn and do inference in as the number of variables grows. This paper therefore presents a discrete model that balances the flexibility of interdependence modeling with the speed of inference under assumptions of independence. The algorithm

considers the influence of observed responses on a given item as independent, *subject to the learned dependency map*. In other words, redundant evidence (from a prerequisite of an observed prerequisite, for example) is filtered out, as is irrelevant evidence (evidence from items not transitively connected to the query item by the dependency map). This combines some of the best of Bayesian Networks (expressive capability and dependency modeling) with IRT (fast inference due to independence assumptions). This algorithm is called Directed Item-Dependence And Confidence Thresholds (DIDACT). Construction of the dependency map proceeds by 4 steps.

(1) Prepare statistics on all pairs of test items. How many students are correct on both, only the first, only the second, or neither?
(2) Sort prospective edges for the dependency map by the mutual information between that pair of questions.
(3) For each prospective edge, determine if it is an equivalence relation, prerequisite relation, or other.
(4) Add equivalence and prerequisite relations according to their sorted order, using a DAG structure over equivalence classes to enforce non-circularity of the dependencies.

For Step 3, a globally estimated guess parameter g is used to construct a test for the different relations. Let $\hat{a}$ be the estimated proportion of students who answer both questions incorrectly and let $\hat{b}$ and $\hat{c}$ be the estimated proportion who answered one question correctly or the other, respectively. If, with probability at least γ, $b > \frac{g}{1-g}a$ and $c > \frac{g}{1-g}a$, there is no relation. If exactly one of the inequalities is true with confidence greater than γ, then there is a directed relationship (one of the things can be known without the other, but not the other way around). Finally, if neither is greater, then the two are treated as equivalent.

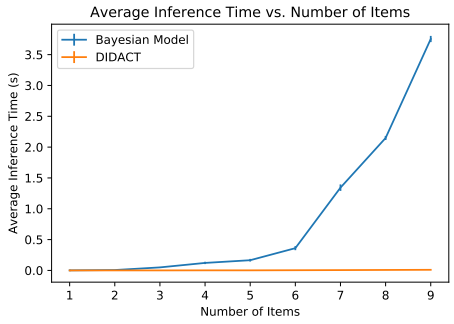

Figure 1: Exact Bayesian Inference quickly grows intractable, motivating the efficient DIDACT algorithm.

Doing inference with DIDACT is likewise straightforward. Given a vector of previously observed answers:

(1) For each node x, construct a partial reduction (all observed nodes with a transitive dependence on x that do not have another observed node on any of their paths to x).
(2) Treat all observed variables as exerting independent influences on x. Take the product of their conditional likelihoods for $x = 1$ and $x = 0$ (we use Bayesian pseudo-counts to prevent probabilities of 1 or 0), multiply by the base answer rate for x, and then normalize over the two possible outcomes.

Finally, DIDACT uses a myopic active learning (item selection) rule dependent on γ—given its current predictions, see which question will increase its expected score the most in the next round. If the expected increase is non-positive, stop asking questions and predict the responses to all of the remaining questions. Although DIDACT is just one possible combination of dependency inference and independence assumptions, the plots in Figures 1 and 2 show that it is very fast and fairly accurate.

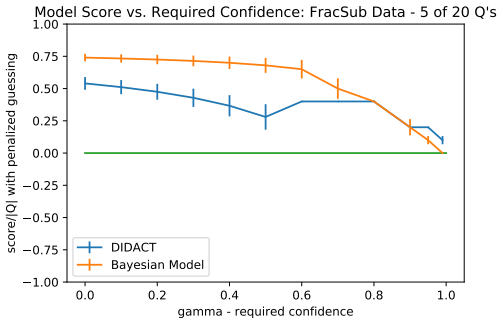

Figure 2: Although DIDACT is far from perfect, it still achieves good performance on real data, and with a much shorter runtime than exact Bayesian Inference.

4 EXPERIMENTAL RESULTS

In this section, the results of using an IRT-based model and using DIDACT are compared. Python code to reproduce the experiments in this paper is available online[1].

4.1 We Use Human Data

Results below are based on two publicly available real-world datasets: the FracSub dataset (Figure 3); and the SAT dataset (Figure 4). The FracSub dataset includes graded responses from 536 students to 20 middle-grade math questions and was first published in conjunction with [21].[2] The SAT dataset consists of responses from 296 students to 40 questions across multiple subject areas, and was first published in conjunction with [9], and is available through the adaptive testing repository made available by Vie[3], which we also use for our IRT baselines.

4.2 Experiments show Benefits from Modeling Interdependence

On the FracSub dataset (Figure 3), DIDACT and IRT begin with very similar performance, but the increasing γ shows that DIDACT has more accurate estimates of the likelihood of inferred answers. For convenience, a baseline algorithm is also plotted. The Base Rate algorithm estimates the base likelihood (without seeing any other answers) of each item being correctly answered. As long as an item's likelihood (or its complement) exceed γ, that item's response is inferred. Otherwise, the item is specifically queried.

[1] https://sam-saarinen.github.io/artifacts/
[2] http://staff.ustc.edu.cn/ qiliuql/data/math2015.rar
[3] https://github.com/jilljenn/qna

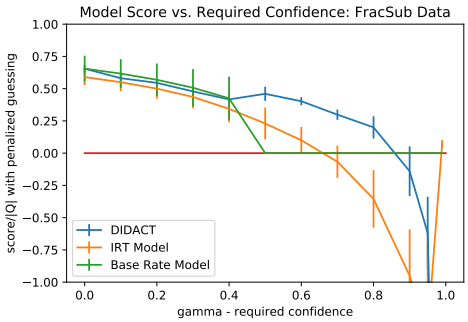

Figure 3: Adaptive Testing Performance subject to required confidence threshold γ on the FracSub dataset.

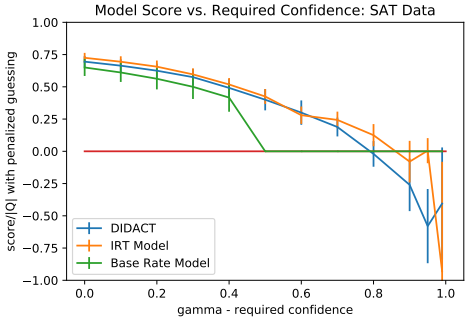

Figure 4: Adaptive Testing Performance subject to required confidence threshold γ on the SAT dataset.

DIDACT achieves significantly higher accuracy once the base rate is no longer informative, although both models overestimate their own accuracy, as revealed at γ very close to 1.

In contrast, on the SAT dataset (Figure 4), the performance of these two models are roughly equal. (If anything, DIDACT performs slightly worse, although no statistically significant conclusion can be drawn given the error bounds.) What accounts for the difference in results between these two datasets? We note that the FracSub dataset involves many questions that are closely related semantically, whereas the SAT dataset includes questions from multiple unrelated subject areas. This suggests the following hypothesis: on closely related questions, question inter-dependence violates the conditional independence assumption of IRT, leading to worse performance than when the questions are nearly independent.

To test this hypothesis, experiments were run on two synthetic datasets. One is based on a prerequisite structure with very high interdependence; 10 items are placed in a single chain of prerequisite dependencies. (See Figure 5.) In the other dataset, (Figure 6) 40 items are placed in a large sparse DAG structure where many items do not have any transitive relationship. The results align with expectations. On the Chain Dataset, DIDACT performs excellently, querying only 2-4 items (on the order of $\log_2 10$) across all levels of γ. This performance is possible only because DIDACT explicitly captures the transitive dependence relationships between items. On the same dataset, IRT is forced to query many more nodes and suffers from inaccurate probabilities (revealed as γ approaches 1). In contrast, both models achieve good (and nearly indistinguishable) performance on the Broad DAG dataset, where the conditional independence assumption of IRT is a reasonable simplification of the true structure of the data.

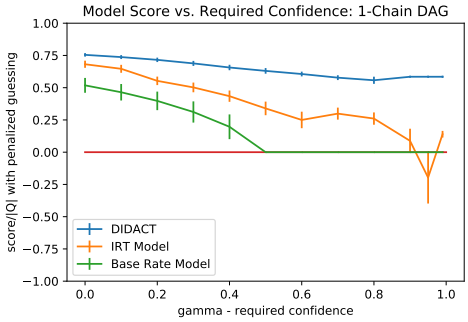

Figure 5: Adaptive Testing Performance subject to required confidence threshold γ on a synthetic dataset where dependencies form a single chain.

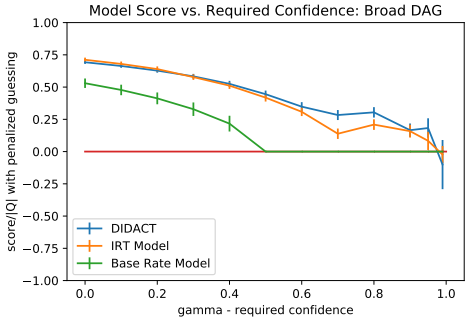

Figure 6: Adaptive Testing Performance subject to required confidence threshold γ on a synthetic dataset where dependencies form a broad but connected DAG.

These results suggest two interesting findings: first, for closely-related questions, models that are able to capture the interdependence of test items have higher predictive power; second, this phenomenon may not be discoverable from data based on comprehensive or broad assessments, because in these settings the two models are indistinguishable.

5 LEVERAGING INTERDEPENDENT MODELS

Part of why IRT models (such as the Rasch model), have been so popular over the last decades is their auxiliary uses based on interpretation of the model. For example, questions can be ranked based on how well they fit the model defined by the other questions (a form of internal validity and the basis of measures like Cronbach's Alpha). Students can be evaluated. And, questions (and related topics) can be ordered by their difficulty, leading to a natural

curriculum. The goal of this section is to illuminate how some of these use cases can benefit from modeling the interdependence of assessment items.

5.1 How to Make Exams More Reliable

Large assessments can often be made more reliable by removing questions that have little relevance to the rest of the exam. In the ideal of the adaptive test setting, the minimal number of questions are asked to accurately predict responses to the remaining questions, so a natural way of ranking items (represented as random variables X_i) is by the following:

$$R(X_i) = \sum_j I(X_i, X_j),$$

where $I(X_i, X_j)$ is the mutual information between X_i and X_j. Note that this score includes the amount of mutual information the variable has with itself, which is just the entropy of the random variable $H(X_i) = I(X_i, X_i)$. It slightly favors questions that are neither too easy nor too hard for most students.

Given a means of ranking questions, assessments can be designed subject to budget constraints for a particular γ. This goal can be accomplished by adding questions in order of decreasing rank until the mean score at γ begins to decrease.

5.2 How to Evaluate Students

Student abilities can be represented as a vector indicating whether the student has mastered each item. Given a dependency map, this vector space has a partial ordering that captures possible learning trajectories for each student. It also allows for fine-grained student diagnostics - perhaps the student isn't lacking practice, but specific prerequisite knowledge that would allow them to succeed. Although there are many possible ways to collapse the student mastery vector into a single grade or score, the fine-grained vector may hold more utility for practical classroom use.

5.3 How to Infer a Curriculum

Just as student ability vectors define a partial order over students, the dependency map defines a partial order over content. By the assumptions of the model, content appearing in the dependency map cannot be mastered before the content it is dependent on. Thus, all prerequisite topics should occur in a curriculum before the topic that depends on them.

6 CONCLUSION

This paper provided empirical evidence that assessments involving closely related items are likely to benefit from interdependence modeling. To facilitate these experiments, a novel evaluation framework was introduced that explicitly navigates the tradeoff between data-efficiency and accuracy in adaptive testing. Additionally, a novel algorithm for interdependence modeling, DIDACT, was introduced, which achieves high performance while remaining computationally efficient. Finally, these results were connected to related educational problems, including assessment creation, adaptive pedagogy, and curriculum design. These results can be applied directly in future work expanding the use of dependency modeling in adaptive testing, which may be of particular use when assessment items come from nontraditional sources or the pool of items changes over time. Future work may consider how to extend these models to more general models of assessment than binary correct/incorrect items.

ACKNOWLEDGMENTS

This work was partially funded by DARPA Award FA8750-19-2-1006.

REFERENCES

[1] Giovanni Adorni, Chiara Alzetta, Frosina Koceva, Samuele Passalacqua, and Ilaria Torre. 2019. Towards the Identification of Propaedeutic Relations in Textbooks. In *International Conference on Artificial Intelligence in Education*. Springer, 1–13.
[2] Jaroslav Bayer, Hana Bydžovská, and Jan Géryk. 2012. Towards course prerequisites refinement. *IMEA 2012* (2012), 4.
[3] Anthony F Botelho, Seth A Adjei, and Neil T Heffernan. 2016. Modeling Interactions across Skills: A Method to Construct and Compare Models Predicting the Existence of Skill Relationships. *International Educational Data Mining Society* (2016).
[4] Emma Brunskill. 2011. Estimating Prerequisite Structure From Noisy Data.. In *EDM*. Citeseer, 217–222.
[5] R Philip Chalmers et al. 2016. Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software* 71, 5 (2016), 1–39.
[6] Haw-Shiuan Chang, Hwai-Jung Hsu, and Kuan-Ta Chen. 2015. Modeling Exercise Relationships in E-Learning: A Unified Approach.. In *EDM*. 532–535.
[7] Yetian Chen, José P González-Brenes, and Jin Tian. 2016. Joint Discovery of Skill Prerequisite Graphs and Student Models. *International Educational Data Mining Society* (2016).
[8] Yang Chen, Pierre-Henr Wuillemin, and Jean-Marc Labat. 2015. Discovering Prerequisite Structure of Skills through Probabilistic Association Rules Mining. *International Educational Data Mining Society* (2015).
[9] Michel Desmarais et al. 2011. Conditions for effectively deriving a q-matrix from data with non-negative matrix factorization. In *4th international conference on educational data mining, EDM*. 41–50.
[10] Soo-Yun Han, Jiyoung Yoon, and Yun Joo Yoo. 2017. Discovering skill prerequisite structure through Bayesian estimation and nested model comparison.. In *EDM*.
[11] Pan Liao, Yuan Sun, Shiwei Ye, Xin Li, Guiping Su, and Yi Sun. 2017. Predicting learners' multi-question performance based on neural networks. In *2017 International Conference on Behavioral, Economic, Socio-cultural Computing (BESC)*. IEEE, 1–6.
[12] Danny Lynch and Colm P Howlin. 2014. Real world usage of an adaptive testing algorithm to uncover latent knowledge. In *7th international conference of education, research and innovation (ICERI2014 proceedings). IATED, Seville, Spain.* 504–511.
[13] Sein Minn, Yi Yu, Michel C. Desmarais, Feida Zhu, and Jill Jenn Vie. 2018. Deep Knowledge Tracing and Dynamic Student Classification for Knowledge Tracing. *Proceedings - IEEE International Conference on Data Mining, ICDM* 2018-Novem (2018), 1182–1187. https://doi.org/10.1109/ICDM.2018.00156 arXiv:arXiv:1809.08713v1
[14] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas Guibas, and Jascha Sohl-Dickstein. 2015. Deep knowledge tracing. *Advances in Neural Information Processing Systems* 2015-Janua (2015), 505–513.
[15] Martin Plajner. 2016. Student Skill Models in Adaptive Testing. *Journal of Machine Learning Research* 52 (2016), 403–414. http://www.jmlr.org/proceedings/papers/v52/plajner16.html
[16] Martin Plajner. 2017. Probabilistic Models for Computerized Adaptive Testing. (2017). arXiv:1703.09794 http://arxiv.org/abs/1703.09794
[17] Sam Saarinen, Shriram Krishnamurthi, Kathi Fisler, and Preston Tunnell Wilson. 2019. Harnessing the Wisdom of the Classes: Classsourcing and Machine Learning for Assessment Instrument Generation. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*. ACM, 606–612.
[18] Jill Jenn Vie. 2016. Adaptive Testing Using a General Diagnostic Model. 2 (2016), 640–643. https://doi.org/10.1007/978-3-319-45153-4
[19] Jill-Jênn Vie. 2018. Deep Factorization Machines for Knowledge Tracing. (2018), 370–373. https://doi.org/10.18653/v1/w18-0545 arXiv:arXiv:1811.03388v2
[20] David J Weiss. 1985. Adaptive testing by computer. *Journal of consulting and clinical psychology* 53, 6 (1985), 774.
[21] Runze Wu, Qi Liu, Yuping Liu, Enhong Chen, Yu Su, Zhigang Chen, and Guoping Hu. 2015. Cognitive modelling for predicting examinee performance. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.