International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

Predicting Sub-topic-level Weak Areas of Competitive Exam Aspirants from Previous-Year Paper Responses: A Review

Ms. Mansi Patil, Dr. Rajashree Patil, Dr. Vishal Waghmare, Dr. Irfan Mujawar
Department of Computer Science, Vivekanand College, Kolhapur, India

Abstract: Students preparing for competitive examinations such as UPSC, MPSC, SSC CGL and IBPS usually receive subject-level scores from mock tests, which do not show the specific sub-topic in which they are weak. Predicting sub-topic-level weak areas from previous-year paper responses is closely related to knowledge tracing and cognitive diagnosis, two research lines that estimate a learner's mastery of individual skills from response data. This paper reviews 23 studies on knowledge tracing, educational data mining and computerized adaptive testing. They cover probabilistic models, regression-based models, recurrent, memory-augmented and attention-based neural models, exercise-content-aware models, interpretability work and adaptive testing. The studies are compared on the problem addressed, the method used and their relevance to sub-topic-level diagnosis. The review shows that per-skill mastery estimation is well established; that simple extended or logistic-regression baselines are often competitive with deep models; that data preparation strongly affects reported gains; and that sparse response data favour attention-based, content-aware and interpretable models. In the studies reviewed we did not find a peer-reviewed evaluation on previous-year papers of Indian competitive exams; this is stated as a gap observed in a limited search, not as a proven fact. The paper closes with a suggested modelling pipeline and future research directions.

Keywords: knowledge tracing, educational data mining, student performance prediction, competitive exams, sub-topic analysis, computerized adaptive testing

I. Introduction

Millions of students in India prepare for government recruitment and civil service examinations every year. Mock tests and previous-year papers are an important part of this preparation. After a mock test, a student usually sees a total score, a rank and the accuracy for each subject. Commercial platforms increasingly advertise AI-based mock tests and performance analysis for these exams [1], [2]. A subject-level report tells a student that Quantitative Aptitude is weak, but it does not tell whether the difficulty lies in Percentage, Algebra or Geometry. Without that information, revision time is spread over the whole subject instead of the few sub-topics that matter.

Researchers in educational data mining have studied for decades how to estimate what a student knows from the answers given so far. This task is called knowledge tracing [3]. Early models kept a probability that a student had learned each skill [3]. Since the introduction of deep knowledge tracing [4], a large family of neural models has been proposed, and several surveys now organise the field [5], [6], [7]. In a parallel line of work, computerized adaptive testing (CAT) uses cognitive diagnosis models to choose questions and to estimate ability [8].

The problem considered here, finding weak sub-topics from the responses given on previous-year papers, can be seen as per-skill mastery estimation where each sub-topic plays the role of a skill. The literature on knowledge tracing, however, is built mostly on data from school and online-course platforms, and the conditions of competitive-exam preparation differ from those platforms. This review therefore asks three questions. (1) Which families of models have been proposed to estimate mastery at the skill level? (2) What does the evidence say about their accuracy, stability and interpretability? (3) Which of them suit the sparse, exam-style data of competitive-exam aspirants, and what gaps remain?

The main contributions are: (i) a structured summary of 23 studies in terms of problem, method, result and relevance to sub-topic diagnosis; (ii) a taxonomy that groups the approaches into seven families; (iii) a discussion of how the findings apply to competitive-exam data; and (iv) a list of research gaps and a suggested pipeline for future work.

II. Review Methodology

This is a narrative review, not a full systematic review. The studies were identified through searches on Google Scholar and arXiv using terms such as knowledge tracing, student performance prediction, educational data mining and computerized adaptive testing, and by following the reference lists of recent surveys. Studies were kept if they (a) model a learner's knowledge or ability from response data, (b) report a method and an empirical result, or (c) are surveys that organise the area. Foundational papers and recent papers were both included so that the development of the field can be followed.

In total 23 studies were read: 8 on the foundations of knowledge tracing, 6 on attention-based and improved models, 6 surveys or overviews, and 3 on computerized adaptive testing. For each study the problem, the method, the main reported result and the link to sub-topic diagnosis were recorded. Two industry sources [1], [2] are cited only as background on current practice, because they are not peer reviewed. A few classical references on regression-based and cognitive diagnosis models [9], [10], [11] were added to explain terms used by the reviewed studies. Because the selection was purposive and the search was limited, studies that are relevant may have been missed, and statements about gaps in the literature should be read with that in mind.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

III. Background and Problem Formulation

A. Sub-topic weak areas as skill mastery

Let each question in a paper be tagged with the sub-topic (skill) it tests, and let a student's response to a question be recorded as correct or wrong. The question-to-skill mapping is often called a Q-matrix in the psychometric literature; knowledge tracing machines also use this mapping directly [12]. The prediction task is to estimate the probability that a student answers a given question correctly. A weak area is then a sub-topic where the estimated mastery is low. Since mastery is not observed directly, the quality of an estimate is judged indirectly, usually by how well it predicts the next response.

B. Why subject-level reports are not enough

The first knowledge tracing systems already worked skill by skill: the ACT Programming Tutor kept a probability that a student had learned each production rule and continued giving exercises until the rule was mastered [3]. Subject-level scores hide this structure. Two students with the same Maths score can have different weak sub-topics, and the useful revision advice differs for them.

C. Data conditions in competitive-exam preparation

The data of an aspirant are different from the data of a tutoring system. A student typically attempts a few full papers, each with a limited number of questions spread across several subjects, and each question is answered once. Long, repeated practice sequences on the same skill, which many sequence models assume, are usually not available. Sub-topic labels are not provided by the exam bodies and have to be assigned from the syllabus. These conditions make sparsity and tagging quality central concerns, and they guide the comparison in Section V.

IV. Families of Approaches

Fig. 1 shows the chronology of the main studies reviewed, and the subsections below describe each family.



Fig. 1. Chronology of the main studies reviewed (entries are equally spaced; the interval between 1995 and 2015 is not to scale).

A. Probabilistic models

Bayesian knowledge tracing (BKT) treats knowledge of a skill as a hidden state that changes after each answer, and the ACT Programming Tutor used it to decide what to teach next; it was fairly successful at predicting test performance [3]. A limitation is that skills are modelled separately, so BKT cannot show how skills are related. Käser et al. addressed this with dynamic Bayesian networks that model many skills together, and reported better predictions than earlier methods on five large datasets covering maths, spelling and physics [13]. For sub-topics, the idea that skills are connected is natural, because Percentage, Ratio and Profit and Loss share prerequisites.

B. Regression and factorization models

Performance factors analysis uses counts of earlier successes and failures in a logistic model [9]. Knowledge tracing machines generalise this idea with factorization machines, which build higher-dimensional student models, use the question-to-skill mapping and the number of attempts, and are meant for sparse data [12]. A recent comparison of nine light models on two datasets found that multi-task and temporal models did best, but that plain logistic regression stayed competitive [14]. These results support using a regression baseline when data are small.

C. Recurrent and memory-augmented neural models

Deep knowledge tracing (DKT) applied recurrent neural networks to the sequence of a student's answers and improved prediction on several datasets, without the hand-coded knowledge that earlier models needed [4]. DKT does not directly show how good or weak a student is at each concept. The dynamic key-value memory network (DKVMN) added a static key memory for concepts and a dynamic value memory for the mastery of each concept, which yields a mastery estimate for every concept [15]. Two problems of DKT were later documented: its prediction for a skill can drop even when the student answered correctly, and its predictions change too much over time. Adding two regularization terms to the loss reduced both problems without harming the main task [16]. A later model combined a convolutional network, to find spatial features of the exercise sequence, with a bidirectional LSTM for temporal features, and predicted better than DKT and CKT on five public datasets [17]. A survey of loss-augmented knowledge tracing reviews such extra loss terms, including regularization and contrastive learning, and notes that the loss function affects how well a model trains [18].

D. Attention-based models

Recurrent models do not generalise well on sparse data. The self-attentive knowledge tracing model (SAKT) selects the relevant past activities with self-attention and predicts from only those, improving AUC by 4.43 percent on average [19]. Context-aware attentive knowledge tracing (AKT) combines attention with components from psychometrics, including the Rasch model, reached AUC up to 6 percent better than earlier models, and showed in case studies that it is interpretable [20]. Attention models are of interest here because an aspirant answers few questions and only some of them are relevant to a given sub-topic.

E. Exercise-content-aware models

Most of the models above use only the identity of the question and the answer, and ignore the text of the exercise. The exercise-enhanced recurrent neural network (EERNN) uses a bidirectional LSTM to learn from the exercise text and an RNN to track the student state, in a Markov and an attention version [21], [22]. Experiments on large real data showed that it works, and that it also helps with the cold-start problem for new students and new exercises [21]. For previous-year papers this is relevant because a question's wording may carry sub-topic information beyond the tag, and a student may have very few answers.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

F. Interpretability

Accurate deep models are often hard to explain with learning theory. Minn and Nkambou added human-understandable features derived from the interaction data, especially features about learning ability, and reported improved accuracy while staying consistent with cognitive theory [23]. For a student-facing weak-area report, explanation matters as much as accuracy, because the output has to be trusted and acted on.

G. Cognitive diagnosis and adaptive testing

Cognitive diagnosis models such as DINA relate responses to a Q-matrix of required skills [10], and neural versions have been proposed [11]. Computerized adaptive testing builds on such models, and a survey divides a CAT system into four parts: the measurement model, question selection, question bank construction and test control [8]. Existing CAT systems depend on one specific cognitive model and may not give both good and varied questions; the model-agnostic framework MAAT adds quality, diversity and importance modules [24], and EnCAT combines several diagnosis models for stability and learns question content to avoid similar questions [25]. CAT is relevant to a later stage of this work, when a platform chooses which sub-topic questions to ask next.

V. Comparative Analysis

Table I summarises the 23 studies. Table II groups them by family and lists the main strength and the main limitation as reported in the studies.

TABLE I. SUMMARY OF THE REVIEWED STUDIES

Study	Method and main result	Relevance to sub-topic diagnosis
Corbett and Anderson, 1995 [3]	ACT Programming Tutor keeps a probability that each rule is learned and gives exercises until mastery. Fairly successful at predicting test performance.	Base of the skill-by-skill idea.
Piech et al., 2015 [4]	DKT: recurrent network on the answer sequence. Large gains on several datasets.	Main deep learning method of the field.
Zhang et al., 2017 [15]	DKVMN: static key memory for concepts and dynamic value memory for mastery. Gives mastery per concept.	Per-concept mastery, as needed per sub-topic.
Käser et al., 2017 [13]	Dynamic Bayesian networks model many skills together. Better than earlier methods on five datasets.	Sub-topics may be related.
Khajah et al., 2016 [26]	Four patterns missed by BKT (recent answers, question order, skill similarity, student ability). Extended BKT matched DKT.	A simple model is a fair baseline.
Xiong et al., 2016 [27]	Rebuilt DKT and compared with PFA and BKT. On properly prepared data DKT had no overall edge over PFA; a data quality problem was found in a public dataset.	Data preparation matters.
Yeung and Yeung, 2018 [16]	Two regularization terms reduce inconsistent and unstable DKT predictions.	DKT predictions can be unstable.
Vie and Kashima, 2019 [12]	Factorization machines with question-to-skill mapping and attempts, for sparse data.	Mapping is like sub-topic tags.
Pandey and Karypis, 2019 [19]	SAKT: self-attention over relevant past activity. AUC up 4.43 percent on average.	Suited to few answers per student.
Ghosh et al., 2020 [20]	AKT: attention with Rasch-based components. AUC up to 6 percent better; interpretable.	Links psychometrics and deep learning.
Liu et al., 2021 [22]	EKT: exercise text via BiLSTM plus student state tracking.	Question text can help.
Exercise-enhanced model, 2018 [21]	EERNN idea tested on large real data; handles cold start.	Helps when answers are few.
DKT-STDRL, 2023 [17]	CNN for spatial and BiLSTM for temporal features. Better than DKT and CKT on five datasets.	Recent deep model.
Minn and Nkambou, 2025 [23]	Human-understandable features added. Higher accuracy, consistent with cognitive theory.	Supports adding student background features.
Song et al., 2022 [5]	Survey that groups and compares deep learning KT models.	Overview for related work.
KT survey, 2021 [6]	Three basic model types, variants, applications; open-source libraries EduData and EduKTM.	Reference list and code libraries.
EDM survey, 2023 [7]	Four areas: knowledge tracing, behaviour detection, performance prediction, personalized recommendation.	Places the topic within EDM.
Shukurlu, 2025 [18]	Survey of deep KT trained with extra loss terms, with benchmarks.	Background on training.
Das Adhikary et al. [14]	Taxonomy of nine model families; nine light models tested on ASSISTments and DigiArvi. Multi-task and temporal models best; logistic regression competitive.	Supports a regression baseline.
KT overview [28]	Short overview of promising deep KT models for early prediction of student outcomes.	Easy background reading.
Liu, Zhuang et al., 2024 [8]	Survey of CAT: measurement models, question selection, bank construction, test control.	Adaptive testing background.
Bi et al., 2021 [24]	MAAT: model-agnostic CAT with quality, diversity and importance modules.	How questions are chosen.
Ning et al., 2021 [25]	EnCAT: several diagnosis models for stability; question content to avoid similar questions.	Another question selection method.

TABLE II. FAMILIES OF APPROACHES: STRENGTHS AND LIMITATIONS REPORTED

Family	Main strength	Main limitation
Probabilistic (BKT, DBN) [3], [13]	Explicit mastery probability per skill; DBN models skill relations.	BKT treats skills separately and cannot show how they are related.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

Family	Main strength	Main limitation
Regression and factorization [12], [14], [27]	Use the question-to-skill mapping; competitive with deep models on prepared data.	One-dimensional parameters and sparse data limit earlier models.
Recurrent deep models [4], [16], [17]	Large gains on several datasets; no hand-coded knowledge.	No direct per-concept view; inconsistent and unstable predictions.
Memory networks [15]	Mastery for every concept.	Needs a concept (skill) definition for each question.
Attention models [19], [20]	Work with sparse data; psychometric parts give interpretability.	Deep models are in general hard to interpret.
Content-aware models [21], [22]	Use exercise text; handle cold start.	Earlier models ignore text and rely on answer records only.
Interpretable features [23]	Accuracy and consistency with cognitive theory.	Features come from interaction data that exam data may lack.
Cognitive diagnosis and CAT [8], [24], [25]	Question selection and ability estimation.	Dependence on one unstable cognitive model; similar questions chosen.

A. Main findings

Four findings run across the studies. First, mastery estimation at the level of individual skills is well established, from BKT [3] to memory networks [15], so a sub-topic-level report does not need a new kind of model. Second, the gains of deep models over simple ones are not always large: an extended BKT matched DKT [26], DKT had no overall edge over PFA on properly prepared data [27], and logistic regression stayed competitive in a recent comparison [14]. Third, reported results depend on data preparation, as shown by a data quality problem found in a public dataset [27], and on prediction stability [16]. Fourth, sparse data motivate attention-based [19], content-aware [21] and interpretable [20], [23] models.

VI. Applicability to Competitive-Exam Data

A. Suitability of the model families

The following points are our reading of the reviewed studies for the setting described in Section III-C; they are reasoned arguments, not experimental results. Sequence models such as DKT rely on long response histories, and a student who answers a handful of questions per sub-topic in one sitting gives little history. Regression and factorization models that use the question-to-skill mapping directly fit this situation better, and they offer a baseline that is cheap to train and easy to explain. Attention models may help where each student contributes few answers, and content-aware models may help where the tag alone is noisy. Adaptive-testing methods become useful only after a diagnosis model is in place.

B. A suggested pipeline

A pipeline consistent with the reviewed evidence has five steps.

- Collect responses to previous-year papers and tag each question with subject and sub-topic using the exam syllabus, with a second person checking a sample of tags.
- Fit a subject-level model and a regression or factorization sub-topic model with student background features [12], [14], so that the benefit of sub-topic information can be measured.
- Add an attention or content-aware model if the baseline leaves room for improvement [19], [20], [21].
- Split data by student, so that test students are unseen, and report AUC, accuracy and F1 score for every model, as recommended by the careful comparisons reviewed above.
- Convert predicted mastery into a ranked list of weak sub-topics, with simple features that a student can understand [23].

C. Defining a weak area

None of the reviewed studies tests a definition of weak area for exam preparation. A practical definition is a sub-topic whose predicted probability of a correct answer is below a chosen threshold, reported together with the number of questions behind the estimate. A sub-topic with few questions gives an uncertain estimate, and a report should say so instead of ranking it as confidently as a well-supported one.

VII. Research Gaps and Future Directions

From the 23 studies and the limited search described in Section II, the following gaps were observed.

- Datasets in the reviewed studies come mainly from school and online-learning platforms. We did not find a peer-reviewed evaluation of sub-topic-level prediction on previous-year papers of Indian competitive exams. The claims made by commercial platforms are product descriptions that have not been independently evaluated.
- There is no shared, openly tagged dataset of exam responses with sub-topic labels, so results cannot be compared across studies.
- The effect of tagging errors on weak-area prediction is not studied, and the cost of tagging by hand is a practical barrier.
- Methods designed for many repeated attempts have not been tested when each student answers each question only once.
- Explanations that are useful to aspirants, and not only accurate, need user studies in addition to offline metrics.
- Privacy and consent practices for collecting student responses in exam-preparation settings need to be described clearly in future work.

VIII. Limitations of this Review

The review is narrative, and the search was limited to a small set of sources, so it is not exhaustive. Several included studies are arXiv preprints, whose results may change after peer review. The results quoted are those reported by the original authors, on different datasets and metrics, and cannot be compared directly across studies. The two industry sources are not peer reviewed. The applicability discussion in Section VI is based on

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

reasoning from the literature and has not been tested on exam data.

IX. Conclusion

This paper reviewed 23 studies from knowledge tracing, educational data mining and computerized adaptive testing, to see how sub-topic-level weak areas of competitive-exam aspirants can be predicted from previous-year paper responses. The review found that per-skill mastery estimation is mature, that simple baselines are often competitive, that data preparation and stability affect reported results, and that sparse data favour regression, attention-based, content-aware and interpretable models. We did not find a peer-reviewed evaluation on Indian competitive-exam data, which is a gap that future empirical work can address by comparing subject-level and sub-topic-level models on a tagged dataset of previous-year paper responses.

References

[1] GetVidyaAI, "Why India's competitive exam students are switching to AI mock tests," getvidya.in, May 2026 (industry blog, not peer reviewed).

[2] TechStory, "How Law Prep Tutorial (LPT) is using AI to personalise CLAT preparation," techstory.in, 2026 (industry article, not peer reviewed).

[3] A. T. Corbett and J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge," User Modeling and User-Adapted Interaction, vol. 4, no. 4, pp. 253–278, 1995.

[4] C. Piech et al., "Deep knowledge tracing," in Advances in Neural Information Processing Systems (NeurIPS), 2015.

[5] X. Song et al., "A survey on deep learning based knowledge tracing," Knowledge-Based Systems, vol. 258, 2022, doi: 10.1016/j.knosys.2022.110036.

[6] "A survey of knowledge tracing: Models, variants, and applications," arXiv:2105.15106, 2021.

[7] "A comprehensive survey on deep learning techniques in educational data mining," arXiv:2309.04761, 2023.

[8] Q. Liu, Y. Zhuang et al., "Survey of computerized adaptive testing: A machine learning perspective," arXiv:2404.00712, 2024.

[9] P. I. Pavlik Jr., H. Cen and K. R. Koedinger, "Performance factors analysis — A new alternative to knowledge tracing," in Proc. Artificial Intelligence in Education (AIED), 2009.

[10] J. de la Torre, "DINA model and parameter estimation: A didactic," Journal of Educational and Behavioral Statistics, vol. 34, no. 1, pp. 115–130, 2009.

[11] F. Wang et al., "Neural cognitive diagnosis for intelligent education systems," in Proc. AAAI Conf. on Artificial Intelligence, 2020.

[12] J.-J. Vie and H. Kashima, "Knowledge tracing machines: Factorization machines for knowledge tracing," in Proc. AAAI Conf. on Artificial Intelligence, 2019.

[13] T. Käser et al., "Dynamic Bayesian networks for student modeling," IEEE Trans. Learning Technologies, vol. 10, no. 4, 2017.

[14] P. Das Adhikary, J. Metsämuuronen, M.-J. Laakso and J. Heikkonen, "Knowledge tracing models in educational data mining: Historical evolution, categorization, and empirical evaluation," IEEE Access, 2026, doi: 10.1109/ACCESS.2026.3678846.

[15] J. Zhang, X. Shi, I. King and D.-Y. Yeung, "Dynamic key-value memory networks for knowledge tracing," in Proc. 26th Int. Conf. on World Wide Web (WWW), 2017.

[16] C.-K. Yeung and D.-Y. Yeung, "Addressing two problems in deep knowledge tracing via prediction-consistent regularization," arXiv:1806.02180, 2018.

[17] "DKT-STDRL: Spatial and temporal representation learning enhanced deep knowledge tracing," arXiv:2302.11569, 2023.

[18] Shukurlu, "Survey of loss augmented knowledge tracing," arXiv:2504.15163, 2025.

[19] S. Pandey and G. Karypis, "A self-attentive model for knowledge tracing," arXiv:1907.06837, 2019.

[20] A. Ghosh, N. Heffernan and A. S. Lan, "Context-aware attentive knowledge tracing," in Proc. KDD, 2020, arXiv:2007.12324.

[21] "Exercise-enhanced sequential modeling for student performance prediction," in Proc. AAAI Conf. on Artificial Intelligence, 2018.

[22] Q. Liu et al., "EKT: Exercise-aware knowledge tracing for student performance prediction," IEEE Trans. Knowledge and Data Engineering, vol. 33, no. 1, 2021.

[23] S. Minn and R. Nkambou, "Enhanced interpretable knowledge tracing for students' performance prediction with human-understandable feature space," arXiv:2509.18231, 2025.

[24] H. Bi et al., "Quality meets diversity: A model-agnostic framework for computerized adaptive testing," arXiv:2101.05986, 2021.

[25] Ning et al., "Stable and diverse: A unified approach for computerized adaptive testing," in Proc. IEEE CCIS, 2021, doi: 10.1109/CCIS53392.2021.9754532.

[26] M. Khajah, R. V. Lindsey and M. C. Mozer, "How deep is knowledge tracing?," arXiv:1604.02416, 2016.

[27] X. Xiong et al., "Going deeper with deep knowledge tracing," in Proc. Educational Data Mining (EDM), 2016.

[28] "Deep learning for knowledge tracing in learning analytics: An overview," CORE, core.ac.uk/works/46870975.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)