Auto Generation of Diagnostic Assessments and their Quality Evaluation

Soma Dhavala, Chirag Bhatia, Joy Bose, Keyur Faldu, Aditi Avasthi
Embibe Inc.
Bangalore, India
{soma.dhavala, chirag.bhatia, joy.bose, k, a}@embibe.com

ABSTRACT

A good diagnostic assessment is one that can (i) discriminate between students of different abilities for a given skill set, (ii) be consistent with ground truth data and (iii) achieve this with as few assessment questions as possible. In this paper, we explore a method to meet these objectives. This is achieved by selecting questions from a question database and assembling them to create a diagnostic test paper according to a given configurable policy. We consider policies based on multiple attributes of the questions such as discrimination ability and behavioral parameters, as well as a baseline policy. We develop metrics to evaluate the policies and perform the evaluation using historical student attempt data on assessments conducted on an online learning platform, as well as on a pilot test on the platform administered to a subset of users. We are able to estimate student abilities 40% better with a diagnostic test as compared to baseline policy, with questions derived from a larger dataset. Further, empirical data from a pilot gave an 18% higher spread, denoting better discrimination, for our diagnostic test compared to the baseline test.

Keywords

Diagnostic Test Paper, Question Paper Generation, Item Response Theory, Quality Evaluation

1. INTRODUCTION

Learning theory is an important field of research, which incorporates insights from such diverse fields as psychology, pedagogy, neuroscience, and computing to model how well a student learns the taught information. Insights from learning theory are applicable in a wide variety of applications, such as creating intelligent tutor systems and learning platforms, designing courses, designing test papers for exams, and teaching a learner a skill. A prerequisite for any of these activities is to diagnose the current skill level of a new student. This is akin to the *cold start problem* in recommender systems. One proven technique to assessing the current skill level of a new student is to use a set of assessment challenges, most commonly taking the form of a test paper. A good test paper is one that has specific characteristics in terms of *accuracy* and *discrimination*: The test paper should be able to *accurately* diagnose the ability level of a student for the skill set being evaluated, and it should be able to *discriminate* between students of different abilities. Additionally, it should be able to meet these objectives using as few questions as possible.

Soma Dhavala, Chirag Bhatia, Joy Bose, Keyur Faldu and Aditi Avasthi "Auto generation of diagnostic assessments and their quality evaluation" In: *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, Anna N. Rafferty, Jacob Whitehill, Violetta Cavalli-Sforza, and Cristobal Romero (eds.) 2020, pp. 730 - 735

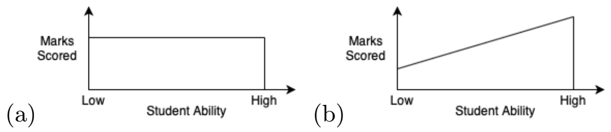


Figure 1: (a) A test in which students of different abilities perform similarly, i.e. get similar scores, is not a good test (b) A better test which can discriminate between students of different ability

If a test paper has questions that many, or all, students answer equally correct or wrong, it will not provide any meaningful information about students. An ideal test paper would reflect student performance such that students with low ability level would get fewer questions correct (lower marks scored) while students with high ability level would get more questions correct (higher marks scored) Fig. 1 illustrates both types of test papers.

In this paper, we present an approach to select questions from a question bank, using configurable policies, that meet the above criteria. We use the selected questions to create a test paper. We then evaluate the generated test paper as per the criteria of accuracy and discrimination, and thus decide on the *goodness* of the policy. Finally, we validate the generated test paper with the best policy on a pilot study of students attempting the test paper. The rest of the paper is organized as follows. Section 2 looks at related work in test paper generation. Section 3 describes our approach to model the problem. Section 4 outlines multiple policies to select questions to compose a test paper. Section 5 discusses the quality evaluation criteria. Section 6 discusses and analyses the results on the simulated and pilot test papers. Finally, Section 7 concludes the paper and presents directions for future work.

2. RELATED WORK

Cen et al. [1] described the architecture of an automated test generation system, using random selection and other strategies to generate questions. They focused on the architecture and not on the effectiveness of the selected questions in diagnosing student ability. A number of studies have been performed on the effectiveness of adaptive test generation, using algorithms to select test questions dynamically from a given pool. Linacre [2] surveyed computer adaptive testing (CAT) in relation to its history and advantages such as needing fewer questions and a shorter time frame than classical tests to diagnose a student's skill level. The questions are selected from a question database, and models such as the Rasch model (a variant of the popular item response theory (IRT) model [3]) are used. CAT starts by presenting questions with average calibrated difficulty at first, then increasing or decreasing the difficulty level of subsequent questions depending on whether the student got the answer right or not. This continues until the system has reached a good estimate of the student's true ability. CAT testing has limitations such as restrictions on re-calibration if the student changes their mind about a previous answer. Another limitation is that the calibration methodology is based on a single parameter, that of difficulty, and not other parameters such as behavior. Kingsbury [4] suggested an approach to improve the adaptive calibration process in a CAT test by considering the student's momentary trait level estimate, in addition to item difficulty, while selecting questions. Also, the estimated difficulty of each question, initially tagged by experts, is continually calibrated based on how many students have answered correctly in the tests given. They found this approach yielded better results in estimating the difficulty of an item. Makransky [5] compared calibration strategies for test questions, including a random strategy and a strategy where the questions are calibrated at the end of a phase or multiple phases, in order to estimate the item difficulty accurately. They implemented the strategies on 1PL and 2PL models of IRT, and found that a continuous updating strategy performed best. Wim [6] surveyed student ability estimation as well as item selection for CATs, using models such as Maximum Likelihood and Bayesian criteria to estimate ability and mean absolute error as the evaluation parameter. Our paper also uses similar models, and additionally realtime data of administered tests to evaluate the accuracy of the models as well as the discrimination ability.

Some researchers have studied factors other than item difficulty when selecting questions. Liu et al [7] found that behavioral factors such as test-taking motivation in students can play an important role in determining learning outcomes. Similarly, Tsaousis [8] suggested a variant of the IRT model in which behavioral parameters like item response time can be incorporated. In another study on behavior as a factor, Jaworski [9] discussed the calibration of control questions in a personalized polygraph test, using emotion and behavior as parameters in selecting the questions. Daroudi et al. [10] surveyed reinforcement learning as a strategy to model the sequencing of instructions in order to maximize learning.

3. PROBLEM FORMULATION

For our analysis, we use a question database taken from Embibe, an online learning platform, along with responses from a set of students on each question. The student's ability is a latent variable, which when estimated with statistically adequate data samples gives a better estimation of the ground truth. For this paper, we consider the ability derived from a larger dataset (in this case, the question database) as ground truth, and abilities derived from a single test as the predicted abilities. For each question in the database, we have the following parameters: Discrimination factor, Difficulty level, Chapter number (represents the chapter number in the syllabus which the question comes from) and Student behavior data for the question. For each student, we have the Ability and Discrimination factor parameters (from the fitted IRT model). The difficulty level and chapter number of each question are annotated by human experts. The anonymized data related to the student responses is collected by the platform.

Out of this ground truth dataset, our objective is to select a subset of questions to assemble into a test paper, which meets the criteria such as best discriminative ability and best match of the identified student ability with the ground truth.

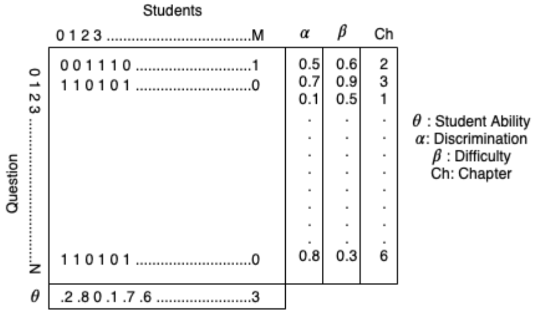


Figure 2: Ground truth dataset of questions taken of a learning platform, with IRT parameters and chapter information

Fig. 2 illustrates the ground truth dataset of questions, along with data on the correctness of students' past responses on each question (whether they answered the question correctly or not). Out of this matrix, we select a small subset of exam questions that can discriminate between students of different abilities.

As per the Item Response Theory (IRT) model, for each question we have a measure of its difficulty and discriminative ability, as well as a measure of the student ability for each student. The standard IRT model gives a relation between the ability and the difficulty, based on one or more parameters and predicts the likelihood that the student will answer that question correctly. We use the 2PL IRT model to calibrate and evaluate our generated test papers.

As per the 2PL IRT model, the probability or likelihood of the student answering a question correctly is given by the following equation:

$$P(X=1|\theta,\alpha,\beta) = \frac{e^{\alpha(\theta-\beta)}}{1+e^{\alpha(\theta-\beta)}} \tag{1}$$

Here, θ represents the student's skill/ability level, α represents the discrimination factor of the question, β represents the difficulty level of the question and P represents the probability that the student will answer correctly.

We infer the IRT model parameters (α, β, θ) from our ground truth dataset by fitting a fully connected deep neural network (modeled using Keras [11] library). The inputs to the neural network are one-hot encodings of the student and question vectors, and the output is the correctness of the student's response for that question, which is a binary value. The IRT parameters are estimated by fitting the neural network using Binary Cross Entropy (BCE) loss. The fitted model is scalable and can handle missing data and imbalanced classes very well. Fig. 3 shows the architecture of the deep neural network for 1PL IRT model. Other IRT models can be realized using the same template.

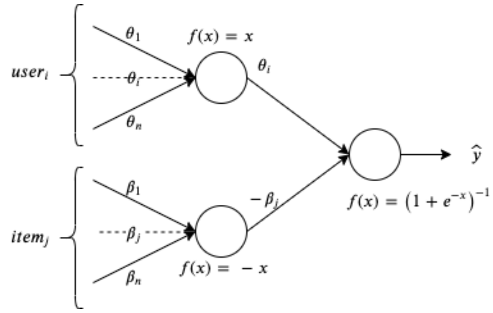


Figure 3: Neural network architecture for estimating 1PL IRT parameter values.

Our problem of selecting an optimal set of questions to form a test, following various constraints, can be modeled in the following manner: Let A be a K-dimensional tensor of size $n_1, n_2, \ldots, n_K$. Each entry of this tensor is either 0 or 1 indicating whether a question with a particular set of attributes was sampled or not. Each dimension of this tensor represents a question attribute, such as a *chapter number*, *difficulty level*, etc. For example, let us say we are interested in creating a test such that four chapters are to be covered, with difficulty levels ranging from 1-10. Then we have $n_1 = 4, n_2 = 10$. Here $A[i, j] = 1$ means that we select a question from chapter i with difficulty level j. We can then set constraints on this tensor to reflect some desired characteristics. For example, the following constraint says that there has to be at least one question from each chapter.

$$\sum_j A[i, j] \geq 1$$

Likewise, we can say that difficulties should follow a certain distribution. Let d_j be the number of questions we like to have whose difficulty level is j. Then,

$$\sum_i A[i, j] = d_j$$

Now we can count how many times the above condition is not met, as a way to measure the quality of the assignment/sampling. Using this, we can form an objective function that evaluates how well the chosen test reflects the above loss, which simply counts the number of disagreements.

$$\min \sum_j I(\sum_i A[i, j] \neq d_j)$$

The above objective function is zero when conditions are met exactly (hard constraint). We can generalize this idea to include constraints about all the question attributes (that are factor variables). Let there be n_k levels for the $k - th$ dimension of the tensor A. These levels represent, for each attribute, the range of values that attribute can take. Let $d_{k(i)}$ be the number of questions needed where the question's $k - th$ attribute has level $c_{k(i)}$. Notice that different attributes can have different number of levels.

$$\min \sum_{k=1}^{K} \lambda_k \sum_{i=1}^{n_k} I(\sum A^k[i] \neq d_{k(i)})$$

Here $\sum A^k[i]$ means that, we select the $k - th$ dimension of the tensor, and its $i - th$ cube, and summing along the cube. In particular, when $\forall_{k(i)} d_{k(i)} = 1$ then Latin HyperCube sampling can be used. The above objective can also be used as a fitness function in genetic algorithms or other search techniques, both stochastic and deterministic, to allocate questions to a test paper. λ_k is a weight parameter which we can tune, for our purposes in this paper we set all the values of λ_k to be equal.

The above objective function, which can be coupled with other IRT based test design objectives, is dealing with domain constraints. Test designs that consider the variance-covariance matrices of parameters in the IRT are also widely used[12]. In particular, the relationship between the item difficulty, discrimination and ability has been addressed from a D-optimality sense. Based on those insights, we formulate a theorem along with proof as below. This is used to develop one of our question selection policies.

Theorem 1. *In a 2PL IRT model, when the difficulty of an item is close to the ability of the person, an item with high discrimination will have high information, and is locally D-optimal.*

Proof. The Item Information function for the 2PL IRT model introduced earlier is given as:

$$I(\theta; \alpha, \beta) = \frac{\alpha^2 e^{\alpha(\theta - \beta)}}{(1 + e^{\alpha(\theta - \beta)})^2}$$

The above equation can be rewritten as:

$$I(\epsilon; \alpha) = \frac{\alpha^2 e^{\epsilon\alpha}}{(1 + e^{\epsilon\alpha})^2}$$

where $\epsilon = \theta - \beta$. Let us consider another item with higher discrimination $\alpha' = \alpha + \delta, \delta > 0$, but with difficulty close to the ability. Then,

$$\lim_{\epsilon \to 0} \frac{I(\epsilon; \alpha')}{I(\epsilon; \alpha)} = \left(\frac{\alpha + \delta}{\alpha}\right)^2 > 1$$

Hence, an item with high discrimination will have higher asymptotic relative efficiency, when the difficulty is in the neighbourhood of the ability. We can claim that such a policy is D-optimal. □

4. TEST PAPER GENERATION

In order to generate a test paper, we propose a set of candidate policies to select questions from the ground truth dataset and assemble the selected questions to form a test paper. All policies assume that the syllabus is covered adequately, i.e. questions are selected from each area of the syllabus. Based on theorem 1 and [12], we select questions with a mix of difficulty levels. We evaluate these policies as per their effectiveness in distinguishing students. To evaluate each policy, we measure parameters such as spread of the scores obtained by different students on the test and how well the diagnosed abilities of the students correspond with the ground truth abilities (by computing the Mean Square Error, Spearman's Rank Correlation and a scatter plot of diagnosed ability vs. true ability). We then choose the best policy to generate a test paper to validate our model by testing on a real world group of students using the same online learning platform where we sourced the ground truth dataset. Fig. 4 shows a flowchart illustrating this method.

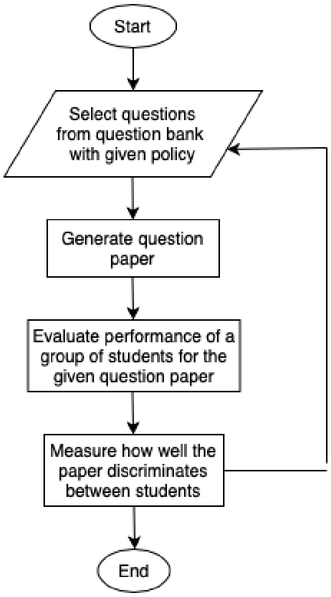

Start
Select questions from question bank with given policy
Generate question paper
Evaluate performance of a group of students for the given question paper
Measure how well the paper discriminates between students
End

Figure 4: Flowchart showing a method to generate test papers from a question bank by selecting questions using a configurable policy, and evaluating how well the policy diagnoses different kinds of students

The candidate policies are described in the following subsections.

4.1 Baseline policy π_{BSP}

As a baseline, we select N questions from the ground truth dataset, by randomly selecting over other question attributes after ensuring a mix of difficulty levels and syllabus coverage. This selection of questions becomes our standard baseline - **BSP**.

4.2 Discrimination only policy π_{DOP}

We use the discrimination parameter values inferred from the fitted 2PL IRT model. We select questions with a mixture of difficulty levels, and the highest values of the discrimination factor for each given difficulty level. We select N questions from the ground truth dataset, ensuring syllabus coverage (at least one question from each chapter), but also ensuring that the overall discrimination factor of the questions is maximized. This policy, **DOP**, ensures that high discrimination questions are selected, at any given difficulty level.

4.3 Discrimination+behavior policy π_{DBP}

In this policy, we incorporate behavior parameters along with discrimination, difficulty and syllabus coverage, while selecting questions. Behavior parameters refer to the student behavior when taking the test, captured by the learning platform. These include parameters such as number of questions that are likely to be answered too fast and incorrectly, or questions that are answered too slow but correctly, among others. The questions are tagged as per which parameters are mostly manifested by students answering that question and the top questions from each parameter are selected. This policy, **DBP**, ensures that high discrimination questions as well as student behavior are taken into account.

5. QUALITY EVALUATION CRITERIA

In order to evaluate the generated test papers, we use two criteria: *accuracy* and *discrimination*. *Accuracy* refers to how closely the diagnosed ability using the student responses to the test paper corresponds to the actual ability of the students. We use the RMSE between the ground truth and the inferred ability as a measure of the accuracy. The rank correlation between the ground truth rank and the estimated rank, and scatter plot between the inferred and ground truth ability, also indicate the accuracy.

Discrimination measures how successful the test paper is in discriminating between students of different abilities. We evaluate the accuracy and discrimination for the generated test papers on a subset of M students (evaluation student set) from our ground truth dataset. We use the spread and distribution of scores as a measure of the discrimination.

Evaluation using RMSE
Using the IRT model, we predict the probability of each student in the evaluation set answering the questions correctly, and compute the average ability from the scores of the students if they were to attempt the generated test paper. We also determine the ground truth ability of each student from the IRT model. Finally, we compute the root mean squared error (RMSE) between the ground truth ability and inferred ability to get a measure of the accuracy.

Evaluation using Spearman's ρ
Here we sort the abilities of students obtained from the ground truth data and from the generated test, and determine the rank correlation ρ between the two ranks.

Evaluation using scatterplots
We plot the abilities of students, inferred from the ground truth, against the diagnosed abilities from the generated test papers. The degree of scatter gives an indication of how much the ability matches the inferred ability.

6. RESULTS AND DISCUSSION

We have an initial ground truth dataset, obtained from the online learning platform, of close to 1300 questions and 1700 students along with the responses for each of the students on each question, along with the derived IRT parameters. From the dataset, we filter those students who have attempted less than 25% of the questions in each paper, so that we have sufficient data to estimate their abilities.

6.1 Simulated tests

We choose 75 questions from the ground truth dataset for each policy, in effect simulating a test of 75 questions. In selecting the questions, we ensure syllabus coverage. Table 1 shows various test statistics.

Table 1: Comparison of test results from simulated tests generated by the three policies

	BSP	**DOP**	**DBP**
No. of students	312	312	312
No. of questions	75	75	75
Max. score possible	300	300	300
Max. score achieved	188	251	218
Min. score achieved	-22	-17	-23
Score at 95th percentile	118.4	148	144.5
Score at 5th percentile	3	4	0
Avg. score achieved	60.80	79.25	77.18

Fig. 5 shows the comparison in spread of student scores for the simulated tests on test papers generated using the three policies.

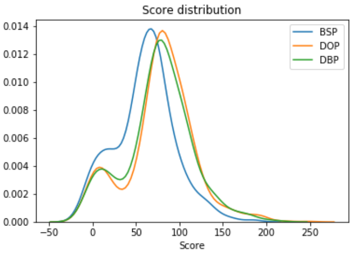


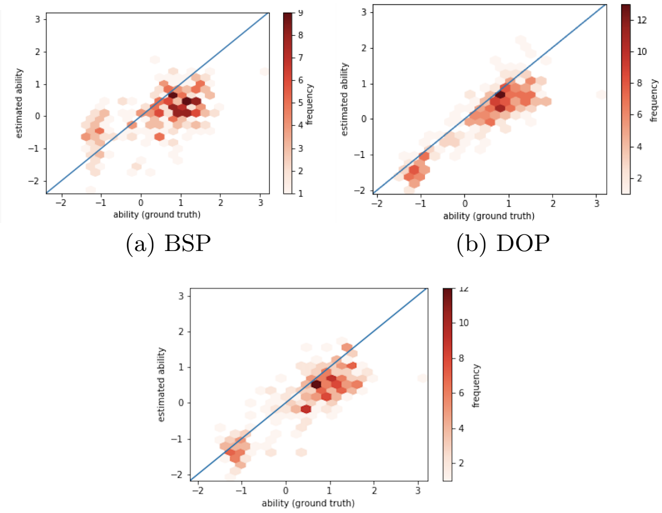
Figure 5: Comparison of the spread in score obtained from the simulated tests generated using following policies: BSP, DOP and DBP.

For each of the test papers selected using different policies, we evaluate the accuracy and discrimination as mentioned in the previous section. We also calculate the ability of each student from the remaining questions in the ground truth dataset, which are not included in any of the generated test papers.

Table 2: Comparison of RMSE (inferred ability and ability from ground truth) and rank correlation ρ in tests generated by different policies

Policies	**RMSE**	**Rank corr ρ**
BSP	0.844	0.59
DOP	0.549	0.83
DBP	0.615	0.788

We find that the DOP test gives 24.8% better spread of scores (score at 95^{th} percentile of students - score at 5^{th} percentile), as compared to the BSP baseline. DBP test gives 25.2% better spread. The mean squared error for the inferred ability of the students compared to the ground truth ability is 0.844 for the BSP, 0.549 for the DOP and 0.615 for DBP. Table 2 shows the comparison between the policies with respect to root mean square error (RMSE) and Spearman rank correlation. We obtain a 40% higher correlation for the DOP policy as compared to BSP.



(a) BSP (b) DOP

(c) DBP

Figure 6: Scatterplots of the abilities of the generated test papers, against the ground truth abilities. Degree of scatter is highest for the BSP paper

Fig. 6 shows a scatterplot of the abilities of the student from the test papers using the three policies, plotted against their ground truth abilities. We can see that the paper generated using BSP policy has the highest degree of scatter and the DOP paper has the lowest, i.e. it most closely matches the ground truth abilities of the student.

6.2 Analysis of the simulated test results

Comparing the policies from the score distribution in the generated test papers, we can see that the DOP and DBP policy give a better spread of scores than BSP, meaning they are better in discriminating between students of different abilities. Tests generated by both DOP and DBP policies also had a higher rank correlation than the BSP test, meaning we get a better accuracy at diagnosing the ability of the students.

The DBP test had a lower spread and lower rank correlation as compared to the DOP test. This could be because we only used the standard 2 PL model of IRT, without any modifications to include behavior parameters. Moreover, behavior parameters, such as time spent not attempting questions, give a more holistic view of how the student performs in a test (such as indicating the confidence level of the student) than simply the academic performance i.e. how many questions the student answered correctly. Perhaps future test papers could be designed in a way that takes into account these factors when computing the student's score.

6.3 Pilot test

To further validate our model, we conducted a pilot study as follows: We selected a group of M students and asked each student to attempt two test papers, using the same online

Table 3: Comparison of pilot test results generated by BSP policy and DBP policy

	BSP	**DBP**
Number of students	98	99
Number of questions	30	30
Max/Min/Total score	92/0/120	111/-1/120
Score at 95th/ 5th percentile	76/5	85/1
Average score	43.94	42.56

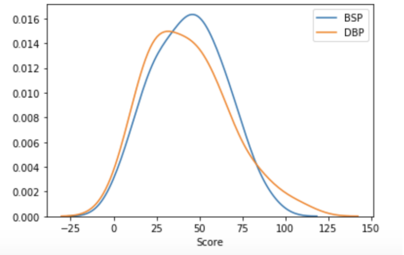
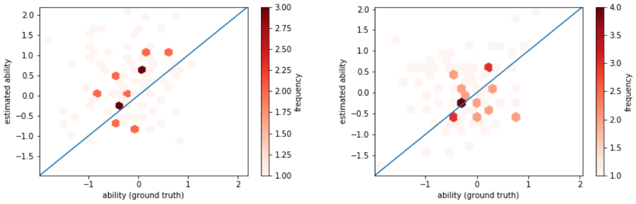


Figure 7: Scores distribution of students in the DBP generated pilot test vs BSP pilot test

learning platform we used for the earlier test generation. For the first paper, we generated the questions using BSP policy and for the second, we generated the questions using DBP policy. We then compared the spreads of scores for these test papers. The results are shown in table 3.

On the pilot test papers generated using the two policies, we found that the DBP test gives 18% higher spread of scores (95^{th} percentile score - 5^{th} percentile score), as compared to the BSP test. The mean squared error for the inferred ability of the students compared to the ground truth ability was 1.08 for the BSP test, and 0.86 on DBP. This is 20% less RMSE for DBP compared to BSP. From the scatterplots in fig. 8 for the inferred vs ground truth ability, we can further confirm that the degree of scatter is lower in the DBP pilot test and higher in the BSP pilot test paper. This confirms that the DBP test paper more accurately reflects the ability of the student, and is also better at discriminating between students of different abilities. The spread of scores in the DBP is better than that of the BSP policy. This validates our findings from the simulated tests, where also we obtained a better spread for the diagnostic policies (DOP and DBP). Moreover, the higher accuracy of the inferred ability for the DBP pilot test is confirmed by a lower value of the RMSE and lesser degree of scatter compared with BSP.



(a) BSP (b) DBP

Figure 8: Scatterplots of the inferred abilities of the pilot test papers, against the ground truth abilities

7. CONCLUSION AND FUTURE WORK

In this paper, we proposed a few policies to generate test papers by selecting a list of questions from a question database. We validated the policies by a pilot test of test papers generated using two policies. We found that the policy of selecting questions based on highest discrimination ability for a given difficulty level yielded the best results.

In future, we intend to extend the IRT model to include behavioral parameters and further validate our method of selecting policies with more candidate policies and a larger sample size of students.

8. ACKNOWLEDGMENTS

The authors express their gratitude to Anwar Sheikh and his team for helping us conduct the pilot study.

9. REFERENCES

[1] Guang Cen, Yuxiao Dong, Wanlin Gao, Lina Yu, Simon See, Qing Wang, Ying Yang, and Hongbiao Jiang. A implementation of an automatic examination paper generation system. *Mathematical and Computer Modelling*, 51, 2010.

[2] John Michael Linacre. Computer-adaptive testing: A methodology whose time has come. *Development of computerized middle school achievement test*, 69, 2000.

[3] Frank B. Baker. *The basics of item response theory.* ERIC, USA, 2001.

[4] G. Gage Kingsbury. Adaptive item calibration: A process for estimating item parameters within a computerized adaptive test. *GMAC conference on computerized adaptive testing*, 2009.

[5] Guido Makransky. An automatic online calibration design in adaptive testing. *Journal of Applied Testing Technology*, 11(1), 2014.

[6] Wim J. van der Linden and Peter J. Pashley. Item selection and ability estimation in adaptive testing. *Elements of adaptive testing, Springer*, 2009.

[7] Ou Lydia Liu, Brent Bridgeman, and Rachel Adler. Measuring learning outcomes in higher education. *Educational Researcher*, 41, 2012.

[8] Sideridis GD Tsaousis, I and AA Sadaawi. An irt–multiple indicators multiple causes (mimic) approach as a method of examining item response latency. *Frontiers in psychology*, 9, 2018.

[9] Ryszard Jaworski. Personalization and calibration of the control question in the control question test. *Journal of Forensic Identification*, 61(5), 2011.

[10] Shayan Doroudi, Vincent Aleven, and Emma Brunskill. Where's the reward?: A review of reinforcement learning for instructional sequencing. *International Journal of Artificial Intelligence in Education*, 2019.

[11] Keras documentation. https://keras.io, 2015.

[12] Ronald K Hambleton and Wim J Linden. *Handbook of modern item response theory. Volume two: Statistical tools.* CRC Press, USA, 2016.