Implementation of computer adaptive testing based on adaptive recommendation systems

Jufeng Yang
Supervised by: Jin Zheng

July 2, 2023

Abstract

The way people live and work has changed dramatically over the past few years as a result of the epidemic. Much of the work, learning, testing and other social behaviour that used to take place offline has been shifted to online. As a result, a large number of platforms to support online work and learning have gained significant support in the last few years, while a large number of online competency assessment platforms have started to develop rapidly. Based on a large dataset of listening tests, this project will apply NLP techniques, combined with IRT(Item response theory) algorithms, to enable users to test their abilities and recommend tests questions that reinforce their weaknesses. The project report will present a new IRT algorithm and a new readability formula. To implement the listening ability test, I will design and build a simple GUI that will be used by the user to answer the test questions and then the backend will validate the user's answers and assess the user's ability. After completing the final questions, the platform will push out questions that reinforce the user's ability. The project will ultimately result in a complete and independent platform, based on a new IRT algorithm, and a readability formula that will enable accurate question recommendation and assessment of the user's true English listening ability level. The ultimate in accurate weakness assessment and fixing of user weaknesses. This will be the prototype of a complete learning support platform, based on which additional learning materials and assessment questions can be added to the database, ultimately allowing for a comprehensive, extensive knowledge-based learning support platform.

Ethics statement: This project fits within the scope of the blanket ethics application, as reviewed by my supervisor Jin Zheng.

I have completed the ethics test on Blackboard. My score is 12/12.

1 Project Plan

This section focuses on the background of the project, an overview of CAT (Computer Adaptive Testing), the challenges encountered in implementing CAT, the data, and the purpose of the project.

1.1 Background

In the last two years, the epidemic has prevented people from attending work properly, completing their studies properly or taking exams properly. However, based on the mature environment of the Internet, all these things that cannot be done properly will be solved by the software tools that have been developed, which has led to the rapid development of digital technology during this period [1]. It is also for this reason that online examinations are becoming popular and generally accepted. The recent popularity of the multi-neighbouring British Proficiency Test has been recognised by a large number of colleges and universities, and this has therefore led to the application level acceptance of computer adaptive testing techniques.

The core recommendation algorithm that I have chosen to implement for the adaptive recommendation function in the current project is the Item response theory (IRT) algorithm. The basic algorithm of IRT was originally proposed in the 1950s and 1960s by Frederic Lord and other psychometricians whose original aim was to develop a method for assessing respondents consisting of different test items [2].IRT is a statistical model which incorporates two features [2].

- The IRT assesses test-takers according to their abilities or potential characteristics and gives them a score
- The performance of an item is assessed by its characteristic curve.

Other technical points or algorithms will be described in detail in subsequent sections. These include modified and corrected versions of the IRT and more complex applications of the model (MIRT, T-MIRT).

The rise of adaptive recommendation technology has also led to a new idea for ability assessment tests, the Computerised Adaptive Test(CAT). In fact, the concept was earlier called 'customised testing', proposed by Lord and 1970, the authors of IRT [3]. is that, based on an adaptive recommendation model, a computer selects questions in a database that are appropriate to the ability of the test taker to assess ability. The practice of this technique would no longer limit the assessment to those with very high and weak abilities, and the test could be floated according to ability, which is the biggest improvement over traditional tests.

1.2 CAT Overview

The CAT is tied to the GUI, which is the platform on which the user takes the test and the means by which the user's findings are transmitted to the

backend. The CAT is divided into several steps, as shown in Figure 1, including an initial proficiency test, test selection, a new proficiency test, determination of whether the stop test criteria, ending or returning to the ability test, determining weaknesses, and recommending training tests.

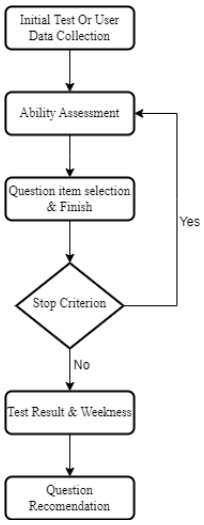

Figure 1: CAT Flowchart

Initial Test or User Data Collection In this step, CAT can select a series of questions of varying difficulty from the dataset and form an initial proficiency estimate based on the user's answers. Alternatively, CAT can generate a questionnaire asking some questions related to user's English proficiency level to generate an initial ability value.

Ability Assessment In the initial step, an initial proficiency estimate is generated based on the initial test or user survey. In a subsequent step, if the conditions for stopping the assessment are not met, a new assessment of the student's ability will be made based on the new test answers.

Question item selection & Finish The backend algorithm will make a selection of test questions based on the user's current ability. The tester will complete a new test in this section and the backend will determine if the stopping criteria has been met, if not, the step will be repeated again as shown in Figure 1.

Stop Criterion Used to determine if the test can be ended, if not, the test will continue, if so, the test result and the ability that the tester is not good at will be returned.

Test Result & Weakness Complete all tests, return testers' results and testers' drawbacks.

Question Recommendation Based on the weaknesses and the final ability assessment, test questions from the dataset are recommended that are suitable for the test taker's training.

All test data were processed by NLP to remove noise and used for topic modeling for subsequent test recommendation. In addition, the difficulty of the questions was assessed by comparing them with various readability formulas [4, 5, 6, 7, 8], as well as the impact factor of the difficulty of the English listening questions [9]. A final difficulty score was assigned to all questions.

The core question recommendation algorithm [10] is based on the test taker's answers, ability, item topic, difficulty of the question, by calculating expectations and applying logistic regression. The algorithm is described in detail in the literature review that follows.

The core algorithm for assessing the ability of testers is based on IRT, with improvements made to IRT to make it more relevant to the current project, and several improvements to IRT have been made, including those based on MIRT [11], T-SKIRT [12], T-BMIRT [13] and others. In the current project, however, new IRT algorithms will be constructed by combining and improving several models.

1.3 CAT Challenge

Topic modeling Topic models do not capture the specific content of topics, but only the clustering of an index of similar topics. The biggest challenge is to identify the weaknesses of the test takers and to provide accurate recommendations of relevant topics and questions of relevant difficulty for these weaknesses.

Accurate test recommendations After the relationship between the questions based on the topic modelling, it is a difficult task to assess the weaknesses of the test takers in relation to their answers. Furthermore, it is difficult to select the recommended test questions based on the weaknesses of the test takers.

Question Difficulty Level A creative formula for assessing English listening difficulty was generated based on existing formulas for assessing the readability of texts, combined with factors affecting English listening difficulty.

IRT Algorithm Inovation The innovation of the IRT algorithm to make it more relevant to the current project is a difficult task. The difficulty of the algorithm lies in the mathematical principles, the IRT algorithm is the basic form of the algorithm and this does not need to be changed, but building on the basic IRT algorithm and adding more other mathematical relationships is what makes IRT a more appropriate algorithm.

1.4 Data

All the data to support the running of the project came from my supervisor, Jin Zheng. The data set consists of a large amount of text, each dictionary contains a key and a value, the key is the content of the listening material and the value is a dictionary, the key of the dictionary is the question and the value is the option. The text of the listening material is converted into speech, which is saved locally, and there is a play button in the GUI to support the playback of the generated speech. The questions and options are also displayed in the GUI for the tester to choose from.

There is some noise in the data and after a brief analysis of the text, a regular formula is used to remove some of the text content that is not relevant to the speech. The option text is also sliced and diced and the options are kept in a logical list. Ultimately, the listening material, the listening questions, and the question options are stored in separate lists, thus forming a new dataset.

1.5 Project Aim

The project is based on techniques such as IRT, transformer, Readability formulas and more, and adds more complexity to IRT and Readability formulae to achieve a platform that provides accurate results of proficiency tests and weakness fixes.

Using appropriate Topic modelling, such as LSA [14], LDA [15], PAM [16], etc., an accurate classification of topics is obtained, and in the final session of weakness repair, this allows the test taker to know which area of conversation is unfamiliar in reality, and therefore to focus on the relevant topic. Once the relationship between the topics has been obtained, an accurate recommendation will be obtained based on the test taker's answers, combined with a recommendation algorithm which will be described in detail in the literature review in the follow-up. The combination of these two processes will ensure the strong functionality and accuracy of the assisted learning platform.

The innovation of a Readability Formula allows for accurate determination of the difficulty of the listening questions. This is a key part of the system, as a high degree of accuracy in difficulty ensures the accuracy of all the content that follows. The development of the Readability Formula was one of the key aims of the project.

Another important task was to innovate an accurate and suitable IRT algorithm for the project at hand. A high performance IRT core algorithm is necessary when the difficulty of the entities in the database possesses a high degree of accuracy. The final conclusion is an accurate assessment of the test taker's weaknesses and ability scores based on the available questions and obtaining the test taker's ability scores.

Finally, the final output of the project will rely on a stand-alone CAT platform based on a GUI that will enable testers to complete the English proficiency test, which is the key pathway for all data collection and data delivery. The back end of this will be a data processing, predictive code base for this performance.

2 Literature Review

This section covers some of the key theoretical underpinnings that will help the project progress. It covers the model of IRT and its extensions, the structure of CAT, and the idea of implementing an adaptive recommender system. Finally there are relevant technical points about (NLP) natural language processing.

2.1 IRT and its extension algorithm

T-BMIRT is a temperal mutil-dimensional IRT model. Jiankun and Wenjun [13]in their paper refer to a variety of IRT models and the flaws and advantages of their respective counterparts. Among them are IRT (Item Response Theory), MIRT (Mutil-diemension IRT), T-MIRT (Temperal MIRT), and the innovative model of the thesis, T-BMIRT (Temperal-Blended MIRT). These models led me to look for directions to implement adaptive recommendation algorithms for learning resources. The thesis also deals with the mathematical principles related to IRT, which have different mathematical principles depending on the single dimension and the multidimensional dimension. The thesis mentions that an important task in online education is the estimation of the students' abilities and the characteristic components of the educational content, therefore the test or validation method of online exams is the most important assessment method for the model of the project. Furthermore, the model used in the project takes into account time, so that T-BMIRT does not only consider the user's interaction with the test, but also the user's ability to improve by interacting with the learning resources. Ultimately, the user's learning process is used as the raw data to predict changes in the user's ability and the test is used to verify changes in the test taker's ability. By comparing the two results and calculating the AUC and ACC. thus the method used by Jiankun and Wenjun to assess the model is by predicting the AUC and ACC of the student's next response on the test. this article was important in inspiring my current project. the IRT can be used to assess and predict the user's ability and therefore the algorithm is useful for designing a CAT (Computerized Adaptive Test) is highly enlightening.

T-SKIRT is a a temporal, structured knowledge, IRT-based method. Chaitanya and Karlin's model [12] is similar to the T-BMIRT in that it is a time-series model. Unlike the T-BMIRT model, Chaitanya and Yan take into account more details about the student's learning process, forgetting and other factors, and the model can be used to recommend future learning materials through the student's testing. In order to achieve this set of considerations, the authors use 2PO (2 parameter Ogive), which is different from the usual 2PL (2 parameter Logistic) model. Based on 2PO, the authors consider the change in student ability as a Wiener Process, thus introducing the Temporal component. The authors compare the effects of several models in their experiments, among them SPC (Students Percentage Correct), 2PO-IRT, 2PO T-IRT, Factorial MVN Temporal 2PO IRT, Correlated MVN Temporal 2PO IRT (T-SKIRT). He concluded that the accuracy in predicting the student's next response was

as high as 0.7478, with an AUC of 0.8194. Chaitanya and Karlin's report has increased my exploration and understanding of the project. It also increased my understanding of the mathematical principles of the model, which helped in the final implementation process.

MIRT is a mutil-dimension IRT based method. Philip [11] has developed a multidimensional IRT model based on R combined with a sophisticated GUI to implement CAT. this article is also an exposition of the technology and begins by describing the process of implementing MCAT (mutil-dimension CAT) and the mathematical rationale for the key steps involved. Philip describes in detail the mathematical principles associated with the design of these steps, which I will describe in a later section. The steps in the design of the CAT mentioned in this article are similar to those of Douglas and Plinio, which gives me more confidence in the structure of the current project. However, the content of Philip's development is too deeply tied to the GUI and there are not many APIs that can be exploited, so the paper can only provide structural and theoretical insights.

Marginal maximum likelihood(MML) estimation of IRT HARUHIKO's [17] main thesis is a study on the design of MML to estimate the equivalence coefficients of IRT. Haruhiko's approach is similar to that of Yoyoda in that both use the MML method but the only difference is in the equivalence design. The paper also provides a very detailed mathematical rationale for the MML model, as well as the mathematical rationale for the MML estimation method. The hint for me in this paper is in the innovation points that provide food for thought on how to improve the model performance of IRT or improve the accuracy of the prediction of candidate ability.

2.2 CAT structure and Adaptive recommendation idea

IRT based CAT In this thesis, Quan [18] describes the process of implementing CAT and the mathematical principles of the corresponding algorithms that are needed to implement all the key components of CAT. The aim of the thesis project was to identify children with autism and intellectual disability, in which he found the three-parameter unidimensional model to be well suited to the project, and he also found in his experiments a very suitable selection strategy and a marginal maximum likelihood estimation method among existing parameter estimation algorithms. In the course of his experiments, the author discusses a wide range of ability estimation methods and selection strategies, including estimation methods such as great likelihood estimation, EAP, MAP, etc., as well as extremely informative methods, Bayesian selection methods, 0-1 integer linear programming methods, outlier weighting methods, and other selection methods. In addition, he also sets out relevant selection strategies of his own choosing. Quan's project process and the mathematical principles mentioned in the paper helped me understand the IRT model and helped me to construct the model and anticipate the project process.

Adaptive Recommendation in English Fragmented Reading The main objective of [19] is to build a system that analyses user behaviour, habits, preferences, weaknesses and other factors to recommend fragmented English learning resources to users; Chen and Wang believe that fragmented learning is an area that should be explored in an era where entertainment is prevalent and people are using fragmented time to learn. The system provides an initial self-selection scheme, allowing users to select their preferred learning areas, their ability level and other factors, and this information will be used to initially determine the appropriate recommended content. The system will analyse the user's behaviour according to the algorithm and change the recommended content to better match the user's current preferences and abilities. Once the data layer has reached a certain level of richness, the user's initial recommendations will not simply be pushed to the user's choices. Wang and Chen apply a collaborative filtering method that matches the user's choices with those that match theirs, and uses the matched user's recommendations to recommend new users, which will improve the accuracy of the recommended content. The direction of this thesis has helped to define my direction and broaden the areas I need to explore. For my current dataset, I will not just complete a CAT test, I will split the dataset into 2 parts based on topic and difficulty, one part for training candidates and the other for testing. The test results will then be used to recommend training topics that the user is not good at.

Application and Simulation of CATs Through the Package catsim [20] is a technical paper in which Douglas and Plinio develop a Python library based on IRT, called catSIM, which performs the key steps of CAT (Computerized Adaptive Test) design, IRT model, selection strategy, termination strategy, etc. In addition to the key technical explanations, the most critical part of the paper is the structural design of the CAT. Douglas and Plinio state that a CAT consists of an initial strategy, an IRT capability estimate, a selection strategy, and a termination condition, in that order. The final composition is for a user assessment. The paper also explains in detail the mathematical principles of the relevant metrics and models, such as the four-parameter logic model, ICC (Information characteristics curve), SEE (standard error of estimation), reliability, etc. The CAT architecture provided in the thesis, together with the associated code and mathematical principles, has been a great inspiration for the development of my project. In my work, I will need to innovate the algorithms and the overall output of the CAT, so the existing thesis will not be able to meet my needs, and this thesis will be the cornerstone of my innovation.

Recommendation for English multiple-choice cloze questions based on expected test scores This thesis [10] creates a test recommendation algorithm for multiple-choice questions in the context of the Test of English for International Communication (TOEIC). In the paper, Iwata el at recommends test questions based on the user's initial test results and the user's learning log to help the user improve on specific aspects of the topic. The use of a greedy

algorithm to recommend questions based on the expectation of maximising test scores in the thesis will allow for a flexible number of questions to be handled.

$$
\begin{aligned}
E(z|x) &= \sum_{i:x_i=-1} S_i P(i) P(y_i = 1 | x_i = 1, z) \\
\hat{j} &= \arg \max_{j:z_j=0} E(z^{+J}|z)
\end{aligned}
\tag{1}
$$

where S_i represents the score allocated to question i, $P(i)$ represents the probability that question i is asked in future test $P(i) + P(\bar{i}) = 1$ in which $P(\bar{i})$ represents the probability that question i is not asked in future tests, $P(x_i = -1)$ represents the probability that question i is incorrectly answered before the studying phase, and $P(y_i = 1|x_i = -1, z)$ represents the probability that question i is correctly answered after the studying phase when question i is incorrectly answered before the studying phase and questions z are recommended in the studying phase.

In addition, the authors have also refined the model by using logistic regression to predict the probabilities in the above equation $P(y_i = 1|x_i = 1, z)$ and then using maximum likelihood estimation to predict the unknown parameters in the logistic regression equation.

$$
\begin{aligned}
P(y_{ni} = 1 | x_{ni} = & 1, z_n) = \frac{1}{1 + \exp(-(\mu_i + \theta_i^\top z_n)}) \\
L(\Theta) = & \sum_{n \in N} \sum_{i \in V} (I(x_{ni} = -1 \wedge y_{ni} = 1)(\mu_i + \theta_i^\top z_n) - \\
& I(x_{ni} = -1) \log(1 + \exp(-(\mu_i + \theta_i^\top z_n))))
\end{aligned}
\tag{2}
$$

where μ_i and $\theta_{ij} = (\theta_{ij})_{j \in M}$ are unknown parameters. $\Theta = \{\mu_i, \theta_{i i \in V}\}$. The value of the output of the logistic regression is calculated as the key part of the expectation. Finally, the paper also tested backtesters who had studied the recommended test and obtained the result that the testers' scores were improved. Therefore, they believe that this recommendation method is promising and will be an effective learning aid.

2.3 NLP tech in readability formula and topic model

ATOS' Readability Formula The School Renaissance Institute [21] believes that readability formulas are used to assess the difficulty of a book in order to help children choose books that are more appropriate for their reading ability. The report briefly explains a number of reading ability formulas including the Dale-Chall Readability Formula, Degrees of Reading Power Values (DRP), Flesch-Kincaid Formula, Fry Index, Lexile Framework. Among other things, the detailed theory of ATOS is explained in detail in the report. The above formulas share a common flaw in that they all use semantic or syntactic measures of difficulty, but they will all give similar results for general books. However, for some specific areas of books they will be more misleading. The variables used

in ATOS are 'Words per sentence', 'Average grade level of words', 'Characters per word". In fact, many other readability formulas use the number of syllables to judge the difficulty of a word. In ATOS, the number of syllables is not used, but rather the number of characters per word. This is similar to the Coleman-Liau Readability Formula. This paper gave me a first insight into the partial readability formula.

The Coleman-Liau Index Since early computer programming was not well developed and obtaining the syllables of a word was much more complicated than obtaining the characters of a word, Coleman and Liau [5] thought that the criteria for assessing the difficulty of a word could be changed from the number of syllables to the number of characters. In addition, they believe that it is feasible to use this formula to set readability for public school reading materials.

$$\begin{aligned} Gradelevel &= -27.4004 \cdot estimatedcloze\% + 23.06395 \\ Estirlatedcloze\% &= 141.8401 - .214590 \cdot L + 1.079812 \cdot S \end{aligned} \tag{3}$$

where estimated cloze % = percentage of deletions that can be filled in by a college undergraduate, L = number of letters per 100 words, and S = number of sentences per 100 words. multiple R of this formula is .92.

Determining the Difficulty Level of Listening Tasks In this paper Zohre [9] mentions a large number of factors that influence the difficulty of a listening task. They include the nature of the input, the nature of the assessment task, the listening instruction, and the individual listener factor. The nature of the input relates to the nature of the listening material. There are 11 aspects to the design but I think the factors that are more important to the current data are the difficulty of the listening material, the speed of speech, the familiarity of the topic and the format of the listening material (live or recorded). The nature of the assessment task relates to the type of topic, the amount of listening material, the context, and for both the current data and my design, the type of topic is a choice, but the number of topics is inconsistent. The amount of listening material is consistent and the number of times the listening material is played is consistent. The listening instruction and individual listener factors are both individual ability factors for the test and are not considered in this item. The paper mentions listening material difficulty, topic familiarity, and number of listening questions as the most important influencing factors in the current project.

Evaluating Text Complexity and Flesch-Kincaid Grade Level In this paper [4], Solnyshkina et al. present an automated tool for T.E.R.E., which involves five text complexity parameters that can be used to determine the complexity and readability of a text. The five parameters are narrativity, syntactic simplicity, word concreteness, referential cohesion and deep cohesion. Solnyshkina based the results on eight texts with similar scores given by the Flesch-

Kincaid Grade and compared these scores. and compared the way in which the five text complexity parameters of these eight texts were related to each other. The narrativity depends on the average number of verbs in each phrase, the presence of common words and the overall story-like structure. Syntactic simplicity depends on the average number of subordinate clauses in the sentence, the number of words preceding the main word in the sentence. Word concreteness depends on the ratio of abstract words to concrete words in the text. Solnyshkina's findings suggest that the grammatical simplicity of T.E.R.E has little to do with the Flesch-Kincaid readability formula, but that the abstractness/concreteness of T.E.R.E has a relatively strong correlation. In addition, the narrativity of the text tended to be inversely related to the deep cohesion of the text. Referential cohesion and deep cohesion did not correlate with the Flesch-Kincaid Grade Level.

Combining Readability Formulas and Machine Learning for Reader-oriented Evaluation of Online Health Resources Liu et al.[8] argue that traditional readability formulas are an objective, user-independent way of evaluating text. Liu et al. tested a total of nine ML models, namely XGBoost Tree, Random Trees, Bayes Net. Random Forest, C&R Tree, C5.0, CHAID, Quest and Neural Net. In this report, readability was divided into two levels: high readability and low readability, and the models were then trained to dichotomise the data. From the training results, random forests had the best performance on the training set and random trees showed the best performance on the test set. The paper's approach combines traditional readability formulas and ML models, which can, to some extent, re-evaluate or complement traditional methods to achieve results similar to human evaluation.

BERTopic: Neural topic modeling with a class-based TF-IDF procedure Grootendorst [22] has developed a new technique called BERTopic, a topic model, to extend this process by developing a coherent topic model. Extending this process by developing a variant of the class-based TF-IDF, BERTopic generates topic representations in three steps. First, each document is transformed into its embedding representation using a pre-trained language model. Then, the dimensionality of the resulting embeddings is reduced to optimise the clustering process before these embeddings are clustered. Finally, a custom class-based variation of the TF-IDF is used from the clustering of documents. The model's topic representation is modelled based on the documents in each cluster, where each cluster is assigned a topic. Alternatively, BERTopic supports dynamic topic modelling. By applying the different topic modelling models to the three data datasets, both BERTopic and CTM were found to have very good performance. However, BERTopic is not able to handle single documents with multiple topics very well. However, for the data in the current project, there is no significant negative impact.

References

[1] R. De', N. Pandey, and A. Pal, "Impact of digital surge during covid-19 pandemic: A viewpoint on research and practice. int j inf manage," *Int J Inf Manage*, vol. 55, 2020.

[2] C. Zanon, C. Hutz, and H. Yoo, "An application of item response theory to psychological test development," *Psicol. Refl. Crít.*, vol. 29, no. 18, 2016.

[3] W. D. Way, "Practical questions in introducing computerized adaptivetestingfork-12assessments," Pearson Educational Measurement, Tech. Rep., April 2006.

[4] M. I. Solnyshkina, R. R. Zamaletdinov, L. A. Gorodetskaya, and A. I.Gabitov, "Evaluating text complexity and flesch-kincaid grade level," *Social Studies Education Research*, vol. 8(3), pp. 238–248, 2017.

[5] M. Coleman and T. L. Liau, "A computer readability formula designed for machine scoring," *Journal of Applied Psychology*, vol. 60(2), pp. 283–284, 1975.

[6] K. Grabeel, J. Russomanno, S. Oelschlegel, E. Tester, and R. Heidel, "Computerized versus hand-scored health literacy tools: a comparison of simple measure of gobbledygook (smog) and flesch-kincaid in printed patient education materials," *J Med Libr Assoc*, vol. 106(1), pp. 38–45, 2018.

[7] E. Smith and R. Senter, "Automated readability index," AEROSPACE MEDICAL RESEARCH LABORATORIES, Tech. Rep., November 1967.

[8] Y. Liu, M. Ji, S. Lin, M. Zhao, and Z. Lyv, "Combining readability formulas and machine learning for reader-oriented evaluation of online health resources," *IEEE Access*, vol. 9, pp. 67 610–67 619, May 2021.

[9] Z. Mohamadi, "Determining the difficulty level of listening tasks," *Theory and Practice in Language Studies*, vol. 9, no. 6, pp. 987–994, June 2013.

[10] T. Iwata, T. Kojiri, T. Yamada, and T. Watanabe, "Recommendation for english multiple-choice cloze questions based on expected test scores," *KES Journal*, vol. 15, pp. 15–24, 03 2011.

[11] R. P. Chalmers, "Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications," *Journal of Statistical Software*, vol. 71, no. 5, p. 1–38, 2016.

[12] C. Ekanadham and Y. Karklin, "T-skirt: Online estimation of student proficiency in an adaptive learning system," *arXiv:1702.04282v1*, Feb 2017.

[13] H. Jiankun and W. Wenjun, "T-bmirt: Estimating representations of student knowledge and educational components in online education," in *2017 IEEE International Conference on Big Data (Big Data)*, 2017, pp. 1301–1306.

[14] P. Foltz, "Semantic processing: Statistical approaches," in *International Encyclopedia of the Social and Behavioral Sciences*, N. J. Smelser and P. B. Baltes, Eds. Oxford: Pergamon, 2001, pp. 13 873–13 878.

[15] D. M. Blei, A. Y. Ng, and J. Michael I., "Latent dirichlet allocation," *Journal of Machine Learning Research*, pp. 993–1022, 2003.

[16] W. Li and A. McCallum, "Pachinko allocation: Dag-structured mixture models of topic correlations," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 577–584.

[17] O. HARUHIKO, "Marginal maximum likelihood estimation of item response theory (irt) equating coefficients for the common-examinee design," *Japanese Psychological Research*, no. 2, pp. 72–82, 2001.

[18] M. Quan, "Research on item response theory based computerized adaptive testing system," Master's thesis, Shanghai Jiao Tong University, 2009.

[19] J. Chen and H. Wang, "Adaptive algorithm recommendation and application of learning resources in english fragmented reading," *Complexity*, 2021.

[20] M. DDR and J. PTA, "Application and simulation of computerized adaptive tests through the package catsim," *arXiv:1707.03012v2*, Jul 2018.

[21] "The atos' readability formula for books and how it compares to other formulas," School Renaissance Inst., Inc., Madison, WI., Tech. Rep., Jul 2000.

[22] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv:2203.05794v1*, Mar 2022.

A Project Timeline

The tasks to be carried out and the technical points that have to be tackled are shown below:

- Week 1 : Project Preparation
  - Confirmation of directions: To confirm project direction with supervisor.
  - Find Data: To find the meta data support my whole project.
  - Explore data: know what is the data looks like(listening materials and questions and choices)
  - Simple GUI Design: To design how GUI looks like.
- Week 2: To do some literature in 4 aspects, which include algorithms, structure, procedure, and other more theory support.
- Week 3: To remove all noises existed in data and split data to three parts(listening materials, question and choices). Then, try to get the difficulty level of listening materials and questions.
- Week 4: To implement IRT algorithms and apply IRT model with GUI. Go for an insight into maximum likelihood estimation and how it can be applied to IRT models.
- Week 5: To implement Temperal algorithms or Find more effective recommendation algorithms to recommend question which can improve ability of tester.
- Week 6: Algorithms innovation in IRT algorithms and difficulty level formulas.
- Week 7: Using topic modeling to find relationship between questions and implement the recommendation system.
- Week 8: Validation and test whole CAT system.
- Week 9: Material collation and review whole literature and finish the first 2 chapter of final report.
- Week 10: Finish whole report.
- Week 11: Prepare presentation slide and do a oral presentation record.
- Week 12: Do final adjustments, submit report, and prepare for final q/a session.

Gantt Chart :

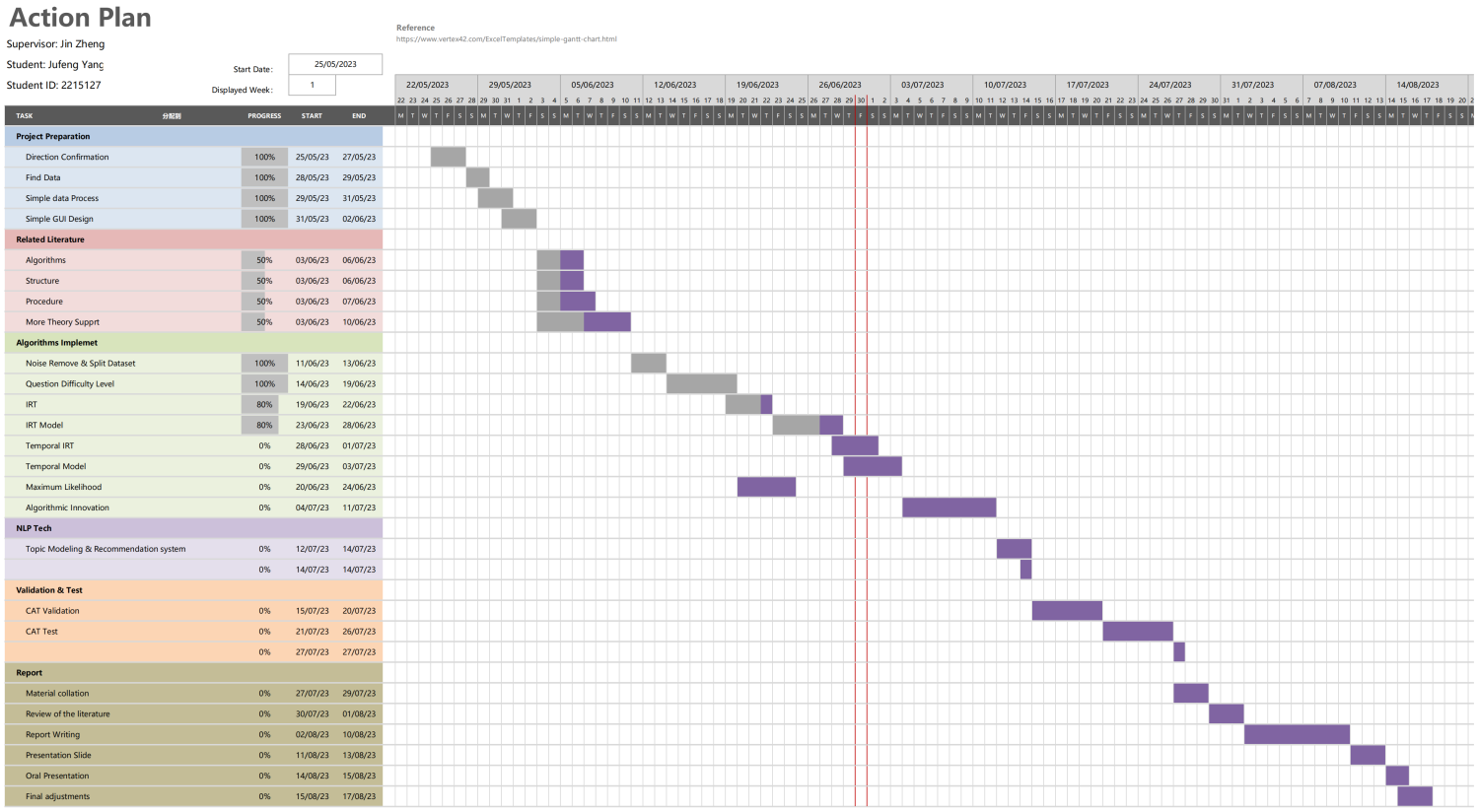

Figure 2: Gantt Chart Plan

B Risk Assessment

Risks arise mainly from unpredictable outcomes and unforeseen difficulties in the progress of the project. Although, as the project progresses, the results become clearer and the difficulties are gradually overcome. However, we still need to consider the risks that we will encounter in the future and warn ourselves of them in advance. This will also reduce the likelihood of risks arising and increase the ability to deal with them in a timely and appropriate manner.

Table 1: Risks and Mitigations

Risk	Likelihood	Severity	Mitigation
Computer problem (stolen or broken)	Medium	Extremely High	Backup all work using the cloud.
Unable to get an accurate difficulty level	Medium	High	Read more literature, practice more methods and seek the help of a mentor.
Unable to get accurate test recommendations	Medium	High	Do more research, use rigorous algorithms and decision-making processes.
Inability to collect sufficient test data	Medium	High	Wrapping the code to make it easier for more people to install, run, or implement online functionality..
Serious errors in the course of the project	Low	Extremely High	Consider each step of the project carefully and refer to more published literature.
Unresolvable bugs, or code architecture problems	Low	Extremely High	Ask for help, structure my code carefully and make changes to errors in a timely manner.