Mathematics Diagnostic Test Project

Chito Wong

October 10, 2018

Contents

Executive Summary

Chapter 1

Introduction

The mathematics diagnostic test (MDT) is applied to new students who are enrolled in courses taught by the School of Mathematics. In 2011, the first generation of MDT (MDT1) was created using Maple T.A., which consists of 32 questions. In 2012, the test was reduced to the form of 20 questions and used in 2012 through 2016. This test is referred as MDT2. Last year, MDT was moved to the STACK online assessment system and replaced Maple T.A. with Moodle [8]. With this new system, the validity of a student's response is available to access, along with the correctness of the response [21].

Over the summer in 2017, two students investigated the data collected on MDT2 in 2013 through 2016, provided suggestions of dropping or adding questions to MDT [3]. This informed the development of a new version of the test, MDT3. The aim of the project this year is to evaluate the performance of MDT3.

1.1 Changes of questions in MDT3

MDT2 and MDT3 both include 20 questions and each question is scored out of 5 marks. For convenience, the questions from MDT2 are labelled as 'Q1', 'Q2', etc. ('Q' is short for 'Question'), while the questions from MDT3 are labelled as 'N1', 'N2', etc. ('N' is short for 'New').

There are three major changes in MDT3:

- Based on the suggestions from the report last year, Q2, Q8 and Q11 are removed from MDT, which are replaced by three new questions [3].
- Q4 and Q12 exist in MDT3 with different marking schemes. The new marking schemes are not convertible into the old marking schemes from the previous tests, so they are also treated as new questions.
- The questions are in different orders.

For the sake of simplicity, the questions in MDT3 are relabelled, where questions which remain unchanged from MDT2 use the same labels as in MDT2 with prefix 'Q' and the other questions keep the labels with prefix 'N'. Table 1.1 shows both labels and the changes of the questions. The letters in column 'Type' are the classifications of the questions using taxonomy [5], where Type-A questions require routine procedures, Type-B questions are examining existing knowledge in new situations and Type-C questions are logical questions.

1.2 Data exploration and cleaning

Before analyses are performed on the data retrieved, we should first explore the data, getting basic information of the data and remove any outlier that results in noise contamination.

Label	Relabel	Type	Changes (if any)
N1	Q1	A	
N2	Q3	B	
N3	N3	A	Q4 with different marking schemes
N4	Q5	A	
N5	Q6	A	
N6	Q7	B	
N7	N7	B	New question
N8	Q9	A	
N9	Q10	B	
N10	N10	B	New question
N11	N11	C	New question
N12	N12	A	Q12 with different marking schemes
N13	Q13	A	
N14	Q14	B	
N15	Q15	A	
N16	Q16	A	
N17	Q17	A	
N18	Q18	A	
N19	Q19	B	
N20	Q20	B	

Table 1.1: Labels of questions and corresponding types using taxonomy

1.2.1 Raw data

The raw data of the results of MDT2017 contains the scores and the responses of 926 students in each question. Instead of showing only numerical numbers, the mark of a student scoring in a question may also be a dash (-), indicating the response of the student is not a valid answer. The data also contains the actual response to the questions and the time each student spent in the test.

The raw data is converted to a numerical version where all dashes are treated as 0 marks for convenience of doing exploratory data analysis. Table 1.2 shows the summary statistics of the total scores in MDT3, including mean, standard deviation, maximum and minimum. The mean of the total scores is skewed to the left, compared with the median, implying there may exist a large number of students who got low scores in MDT.

	Value
Min.	0.00
1st Qu.	57.75
Median	72.88
Mean	69.22
3rd Qu.	84.50
Max	100.00
Stand. Dev.	20.42

Table 1.2: Summary statistics of raw data of MDT3

The histogram of the total scores (Figure 1.1) shows that the students scoring 0 to 40 are approximately uniformly distributed. These statistics suggest that unreliable data may be contained in the data set, which needs to be removed.

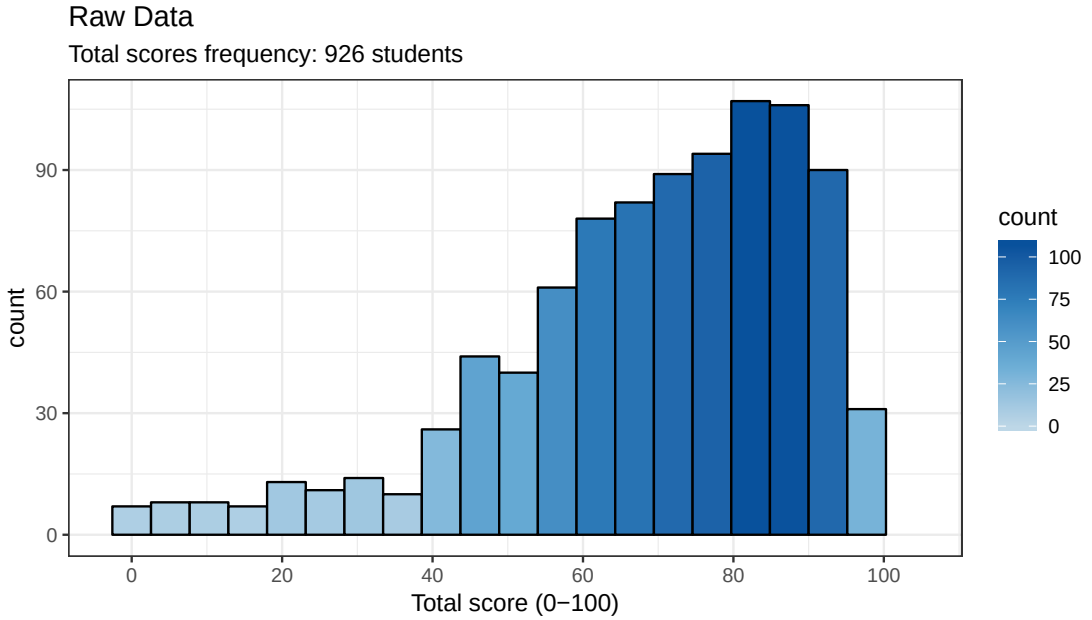

Figure 1.1: Histogram of total test scores of raw data

The raw data includes some students more than once, due to some students being allowed a second attempt following technical problems with their first attempt. In such cases, we disregard their worst result.

1.2.2 Cleaning data with previous technique

The method of data cleaning in the analysis of the last year [3] is

1. For students who took the test more than once, consider the attempt with the highest scores only and remove the others;
2. Eliminate the students who scored three or more zeros in the 5 easiest questions in the second-half of the test; and
3. Add the students scoring more than 30 marks in total back to the sample.

The aim here is to remove non-serious attempts at the test.

For consistency, we apply the same cleaning technique on the data from MDT3. Table 1.3 shows the facility of each question by calculating their mean using the data without duplicated students. The five questions with the least facility (or highest mean score) are N11, N12, Q16, Q17 and Q18.

	Q1	Q3	N3	Q5	Q6	Q7	N7	Q9	Q10	N10
Difficulty	4.44	3.60	4.06	3.94	1.74	3.79	1.89	4.22	3.64	2.01
	N11	N12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20
Difficulty	3.75	4.16	3.71	2.84	3.53	4.44	4.65	3.72	2.88	2.21

Table 1.3: Facility of each question

The number of students removed by using the technique above and the number of students remaining are shown in Table 1.4.

The summary statistics and the histogram are shown in Table 1.5 and Figure 1.2. The difference between the mean and the median of the total scores is reduced and the lump of the students scoring lower than 40 marks in the histogram seems to have been removed properly.

	No.
All students	926
Students without duplicated scores	917
Total records removed	42
Total students removed	33
Students remaining for further analysis	884

Table 1.4: Cleaning the data using the same method as in 2017 report

	Value
Min.	12.50
1st Qu.	60.00
Median	73.94
Mean	71.32
3rd Qu.	85.00
Max	100.00
Stand. Dev.	17.53

Table 1.5: Summary statistics of data after cleaning using previous technique

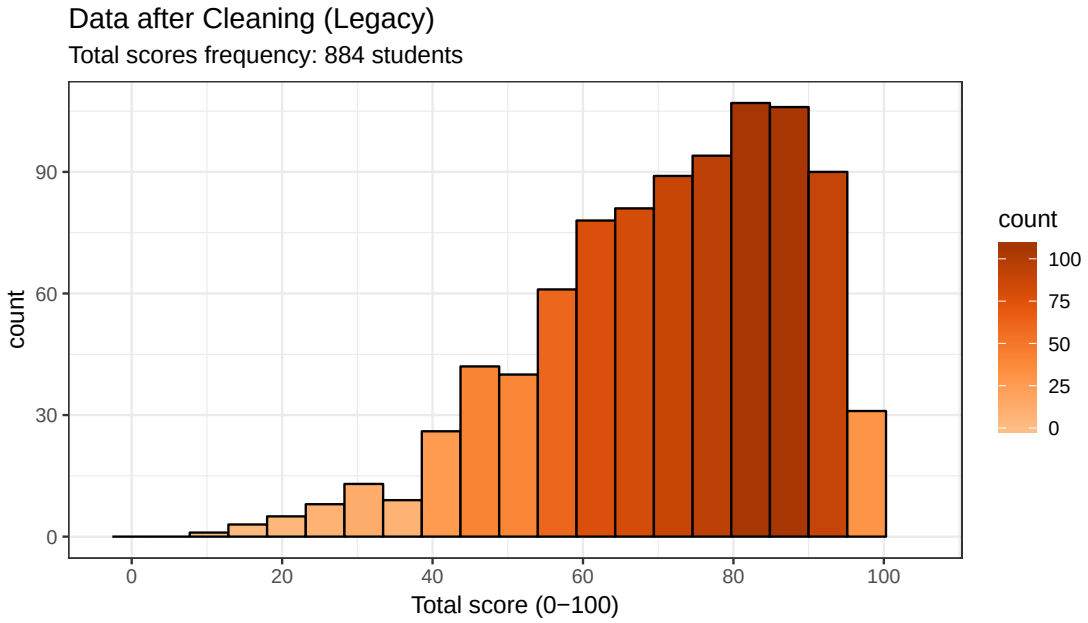

Figure 1.2: Histogram of total test scores of data after cleaning using previous technique

1.2.3 Cleaning data with new technique

Although the technique of removing unreliable data from last year works well on the data from MDT3, there is a problem of the technique: it considers only the marks of the questions but ignores the validity of the responses, which is available for the data from MDT3.

The frequencies of the numbers of valid responses of the test given by the students removed are shown in Table 1.6. There are over a third of the students removed giving more than 10 valid responses in the test, whether or not these were correct.

Number of valid responses	2	4	5	6	7	9	10	11	13	14	15	18
Frequency	2	4	5	5	2	1	2	3	2	5	1	1

Table 1.6: Frequencies of the numbers of valid responses given by the students removed

We therefore modify the data cleaning method to make use of the validity data:

2*. Eliminate the students who gave three or more invalid responses in the 5 easiest questions in the second-half of the test.

The number of students removed by using the new technique are shown in Table 1.7.

	No.
All students	926
Students without duplicated scores	917
Total records removed	30
Total students removed	21
Students remaining for further analysis	896

Table 1.7: Cleaning the data using the new method

The summary statistics in Table 1.8 and the histogram in Figure 1.3 both show satisfying results that the unreliable data are eliminated properly and the distribution seems to be Normal.

	Value
Min.	5.00
1st Qu.	59.50
Median	73.75
Mean	70.61
3rd Qu.	85.00
Max	100.00
Stand. Dev.	18.46

Table 1.8: Summary statistics of clean data of MDT3

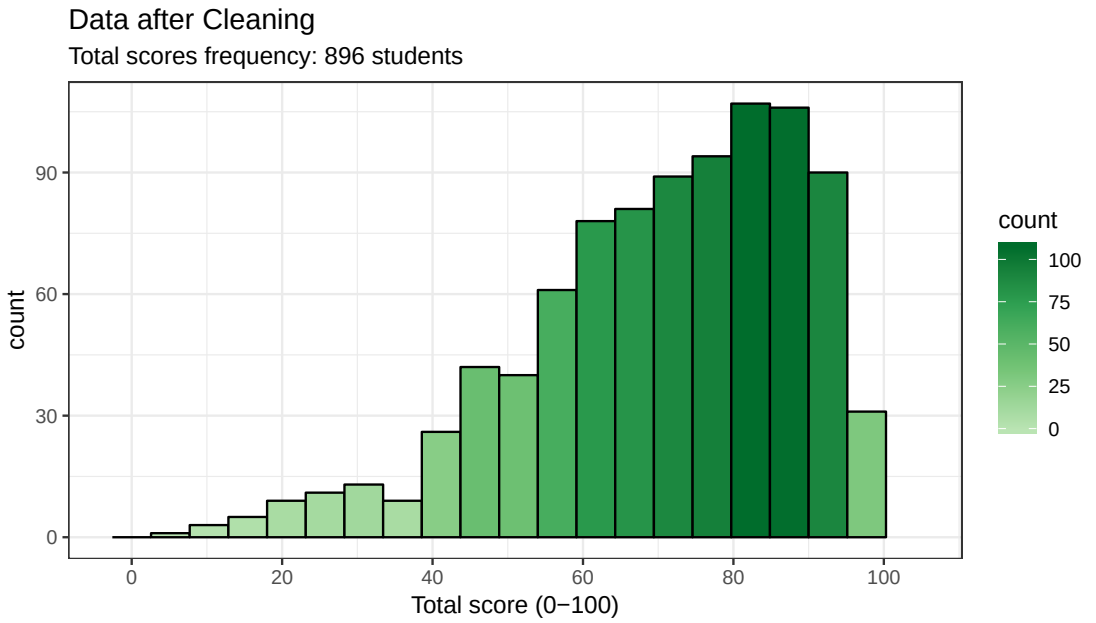

Figure 1.3: Histogram of total test scores of data after cleaning using new technique

The frequencies of the numbers of valid responses of the test given by the students removed by the new technique are shown in Table 1.9. Most of these students did not give more than 10 valid responses, so they can be considered to have failed to complete the test to the best of their ability.

Number of valid responses	2	4	5	6	7	9	10	11
Frequency	2	4	5	5	2	1	1	1

Table 1.9: Frequencies of the numbers of valid responses given by the students who are removed by the new cleaning technique

1.2.4 Comparison of the cleaning methods

From the results of the previous sections, 33 students are removed using the same technique as last year from a data set with 917 students, while 21 are removed using the new technique which considers the validity of responses. This can be seen in the confusion matrix in Table 1.10. A detailed introduction to the confusion matrix and evaluation of binary classifier can be seen in Section 5.4.

	Removed by old method	Kept by old method	Sum
Removed by new method	21	0	21
Kept by new method	12	884	896
Sum	33	884	917

Table 1.10: Confusion matrix based on two cleaning methods

If we consider the new cleaning method to be ideally a perfect classification of reliable and unreliable data and as a reference classification, the accuracy of the old cleaning method can be calculated as

$$ACC = \frac{TP + TN}{TP + FN + TN + FP},$$

where TP (True Positive) are those who are removed by the new method and the old method removes them, FN (False Negative) are those who are removed by the new method but are not be removed by the old method, and TN (True Negative) and FP (False Positive) are analogous to TP and FN. The accuracy of the old method is 98.7% which is quite high, suggesting that two cleaning methods are similar.

We therefore use the new method as the method of cleaning data from MDT3 because it makes the most of the raw data and keeps much sensible data. Although this method cannot be applied to the data from the previous tests because of a lack of validity information, this method should be used consistently on the data from the future tests, unless a better method is discovered.

1.3 Data from MDT2

In this report, analyses performed are using the data of MDT taken in both 2017 and before 2017. There are two different kinds of data of MDT taken before 2017, including a combined data set of MDT2013, MDT2014, MDT2015 and MDT2016, and four data sets of these four years separately.

The combined data set was used in the report last year and will be used in this report. This data set will be cleaned by the old method, i.e. the method used in the analyses last year, for consistency. The tests from year 2013 to 2016 were the same and the data does not contain the year when the students took the test, so only the attempt with the highest total score is considered if a student took the test in more than one year. This is consistent with the analyses conducted last year.

The separated data sets will be used to analyse the growth of ability of the students each year. These data sets will also be cleaned by the old method. However, the students who took the tests in more than on year will be considered to be 'different' and their best results in each year will all be kept. The summary statistics are shown in Table 1.11 and the histogram of the total scores are shown in Figure 1.4, where an increase in the mean of the total scores can be seen, indicating the ability of the new students increases each year.

	MDT2 (2013)	MDT2 (2014)	MDT2 (2015)	MDT2 (2016)
Min.	10.00	13.75	20.50	20.50
1st Qu.	57.50	58.25	61.75	67.75
Median	70.00	71.00	76.00	79.50
Mean	68.96	69.56	72.93	77.15
3rd Qu.	82.50	82.50	87.00	89.00
Max	100.00	100.00	100.00	100.00
Stand. Dev.	17.79	17.53	17.41	15.41

Table 1.11: Summary statistics of cleaned data from year 2013 to 2016

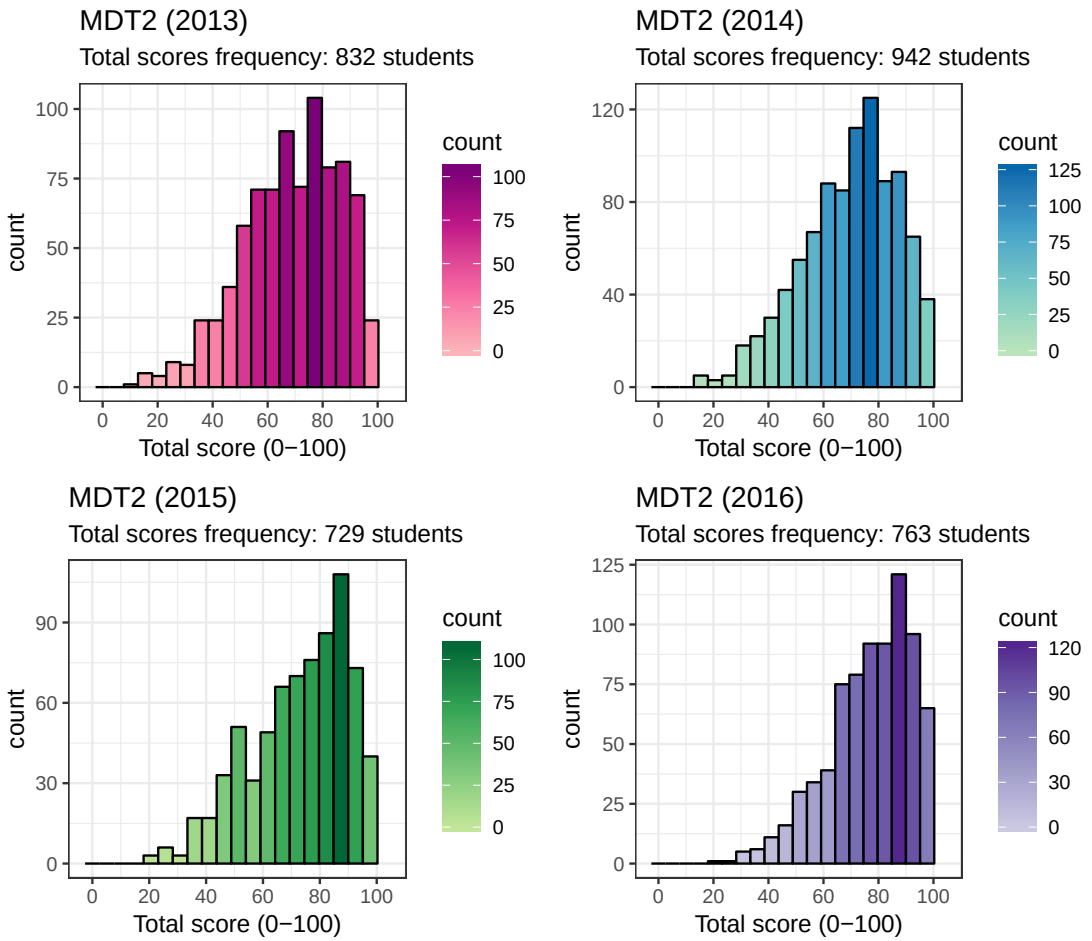

Figure 1.4: Histogram of total test scores of data after cleaning using new technique

Chapter 2

Partial Credit Model

An item response theory model is a method of assessing the relation between categorical observable variables and continuous latent variables [1]. In our case, we theorise a relation between students' mathematical ability (continuous latent variable) and their performance on the MDT, as measured by their question scores (categorical observed variables). A dichotomous item response theory model which has only two alternative responses (e.g. correct or incorrect), is a simple model and can be well-understood and easily implemented. This model was used in the last year. However, our data is polytomous, since partial marks are available.

For item response data with multiple categories, polytomous IRT models need to represent the relation. There are several polytomous IRT models [6], but for our data , the partial credit model (PCM) is appropriate. In `R`, both `ltm` and `mirt` packages can be used to fit PCM.

2.1 Assumptions

As an IRT model, PCM entails the same assumptions as other IRT models. One of the assumptions is the local independence of items, i.e. responding to a question is not related to any other questions. Students are also expected to work independently and not discuss the test with others. The set of latent variables is the only variable that explains the correlation between each pair of observed items.

The dimensionality of the latent variables should also be identified before applying PCM. This is often investigated with factor analysis [4].

2.2 Models

The generalised partial credit model (GPCM) with traditional parameterisation [6] is defined as

$$P_{ik}(\theta) = \frac{\exp \sum_{c=0}^{k} a_i(\theta - \delta_{ic})}{\sum_{r=0}^{m_i} \left[\exp \sum_{c=0}^{r} a_i(\theta - \delta_{ic}) \right]} \text{ for } k = 0, 1, \ldots, m_i, \text{ with } \sum_{c=0}^{0} a_i(\theta - \delta_{ic}) = 0,$$

where $P_{ik}(\theta)$ denotes the probability of responding in category k for item i, given the latent ability θ; a_i are the slope parameters; δ_{ic} are the item-category parameters, representing the relative difficulty between category $c-1$ and c; m_i is the number of categories for item i, excluding 0.

Introducing an integer scoring function T_{ik}, which is the score of each category for item i [14], the model can be rewritten as

$$P_{ik}(\theta) = \frac{\exp(T_{ik} \cdot a_i\theta + d_{ik})}{\sum_{c=0}^{m_i} \exp(T_{ic} \cdot a_i\theta + d_{ic})}.$$

For the GPCM with traditional parameters, the scoring function $T_{ik} = k$ for each item i, i.e. the difference in score between the consecutive categories is fixed at 1. Therefore, the parameters d_{ik} in the alternative model are equivalent to $\sum_{c=0}^{k}(-a_i\delta_{ic})$.

The scoring function T_{ik} can be modified to cooperate with items with non-constant difference between the consecutive categories. For example, if the possible marks for an item are 0, 2, 3 and 5 out of 5, the modified scoring function $\boldsymbol{T}_i$ can be written as $\boldsymbol{T}_i = (0, 2, 3, 5)^T$. `mirt` can fit GPCM with modified scoring function, by adding extra function parameter `gpcm_mat` [4]. However, it only gives parameter estimates but cannot calculate further inferences described below.

The partial credit model (PCM) is a constrained GPCM with equal item slope parameters a_i fixed at 1. This model is called 'Rasch' in both `ltm` and `mirt` packages. Alternatively, the item slope parameters can be assumed equal but estimated in the model '1PL'.

For multidimensional cases with two or more latent traits (sometimes also called factors), the model is defined as

$$P_{ik}(\boldsymbol{\theta}) = \frac{\exp\left(T_{ik} \cdot \boldsymbol{a}_i^T\boldsymbol{\theta} + d_{ik}\right)}{\sum_{c=0}^{m_i}\exp\left(T_{ic} \cdot \boldsymbol{a}_i^T\boldsymbol{\theta} + d_{ic}\right)},$$

where $\boldsymbol{a}_i$ is a vector of item slope parameters on each latent trait and $\boldsymbol{\theta}$ is a vector of latent traits values.

2.3 Inferences

2.3.1 Expected score function

The expected score for each item given uni-dimensional ability θ, is the weighted sum of all the probabilities of responses, where the weights are given by the scoring functions T_{ik} for $k = 0, 1, \ldots, m_i$, i.e.

$$\bar{T}_i(\theta) = \sum_{k=1}^{m_i} T_{ik}P_{ik}(\theta).$$

2.3.2 Information function

The item information function for polytomous IRT [14], including GPCM, with is defined as

$$I_i(\theta) = \sum_{k=1}^{m_i} P_{ik}(\theta)[-\frac{\partial^2}{\partial\theta^2}P_{ik}(\theta)] = a_i^2\sum_{k=1}^{m_i}[T_{ik} - \bar{T}_i(\theta)]^2P_{ik}(\theta).$$

The test information function is the sum of the item information functions, i.e.

$$I(\theta) = \sum_i I_i(\theta).$$

The information functions are used for predicting the accuracy of measuring the latent ability [16]. The standard error of measurement (SEM) is defined as the square root of the reciprocal of the test information function, i.e.

$$SEM(\theta) = \sqrt{\frac{1}{I(\theta)}}.$$

2.3.3 Ability estimates

The ability estimation (or factor score) method used in the report is expected a-posterior (EAP). The EAP scores are defined as

$$\mathrm{EAP}(\theta|\boldsymbol{k}) = \int \theta p(\theta|\boldsymbol{k}) d\theta,$$

where $\boldsymbol{k}$ is the response pattern, $p(\theta|\boldsymbol{k})$ is the posterior probability. The posterior probability is calculated by the Bayes' theorem

$$p(\theta|\boldsymbol{k}) = \frac{p(\boldsymbol{k}|\theta)p(\theta)}{p(\boldsymbol{k})},$$

where the prior distribution $p(\theta)$ is a standard normal distribution.

2.3.4 Goodness of model-fit

The Pearson's χ^2 statistic is defined as

$$\chi^2 = \sum_{r=1}^{n} \frac{[O(r) - E(r)]^2}{E(r)},$$

where r represents a response pattern, n represents the number of different response patterns, $O(r)$ and $E(r)$ are the observed and expected frequencies.

The Pearson's χ^2 statistic is used a goodness-of-fit statistic to test the following hypothesis:

H_0: There is no significant difference between observed and fitted values,

against

H_1: There is a significant difference between observed and fitted values,

where the observed and fitted values are the observed and expected frequencies of each response pattern of the data.

However, the approximation of the χ^2 is not reliable when the expected frequencies are too low. Because we have a large number of items, and several of them have many possible responses, the expected frequencies of many response patterns are less than 1, so a parametric Bootstrap approximation is used to assess the goodness of fit.

The algorithm of the Bootstrap test [19] is implemented as

Step 1 Compute the ordinary statistic χ^2;

Step 2 Simulate new parameters β^* from the multivariate normal distribution $N(\hat{\beta}, \Sigma(\hat{\beta}))$, where $\hat{\beta}$ is the vector of the maximum likelihood estimates of all parameters a_i and δ_{ic}, $\Sigma(\hat{\beta})$ is the variance-covariance matrix of the parameters;

Step 3 Simulate new response patterns using the simulated β^*;

Step 4 Fit the PCM to the simulated response patterns and compute the corresponding statistic χ_i^2; and

Step 5 Repeat Step 2 to Step 4 B times and compute the p-value using $\frac{1 + \sum_{i=1}^{B} \mathbb{1}(\chi_i^2 - \chi^2)}{B+1}$.

Hypothesis testing

After formulating a hypothesis test with two hypotheses H_0 and H_1 which are stated above, the p-value, which is defined as the probability of observing a result at least as extreme as the observed data statistic given that the null hypothesis is true, can be interpreted as [11]

p-value	Interpretation
> 0.1	No evidence against H_0
$0.05 \sim 0.1$	Weak evidence against H_0
$0.01 \sim 0.05$	Moderate evidence against H_0
< 0.01	Strong evidence against H_0

Table 2.1: Interpretation of p-values

2.3.5 Goodness of item-fit

An item-fit statistic $S\text{-}X^2$ proposed by Orlando [15] is used in this report. This item-fit statistic, as well as the traditional item-fit statistics, are calculated by comparing the observed proportion of examinees in each subgroup giving a certain response to each item with those predicted by the model. However, $S\text{-}X^2$ groups the examinees based on the observed data before the model is fitted, unlike the traditional item-fit statistic which groups the examinees by the order of the estimated ability θ.

The statistic $S\text{-}X^2$ has the form

$$S\text{-}X^2 = \sum_{k=1}^{n-1} N_k \frac{(O_{ik} - E_{ik})^2}{E_{ik}(1 - E_{ik})},$$

where N_k is the size of group k, O_{ik} is the observed proportion calculated by the observed data and E_{ik} is the expected proportion.

The goodness-of-fit test on item level is analogous to the goodness-of-fit test of the overall model described above.

2.3.6 Likelihood ratio test

The likelihood ratio (LR) test compares the maximum likelihood attainable if the parameter vector β is under the reduced model ω with the maximum attainable if β is under the full model Ω as

$$LR = L(\hat{\beta}_\omega)/L(\hat{\beta}_\Omega).$$

If the null hypothesis that $\beta_\Omega \setminus \beta_\omega = 0$ is true, $\lambda = -2\log(LR) = -2[l(\hat{\beta}_\omega) - l(\hat{\beta}_\Omega)]$ has the distribution χ^2_{p-r}, where r and p are the number of parameters in β_ω and β_Ω, $l(\beta)$ is the log-likelihood function.

2.3.7 Akaike information criterion

The Akaike information criterion (AIC) is an estimator of the relative qualities between a set of models and is used in model selection. This is a value that rewards goodness of fit and penalises overfitting of a model, because a model with more parameters almost always improves the goodness-of-fit. The best model will be the one with the lowest AIC value.

The AIC is given by

$$AIC = 2k - 2l(\hat{\beta}_\omega),$$

where k is the number of parameters to be estimated in the model.

Chapter 3

Applying PCM to MDT2

Three dichotomous IRT models were fitted to the data on MDT2 from 2013 to 2016 in the summer 2017 project, where the marks of the questions were converted to binary scores [3]. In this chapter, some polytomous IRT models will be fitted to accommodate to the real results of MDT.

3.1 Local independence

In order to satisfy the assumptions of local independence, statistical dependencies between items should be identified.

A Kendall's τ test is the computation of a nonparametric correlation coefficient used to measure the ordinal association between two variables. The test is performed on each pair of items to test the null hypothesis that the items are statistically independent. The Kendall correlation heat map as well as the significance indicators of the test are shown in Figure 3.1. From this we see that the correlation between any two questions are positive and the null hypotheses of all tests are rejected at the 95% significant level.

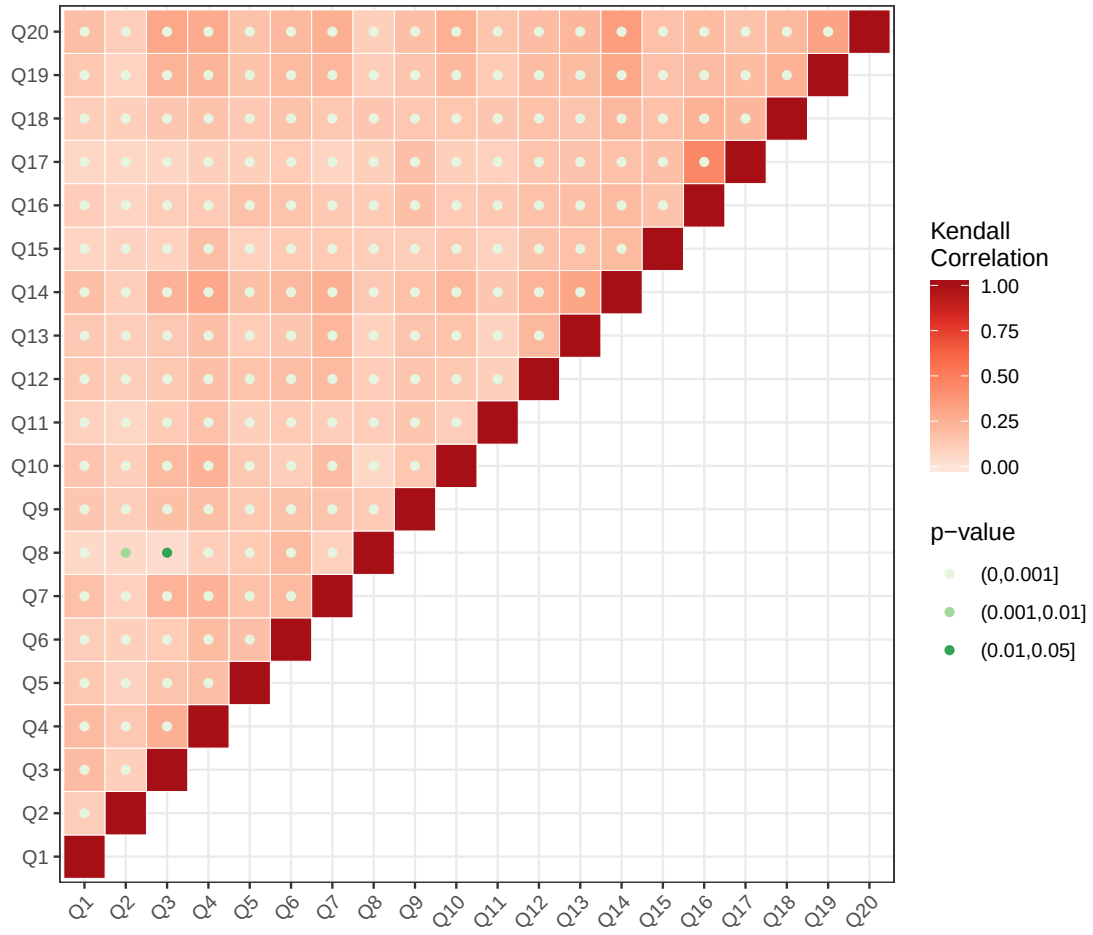

Figure 3.1: Correlation heat map and significance indicators of the Kendall's τ test

The six largest p-values of the significance test are also shown in Table 3.1. All pairs of questions except the pair (Q3, Q8) have p-values lower than 0.01, suggesting the null hypotheses are rejected at the 99% significance level. The null hypothesis of the pair (Q3, Q8) is rejected at the 95% significance level. Therefore, the assumption of local independence that there is an underlying structure is entitled.

Item i	Item j	p-values
Q3	Q8	0.0139
Q2	Q8	0.0015
Q1	Q8	0.0006
Q2	Q17	0.0002
Q8	Q10	0.0002
Q1	Q17	0.0001

Table 3.1: Four largest p-values from Kendall's τ test

3.2 Dimensionality

An exploratory factor analysis (EFA) is essential to identify the latent structure of variables (items) and determine the number of latent variables to use in PCM. The `R` package `psych` is used in this section.

A scree plot of the eigenvalues for a principal axis factor analysis, along with the plots of simulated and resampled data are shown in Figure 3.2. The eigenvalue of factor number 2

decreases greatly and the eigenvalues start to level off from this point. Besides, the difference between simulated data and actual data is minimised at the point of factor number 6. These suggest that an integer between 1 and 6 would be a good choice for the number of factors to be extracted.

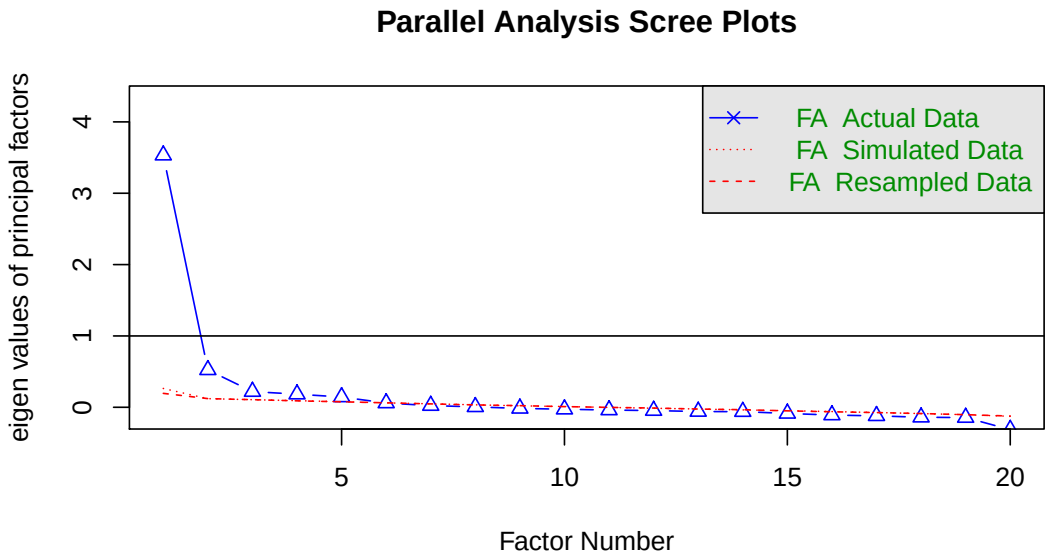

Figure 3.2: Parallel analysis scree plots

A suitable number of factors should result in a simple structure of factor loadings that the loadings should greater than a given threshold (0.25 is used in this report) and an item should not load on more than one factor. We should also validate the model with simple structure by assessing the Root Mean Square of Residuals (RMSR) which is acceptable when it is close to 0, Root Mean Square Error of Approximation (RMSEA) index which shows a good model fit when it is less than 0.05, and Tucker Lewis Index (TLI) which is acceptable when it is greater than 0.9.

3.2.1 One factor

We first carry out EFA with one factor using minimum residual as the factoring method and 'oblimin' as the oblique transformation of the solution, because of the existence of correlation in the factors [17]. With only one factor, this factor can be interpreted as 'general mathematical ability'.

The standardised loadings on the factors and the indicator of loadings greater or equal to 0.25 are shown in Table 3.2. The loadings are also plotted along the Factor 1 axis in Figure 3.3. The variable Q2 is not significant though it has the loading of 0.233 on the factor which is slightly lower than 0.25. It should be noted that the loadings of Type-B questions are generally larger than Type-A, indicating that Type-B questions are better measures of higher mathematical ability.

	Type	MR1	
Q1	A	0.353	*
Q2	A	0.233	
Q3	B	0.421	*
Q4	A	0.537	*
Q5	A	0.356	*
Q6	A	0.405	*
Q7	B	0.456	*
Q8	A	0.256	*
Q9	A	0.378	*
Q10	B	0.414	*

	Type	MR1	
Q11	A	0.309	*
Q12	A	0.432	*
Q13	A	0.448	*
Q14	B	0.608	*
Q15	A	0.357	*
Q16	A	0.421	*
Q17	A	0.366	*
Q18	A	0.411	*
Q19	B	0.499	*
Q20	B	0.550	*

Table 3.2: Loadings of items on factor where '*' indicates loadings greater or equal to 0.25

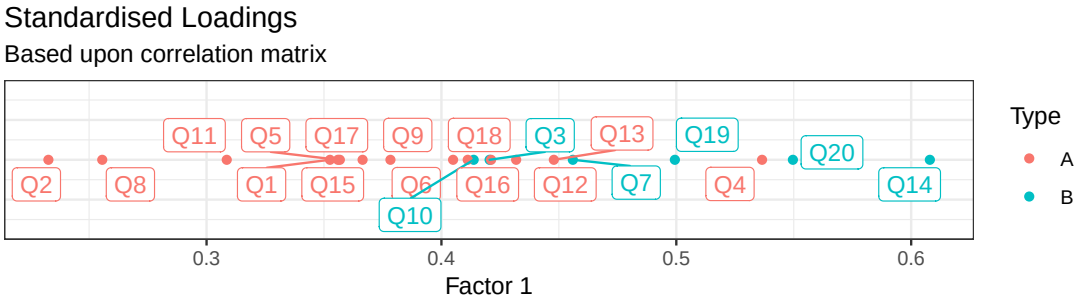

Figure 3.3: Standardised loadings on Factor 1

The RMSR of this model is 0.04. The 90% confidence interval of RMSEA index is (0.041, 0.046). The TLI of factoring reliability is 0.859. All these statistics suggest that this EFA with 1 factor is a reasonable fit.

3.2.2 Two factors

Next, we consider adding an additional factor to EFA to examine whether the results are improved. Table 3.3 shows the factor loadings which are greater or equal to 0.25 only. In this model, three variables (Q2, Q8 and Q15) become non-significant and one variable (Q18) loads on more than one factor. This violates the simple structure we aim to have that items are only single-loading.

	Type	MR1	MR2
Q1	A	0.386	
Q2	A		
Q3	B	0.512	
Q4	A	0.592	
Q5	A	0.327	
Q6	A	0.360	
Q7	B	0.521	
Q8	A		
Q9	A	0.305	
Q10	B	0.441	

	Type	MR1	MR2
Q11	A	0.258	
Q12	A	0.359	
Q13	A	0.374	
Q14	B	0.588	
Q15	A		
Q16	A		0.615
Q17	A		0.688
Q18	A	0.256	0.258
Q19	B	0.446	
Q20	B	0.557	

Table 3.3: Loadings of items on factors which are greater or equal to 0.25 only

Figure 3.4 is the scatter plot of the standardised loadings on Factor 2 against the standardised loadings on Factor 1. The loadings of Type-B questions are again generally higher than Type-A questions. However, Q4, which is a Type-A question, has the highest loading on Factor

1 among all questions. Q16 and Q17 have high loadings on Factor 2 and low loadings on Factor 1, which are the contrary of the rest of the questions. It is worth noticing that Q16 and Q17 are multiple choice questions and require knowledge of differential calculus.

Figure 3.4: Standardised loadings on Factor 2 against Factor 1

3.2.3 More factors

In addition, EFAs with 3, 4, 5 and 6 factors were also performed. However, the structure of variables of these models were not as simple as the model with only 1 factor and there were not large improvement of the statistics including RMSR, RMSEA and Tucker Lewis Index.

Considering the results above, as well as the simplicity of interpreting the factors, we proceed with an one-factor structure, i.e. the overall (undivided) mathematical ability.

3.3 Model selection

A 1PL model, i.e. the model with fixed item slope parameters, is firstly fitted, using the packages `ltm` [19] and `mirt` [4] in `R`. For reference, we refer to this model as Model `MDT2-1PL`. The item response curves (i.e. the item expected score functions), test response curve (i.e. the test expected score functions), item information curves and test information curve are shown in Figure 3.5.

Based on 200 data sets (including 199 bootstrapped samples), the p-value of the goodness-of-fit test is 0.12, which is non significant, suggesting that the 1PL model is an acceptable fit to the data. In Table 3.4 are shown the $S\text{-}X^2$ item-fit statistics and corresponding goodness-of-fit tests on item level. The significance codes are interpreted in the same way as `R` output, i.e.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

These statistics indicate that 12 of the 20 items are not well represented by the estimated 1PL GPCM item parameters.

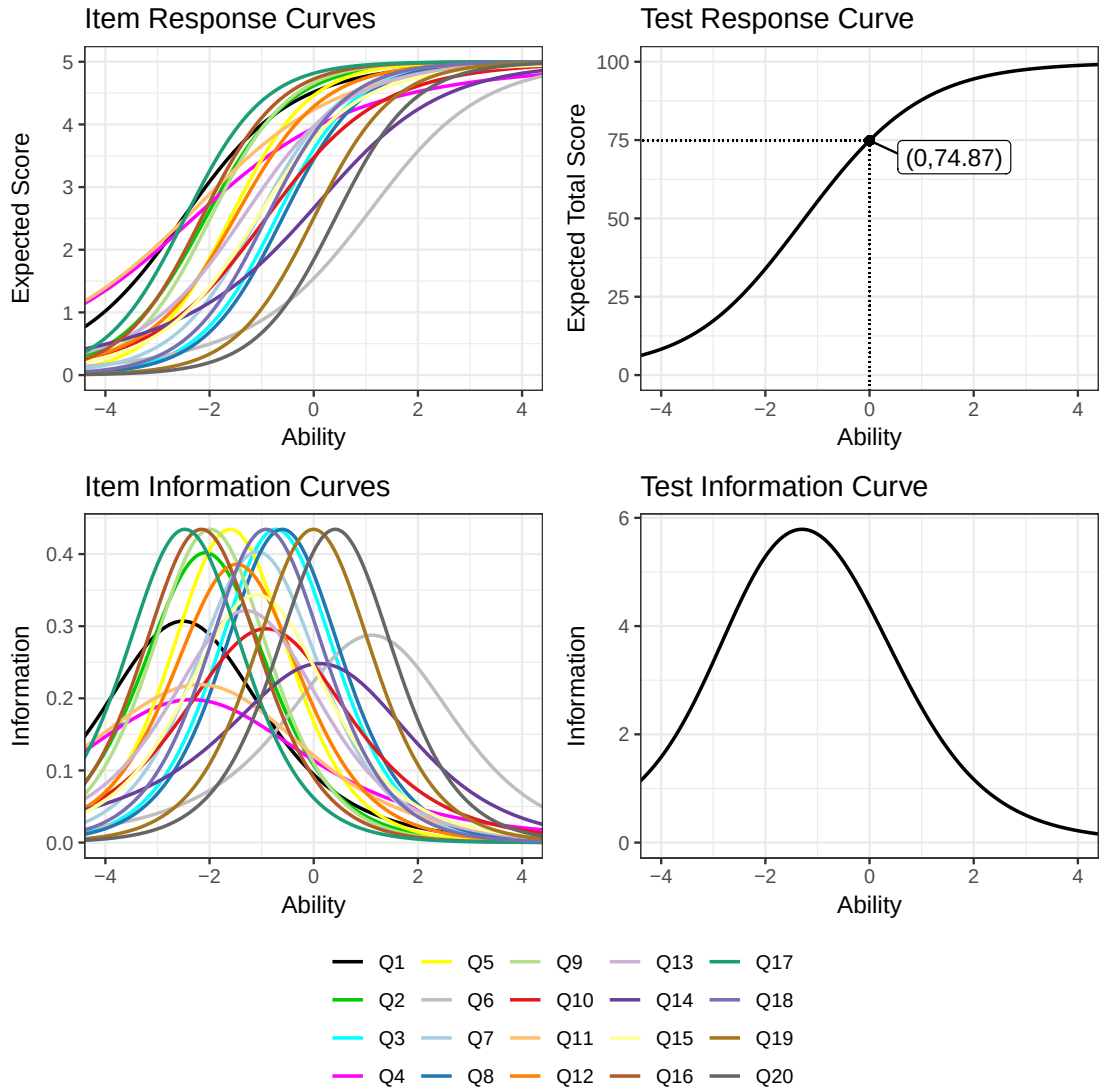

Figure 3.5: Plots of the 1PL model with equal discrimination parameters (Model `MDT2-1PL`)

Item	S-X^2	df	p-value	
Q1	79.2	72	0.262	
Q2	139.9	66	0.000	***
Q3	84.1	37	0.000	***
Q4	352.8	206	0.000	***
Q5	61.3	37	0.007	**
Q6	109.2	67	0.001	***
Q7	74.2	65	0.204	
Q8	293.4	36	0.000	***
Q9	32.5	38	0.723	
Q10	75.4	70	0.308	

Item	S-X^2	df	p-value	
Q11	139.8	114	0.051	.
Q12	123.4	90	0.011	*
Q13	119.2	89	0.018	*
Q14	236.1	110	0.000	***
Q15	284.9	110	0.000	***
Q16	38.7	37	0.392	
Q17	31.4	37	0.726	
Q18	74.8	37	0.000	***
Q19	38.0	34	0.292	
Q20	57.3	32	0.004	**

Table 3.4: Item-fit statistics for Model `MDT2-1PL` using S-X^2 statistics

Based on the finding from the report last year, which used binary scoring, a model with different discrimination parameters provides a better fit to the data than a model with fixed discrimination parameters. A 1PL model is therefore not expected to be a good fit because of the large variation in the discrimination parameters estimated by the 2-parameter dichotomous model.

An unconstrained GPCM (Model `MDT2-GPCM`) is fitted in the next step. The item response curves, test response curve, item information curves and test information curve are shown in Figure 3.6.

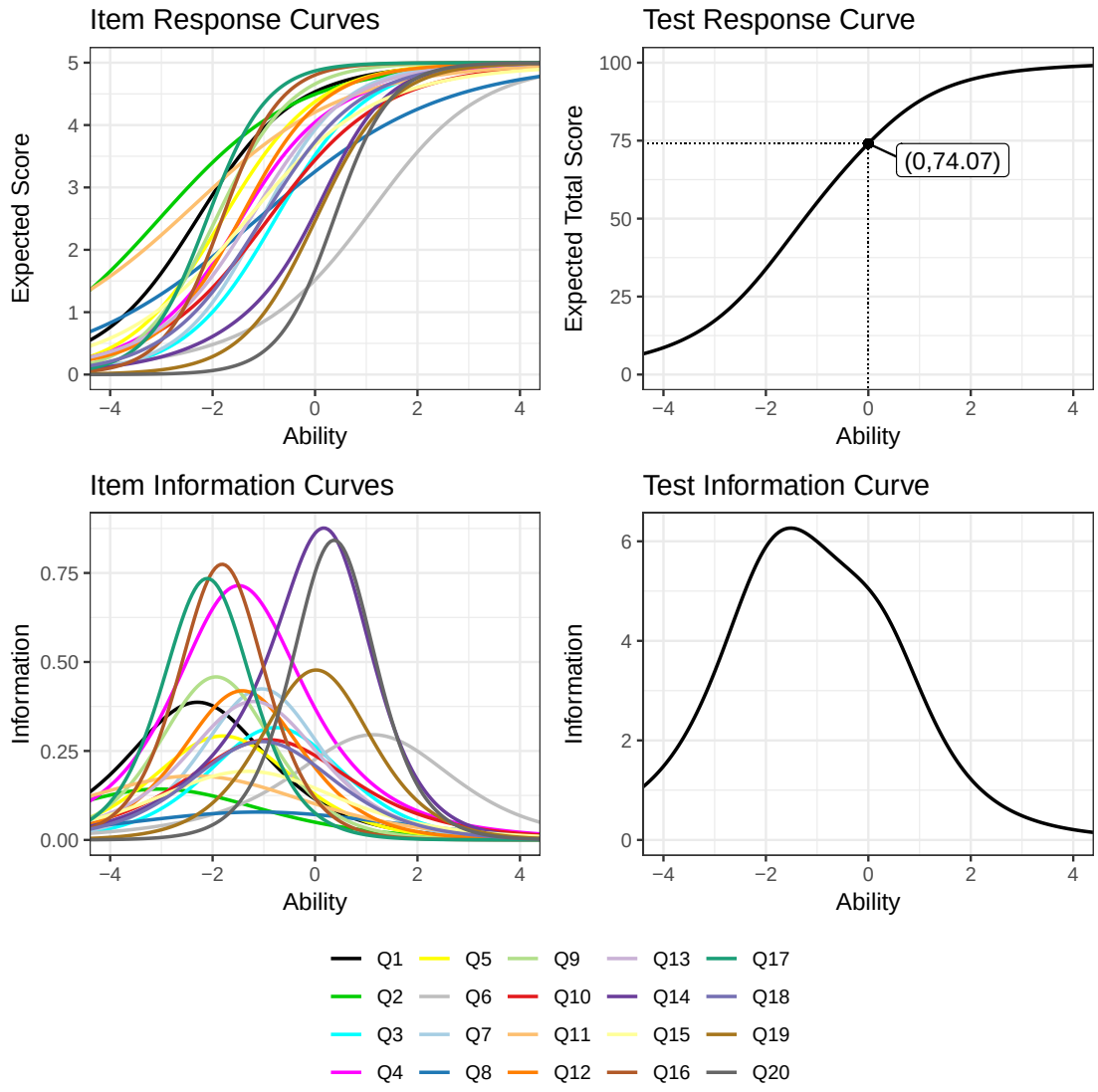

Figure 3.6: Plots of the unconstrained GPCM (Model `MDT2-GPCM`)

Item	S-X^2	df	p-value	
Q1	74.8	71	0.355	
Q2	81.3	69	0.147	
Q3	58.3	37	0.014	*
Q4	230.3	186	0.015	*
Q5	43.1	39	0.300	
Q6	101.7	68	0.005	**
Q7	69.7	65	0.321	
Q8	31.7	39	0.791	
Q9	32.1	38	0.740	
Q10	68.8	70	0.517	

Item	S-X^2	df	p-value	
Q11	129.1	115	0.174	
Q12	123.4	90	0.011	*
Q13	120.4	90	0.018	*
Q14	127.9	102	0.043	*
Q15	214.3	120	0.000	***
Q16	45.9	36	0.124	
Q17	36.0	35	0.420	
Q18	36.3	37	0.503	
Q19	38.5	34	0.271	
Q20	40.6	30	0.093	.

Table 3.5: Item-fit statistics for Model `MDT2-GPCM` using S-X^2 statistics

Based on 200 data sets (including 199 bootstrapped samples), the p-value of the goodness-

of-fit test is 0.14, suggesting an acceptable fit of the model. The S-X^2 item-fit statistics and corresponding goodness-of-fit tests on item level are shown in Table 3.5. There are 7 of the 20 items which are not well represented by the estimated GPCM item parameters.

The likelihood ratio (LR) test statistics are shown in Table 3.6. The p-value suggests the null hypothesis (that item slope parameters $a_i = 0$) is rejected at the 99% significance level, which means that the GPCM is a significant improvement over the 1PL model. In Table 3.6 are also shown the AIC values and other kinds of information criterion used for model selection. The AIC of GPCM is 92414, which is less than the AIC of 1PL model, though the number of parameters increased by 19.

Model	AIC	AICc	SABIC	HQ	BIC	logLik	λ	df	p
`MDT2-1PL`	93133	93135	93278	93242	93437	-46516			
`MDT2-GPCM`	92414	92417	92614	92564	92834	-46138	757	19	0.0000

Table 3.6: Model selection table with likelihood ratio test and AIC comparison

Since all the statistics suggests that Model `MDT2-GPCM` is a better fit to the data than Model `MDT2-1PL`, we proceed with Model `MDT2-GPCM`.

3.4 GPCM

3.4.1 Difficulty and discrimination

When comparing the discriminations and difficulties of the items, the estimations of the coefficients cannot be used directly, because the definitions of the parameters in polytomous IRT models are different from the way they are in dichotomous IRT models [6]. The discrimination of an item depends on the slope parameter a_i and the relative difficulties between consecutive categories δ_{ic}. Alternatively, these characteristics can be accessed from the expected scores $\bar{T}_i(\theta)$ given a sequence of ability values.

In Table 3.7 are shown the expected scores of the students with average ability ($\theta = 0$) and the changes of the expectations of the students within two standard deviations of the mean, i.e. $\theta = \pm 2$. Since the ability θ follows a standard normal distribution, this range of ability covers 95.4% of students. The column 'Facility' shows the expected scores of students average ability. The final column shows the difference in expected scores between the most and least able students, and thus we use this as a measure of discrimination.

	$T_i(-2) - T_i(0)$	Facility	$T_i(2) - T_i(0)$	Discrimination
Q1	-1.65	4.53	0.37	2.02
Q2	-1.09	4.49	0.36	1.46
Q3	-2.52	3.52	1.27	3.78
Q4	-2.27	4.05	0.71	2.99
Q5	-2.14	4.37	0.54	2.69
Q6	-1.04	1.52	2.05	3.09
Q7	-2.77	3.93	0.94	3.71
Q8	-1.36	3.25	1.00	2.36
Q9	-2.27	4.66	0.32	2.59
Q10	-2.05	3.44	1.15	3.20
Q11	-1.22	4.21	0.54	1.76
Q12	-2.57	4.29	0.62	3.19
Q13	-2.40	3.98	0.86	3.25
Q14	-1.98	2.59	2.17	4.15
Q15	-1.83	3.68	0.92	2.76
Q16	-2.71	4.80	0.19	2.90
Q17	-2.14	4.87	0.13	2.27
Q18	-2.41	3.73	1.07	3.48
Q19	-2.18	2.46	2.23	4.41
Q20	-1.62	1.68	3.08	4.70
Sum	-40.21	74.07	20.53	60.75

Table 3.7: Expected scores of students with ability -2, 0 and 2

Using this measure, it can be observed that Q20 is the most discriminating question (with discrimination 4.70) for most students, followed by Q19 (with discrimination 4.41). The least discriminating questions are Q2 (with discrimination 1.46) and Q11 (with discrimination 1.76); in addition, these items have high expected scores for students of mean ability (4.49 and 4.21 respectively), indicating that the expected scores for most of the students in these two questions are greater than 3. This is consistent with the observed proportions of the scores of students (Table 3.8), where a majority of students scored more than 2.5 or even 5 marks.

	Proportion of scores (%)				
	0	1.25	2.5	3.75	5
Q2	9.8	-	3.9	-	86.3
Q11	3.2	4.4	21.0	1.3	70.1
Q19	50.1	-	-	-	49.9
Q20	60.2	-	-	-	39.8

Table 3.8: Proportions of the scores of the students in Q2, Q11 Q19 and Q20

3.4.2 Item information

One of the purposes of fitting IRT models is to estimate the mathematical ability of a student. In order to have accurate estimations, the item information as well as the test information need to be investigated. In addition to the item information curves and test information curves plotted in Figure 3.6, a plot of stacked item information curves is shown in Figure 3.7.

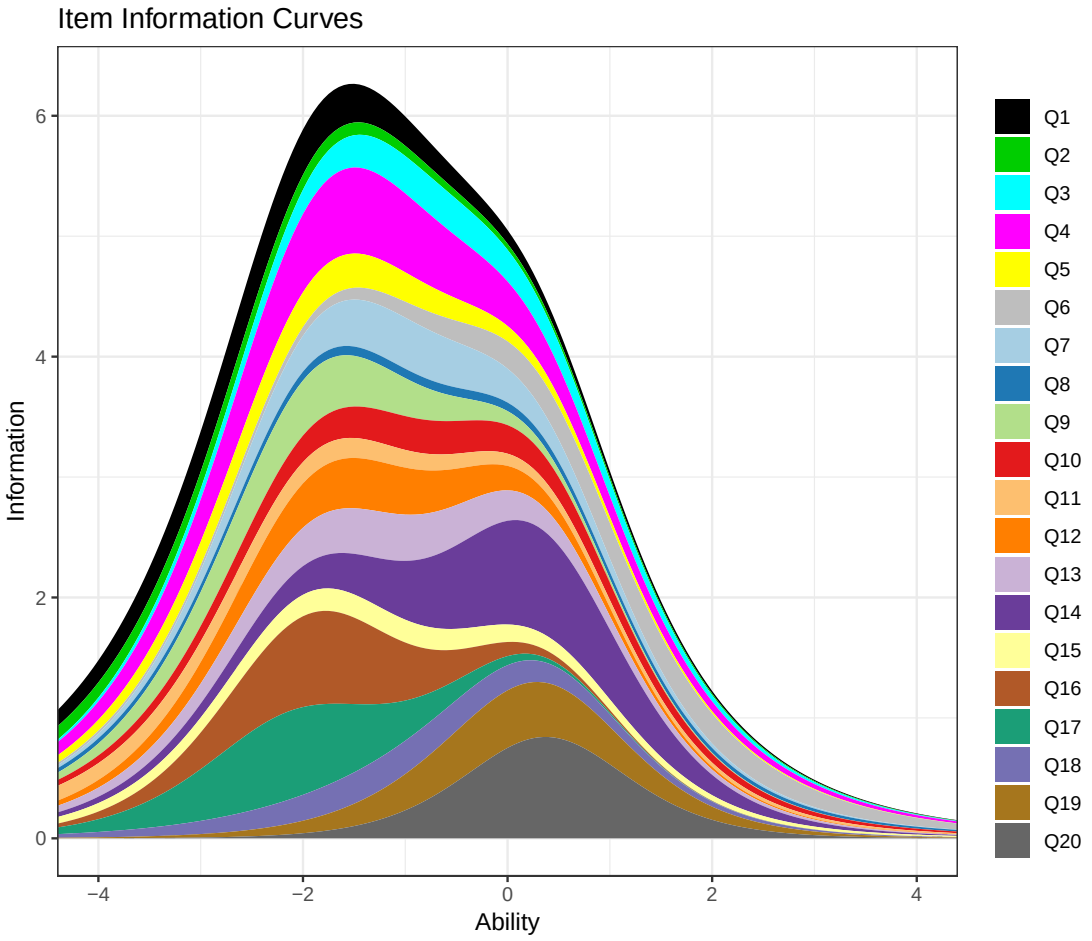

Figure 3.7: Plot of stacked item information curves for Model `MDT2-GPCM`

From the figure and Table 3.9, we see that the test information function attains its maximum at $\theta = -1.52$. The area under the test information curves over the range of negative ability is greater than the information area over the positive ability. We also consider restricting the range of integration to $(-2, 2)$ (i.e. representing 95.4%) of students), shown as 'Major Info Area' in Table 3.9.

From the overall test information shown at the end of Table 3.9, we see that the information area of the below-average students is still twice as large as the area of the above-average students. This indicates that when estimating the abilities of the examinees, the accuracy of measuring above-average students is relatively lower.

The values of ability where the item information functions attain their maximum are also shown in Table 3.9. Only 4 of the 20 items have their maximum at positive ability values, which is one of the contributing factors leading to the negative skew of the test information curve.

		Major Info Area			Total Info Area		
	Max at	Lower	Upper	Total	Lower	Upper	Total
Q1	-2.30	0.49	0.11	0.60	1.35	0.14	1.49
Q2	-3.04	0.17	0.06	0.23	0.70	0.08	0.78
Q3	-0.77	0.57	0.28	0.85	0.79	0.33	1.12
Q4	-1.49	1.21	0.38	1.59	2.15	0.51	2.66
Q5	-1.80	0.46	0.12	0.58	0.95	0.14	1.08
Q6	1.13	0.28	0.55	0.83	0.41	0.93	1.34
Q7	-1.04	0.75	0.26	1.00	1.06	0.29	1.35
Q8	-1.11	0.15	0.11	0.26	0.36	0.20	0.56
Q9	-1.93	0.62	0.09	0.70	1.26	0.09	1.35
Q10	-0.91	0.52	0.30	0.82	0.88	0.40	1.28
Q11	-2.36	0.29	0.13	0.42	1.00	0.19	1.19
Q12	-1.42	0.71	0.17	0.88	1.18	0.20	1.37
Q13	-1.20	0.70	0.25	0.95	1.16	0.30	1.46
Q14	0.17	1.03	1.13	2.17	1.35	1.26	2.61
Q15	-1.32	0.36	0.18	0.54	0.72	0.26	0.98
Q16	-1.81	0.95	0.07	1.02	1.69	0.07	1.76
Q17	-2.11	0.73	0.04	0.78	1.67	0.05	1.71
Q18	-1.03	0.51	0.22	0.73	0.79	0.27	1.05
Q19	0.02	0.60	0.62	1.22	0.68	0.70	1.38
Q20	0.37	0.59	1.13	1.72	0.62	1.22	1.84
Test	-1.52	11.70	6.19	17.89	20.77	7.60	28.37

Table 3.9: The area under each item information curve over different definite integral ranges

Table 3.10 shows the Major Info Area and Total Info Area of each item, ordered by the proportion of the total area. In terms of total item information, Q4 and Q14 have the largest areas and each of them accounts for more than 9% of the total information area, while Q8 and Q2 have the smallest areas and each accounts for less than 3% of the total information area. As shown in Figure 3.7, the area over the range of $\theta < 2$ occupies a large proportion of the total information, but there are only 2.3% of students with abilities which are lower than -2, according to the prior belief that the distribution of ability is Standard Normal Distribution $N(0, 1)$. Therefore, comparing the information area over the range from -2 to 2 is much more meaningful than comparing the area over than whole range of ability.

In terms of major item information, Q2 and Q8 are still the least informative but with even lower proportion of the total information area of the 20 items. Q11 becomes the third least informative item, with 2.33% of the total information area. On the other hand, Q14 becomes the most informative item and accounts for over 12% of the total information. It should also be noted that Type-B questions are generally more informative than Type-A questions. The rank of the major information areas of the items is similar to the rank of the loadings on a one-factor factor analysis model, which is shown in Figure 3.3.

Item	Type	MajorInfo	Prop (%)
Q2	A	0.23	1.28
Q8	A	0.26	1.48
Q11	A	0.42	2.33
Q15	A	0.54	3.02
Q5	A	0.58	3.25
Q1	A	0.60	3.37
Q9	A	0.70	3.92
Q18	A	0.73	4.09
Q17	A	0.78	4.35
Q10	B	0.82	4.58
Q6	A	0.83	4.62
Q3	B	0.85	4.75
Q12	A	0.88	4.90
Q13	A	0.95	5.29
Q7	B	1.00	5.62
Q16	A	1.02	5.71
Q19	B	1.22	6.81
Q4	A	1.59	8.88
Q20	B	1.72	9.64
Q14	B	2.17	12.11

Item	Type	TotalInfo	Prop (%)
Q8	A	0.56	1.97
Q2	A	0.78	2.76
Q15	A	0.98	3.46
Q18	A	1.05	3.71
Q5	A	1.08	3.81
Q3	B	1.12	3.96
Q11	A	1.19	4.18
Q10	B	1.28	4.52
Q6	A	1.34	4.71
Q7	B	1.35	4.77
Q9	A	1.35	4.77
Q12	A	1.37	4.84
Q19	B	1.38	4.87
Q13	A	1.46	5.13
Q1	A	1.49	5.25
Q17	A	1.71	6.04
Q16	A	1.76	6.20
Q20	B	1.84	6.47
Q14	B	2.61	9.20
Q4	A	2.66	9.37

Table 3.10: Tables of item information as percentage of total information for Model `MDT2-GPCM` ordered by proportion

3.4.3 Ability estimates

The scatter plot of the actual MDT scores of the students against their estimated abilities (EAP) is shown in Figure 3.8, along with the corresponding 95% confidence interval of the estimates. The EAP is computed using their responses to each question. Additionally, the test response curve, i.e. the expected total score given ability, is also plotted in the figure for comparison.

It is worth noticing that 76 students scored full marks in the test and their abilities are estimated at 1.86. However, there should not exist a upper limit of the ability in the real world. The upper limit of the questions (5 marks each) and the test (100 marks) restricted students with higher abilities from scoring higher and different marks. This may cause a problem when estimating abilities and may skew the shape of the scatter plot.

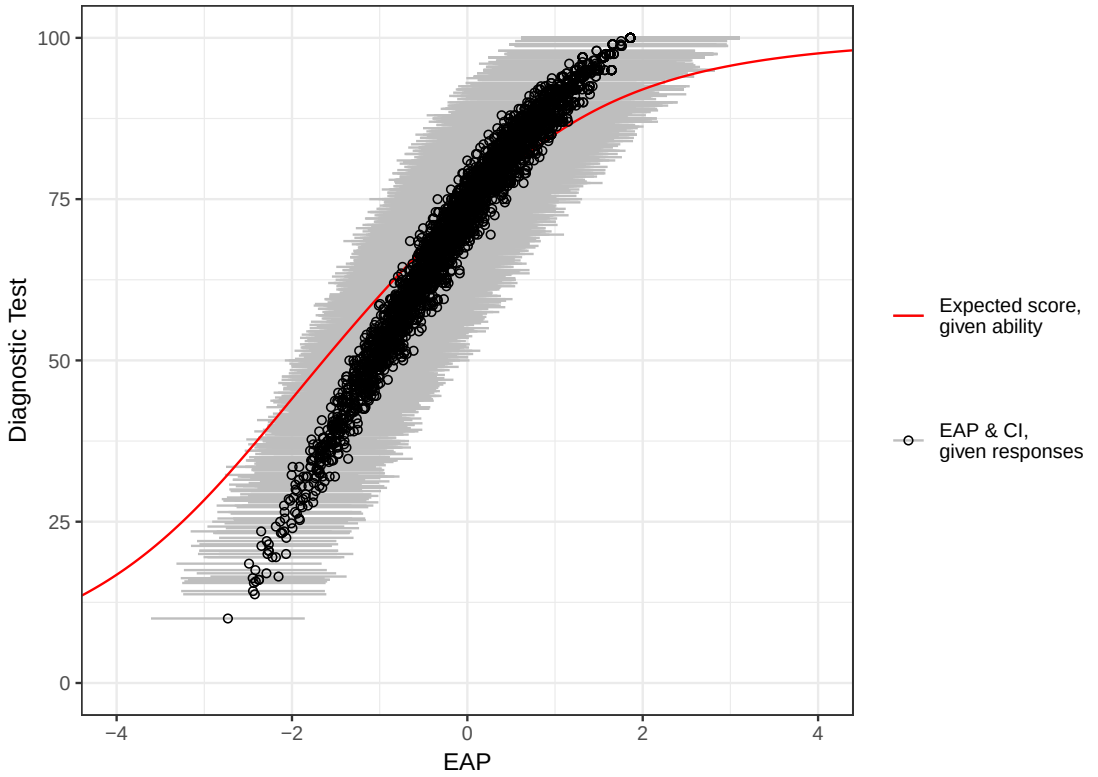

Figure 3.8: Scatter plot of actual MDT scores against EAP and CI, and test response curve

3.4.4 Comparison with dichotomous model

The difficulties and the discriminations estimated by Model `MDT2-GPCM`, a polytomous IRT model, are quite different from the estimates given by the dichotomous models performed last year.

However, our findings are consistent with the previous recommendation to remove Q2, Q8 and Q11. From the total information area in Table 3.10 and the information areas for below-average and above-average students respectively in Table 3.9, these three items cannot give relatively higher standard error of measuring the ability of a student and have the highest accuracy only when measuring the students with abilities lower than -2.

3.5 Academic growth

The academic growth of students from different administrations is always of interest to the examiner. When the entry requirements increase, there would be an upward trend in the abilities of students. In the meantime, the examiner should fine-tune the MDT to address the growth of abilities.

Since all the questions in MDT2 remained unchanged in years 2013 through 2016, it is possible to calibrate the data from a specific year in an IRT model and use its parameters to estimate the abilities of the students taking MDT2 in other years directly from these fixed item parameters. This procedure will place all the students on the same scale of ability.

Choosing the year 2013 as the pivot should be the most proper way, but we need each response to be observed at least once, and this was only the case in 2014. I therefore choose the year 2014 as the pivot and fit the GPCM to it.

The density plot of the estimated abilities of the students using the method of EAP [9] are shown in Figure 3.9, where a trend that the distribution shifts to the right as the years go by can be seen. It should be noted that the density plots are not symmetric with a steeper slope on the right hand side. This may due to the existence of the upper limit of the test scores.

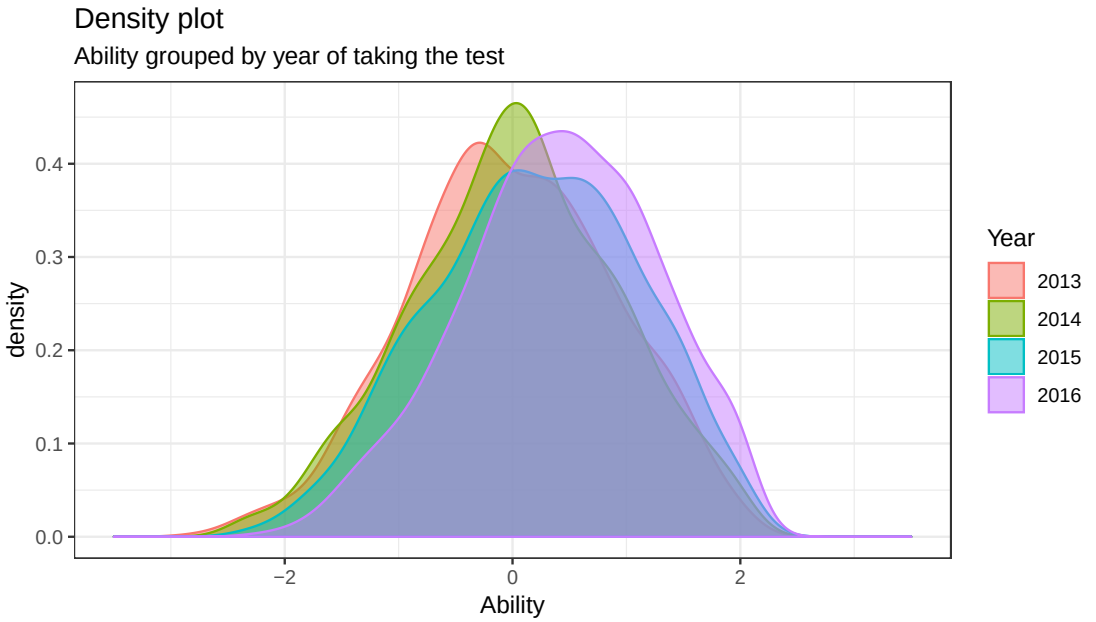

Figure 3.9: Density plot of the expected abilities of students grouped by year of taking the test

In Table 3.11 are shown the basic descriptive statistics of the distributions of the estimated abilities.

	n	mean	sd	min	max	skew	kurtosis
2013	832	-0.01	0.90	-2.75	1.97	-0.09	-0.37
* 2014	942	0.02	0.90	-2.37	1.97	-0.08	-0.35
2015	729	0.20	0.90	-2.28	1.97	-0.15	-0.59
2016	763	0.41	0.85	-2.19	1.97	-0.23	-0.41

Table 3.11: Basic descriptive statistics of the distributions of the estimated abilities (* indicates the pivot test)

The two-sample Kolmogorov-Smirnov (K-S) test is a nonparametric test comparing the empirical distribution functions of two samples and testing the null hypothesis that the two samples are drawn from the same distributions. This test can be used for examining the changes in both location and shape of the distributions of the estimated abilities.

In Table 3.12 are shown the two-sample K-S test statistics and p-values of the data from consecutive years. The non-significant p-value of the test on the data in 2013 and 2014 suggests students from these two years are drawn from the same distribution of ability, while there are significant differences between students from 2014 and 2015, and between students from 2015 and 2016.

	statistic	p-value
2013 vs 2014	0.0542	0.1498
2014 vs 2015	0.1124	0.0001
2015 vs 2016	0.1107	0.0002

Table 3.12: Kolmogorov-Smirnov test statistics and p-values of the data from consecutive years

Chapter 4

Evaluating MDT3

An important objective of this project is to evaluate the performance of the new MDT (MDT3) with different questions based on the suggestions from the report last year.

4.1 Approaches to test equating

A key step in assessing academic growth is test equating, which ensures that all students are placed on a same scale of ability. There are three approaches to IRT equating [9]:

1. Separate calibration with linear transformation, which determines the linear relationship between the difficulty parameters in consecutive years.
2. Fixed common item parameter (FCIP) calibration, which fixes the parameters of the equating items so that these parameters are not estimated in the calibration's future years and remain unchanged, resulting in placing all item parameters on a common metric.
3. Concurrent parameter calibration, where the data for all years are calibrated together in one IRT model.

In this chapter, all these three approaches are used and the performances of these test equating approaches are evaluated.

4.2 Testing of assumptions

Before IRT models are fitted and tests are equated, the assumptions of local independence and dimensionality need to be checked as they are in Chapter 3.

4.2.1 Local independence

The Kendall correlation and the significance indicators of the Kendall' τ test performed on each pair of items are shown in Figure 4.1. It is shown in the figure that the correlation between any pair of questions are positive. The null hypotheses that the questions in each pair are statistically independent are rejected at 95% significant level, except the pair (Q1, Q15).

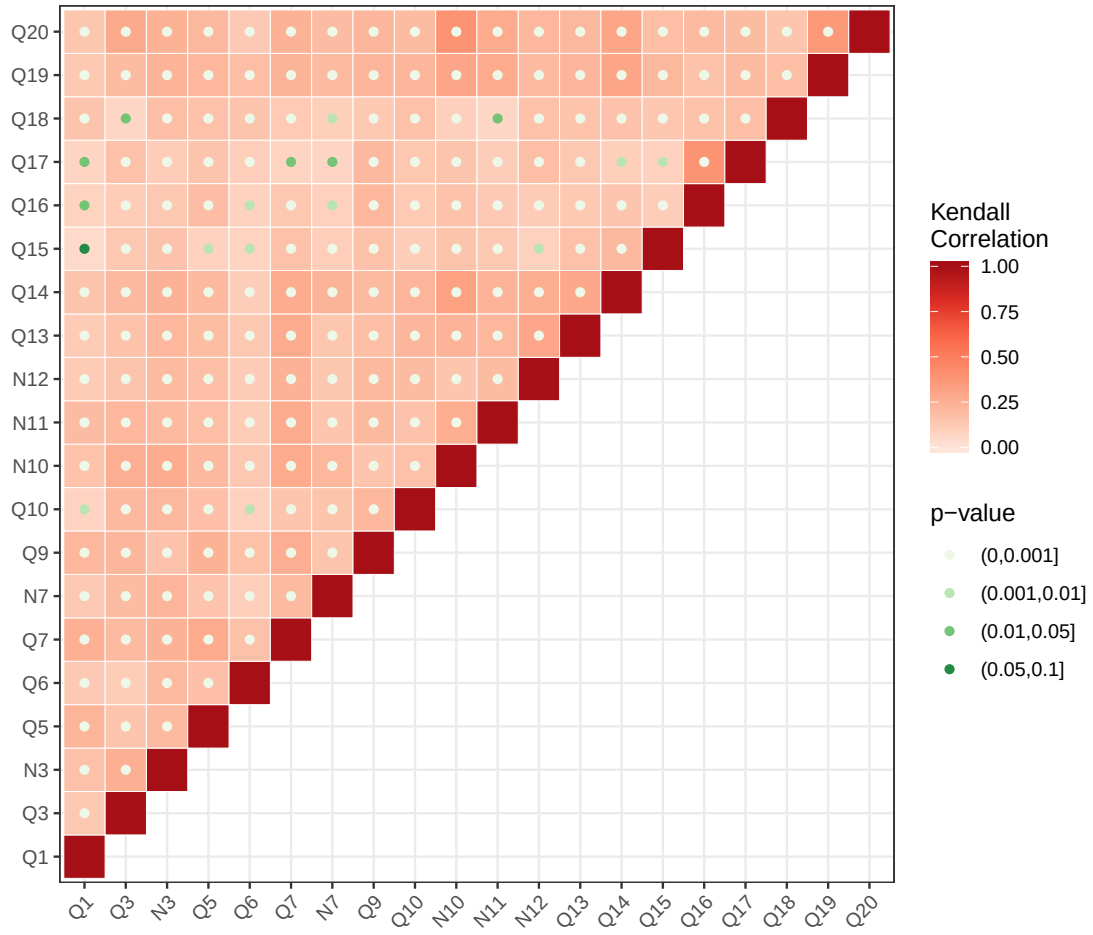

Figure 4.1: Correlation heat map and significance indicators of Kendall's τ test

The six largest p-values of the Kendall's τ test are shown in Table 4.1. The p-value of the test of the correlation between Q1 and Q15 is 0.0907, indicating the null hypothesis cannot be rejected at 95% significance level. However, the significance probability of the same test using the data from previous MDTs is 4.14×10^{-7} and the Kendall correlation between Q1 and Q15 is 0.08 (Figure 3.1), suggesting the correlation exists. According to the mathematics taxonomy provided by the report last year, both Q1 and Q15 are Type-A questions and require routine use of procedures [3].

Item i	Item j	pvals
Q1	Q15	0.0907
Q3	Q18	0.0351
N11	Q18	0.0273
N7	Q17	0.0265
Q1	Q17	0.0160
Q7	Q17	0.0141

Table 4.1: Six largest p-values of Kendall's τ test

4.2.2 Dimensionality

A scree plot of the eigenvalues for a principal axis factor analysis and the plots of simulated and resampled data are shown in Figure 4.2. The eigenvalue of factor number 2 decreases significantly and the difference between simulated data and actual data is minimised at the

point of factor number 7, suggesting that an integer between 1 and 7 would be a good choice for the number of factors to be extracted. This result is similar to the interpretation of the scree plots in Figure 3.2.

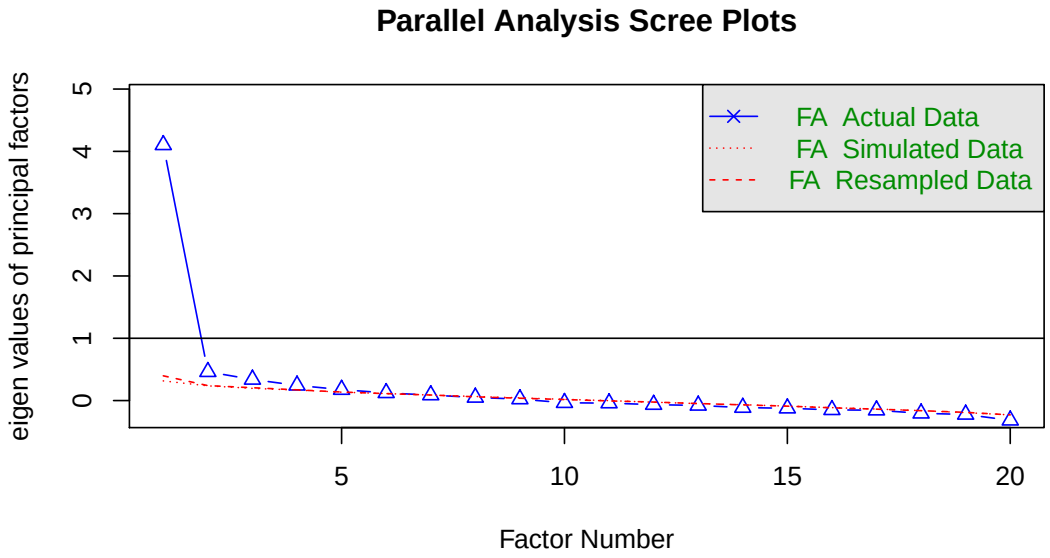

Figure 4.2: Parallel analysis scree plots

In Table 4.2 are shown the standardised loadings of items on the factor of a one-factor exploratory factor analysis (EFA) model. All the loadings of the items are significant, with 0.321 being the smallest factor loading.

	Type	MR1	
Q1	A	0.322	*
Q3	B	0.437	*
N3	A	0.524	*
Q5	A	0.441	*
Q6	A	0.308	*
Q7	B	0.521	*
N7	B	0.375	*
Q9	A	0.459	*
Q10	B	0.436	*
N10	B	0.598	*

	Type	MR1	
N11	C	0.486	*
N12	A	0.467	*
Q13	A	0.495	*
Q14	B	0.575	*
Q15	A	0.328	*
Q16	A	0.352	*
Q17	A	0.338	*
Q18	A	0.321	*
Q19	B	0.536	*
Q20	B	0.556	*

Table 4.2: Loadings of items on factor where '*' indicates loadings greater or equal to 0.25

The loadings are also plotted along the Factor 1 axis in Figure 4.4. Comparing with Figure 3.4, all questions in MDT3 have loadings on the factor higher than 0.3, while two of the questions in MDT2 do not. N7, one of the new Type-B questions in MDT3, has relatively lower loadings than other Type-B questions. The new Type-C question, N11, has higher loading than most of the Type-A questions, but it is not comparable with Type-B questions because there is only one Type-C question in the test and has mid-level of loading.

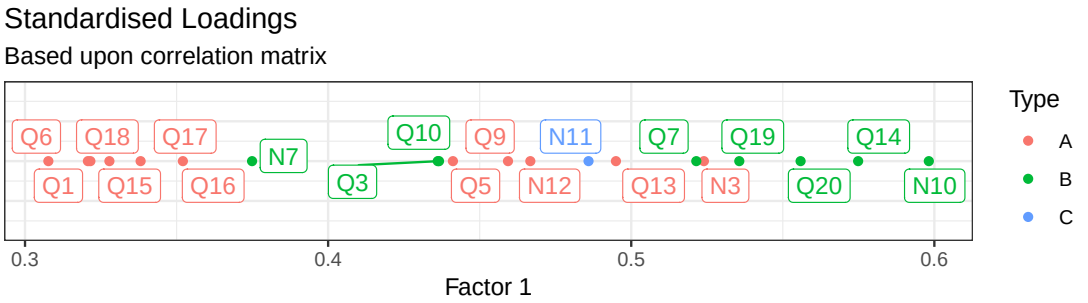

Figure 4.3: Standardised loadings on Factor 1

An additional factor is added to the EFA model in the next step to examine whether the results are improved. The loadings of the items on the two factors which are greater or equal to 0.25 are shown in Table 4.3. In this model, the item Q18 becomes non-significant, which has approximately 0.2 loadings on both factors before they are cut off. Q16 and Q17 have significant loadings only on Factor 2 in contrast to other items, which is the same case as in Table 3.3.

	Type	MR1	MR2
Q1	A	0.329	
Q3	B	0.418	
N3	A	0.500	
Q5	A	0.382	
Q6	A	0.275	
Q7	B	0.575	
N7	B	0.407	
Q9	A	0.355	
Q10	B	0.370	
N10	B	0.643	

	Type	MR1	MR2
N11	C	0.507	
N12	A	0.405	
Q13	A	0.479	
Q14	B	0.624	
Q15	A	0.311	
Q16	A		0.511
Q17	A		0.704
Q18	A		
Q19	B	0.503	
Q20	B	0.532	

Table 4.3: Loadings of items on factors which are greater or equal to 0.25 only

In Figure 4.4 is shown the scatter plot of the standardised loadings on both factors. The items loaded heavily on Factor 1 are more divergent from the items loaded heavily on Factor 2, comparing with the loadings of the questions from previous MDTs showing in Figure 3.4.

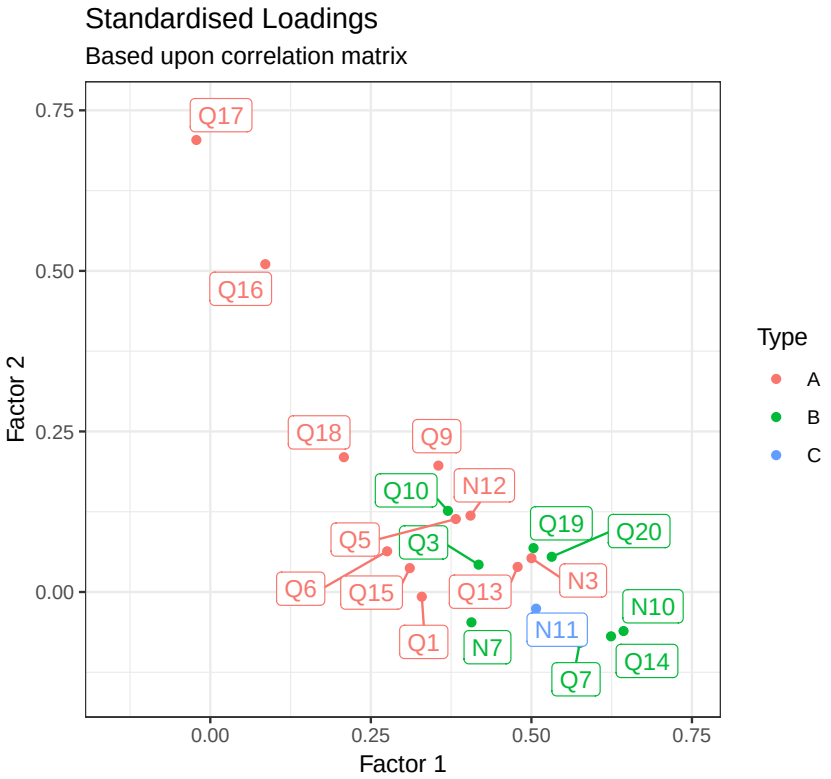

Figure 4.4: Standardised loadings on Factor 2 against Factor 1

The statistics representing the adequacy of the model, including RMSR, RMSEA and Tucker Lewis Index seem to be satisfactory in both one-factor and two-factor EFA models. When the number of factors in the EFA models was increased to 3, 4 or more, the simple-structure (i.e. all items have significant but single loading) of the items does not hold. Therefore, either one-factor or two-factor model should be chosen for further analyses of the data of MDT3.

In Chapter 3, the one-factor structure of the latent ability is chosen because of the simplicity of interpretation. Since the multidimensional PCM functions of `ltm` and `mirt` packages cannot work with non-standard scoring functions (e.g. questions scored 0, 2, 3, 5), it would be difficult to proceed with a multi-factor model. Additionally, our aim is to compare MDT2 and MDT3 and we will continue with the one-factor structure.

However, it is worthwhile and necessary to analyse the two-dimensional model for all data, with self-defined functions for computing inferences with modified scoring functions in the future.

4.3 Separate calibration

Under the equating method of separate calibration with linear transformation, a generalised partial credit model (Model `MDT3-GPCM`) is fitted to to the data from MDT3 separately.

The item response curves, test response curve, item information curves and test information curve based on the untransformed parameters are shown in Figure 4.5. It can be seen that the new question N10 has the highest information at the range of positive ability.

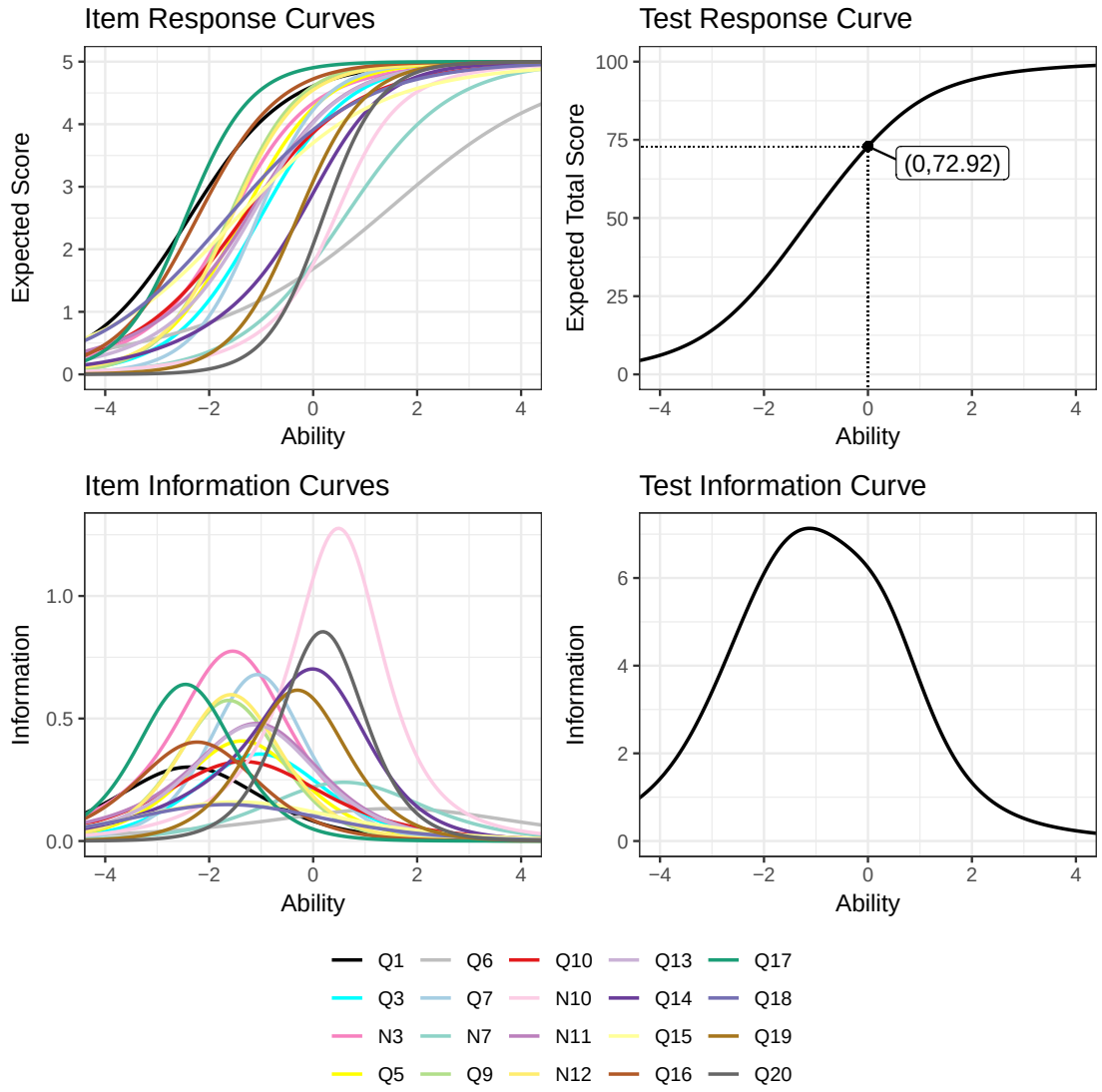

Figure 4.5: Plots of GPCM using separate calibration (Model `MDT3-GPCM`)

To place all the students on the comparable scale of ability and examine how the accuracy of estimating ability changes, the linear relationship between the difficulty parameters d_{ik} of the equating items is determined.

The scatter plot and of the difficulty parameters and the linear regression line with confidence intervals are shown in Figure 4.6. The regression of the difficulty parameters of Model `MDT3-GPCM` d^* on those of Model `MDT2-GPCM` d is estimated by

$$d^* = 0.05 + 0.94d.$$

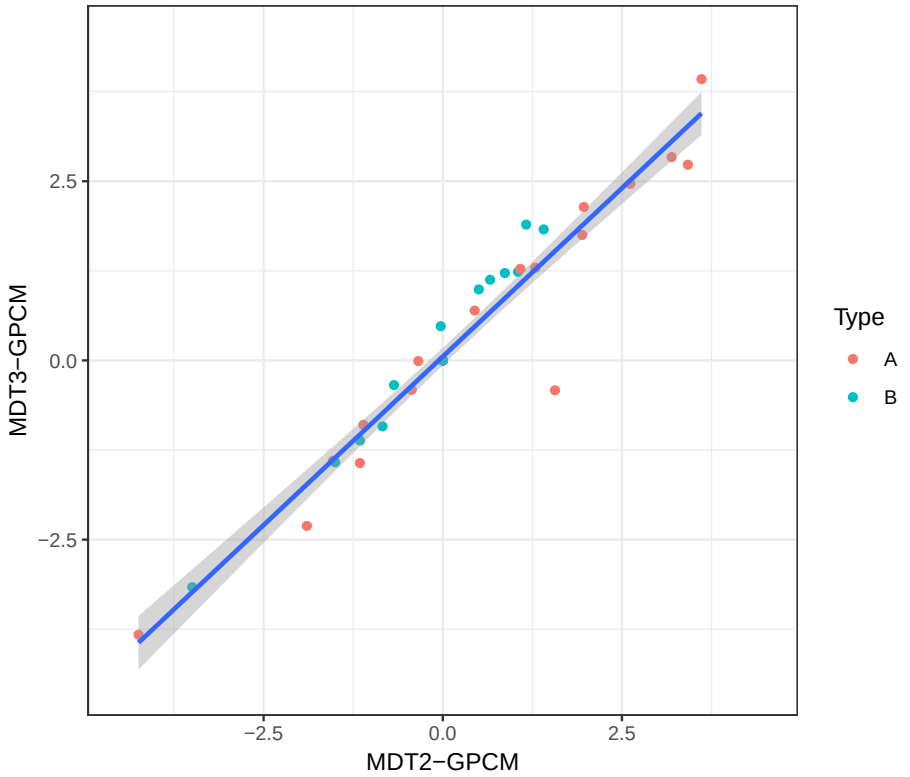

Figure 4.6: Difficulty parameters of Model `MDT2-GPCM` and Model `MDT3-GPCM` and the linear regression line

This linear transformation is applied to the difficulty parameters of all the items from Model `MDT3-GPCM`. The item response curves, test response curve, item information curves and test information curve with linear transformation of the new model (Model `MDT3-SC`) are plotted in Figure 4.7. These plots are now comparable with the plots of Model `MDT2-GPCM` (Figure 3.6), because of the common metric between both calibrations.

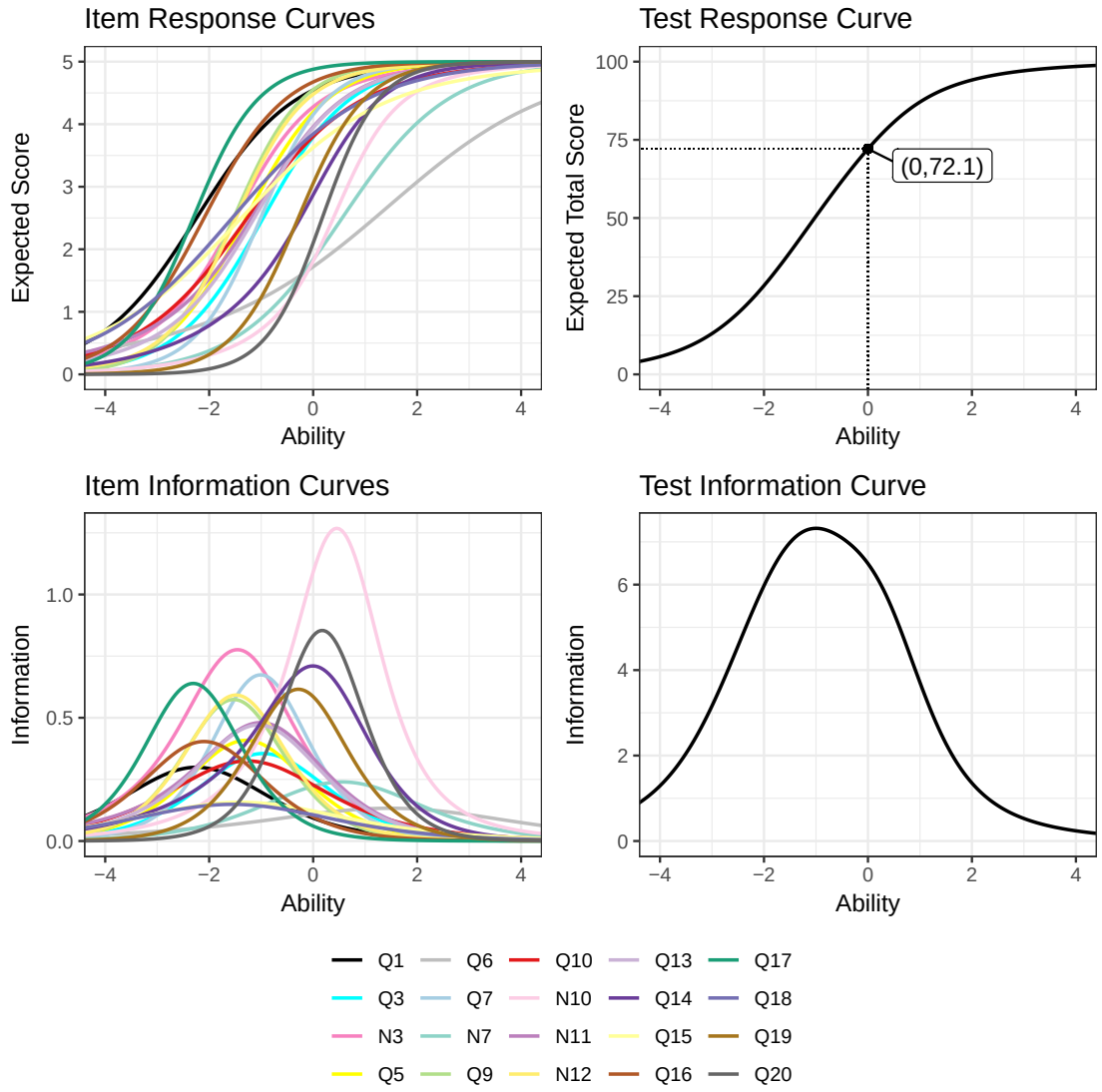

Figure 4.7: Plots of GPCM using separate calibration with linear transformation (Model `MDT3-SC`)

The expected total score of a student with ability 0 after transforming the parameters is 72.1, while the expected total score is 72.92 before transforming, indicating the students from 2017 have higher ability than the students admitted before 2017. However, due to the limited functions in the package `mirt`, the estimated abilities of students are not yet available to be computed.

4.4 FCIP calibration

Under the equating method of fixed common item parameter calibration, the parameters of the equating items in the GPCM on the new test (Model `MDT3-FCIP`) are fixed to the estimated values of parameters in the GPCM on the old test (Model `MDT2-GPCM`). The item and test response curves and information curves are shown in Figure 4.8. In this calibration, the scale of the ability axis, the response and information curves of the equating items, are exactly the same as Model `MDT2-GPCM`.

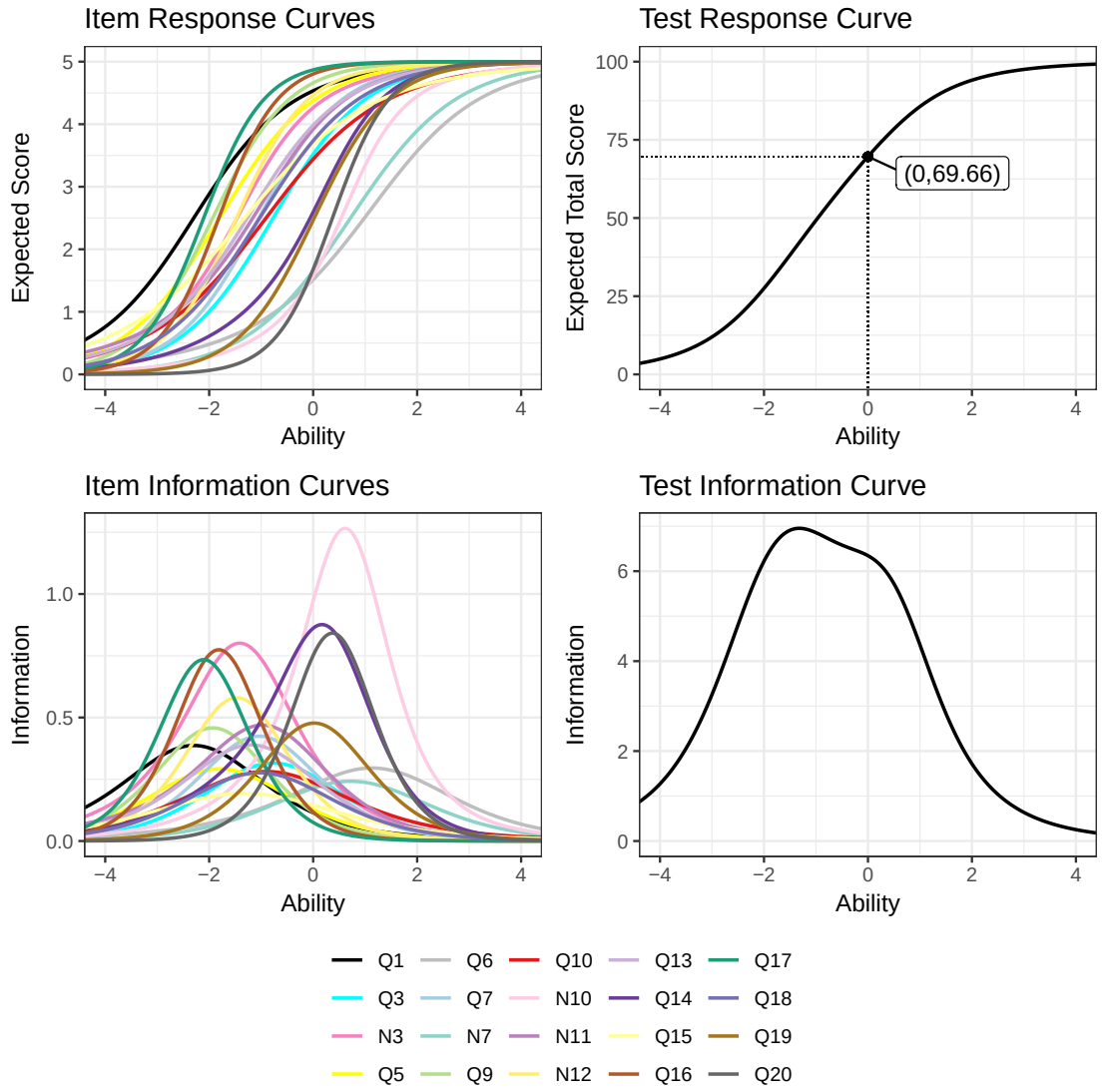

Figure 4.8: Plots of GPCM using fixed common item parameter calibration (Model `MDT3-FCIP`)

The increase of accuracy of estimating the abilities of students is one of the indices that determine the improvement of MDT. According to the definition of item information and standard error of measurement (SEM), a higher item information will lower the SEM and narrow the confidence intervals of the estimated abilities.

The item response curves shown in Figure 4.9 and the expected scores of the students with average ability and the changes of the expectations of the students with ability ± 2 shown in Table 4.4 can be used for comparing the changes in discriminations and difficulties of the new questions and the removed questions.

N3 and N12, which are the same questions as Q4 and Q12 but with different marking schemes, are more discriminating but an average student can score higher marks under the new marking schemes. Additionally, the abilities of a student who score half of the marks are estimated at approximately the same values under both marking schemes. The rest of the new questions, N7, N10 and N11, are much more discriminating and difficult than the questions suggested to be removed.

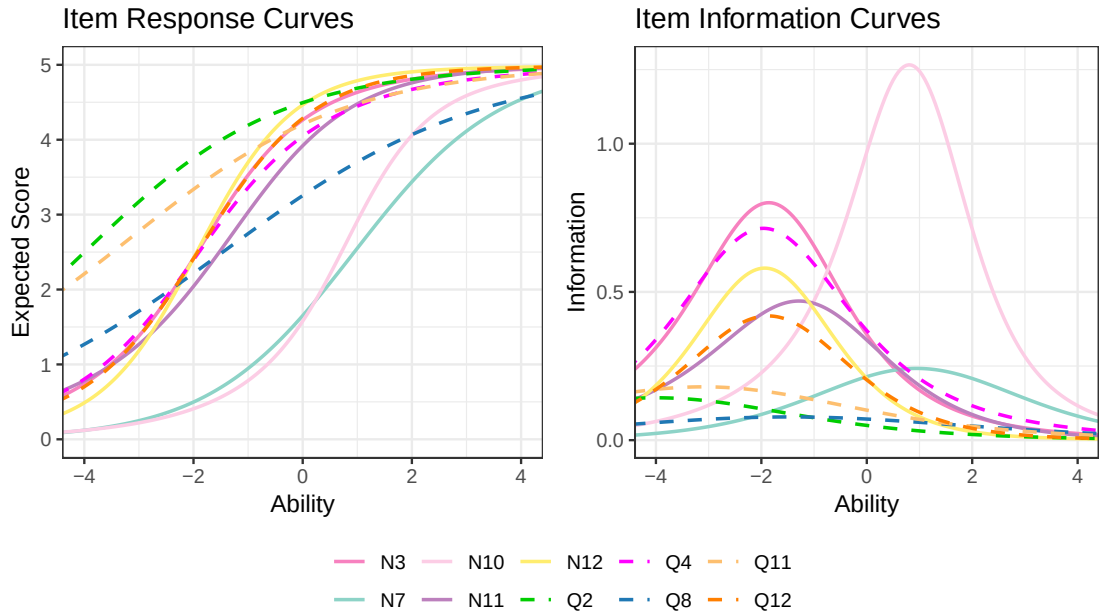

Figure 4.9: Item response curves and item information curves of added items in MDT3 of Model `MDT3-FCIP` and removed items from the previous test of Model `MDT2-GPCM`

	$\bar{T}_i(-2) - \bar{T}_i(0)$	Facility	$\bar{T}_i(2) - \bar{T}_i(0)$	Discrimination
N3 (Q4)	-2.54	4.26	0.61	3.15
N12 (Q12)	-2.88	4.47	0.47	3.35
N7	-1.33	1.65	2.24	3.57
N10	-1.30	1.58	2.87	4.17
N11	-2.40	3.92	0.93	3.34
Sum.add	-10.46	15.88	7.13	17.58
Q4	-2.27	4.05	0.71	2.99
Q12	-2.57	4.29	0.62	3.19
Q2	-1.09	4.49	0.36	1.46
Q8	-1.36	3.25	1.00	2.36
Q11	-1.22	4.21	0.54	1.76
Sum.rm	-8.51	20.29	3.24	11.76

Table 4.4: Expected scores of students with ability -2, 0 and 2

The item information curves shown in Figure 4.9 and the major and total item information areas for the new and removed questions shown in Table 4.5 can be used for comparing the changes in information of the new and removed questions.

The item information functions of N3 and N12 attain their maximum at approximately the same values of ability as Q4 and Q12 respectively. Their information areas over the whole range of ability increase slightly, as well as the areas over the range -2 to 2. However, the information areas for the above-average students drop slightly in these two questions. The new questions have much larger information areas than the questions suggested to be removed, especially N10.

In Figure 4.10 are shown the sums of item information functions of the two different question sets, where the improvement of estimating ability can be seen clearly.

		Major Info Area			Total Info Area		
	Max at	Lower	Upper	Total	Lower	Upper	Total
N3	-1.41	1.32	0.31	1.63	2.21	0.38	2.59
N12	-1.47	0.91	0.15	1.06	1.41	0.17	1.58
N7	0.72	0.26	0.44	0.70	0.32	0.66	0.98
N10	0.61	0.87	1.92	2.80	1.06	2.29	3.35
N11	-0.98	0.84	0.33	1.17	1.37	0.38	1.75
Q4	-1.49	1.21	0.38	1.59	2.15	0.51	2.66
Q12	-1.42	0.71	0.17	0.88	1.18	0.20	1.37
Q2	-3.04	0.17	0.06	0.23	0.70	0.08	0.78
Q8	-1.11	0.15	0.11	0.26	0.36	0.20	0.56
Q11	-2.36	0.29	0.13	0.42	1.00	0.19	1.19

Table 4.5: The area under each item information curve over different definite integral ranges

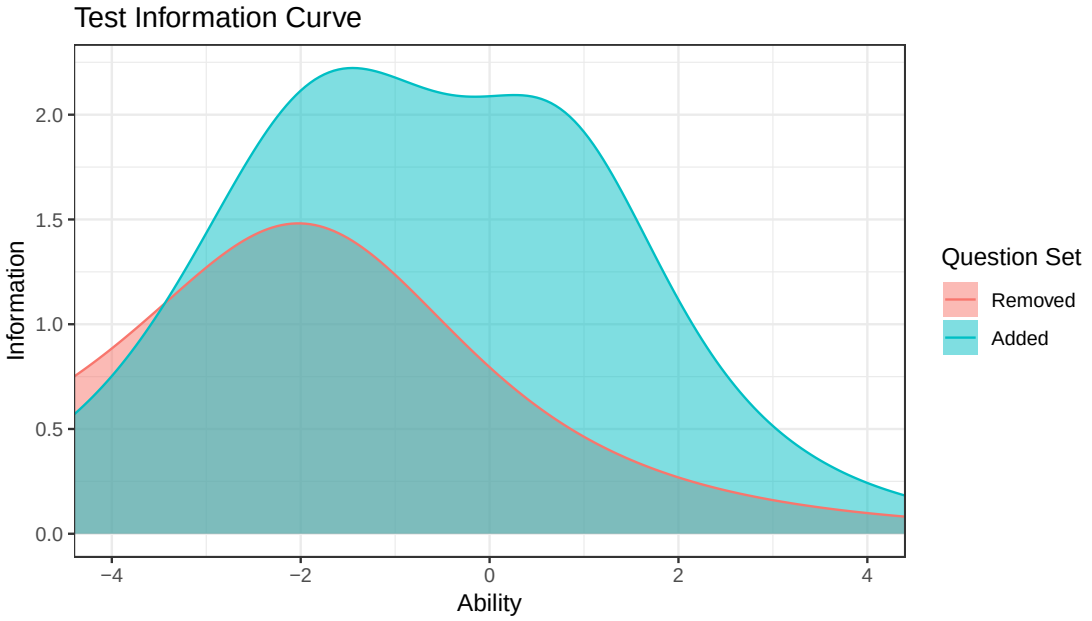

Figure 4.10: Sums of item information functions of different sets of questions

The test information function attains its maximum at $\theta = -1.32$, which is shown in Table 4.6. Comparing with the test information of Model `MDT2-GPCM` in Table 3.9, the area over the range of positive ability increases significantly.

		Major Info Area			Total Info Area		
	Max at	Lower	Upper	Total	Lower	Upper	Total
Test Information	-1.32	13.38	8.50	21.87	21.76	10.31	32.07

Table 4.6: The area under the test information curve of Model `MDT3-FCIP` over different definite integral ranges

In Table 4.7 are shown the major and total item information areas and their proportions of the test information in Model `MDT2-GPCM`. In terms of major information area, the test information area increases by 22.29% in MDT3. The new questions N10 and N11 account for a large proportion of the increase, while N7 is the least informative among the new questions.

The density plots of the estimated abilities of the students taking two different tests using the method of EAP are shown in Figure 4.11, where there is a trend that the distribution of the students examined in 2017 shifts to the right.

Item	Type	MajorInfo	Prop (%)
Q15	A	0.54	3.02
Q5	A	0.58	3.25
Q1	A	0.60	3.37
Q9	A	0.70	3.92
N7	B	0.70	3.93
Q18	A	0.73	4.09
Q17	A	0.78	4.35
Q10	B	0.82	4.58
Q6	A	0.83	4.62
Q3	B	0.85	4.75
Q13	A	0.95	5.29
Q7	B	1.00	5.62
Q16	A	1.02	5.71
N12	A	1.06	5.94
N11	C	1.17	6.54
Q19	B	1.22	6.81
N3	A	1.63	9.11
Q20	B	1.72	9.64
Q14	B	2.17	12.11
N10	B	2.80	15.64
Sum		21.87	122.29

Item	Type	TotalInfo	Prop (%)
Q15	A	0.98	3.46
N7	B	0.98	3.47
Q18	A	1.05	3.71
Q5	A	1.08	3.81
Q3	B	1.12	3.96
Q10	B	1.28	4.52
Q6	A	1.34	4.71
Q7	B	1.35	4.77
Q9	A	1.35	4.77
Q19	B	1.38	4.87
Q13	A	1.46	5.13
Q1	A	1.49	5.25
N12	A	1.58	5.58
Q17	A	1.71	6.04
N11	C	1.75	6.18
Q16	A	1.76	6.20
Q20	B	1.84	6.47
N3	A	2.59	9.13
Q14	B	2.61	9.20
N10	B	3.35	11.82
Sum		32.07	113.04

Table 4.7: Tables of item information of Model `MDT3-FCIP` and corresponding proportions of test information in Model `MDT2-GPCM`

Figure 4.11: Density plots of the estimated abilities of students grouped by the test taken

The mean of the estimated abilities of the students taking MDT3 shown in Table 4.8 are higher than the students taking MDT2. The K-S test of these two sets of estimated abilities suggests that the students are drawn from different distributions of ability (with p-value 0.002).

	n	mean	sd	min	max	skew	kurtosis
MDT2	3248	0.02	0.90	-2.73	1.86	-0.09	-0.48
MDT3	896	0.15	0.90	-2.87	2.04	-0.26	-0.26

Table 4.8: Basic descriptive statistics of the distributions of the estimated abilities

4.5 Concurrent parameter calibration

Under the equating method of concurrent parameter calibration, a generalised partial credit model (Model `MDT3-CP`) is fitted to the data of all the students taking MDT2 and MDT3 together.

The item and test response curves and information curves are plotted in Figure 4.12. The red curves in the test response curve and information curve plots represent the questions from MDT2, while the turquoise curves represent the questions in MDT3. It can be seen that the test response curve of MDT3 is on the right of the curve of MDT2, indicating that the MDT3 is more difficult. The test information curve of the MDT3 is higher than the old test, which is consistent with the findings from Model `MDT3-FCIP`.

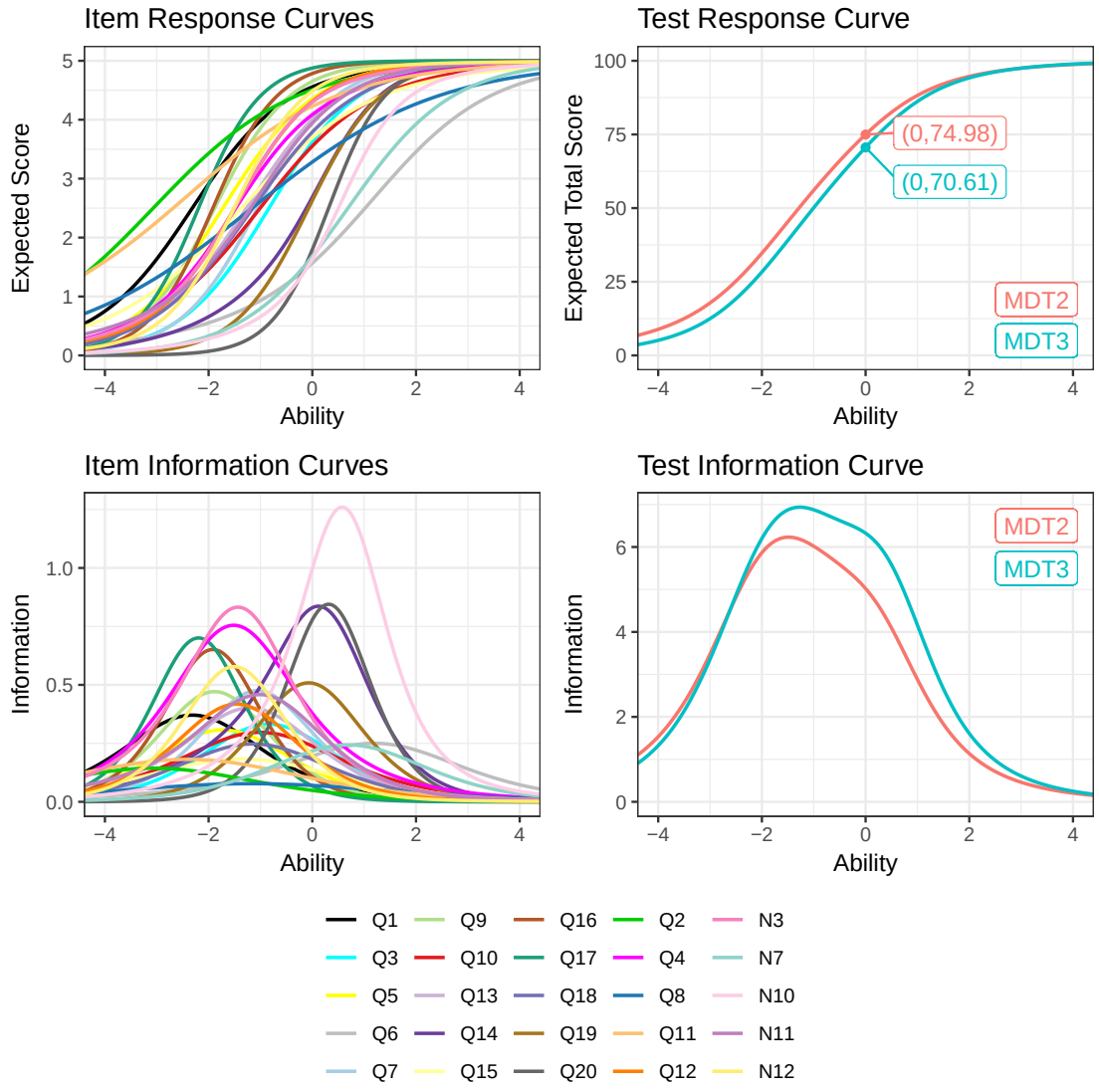

Figure 4.12: Plots of GPCM using concurrent parameter calibration (Model `MDT3-CP`)

The item response curves and the item information curves for the new questions and the removed questions are shown separately in Figure 4.13. It should be noted that the plots are

similar to the plots of Model `MDT3-FCIP` in Figure 4.9.

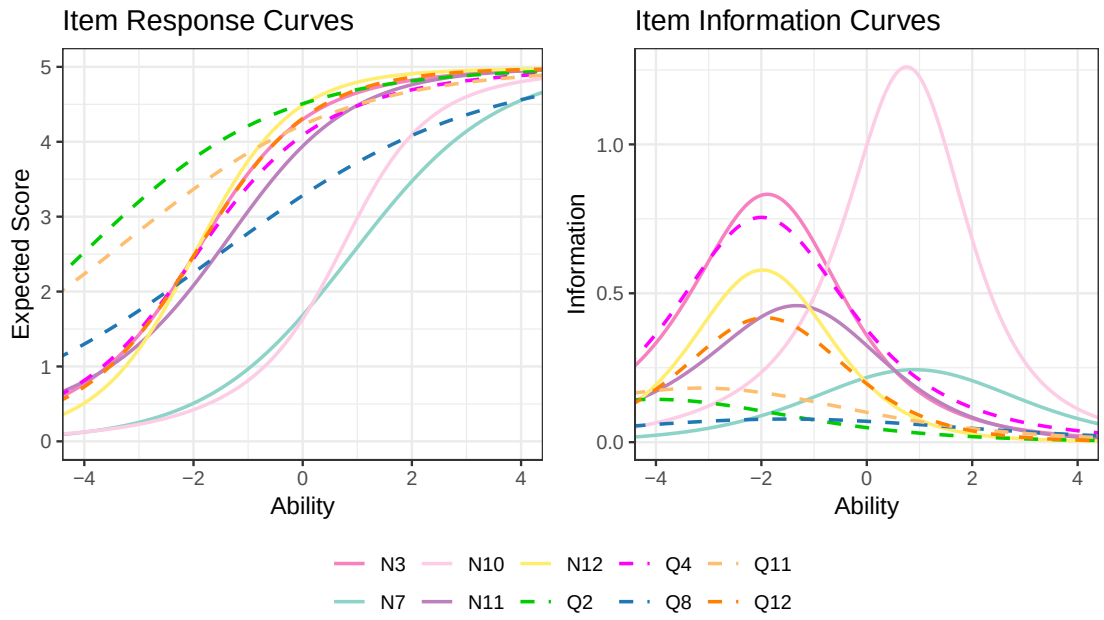

Figure 4.13: Item response curves and item information curves of added items in MDT3 and removed items from the previous test of Model `MDT3-CP`

The major and total information areas and their proportions of the sum of information areas of the 25 questions are shown in Table 4.9, where items in italic style are the questions removed from the test and items in bold style are the new questions.

The density plots of the estimated abilities of students grouped by the test taken are shown in Figure 4.14 and the descriptive statistics are shown in Table 4.10.

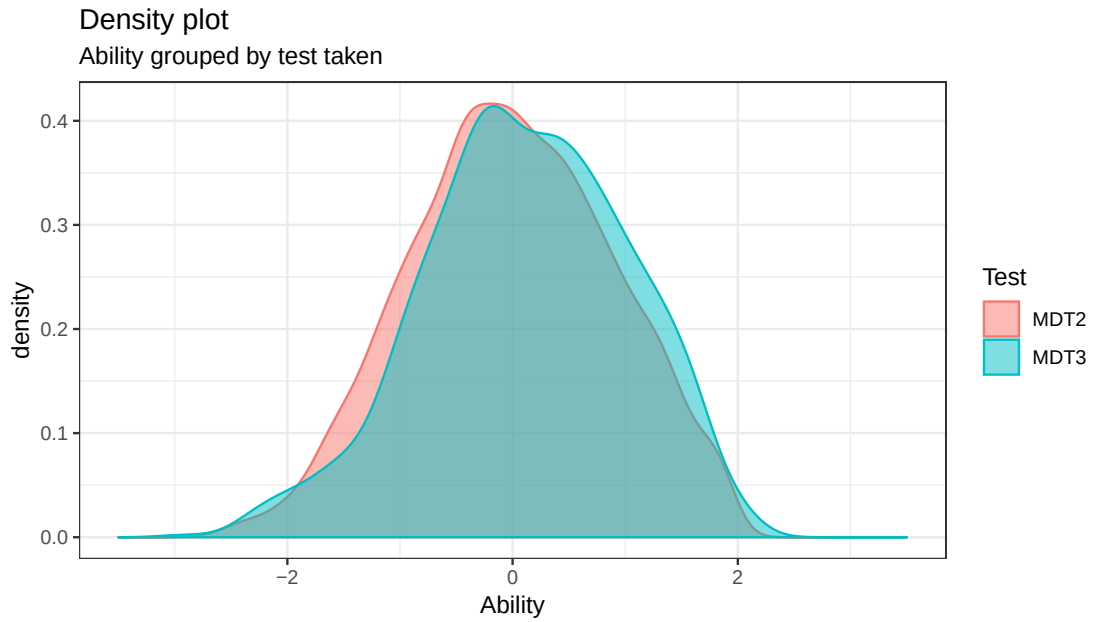

Figure 4.14: Density plots of the estimated abilities of students grouped by test taken

Although the results given by Model `MDT3-FCIP` and Model `MDT3-CP` are similar, there is a fundamental difference between these two models. Concurrent parameter (CP) calibration is

Item	Type	MajorInfo	Prop (%)
Q2	*A*	*0.23*	*0.90*
Q8	*A*	*0.26*	*1.04*
Q11	*A*	*0.41*	*1.66*
Q15	A	0.51	2.06
Q1	A	0.56	2.25
Q5	A	0.62	2.48
Q18	A	0.66	2.62
Q17	A	0.70	2.80
N7	**B**	**0.71**	**2.84**
Q6	A	0.72	2.88
Q9	A	0.73	2.94
Q10	B	0.84	3.36
Q16	A	0.86	3.43
Q12	*A*	*0.86*	*3.45*
Q3	B	0.87	3.49
Q13	A	0.97	3.87
N12	**A**	**1.04**	**4.17**
Q7	B	1.08	4.33
N11	**C**	**1.14**	**4.56**
Q19	B	1.27	5.08
Q4	*A*	*1.65*	*6.60*
N3	**A**	**1.67**	**6.68**
Q20	B	1.73	6.94
Q14	B	2.09	8.38
N10	**B**	**2.80**	**11.19**

Item	Type	TotalInfo	Prop (%)
Q8	*A*	*0.56*	*1.44*
Q2	*A*	*0.79*	*2.04*
Q15	A	0.96	2.48
N7	**B**	**0.99**	**2.56**
Q18	A	0.99	2.58
Q5	A	1.11	2.87
Q3	B	1.15	2.99
Q11	*A*	*1.19*	*3.09*
Q6	A	1.24	3.22
Q10	B	1.32	3.43
Q12	*A*	*1.37*	*3.56*
Q9	A	1.37	3.56
Q1	A	1.40	3.64
Q19	B	1.43	3.70
Q7	B	1.43	3.71
Q13	A	1.48	3.85
N12	**A**	**1.58**	**4.10**
Q16	A	1.61	4.18
Q17	A	1.67	4.34
N11	**C**	**1.72**	**4.47**
Q20	B	1.84	4.77
Q14	B	2.55	6.61
N3	**A**	**2.68**	**6.96**
Q4	*A*	*2.78*	*7.20*
N10	**B**	**3.34**	**8.67**

Table 4.9: Tables of item information of Model 17-FCIP and corresponding proportions of test information in Model PRE17-GPCM (*Italic: questions removed*, **Bold: questions added**)

	n	mean	sd	min	max	skew	kurtosis
MDT2	3248	-0.02	0.90	-2.77	1.82	-0.07	-0.48
MDT3	896	0.11	0.90	-2.92	2.01	-0.25	-0.27

Table 4.10: Basic descriptive statistics of the estimated abilities

the easiest approach to test equating because only one IRT model is fitted to a modified data set. However, the scale of the ability in CP calibration is a combination of the scales of ability of students from different years, which is not equal to any scale of ability from a specific year. If the test from a future year needs to be equated, the scale of the new model is not comparable with the existing one. Therefore, CP calibration is more of a one-off analysis.

4.6 Future MDT

Improving MDT is a long-term task for the School of Mathematics. Based on the analyses in the chapter, a number of new criterion can be summarised, along with the criterion used in the summer 2017 project.

- Item with low Major Info in Model `MDT3-FCIP` and Model `MDT3-CP` should be removed.
- Item which attains the maximum of its information function at positive value of ability should be kept.

Therefore, in accordance with MDT3, we suggest removing Q15 and Q1.

Chapter 5

Prediction of Outcomes

To what extent do the results of mathematics diagnostic test (MDT) predict students' performance in mathematics courses since MDT changed in 2017, the performances of prediction of MDT2 and MDT3 will be compared in this chapter.

5.1 Introduction

The raw data consists of the country of domicile based on fee status group, school, gender and assessment mark and grade in each course, and entry qualifications of the students.

Students will be classified into four different groups of countries of domicile based on their fee status group. The four groups are Scotland, rest of the UK, rest of the European Union and other overseas countries.

The classification of schools of the students are School of Mathematics, School of Informatics, School of Physics, School of Economics and other schools.

In this project, only seven of the courses are considered, including Introduction to Linear Algebra (ILA), Calculus and its Applications (CAP), Proofs and Problem Solving (PPS), Mathematics for Physics 1 (MfP1), Accelerated Algebra and Calculus for Direct Entry (AAC), Accelerated Proofs and Problem Solving (APPS) and Several Variable Calculus and Differential Equations (SVCDE). Within these seven courses, ILA, CAP and PPS are Year 1 mathematics courses, MfP1 is the course for students who need extra mathematics practice, AAC and APPS are mathematics courses for Year 2 Direct Entry students and SVCDE is a Year 2 course for all mathematics students.

According to the raw data of the course results, some special categories of assessment grades of students are recorded. Some students also made more than one attempt on a course. In order to reflect the actual outcomes after one year of study, the data need to be cleaned by the following procedures.

1. Exclude the students with grade NO (non-assessed), WD (withdrawn) or NS (null sit with special circumstances);
2. Keep only the results of the very first attempt in a course for each student;
3. Map grade FF (force fail) onto grade E;
4. Map grade ES (credits awarded with special circumstances) onto grade D; and
5. Treat grade AN (absent) as a grade ranked after grade H.

The entry qualifications of students includes Scottish Qualifications Authority (SQA) national qualifications, General Certificate of Education (GCE) Advanced Level (A-Levels), International Baccalaureate (IB) Diploma Programme and other qualifications. These qualifications are classified into three categories, corresponding to the recommendation of taking extra mathematics introductory courses or not, including

Category 1 Strongly recommended to take extra mathematics courses, if a student has one of the following

- SQA Advanced Higher Mathematics at Grade C or below, or do not have Advanced Higher Mathematics;
- A-Level Mathematics at Grade A AND do not have A-Level Further Mathematics, or have A-Level Mathematics at Grade B or below;
- IB Mathematics SL.

Category 2 Can take extra mathematics courses if a student wants more mathematics practice and has one of the following

- SQA Advanced Higher Mathematics at Grade B;
- A-Level Mathematics at Grade A AND A-Level Further Mathematics at Grade A;
- IB Mathematics HL at Grade H5.

Category 3 Cannot take extra mathematics courses if a student has one of the following

- SQA Advanced Higher Mathematics at Grade A;
- A-Level Mathematics at Grade A*;
- IB Mathematics HL at Grade H6 or above.

Unknown Should discuss with Personal Tutor, if a student has other qualifications.

In the sections below, different subsets of the clean data set may be used to perform analyses on corresponding groups of students.

5.2 Using diagnostic test

According to the report last year, there exists a linear relationship between the three Year 1 courses (ILA, CAP and PPS) results and MDT scores, while the relationship between Year 2 course results and MDT scores is weaker [3]. In this section, the relationship between course results and MDT3 scores will be examined. The relationship between course results and the estimated abilities using GPCM is also investigated, as well as the performance of prediction of using abilities.

5.2.1 Linear regression

The scatter plots of course results against MDT scores are shown in Figure 5.1. The blue lines in the plots are the fitted linear regression lines, with the grey shadow areas as the confidence interval. The correlation coefficients of the linear regressions are annotated. The points represent students who took both the course and MDT.

It can be seen that the correlation coefficients for students taking MDT2 are generally higher than the correlation coefficients for students taking MDT2, implying that MDT3 provides better prediction of Year 1 course results. Within the three Year 1 courses, the correlation coefficients of ILA is the highest, followed by CAP and PPS.

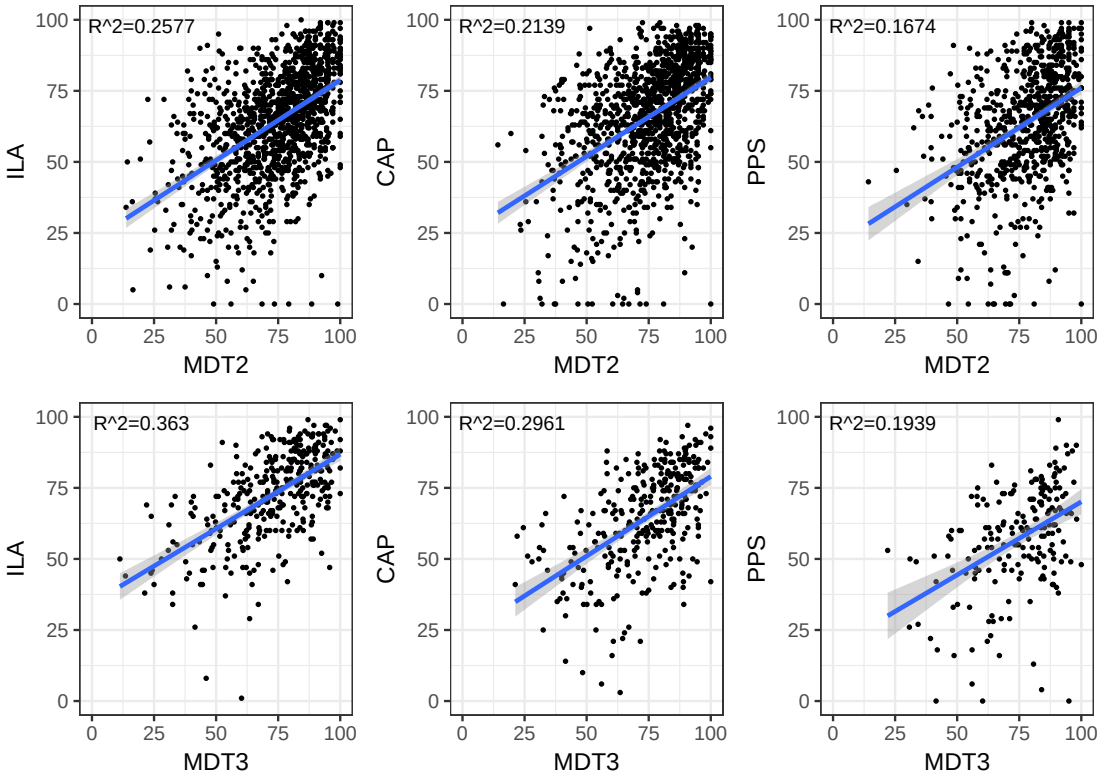

Figure 5.1: Scatter plots of course results against MDT scores and linear regression lines, confidence intervals and correlation coefficients

The corresponding plots of course results against abilities estimated by EAP are shown in Figure 5.2. The correlation coefficients of the linear regressions using estimated ability as the explanatory variable are lower for MDT2 than using MDT scores as the explanatory variable, while they are higher for MDT3. However, as discussed in Section 3.4.3, the estimated abilities come with confidence intervals which are affected by the upper limit of MDT scores, so the correlation investigated may not be accurate, especially for the above-average students.

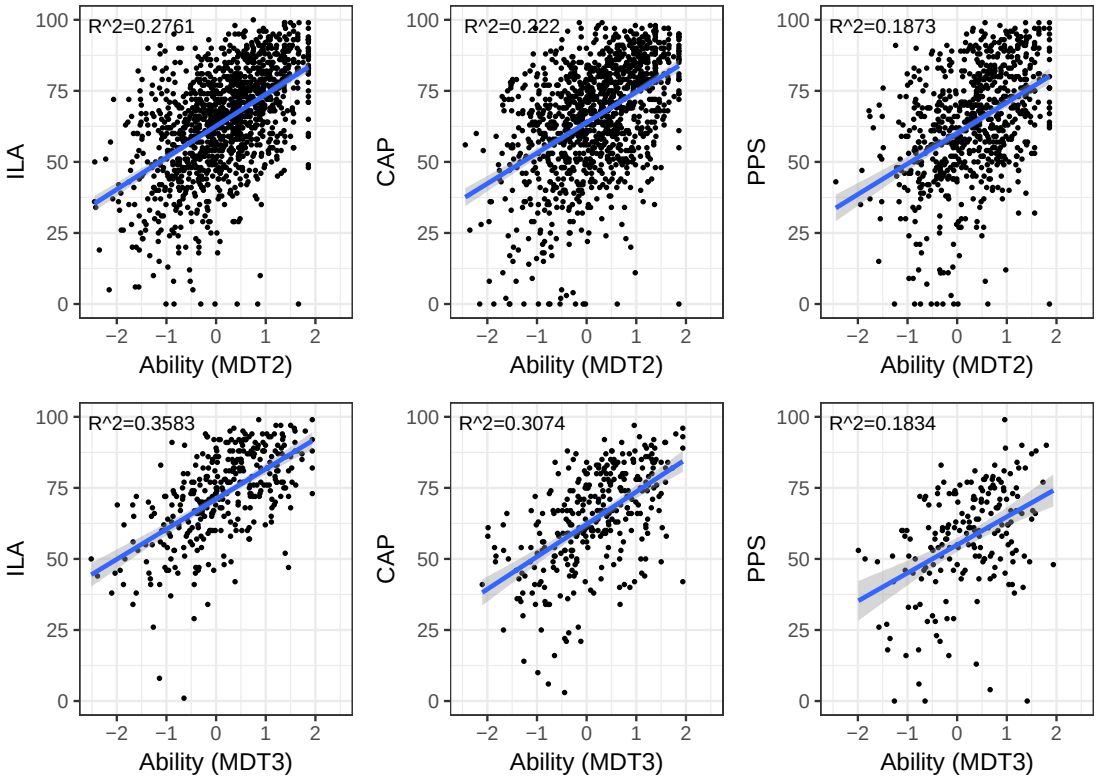

Figure 5.2: Scatter plots of course results against estimated ability (EAP) and linear regression lines, confidence intervals and correlation coefficients

It should be noted that the correlation coefficients are lower than 0.6 whichever explanatory variable is used. A possible reason of this result is that MDT is taken before the teaching weeks start, so the course results may be influenced by other factors during the semester, including but not limited to the courses taught in the lectures, attendance of the tutorials and completion of assignments.

5.2.2 Visual representation

To visualise the relationship between course results and MDT scores, we use stacked bar plots suggested by Bridgeman [2] and change width of the bars to be proportional that can represent the percentage of population in a certain group. Another advantage of using stacked bar plots is that the results can be easily interpreted by people with no knowledge about statistics.

In Figure 5.3, Figure 5.4 and Figure 5.5 are shown the stacked bar plots of the grades of Year 1 courses against different intervals of MDT scores. The population in these plots are the students who took the corresponding courses, ignoring their completion of MDT. The width of each bar represents the proportion of students who have MDT scores within the interval. The frequencies of the tiles are shown in Table 5.1, Table 5.2 and Table 5.3.

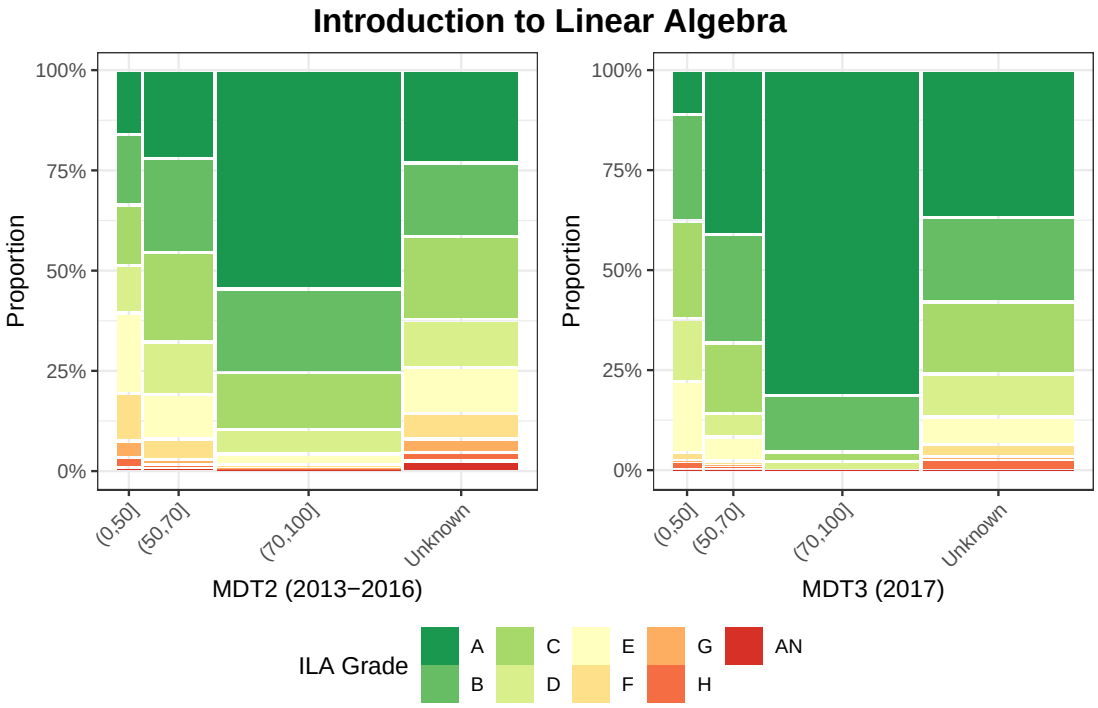

Figure 5.3: Stacked bar plots of ILA grades against MDT scores with proportional width

Introduction to Linear Algebra					
2013-2016	(0,50]	(50,70]	(70,100]	Unknown	Sum
A	19	69	445	118	651
B	21	74	171	94	360
C	18	70	117	106	311
D	14	41	49	61	165
E	24	35	22	59	140
F	14	16	7	32	69
G	5	4	2	17	28
H	3	3	0	11	17
AN	1	2	4	13	20
Sum	119	314	817	511	1761
2017	(0,50]	(50,70]	(70,100]	Unknown	Sum
A	5	35	180	80	300
B	12	23	31	46	112
C	11	15	5	39	70
D	7	5	5	23	40
E	8	5	0	15	28
F	1	1	0	7	9
G	0	0	0	1	1
H	1	1	0	6	8
AN	0	0	0	0	0
Sum	45	85	221	217	568

Table 5.1: Contingency tables of corresponding frequencies

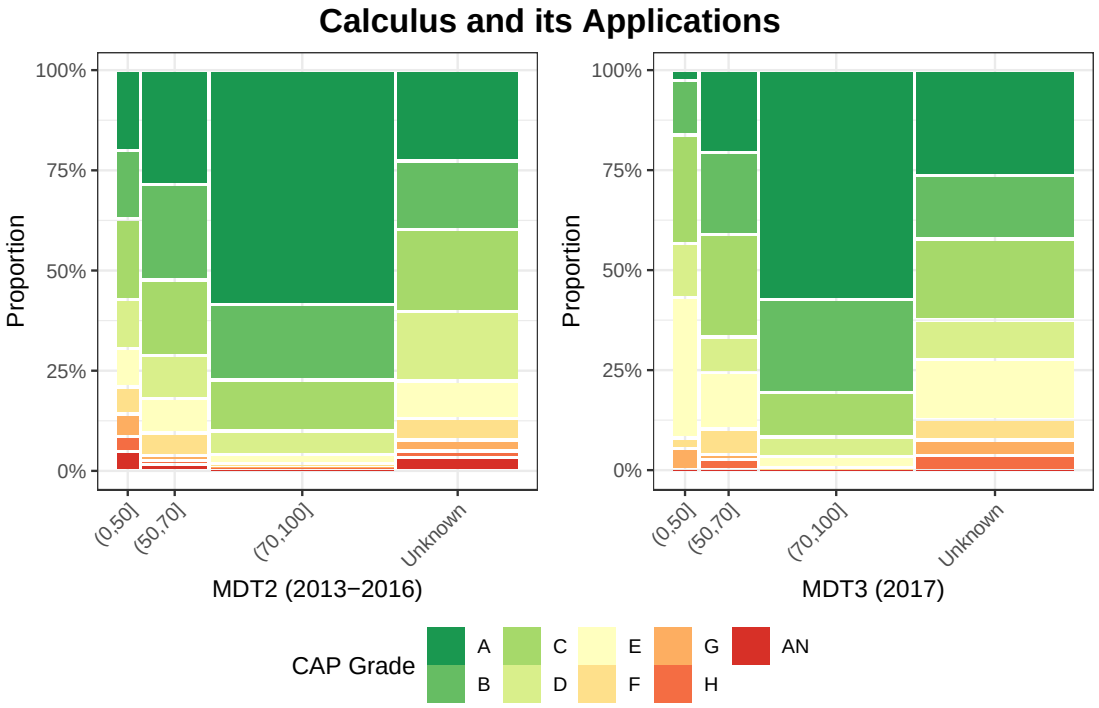

Figure 5.4: Stacked bar plots of CAP grades against MDT scores with proportional width

Calculus and its Applications					
2013-2016	(0,50]	(50,70]	(70,100]	Unknown	Sum
A	21	82	456	118	677
B	18	69	147	89	323
C	21	54	99	107	281
D	13	31	46	90	180
E	10	25	19	49	103
F	7	16	6	28	57
G	6	4	1	14	25
H	4	2	3	9	18
AN	5	5	3	17	30
Sum	105	288	780	521	1694
2017	(0,50]	(50,70]	(70,100]	Unknown	Sum
A	1	16	118	56	191
B	5	16	48	34	103
C	10	20	23	43	96
D	5	7	10	21	43
E	13	11	6	32	62
F	1	5	1	11	18
G	2	1	0	8	11
H	0	2	0	8	10
AN	0	0	0	0	0
Sum	37	78	206	213	534

Table 5.2: Contingency tables of corresponding frequencies

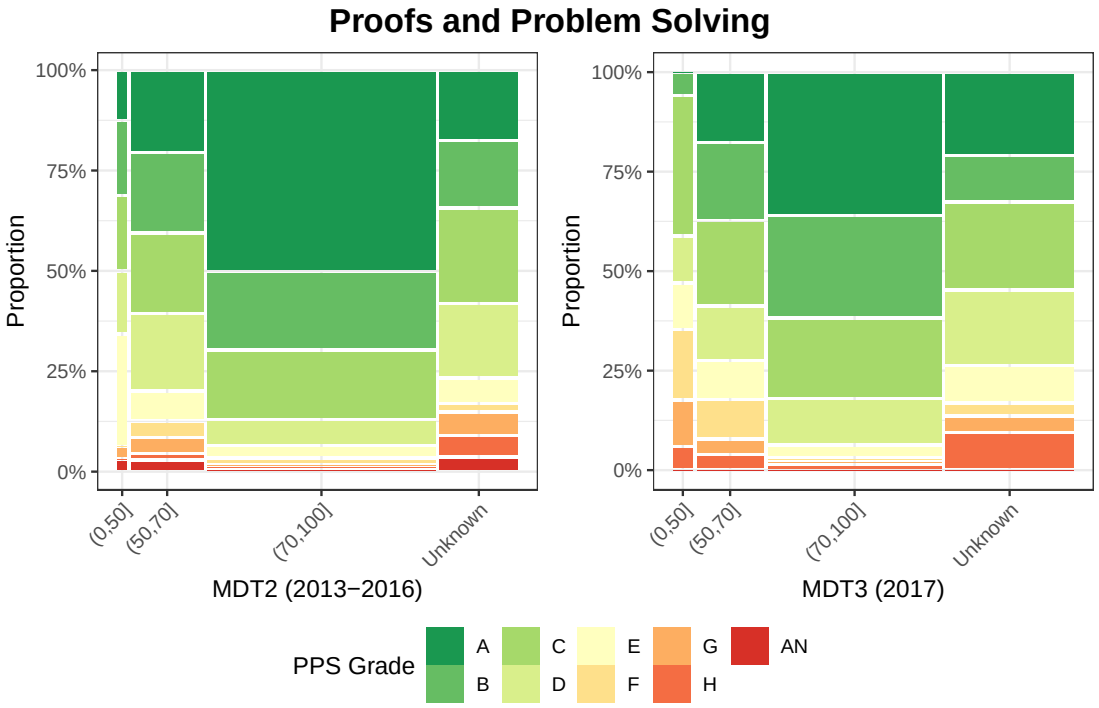

Figure 5.5: Stacked bar plots of PPS grades against MDT scores with proportional width

Proofs and Problem Solving					
2013-2016	(0,50]	(50,70]	(70,100]	Unknown	Sum
A	4	36	268	33	341
B	6	35	105	32	178
C	6	35	92	45	178
D	5	34	35	35	109
E	9	13	17	12	51
F	0	7	9	4	20
G	1	7	3	11	22
H	0	3	2	10	15
AN	1	5	4	7	17
Sum	32	175	535	189	931
2017	(0,50]	(50,70]	(70,100]	Unknown	Sum
A	0	9	46	20	75
B	1	10	33	11	55
C	6	11	26	21	64
D	2	7	15	18	42
E	2	5	4	9	20
F	3	5	1	3	12
G	2	2	1	4	9
H	1	2	2	9	14
AN	0	0	0	0	0
Sum	17	51	128	95	291

Table 5.3: Contingency tables of corresponding frequencies

Considering the students who took all three Year 1 courses, the stacked bar plots are shown in Figure 5.6 and the corresponding frequencies are shown in Table 5.4.

It can be seen that approximately 40% of students who scored below 50 marks in MDT2, got more than 1 Grade D or below in the three Year 1 courses, while the percentage of students

who scored above 70 drops to less than 10%. The situations are similar for the students taking MDT3.

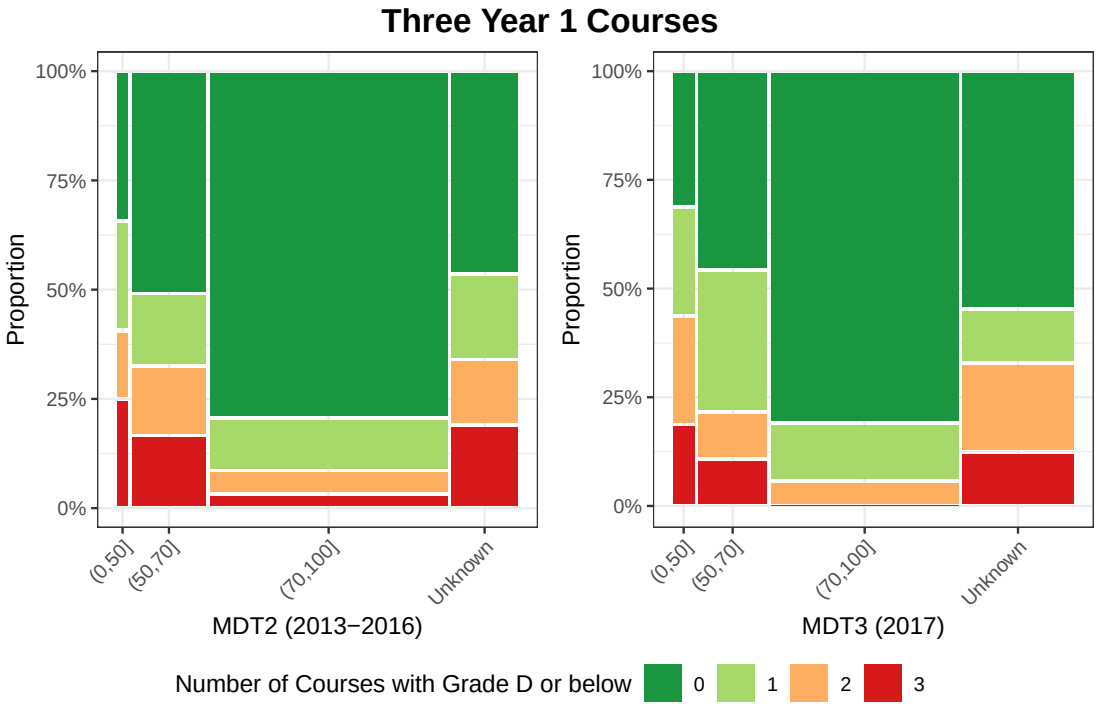

Figure 5.6: Stacked bar plots of the number of courses with Grade D or below against MDT scores with proportional width

Three Year 1 Courses					
2013-2016	(0,50]	(50,70]	(70,100]	Unknown	Sum
0	11	86	415	71	583
1	8	28	63	30	129
2	5	27	28	23	83
3	8	28	17	29	82
Sum	32	169	523	153	877
2017	(0,50]	(50,70]	(70,100]	Unknown	Sum
0	5	21	98	40	164
1	4	15	16	9	44
2	4	5	7	15	31
3	3	5	0	9	17
Sum	16	46	121	73	256

Table 5.4: Contingency tables of corresponding frequencies

Since the three Year 1 courses are compulsory for students in the School of Mathematics, it is worth investigating the relationship between the number of courses with Grade D or below and MDT scores when the population is restricted to School of Mathematics students only.

In Figure 5.7 and Table 5.5 are shown the stacked bar plots and the contingency tables of frequencies for students from School of Mathematics.

It should be noted that the percentages of maths students attaining Grade C or above are slightly lower than the percentages of students from all schools. The report last year suggested that students from School of Mathematics had higher mean scores in MDT than the others [3]. These two findings imply that the factor of schools has higher influence on MDT scores rather than on Year 1 course results.

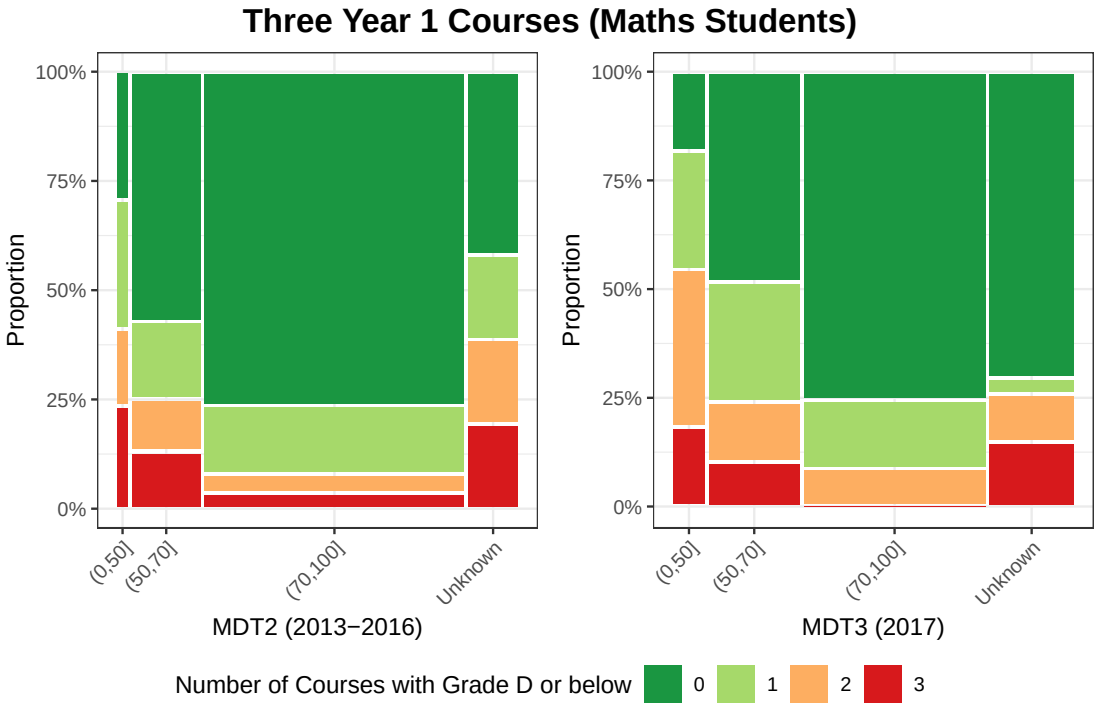

Figure 5.7: Stacked bar plots of the number of courses with Grade D or below against MDT scores with proportional width

Three Year 1 Courses (Maths Students)					
2013-2016	(0,50]	(50,70]	(70,100]	Unknown	Sum
0	5	48	233	26	312
1	5	15	48	12	80
2	3	10	13	12	38
3	4	11	11	12	38
Sum	17	84	305	62	468
2017	(0,50]	(50,70]	(70,100]	Unknown	Sum
0	2	14	43	19	78
1	3	8	9	1	21
2	4	4	5	3	16
3	2	3	0	4	9
Sum	11	29	57	27	124

Table 5.5: Contingency tables of corresponding frequencies

Based on the analyses above, students who scored less than 50 marks in MDT should be strongly recommended to have extra mathematics practice, while those who scored higher than 70 marks do not need extra maths practice. For the students who scored between 50 and 70 marks, a subdivision may be needed.

In Figure 5.8 and Table 5.6 are shown the stacked bar plots and contingency tables of the number of courses with Grade D or below against MDT scores where the range is divided with cut points 50, 60 and 70.

The difference in proportions of the numbers of Grade Ds or below for students who scored between 50 and 60 and students who scored between 60 and 70 varies across the years. In 2017, the percentage of students with 3 Grade Ds or below, given MDT scores between 60 and 70, is even higher than those scoring between 50 and 60 marks in MDT. This may due to the higher difficulties in MDT3 and the small sample size. Considering that for students with 50 to 70 marks in MDT, there are approximately 25% of students with more than 1 Grade D or

below in the three Year 1 courses, given their MDT scores which are between 50 and 60, these students should be suggested to have extra maths practice.

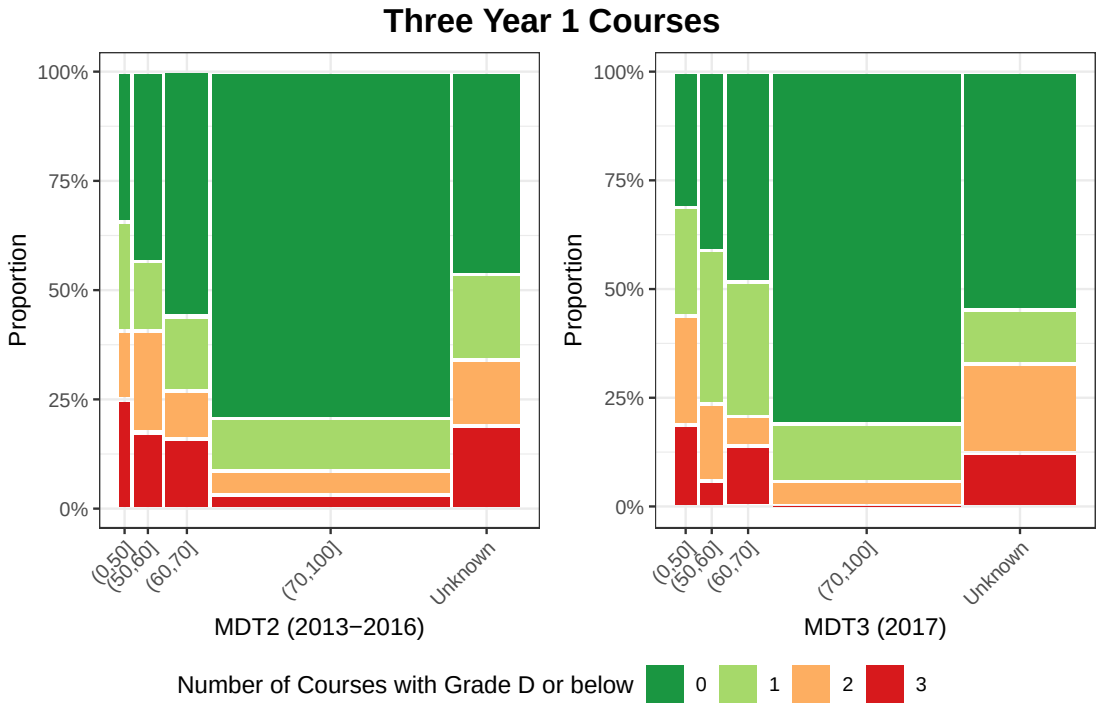

Figure 5.8: Stacked bar plots of the number of courses with Grade D or below against MDT scores with proportional width

Three Year 1 Courses						
2013-2016	(0,50]	(50,60]	(60,70]	(70,100]	Unknown	Sum
0	11	30	56	415	71	583
1	8	11	17	63	30	129
2	5	16	11	28	23	83
3	8	12	16	17	29	82
Sum	32	69	100	523	153	877
2017	(0,50]	(50,60]	(60,70]	(70,100]	Unknown	Sum
0	5	7	14	98	40	164
1	4	6	9	16	9	44
2	4	3	2	7	15	31
3	3	1	4	0	9	17
Sum	16	17	29	121	73	256

Table 5.6: Contingency tables of corresponding frequencies

5.3 Using entry qualifications

The entry qualifications of students may reflect their ability in high school mathematics, which might be correlated to the MDT scores and the Year 1 mathematics course results.

The raw data only consists of the entry qualifications for students from 2017, but some of these students are from other schools and took none of the Year 1 mathematics courses (ILA, CAP, PPS) or Year 2 mathematics courses for direct entry students (AAC, APPS). Therefore, a subset of these students should be used in the analyses.

In Table 5.7 are shown the frequency table of the categories of the students who took at least one of five courses (ILA, CAP, PPS, AAC, APPS). More than 40% of these students have unknown entry qualifications.

Cat1	Cat2	Cat3	Unknown	Sum
120	48	245	295	708

Table 5.7: Frequency of students entry qualifications

Within the students with unknown entry qualifications, over 70% of them are from outside of the UK (Table 5.8).

	EU	Overseas	RUK	Scotland	Sum
Known	31	78	136	168	413
Unknown	90	118	62	25	295
Sum	121	196	198	193	708

Table 5.8: Contingency table of entry qualifications against regions

5.3.1 Visual representation

The stacked bar plots of the grades of Year 1 courses against different entry qualifications with proportional width are shown in Figure 5.9. Comparing with the plots for 2017 in Figure 5.3, Figure 5.4 and Figure 5.5, the proportions of students getting higher grade given Category 1 entry qualifications are higher than those scoring below 50 marks in MDT.

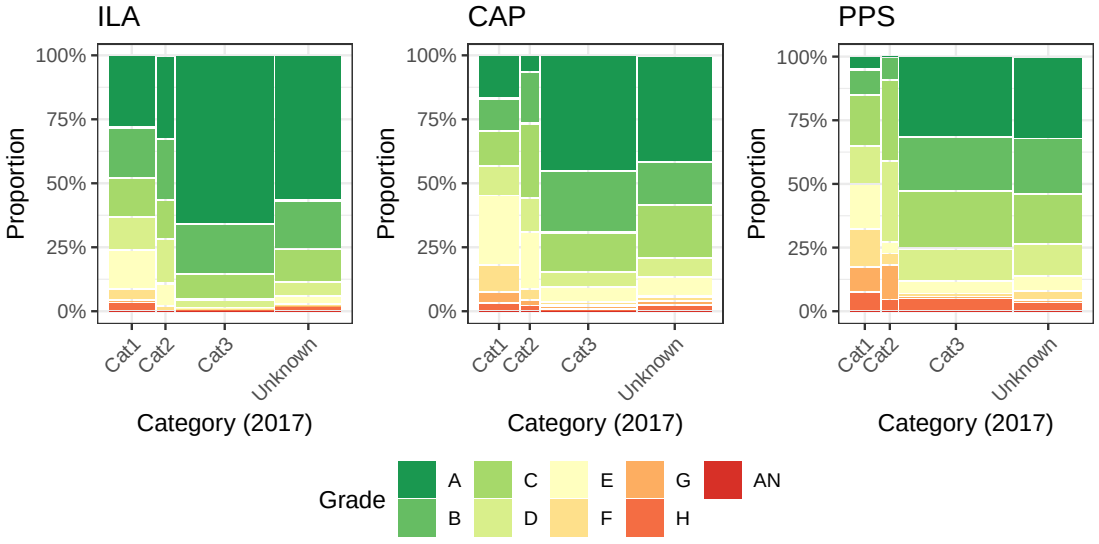

Figure 5.9: Stacked bar plots of grades of Year 1 courses against entry qualifications with proportional width

The corresponding contingency tables of frequencies are shown in Table 5.9, Table 5.10 and Table 5.11.

Introduction to Linear Algebra					
2017	Cat1	Cat2	Cat3	Unknown	Sum
A	33	15	159	93	300
B	23	11	47	31	112
C	18	7	24	21	70
D	15	8	8	9	40
E	18	4	1	5	28
F	5	1	1	2	9
G	1	0	0	0	1
H	4	0	1	3	8
AN	0	0	0	0	0
Sum	117	46	241	164	568

Table 5.9: Contingency table of corresponding frequencies

Calculus and its Applications					
2017	Cat1	Cat2	Cat3	Unknown	Sum
A	16	3	100	72	191
B	12	9	53	29	103
C	13	13	34	36	96
D	11	6	13	13	43
E	26	10	13	13	62
F	10	2	3	3	18
G	4	1	3	3	11
H	3	1	2	4	10
AN	0	0	0	0	0
Sum	95	45	221	173	534

Table 5.10: Contingency table of corresponding frequencies

Proofs and Problem Solving					
2017	Cat1	Cat2	Cat3	Unknown	Sum
A	2	0	45	28	75
B	4	2	30	19	55
C	8	7	32	17	64
D	6	7	18	11	42
E	7	1	7	5	20
F	6	1	2	3	12
G	4	3	1	1	9
H	3	1	7	3	14
AN	0	0	0	0	0
Sum	40	22	142	87	291

Table 5.11: Contingency table of corresponding frequencies

Considering the students who took all three Year 1 courses, the stacked bar plots are shown in Figure 5.10 and the corresponding frequencies are shown in Table 5.12.

It is similar to the findings in Figure 5.9 that the proportion of students getting 0 Grade Ds in the three courses is higher and the proportion of getting more than 1 Grade Ds is higher, for the group of students with Category 1 entry qualifications than the group of students with lower than 50 marks in MDT. The group of students with Category 2 entry qualifications are worse than the group of students with 50 to 70 marks in MDT. The group of students with Category 3 entry qualifications are slightly worse than the group of students with higher than 70 marks in MDT. Interestingly, the students with unknown category of entry qualifications performed even better than those with Category 3, which may due to the higher proportion of the overseas students.

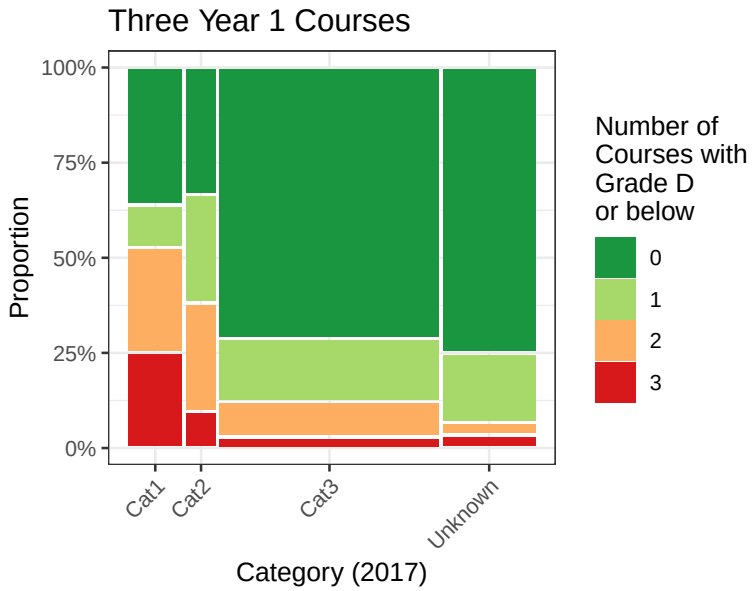

Figure 5.10: Stacked bar plots of the number of courses with Grade D or below against entry qualifications with proportional width

Three Year 1 Courses					
2017	Cat1	Cat2	Cat3	Unknown	Sum
0	13	7	99	45	164
1	4	6	23	11	44
2	10	6	13	2	31
3	9	2	4	2	17
Sum	36	21	139	60	256

Table 5.12: Contingency table of corresponding frequencies

In Figure 5.11 and Table 5.13 are shown the stacked bar plots and the contingency table of frequencies for students from the School of Mathematics. The performances of maths students in these three categories seem to be worse than the three groups using intervals of MDT scores.

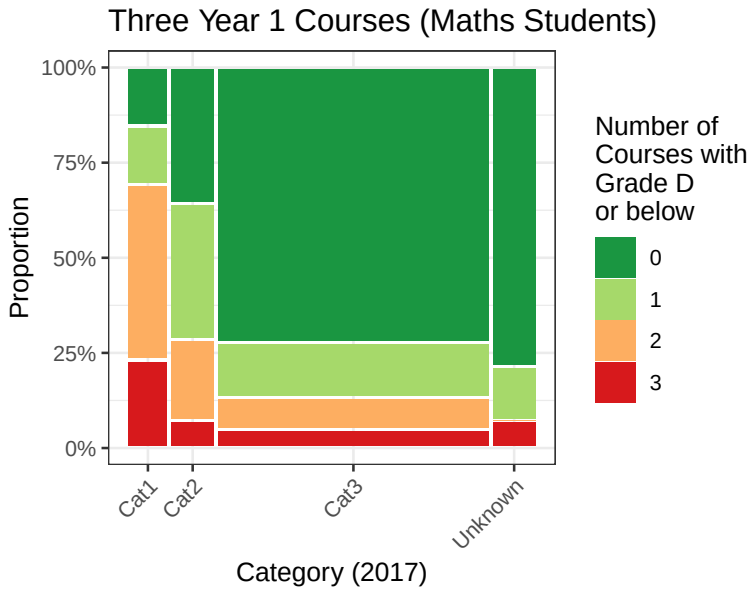

Figure 5.11: Stacked bar plots of the number of courses with Grade D or below against entry qualifications with proportional width

Three Year 1 Courses (Maths Students)					
2017	Cat1	Cat2	Cat3	Unknown	Sum
0	2	5	60	11	78
1	2	5	12	2	21
2	6	3	7	0	16
3	3	1	4	1	9
Sum	13	14	83	14	124

Table 5.13: Contingency table of corresponding frequencies

5.3.2 Comparison with diagnostic test

The relationship between MDT scores and entry qualifications is also of interest. Since the MDT scores or entry qualifications of some students are not recorded in the raw data, only the students who have MDT scores and took at least one of the five courses are considered. In Figure 5.12 and Table 5.14 are shown the stacked bar plots and contingency table of the frequencies of different intervals of MDT scores against different categories of entry qualifications.

It can be seen that only a half of students with Category 1 entry qualifications scored higher than 60 marks in MDT, while there are only approximately 10% of students with Category 3 entry qualifications scoring lower than 60 marks in MDT.

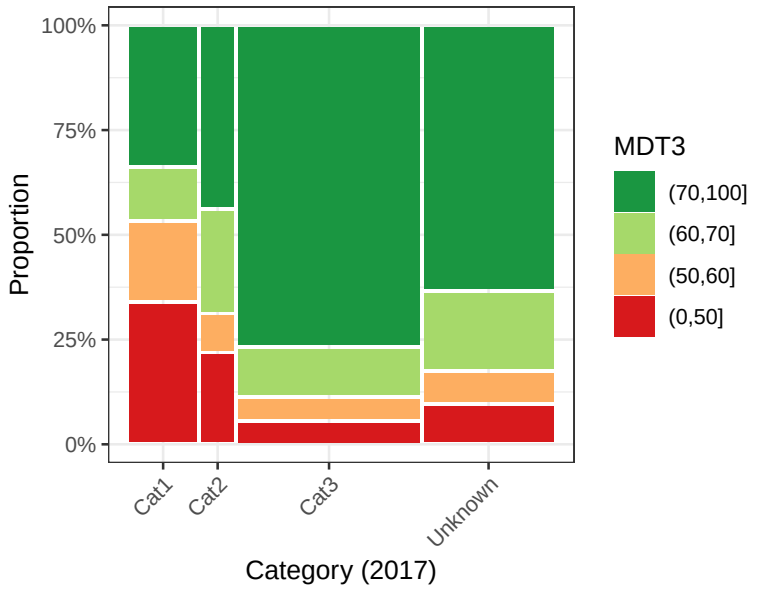

Figure 5.12: Stacked bar plots of MDT scores against entry qualifications with proportional width

2017	Cat1	Cat2	Cat3	Unknown	Sum
(70,100]	21	14	123	73	231
(60,70]	8	8	19	22	57
(50,60]	12	3	9	9	33
(0,50]	21	7	9	11	48
Sum	62	32	160	115	369

Table 5.14: Contingency table of corresponding frequencies

5.4 Combining two

Based on the previous analyses, MDT scores can be used to refine the recommendation to students about whether they need extra mathematics practice.

Table 5.15 shows the detailed categorisations of recommendation to students. The population is the students took all three Year 1 courses.

	Cat1	Cat2	Cat3	Unknown
(70,100]	Cat1+	Cat2+	Cat3+	Unknown
(60,70]	Cat1-	Cat2+	Cat3+	Unknown
(50,60]	Cat1-	Cat2-	Cat3+	Unknown
(0,50]	Cat1-	Cat2-	Cat3-	Unknown
Unknown	Cat1-	Cat2-	Cat3+	Unknown

Table 5.15: Detailed categories of recommendation

In Figure 5.13 and Table 5.18 are shown the stacked bar plot of the number of courses with Grade D or below against different refined categories of recommendation. The population size for each category are relatively small, so the proportions may not be sensible.

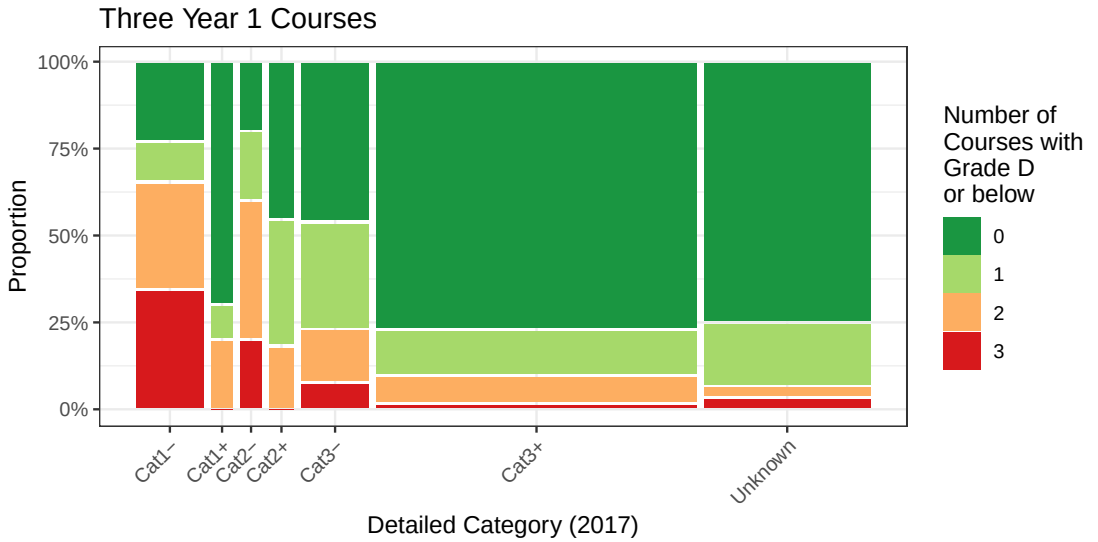

Figure 5.13: Stacked bar plots of the number of courses with Grade D or below against refined categories of recommendation

Three Year 1 Courses								
2017	Cat1-	Cat1+	Cat2-	Cat2+	Cat3-	Cat3+	Unknown	Sum
0	6	7	2	5	12	87	45	164
1	3	1	2	4	8	15	11	44
2	8	2	4	2	4	9	2	31
3	9	0	2	0	2	2	2	17
Sum	26	10	10	11	26	113	60	256

Table 5.16: Contingency table of corresponding frequencies

Table 5.17 shows the refined categories recommendation to students, where 'Cat-' represents 'recommended to take extra maths course' and 'Cat+' represents 'not recommended to take extra maths course'. The population is the students took all three Year 1 courses.

	Cat1	Cat2	Cat3	Unknown
(70,100]	Cat+	Cat+	Cat+	Cat+
(60,70]	Cat-	Cat+	Cat+	Cat+
(50,60]	Cat-	Cat-	Cat+	Cat-
(0,50]	Cat-	Cat-	Cat-	Cat-
Unknown	Cat-	Cat-	Cat+	Unknown

Table 5.17: Refined categories of recommendation

In Figure 5.14 and Table 5.18 are shown the stacked bar plot of the number of courses with Grade D or below against different refined categories of recommendation.

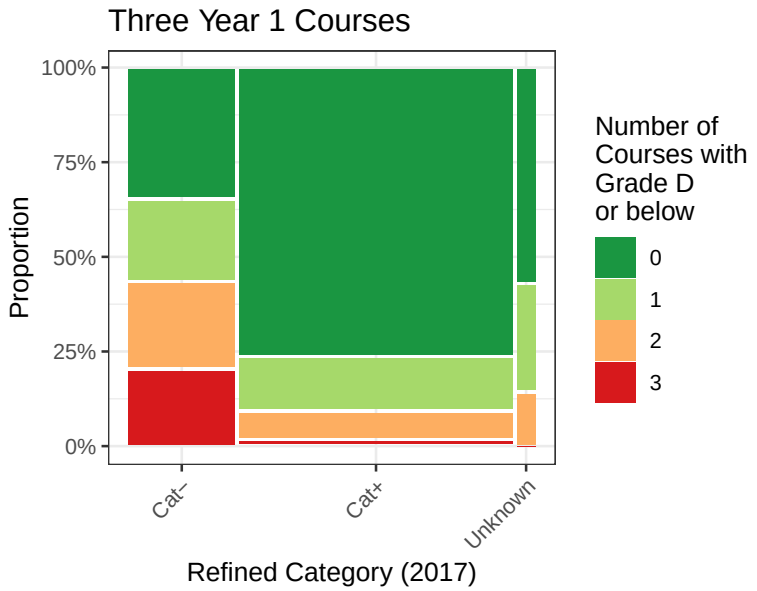

Figure 5.14: Stacked bar plots of the number of courses with Grade D or below against refined categories of recommendation

Three Year 1 Courses				
2017	Cat-	Cat+	Unknown	Sum
0	24	132	8	164
1	15	25	4	44
2	16	13	2	31
3	14	3	0	17
Sum	69	173	14	256

Table 5.18: Contingency table of corresponding frequencies

Binary classification

To evaluate a binary classifier that classifies the elements into two classes, a special kind of contingency table with two dimensions ('true condition' and 'predicted condition') and identical sets of classes ('positive' and 'negative') in both dimensions is used as a visualisation of the performance of the classifier [7]. This special kind of contingency table is called confusion matrix, which is summarised in Table 5.19.

		Predicted Condition	
		Positive	Negative
True Condition	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FN)	True Negative (TN)

Table 5.19: Confusion matrix

In our context, if students with more than 1 Grade D are considered to be (true) at-risk (positive) and the others are considered not to be at-risk, the refined categories can be used as a predictor to the courses result, where Cat- represents predicted positive and Cat+ represents predicted negative.

The accuracy of the classifier, i.e. the proportion of correct classifications, is calculated by

$$ACC = \frac{TP + TN}{TP + FN + TN + FP}.$$

However, accuracy is not a reliable metric and gives misleading results when the numbers of observations in different classes vary greatly [13]. Alternatively, two fundamental statistics that are independent on prevalence (proportion of true conditional positive), Sensitivity (also known as True Positive Rate, TPR) and Specificity (also known as True Negative Rate, TNR) can used as measures of the performance of a binary classifier. Sensitivity or True Positive Rate is calculated by

$$TPR = \frac{TP}{TP + FN},$$

and Specificity or True Negative Rate is calculated by

$$TNR = \frac{TN}{TN + FP}.$$

When we are assessing how the results of MDT predict students' performance in mathematics courses, we are aiming at identifying more students who are at risk (higher TPR) while avoiding decreasing the number of students that are classified into not-at-risk correctly (lower TNR).

The confusion matrix of this classifier that combines MDT results and entry qualifications is shown in Table 5.20.

	Predicted At-Risk	Predicted Not-At-Risk
True At-Risk	30	16
True Not-At-Risk	39	157

Table 5.20: Confusion matrix of refined categories classification

The accuracy of this classification method is calculated by

$$ACC = \frac{TP + TN}{TP + FN + TN + FP} = \frac{30 + 157}{30 + 16 + 39 + 157} = 77.2\%,$$

and the sensitivity is calculated by

$$TPR = \frac{TP}{TP + FN} = \frac{30}{30 + 16} = 65.22\%.$$

5.5 Extra maths courses

In previous years, students from School of Mathematics who scored lower than 65 marks in MDT were suggested to take Mathematics for Physics 1 (MfP1) as the extra mathematics practice. This section is aimed at examining whether MfP1 helped improved the results of other mathematics courses.

The scatter plots of course results in the three Year 1 courses against MDT scores are shown in Figure 5.15. In these plots, students are grouped by their selections of MfP1. The blue points does not form a cluster separated from the red points, implying that taking MfP1 might not do great help to students.

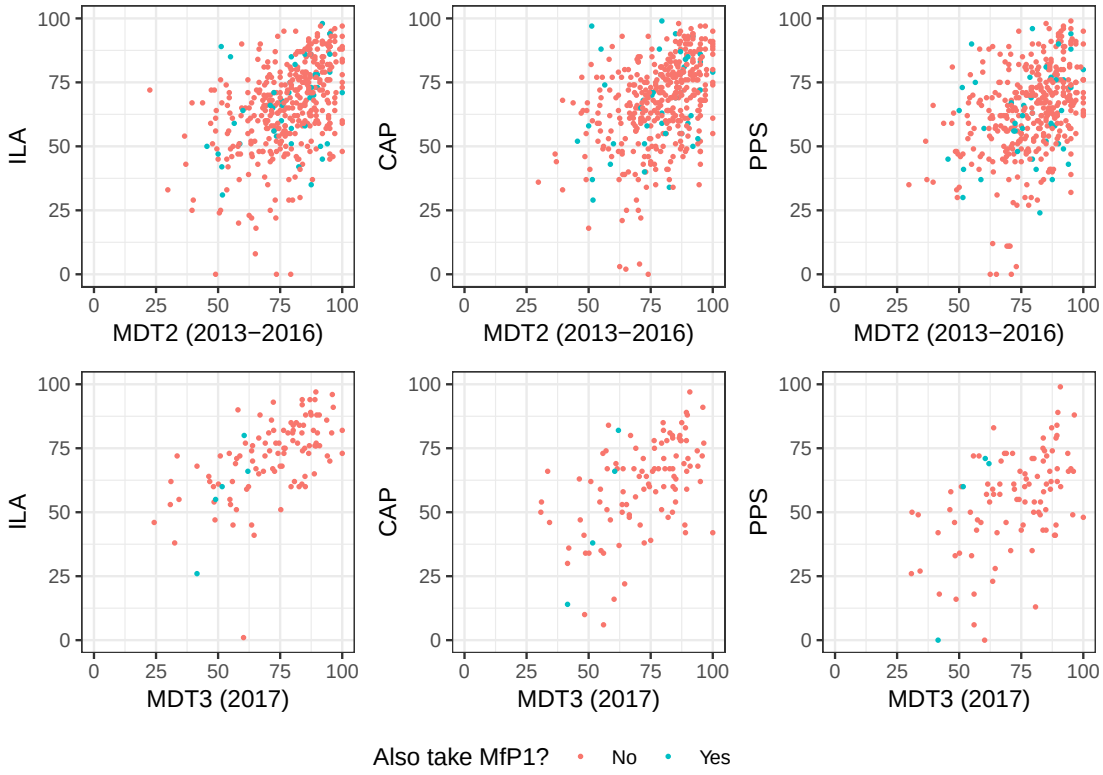

Figure 5.15: Scatter plots of course results against MDT scores, grouped by the selection of MfP1

In Figure 5.16 are shown the box plots of the course results grouped by MDT scores. Table 5.21 shows the summary table of ANOVA testing the null hypothesis that the mean of the course results for students scoring lower than 65 in MDT are equal. The non-significant probabilities do not reject the null hypotheses, i.e. there are no significant differences between the students taking MfP1 or not.

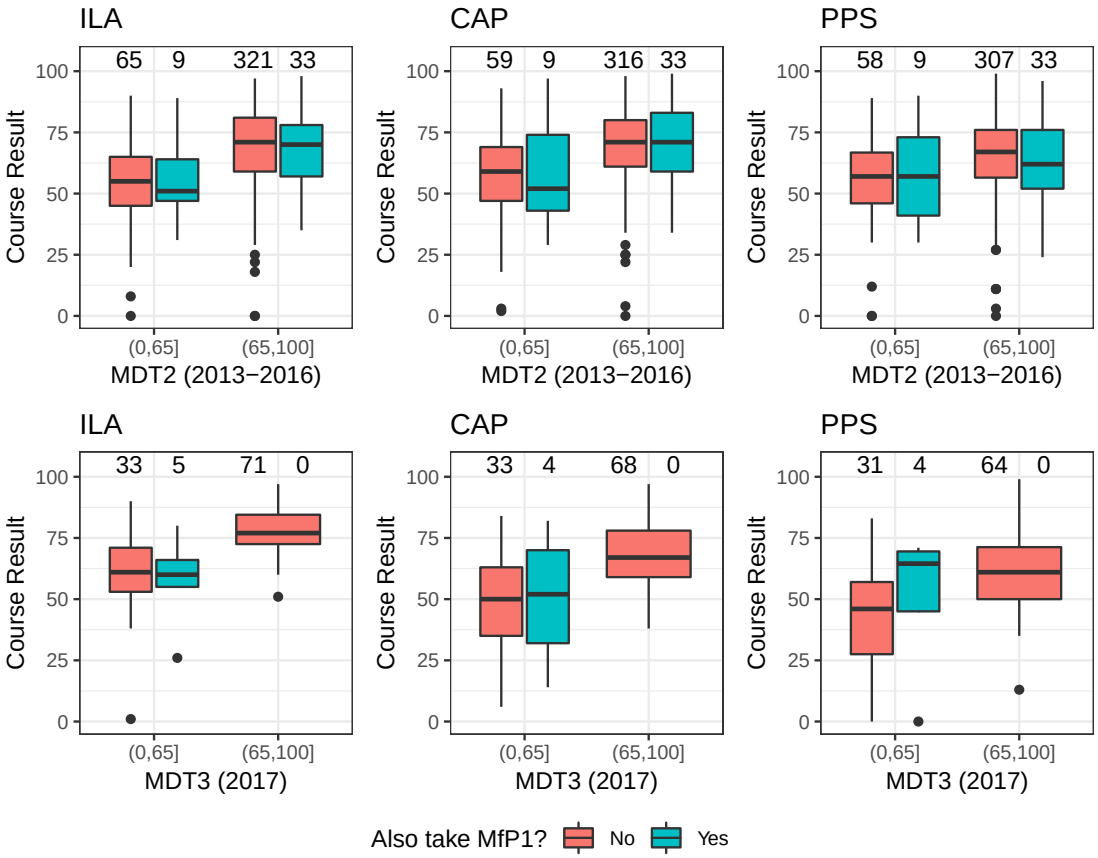

Figure 5.16: Box plots of course results grouped by MDT scores

2013-2016	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ILA.MfP1	1	209	208.87	0.661	0.419
Residuals	72	22754	316.03		
CAP.MfP1	1	33	33.33	0.088	0.768
Residuals	66	25064	379.75		
PPS.MfP1	1	35	34.79	0.098	0.755
Residuals	65	23121	355.71		
2017	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ILA.MfP1	1	16	16.23	0.061	0.807
Residuals	36	9613	267.01		
CAP.MfP1	1	10	9.55	0.024	0.879
Residuals	35	14226	406.45		
PPS.MfP1	1	197	196.72	0.428	0.517
Residuals	33	15166	459.57		

Table 5.21: ANOVA table

5.6 Predictive modelling

In Section 5.4, we make up the correlation between MDT scores and students' entry qualifications and use it as the feature of the binary classification that classify students in to 'At-Risk' group and 'Not-At-Risk' group. As suggested by Marbouti [12], it is possible to identify students who are at-risk by using predictive modelling techniques from the field of machine learning.

In this section, we introduce several machine learning techniques and use them to forecast the performance of students in Year 1 courses. Some `python` packages including `scikit-learn` and `NumPy` are used in the analyses.

5.6.1 Prediction methods

Nine different prediction methods are used to identify at-risk students, which are described below.

- Logistic Regression (LR), which was also used in the report last year, is used to predict the probability of getting over 50% in the Year 1 courses.
- Naive Bayes (NB) Classifier, is a simple classifier based on Bayes' theorem with assumption of independent between the variables. This assumption is violated in this project, but this classifier still works when the dependencies of variables from each other are similar [22]. In this section, the distribution of each variable is assumed to be multinomial, because some the variables are categorical variables.
- K-Nearest Neighbour (KNN), is a non-parametric classifier, which classifies students based on their nearest neighbours.
- Multi-Layer Perceptron (MLP), is a kind of artificial neural network. In this section, two hidden layers with half of the number of observations as hidden nodes are used, as recommended by Marbouti [12].
- Decision Tree (DT), is a non-parametric classifier classifying data with a set of if-then-else decision trees. In this section, a depth of 10 for the trees are chosen, because very fine splits are made involving very few data points when the trees are grown deeper and might overfit the data.
- Random Forest (RF), is constructed with multiple decision trees.
- Linear Support Vector Machine (SVM), finds a hyperplane separating the classes.
- Linear Discriminant Analysis (LDA), uses a maximum likelihood discrimination rule based on linear combinations and separates the classes with linear decision surface.
- Quadratic Discriminant Analysis (QDA), uses a maximum likelihood discrimination rule based on quadratic combinations and separates the classes with quadratic decision surface.

To avoid accidentally over-fitting a model to the available data and generalising a poor fit to future data, we hold out a part of data as validation set and use the remaining part as training data. This is a statistical procedure called cross-validation. In particular, we use K-fold cross-validation for our data which is in short supply. The data is split into K parts in the K-fold cross-validation procedure. Each model is fitted K times with a different part of data used as a validation set and the remaining $K-1$ parts used for training. We then select the model with the lowest validation error, when averaged over K folds.

In this section, students' course results are fitted against the features (variables) that are available in the raw data set and significant (which is discussed in summer 2017 project) in the models we fit. These features include MDT mark, fee status group, school and gender. For the students from 2017, extra models which consider entry qualifications are also fitted.

5.6.2 Using MDT

The normalised confusion matrices of different predictive models for each three Year 1 courses, given students admitted before 2017, are shown in Figure 5.17, Figure 5.18 and Figure 5.19, while the entries of the original confusion matrices and the accuracies are shown in Table 5.23,

Table 5.24 and Table 5.25. Additionally, in Figure 5.20 is shown the normalised confusion matrices of models for risk status of students, with the corresponding confusion matrices and accuracies showing in Table 5.26.

The normalised confusion matrix is the marginal proportion of the original confusion matrix with regards to the true condition, which is summarised in Table 5.22.

		Predicted Condition	
		Positive	Negative
True Condition	Positive	True Positive Rate Sensitivity $TPR = \frac{TP}{TP + FN}$	False Negative Rate Miss Rate $FNR = \frac{FN}{TP + FN}$
	Negative	False Positive Rate Fall-out $FPR = \frac{FP}{TN + FP}$	True Negative Rate Specificity $TNR = \frac{TN}{TN + FP}$

Table 5.22: Normalised confusion matrix

It can be seen from the figures and tables that the accuracies of the models range from approximately 70% to approximately 80%. Among the nine models for each case, Naive Bayes Classifiers and K-Nearest Neighbour Classifiers have the highest true positive rates, which are approximately 50%. However, these two models have the lowest true negative rates (around 20%), indicating that more not-at-risk students are classified as at-risk students by these two models than the other.

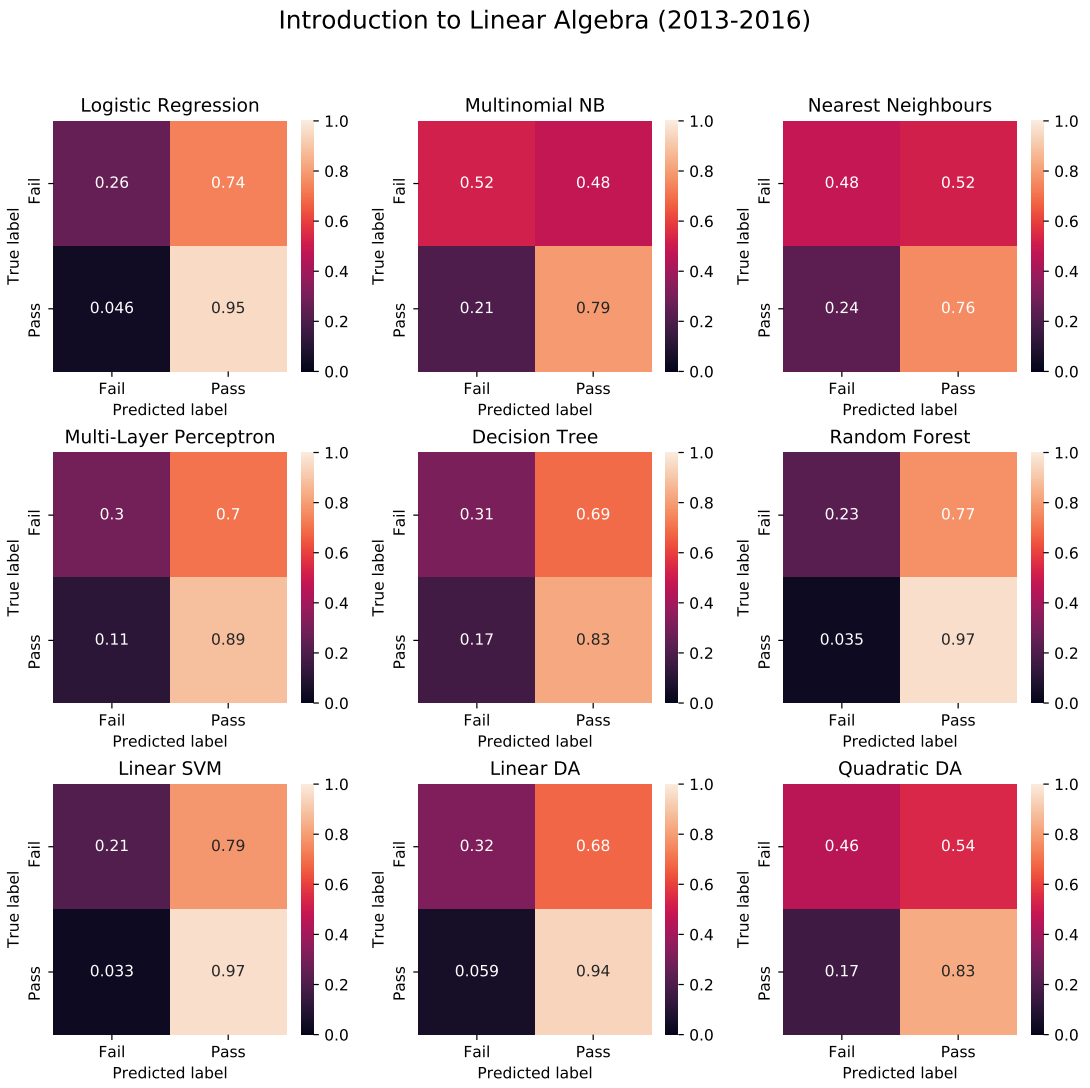

Figure 5.17: Normalised confusion matrices of different prediction methods

	LR	MNB	KNN	MLP	DT	RF	SVM	LDA	QDA
TP	65	128	119	73	76	56	52	79	113
FN	181	118	127	173	170	190	194	167	133
FP	46	207	242	114	173	35	33	59	166
TN	958	797	762	890	831	969	971	945	838
ACC	0.81	0.74	0.70	0.77	0.72	0.82	0.81	0.81	0.76

Table 5.23: Test results and accuracy of predictions

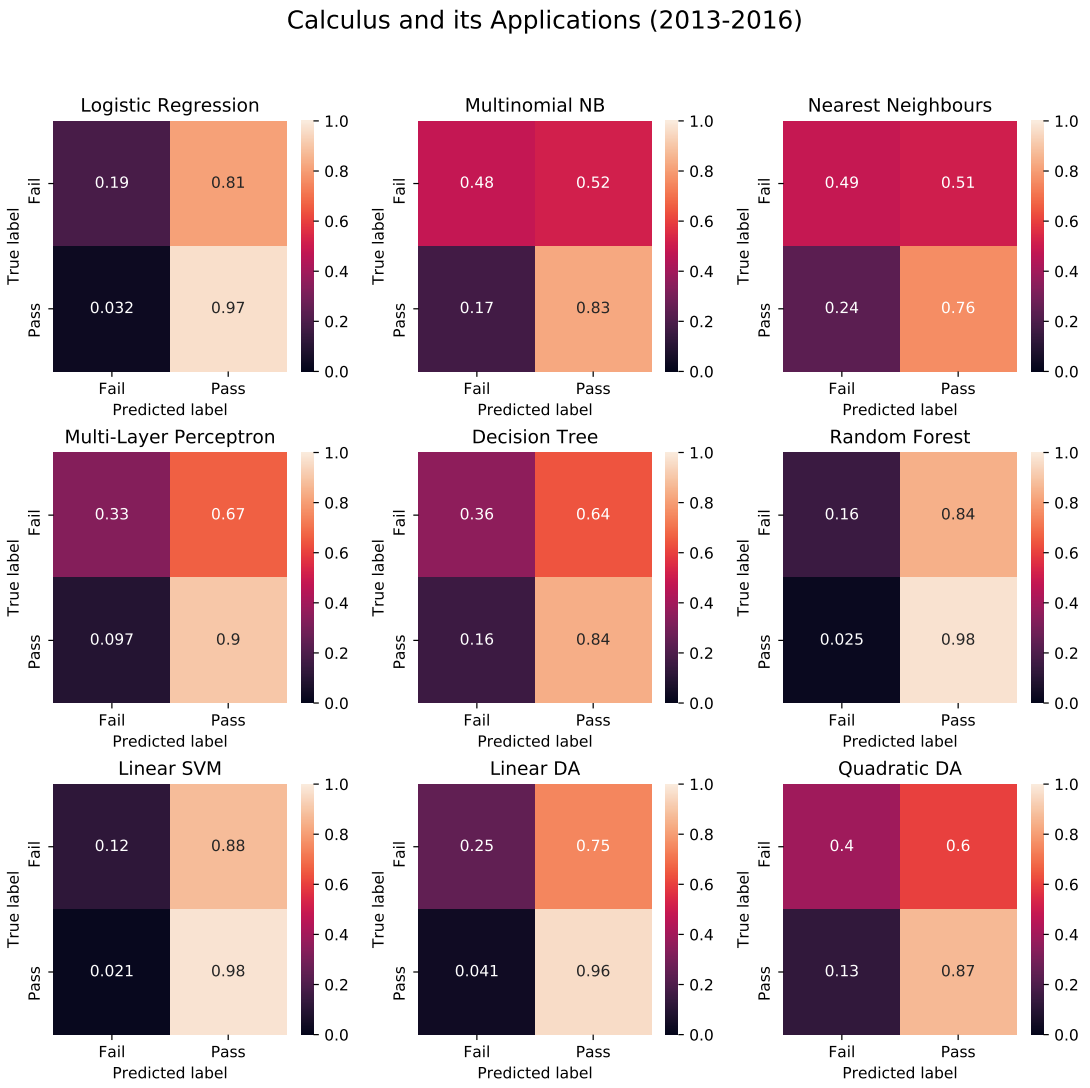

Figure 5.18: Normalised confusion matrices of different prediction methods

	LR	MNB	KNN	MLP	DT	RF	SVM	LDA	QDA
TP	39	99	100	69	74	32	25	52	82
FN	167	107	106	137	132	174	181	154	124
FP	31	167	228	94	154	24	20	40	127
TN	936	800	739	873	813	943	947	927	840
ACC	0.83	0.76	0.71	0.80	0.75	0.83	0.82	0.83	0.78

Table 5.24: Test results and accuracy of predictions

Figure 5.19: Normalised confusion matrices of different prediction methods

	LR	MNB	KNN	MLP	DT	RF	SVM	LDA	QDA
TP	30	76	69	44	51	20	26	43	50
FN	125	79	86	111	104	135	129	112	105
FP	41	101	152	81	97	32	30	51	88
TN	546	486	435	506	490	555	557	536	499
ACC	0.77	0.75	0.67	0.74	0.72	0.77	0.78	0.78	0.73

Table 5.25: Test results and accuracy of predictions

Figure 5.20: Normalised confusion matrices of different prediction methods

	LR	MNB	KNN	MLP	DT	RF	SVM	LDA	QDA
TP	22	57	42	15	30	12	11	29	40
FN	91	56	71	98	83	101	102	84	73
FP	20	100	141	50	92	13	13	30	66
TN	591	511	470	561	519	598	598	581	545
ACC	0.84	0.78	0.70	0.79	0.75	0.84	0.84	0.84	0.80

Table 5.26: Test results and accuracy of predictions

The normalised confusion matrices of different predictive models for each three Year 1 course, given students from 2017, are shown in Figure 5.21, Figure 5.22 and Figure 5.23, while the entries of the original confusion matrices and the accuracies are shown in Table 5.27, Table 5.28 and Table 5.29. Additionally, in Figure 5.24 is shown the normalised confusion matrices of models for risk status of students, with the corresponding confusion matrices and accuracies showing in Table 5.30.

The accuracies of models for ILA are higher compared to the models for students admitted before 2017. However, the contribution to the increased accuracies is mainly from the increase of true negative rates, while the true positive rates decrease. The accuracies and the true

positive rates of the models for other courses are lower than the models for previous years. This may due to the short supply of data in 2017.

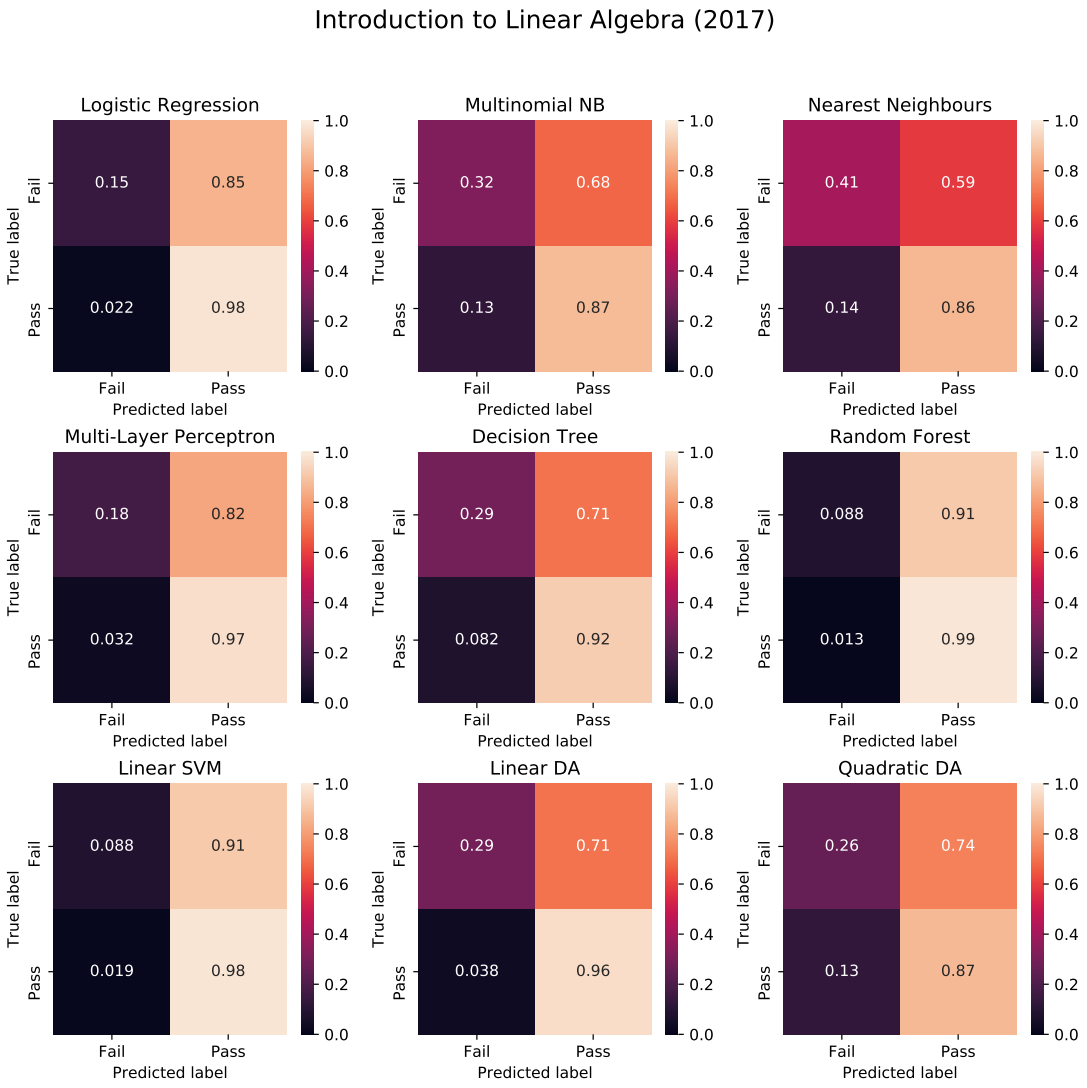

Figure 5.21: Normalised confusion matrices of different prediction methods

	LR	MNB	KNN	MLP	DT	RF	SVM	LDA	QDA
TP	5	11	14	6	10	3	3	10	9
FN	29	23	20	28	24	31	31	24	25
FP	7	40	43	10	26	4	6	12	41
TN	310	277	274	307	291	313	311	305	276
ACC	0.89	0.82	0.82	0.89	0.85	0.90	0.89	0.89	0.81

Table 5.27: Test results and accuracy of predictions

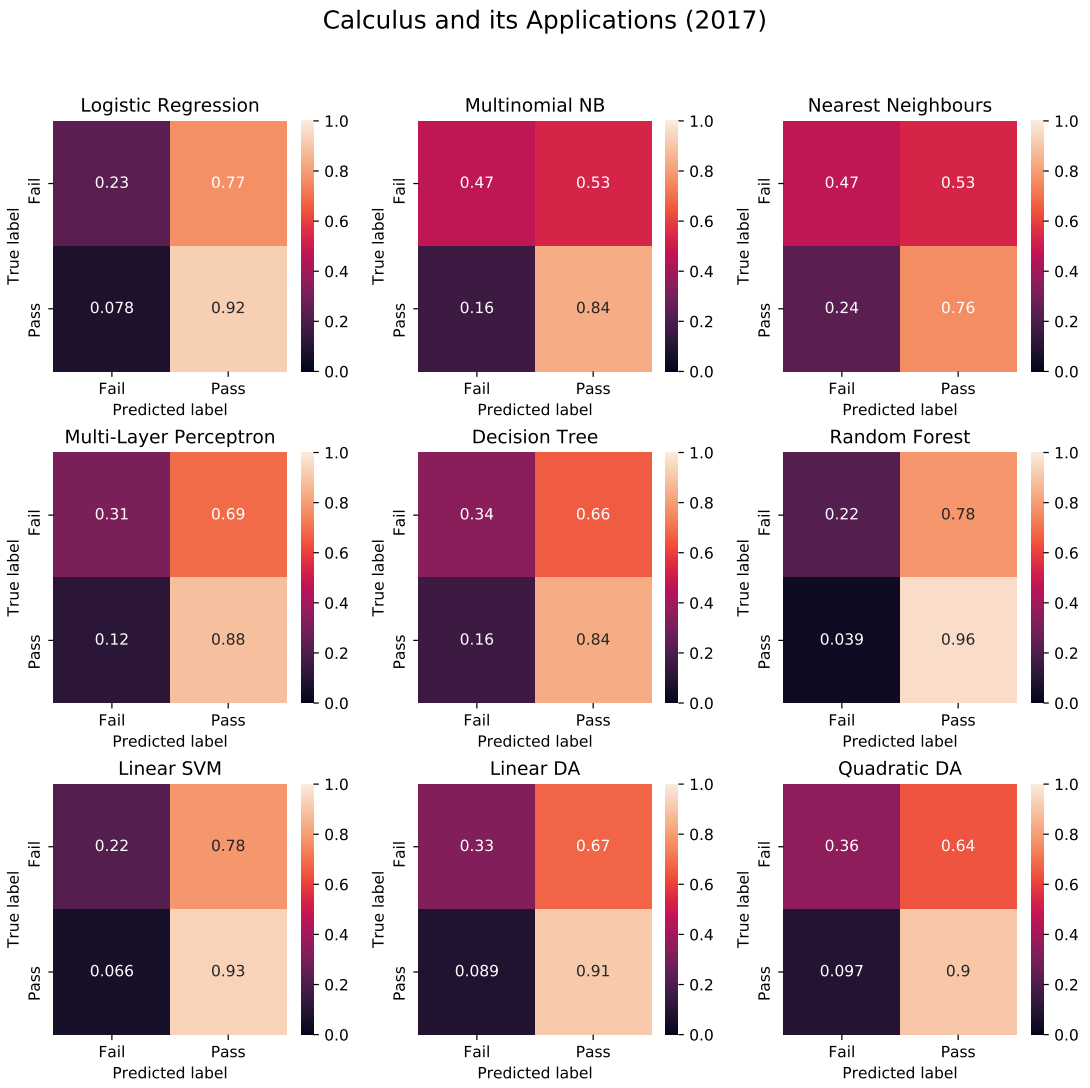

Figure 5.22: Normalised confusion matrices of different prediction methods

	LR	MNB	KNN	MLP	DT	RF	SVM	LDA	QDA
TP	15	30	30	20	22	14	14	21	23
FN	49	34	34	44	42	50	50	43	41
FP	20	42	61	30	42	10	17	23	25
TN	237	215	196	227	215	247	240	234	232
ACC	0.78	0.76	0.70	0.76	0.73	0.81	0.79	0.79	0.79

Table 5.28: Test results and accuracy of predictions

Figure 5.23: Normalised confusion matrices of different prediction methods

	LR	MNB	KNN	MLP	DT	RF	SVM	LDA	QDA
TP	15	21	28	11	21	18	16	15	11
FN	39	33	26	43	33	36	38	39	43
FP	18	25	51	20	38	15	16	17	16
TN	124	117	91	122	104	127	126	125	126
ACC	0.70	0.70	0.60	0.67	0.63	0.73	0.72	0.71	0.69

Table 5.29: Test results and accuracy of predictions

Figure 5.24: Normalised confusion matrices of different prediction methods

	LR	MNB	KNN	MLP	DT	RF	SVM	LDA	QDA
TP	4	6	9	2	6	2	3	8	0
FN	20	18	15	22	18	22	21	16	24
FP	13	15	32	10	27	2	10	14	0
TN	146	144	127	149	132	157	149	145	159
ACC	0.81	0.81	0.74	0.82	0.75	0.86	0.83	0.83	0.86

Table 5.30: Test results and accuracy of predictions

5.6.3 Using both

In Figure 5.25, Figure 5.26, Figure 5.27 and Figure 5.28 are shown the normalised confusion matrices of the models which take students' entry qualifications into account. The corresponding confusion matrices and accuracies are shown in Table 5.31, Table 5.32, Table 5.33 and Table 5.34.

The accuracies and true positive rates of these models are improved, compared with the models whose features do not include entry qualifications of students from 2017. However, the

highest true positive rates of these models is only 50%, which is not as high as expected.

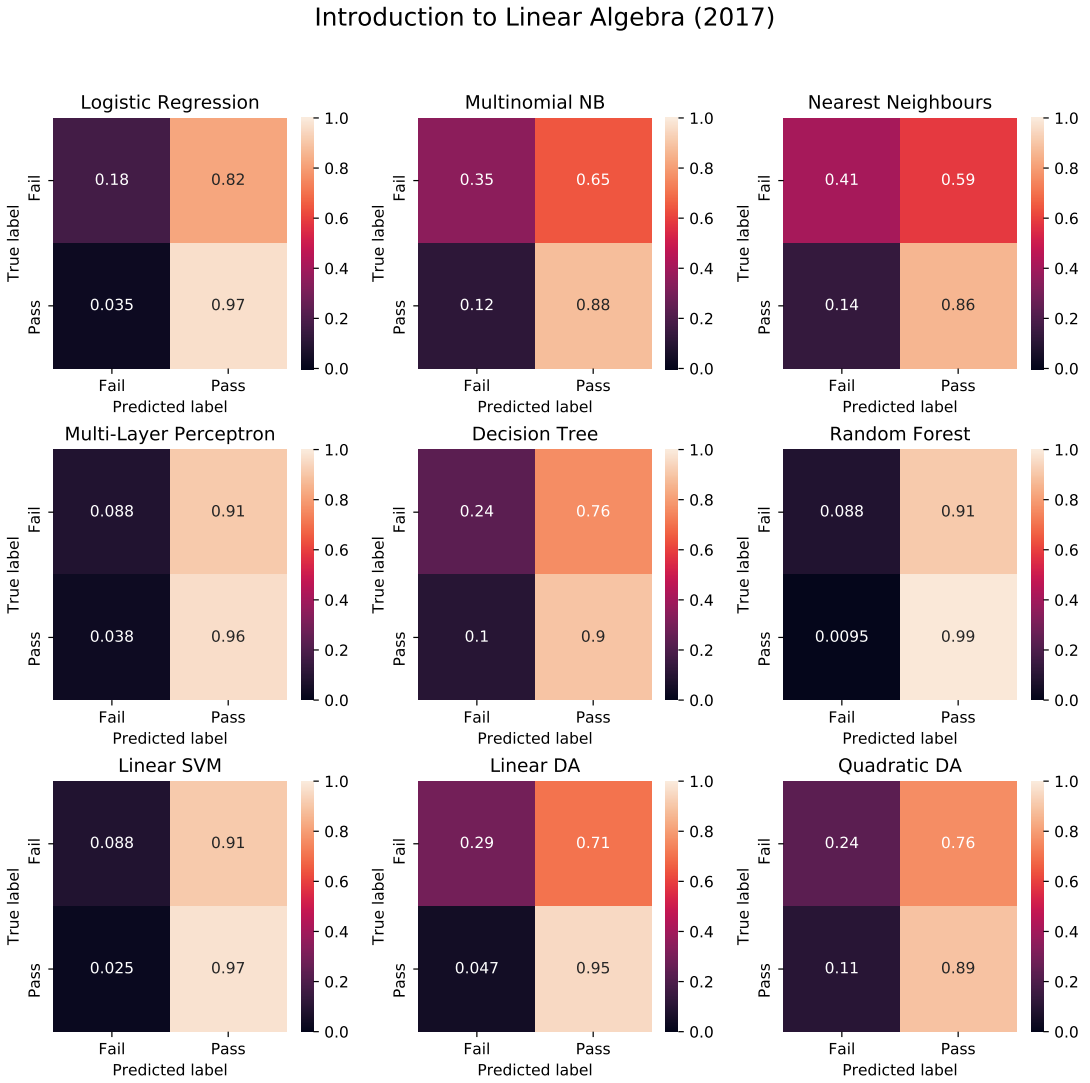

Figure 5.25: Normalised confusion matrices of different prediction methods

	LR	MNB	KNN	MLP	DT	RF	SVM	LDA	QDA
TP	6	12	14	3	8	3	3	10	8
FN	28	22	20	31	26	31	31	24	26
FP	11	38	44	12	33	3	8	15	34
TN	306	279	273	305	284	314	309	302	283
ACC	0.88	0.82	0.81	0.87	0.83	0.90	0.88	0.88	0.82

Table 5.31: Test results and accuracy of predictions

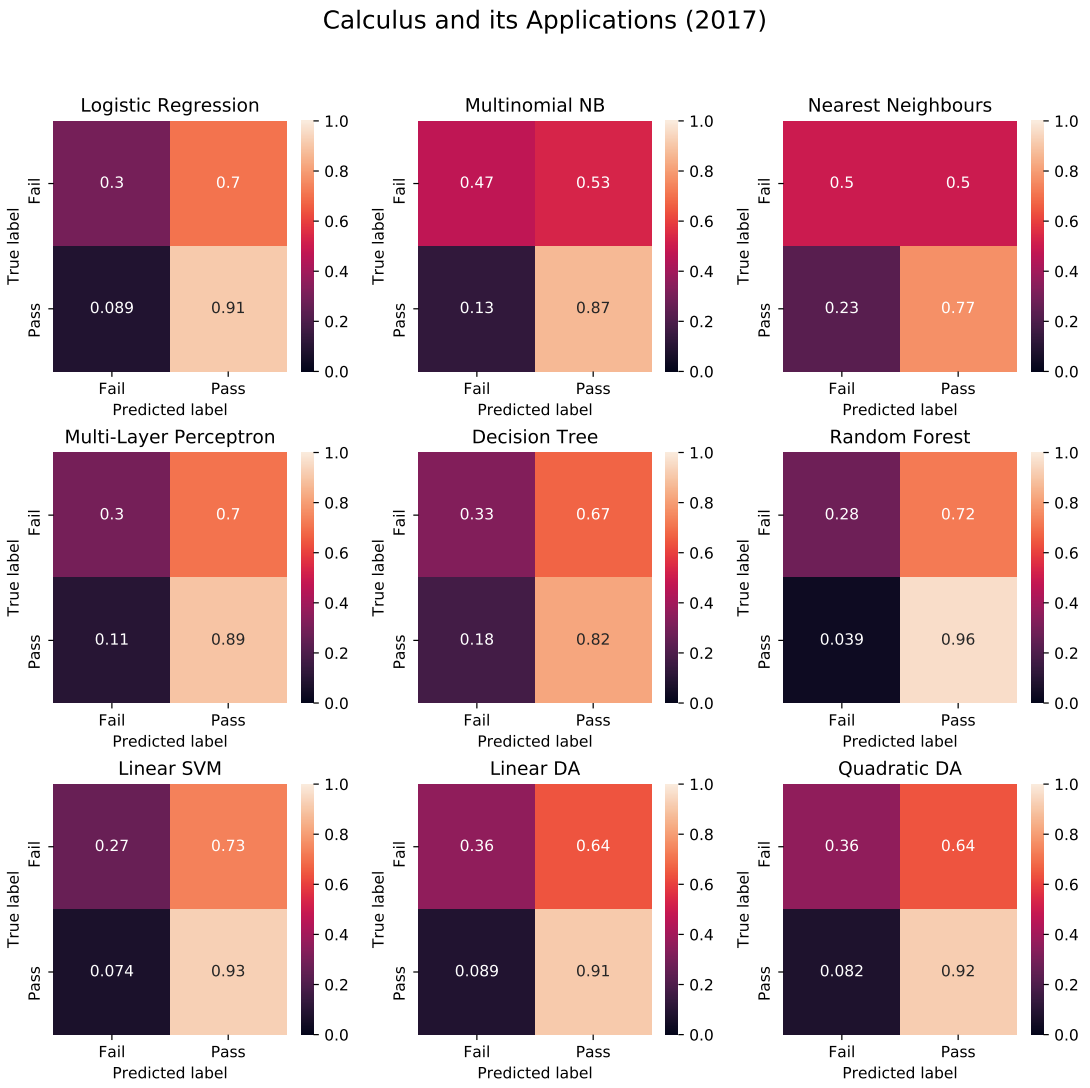

Figure 5.26: Normalised confusion matrices of different prediction methods

	LR	MNB	KNN	MLP	DT	RF	SVM	LDA	QDA
TP	19	30	32	19	21	18	17	23	23
FN	45	34	32	45	43	46	47	41	41
FP	23	34	59	27	46	10	19	23	21
TN	234	223	198	230	211	247	238	234	236
ACC	0.78	0.78	0.71	0.77	0.72	0.82	0.79	0.80	0.80

Table 5.32: Test results and accuracy of predictions

Figure 5.27: Normalised confusion matrices of different prediction methods

	LR	MNB	KNN	MLP	DT	RF	SVM	LDA	QDA
TP	18	22	27	22	20	11	16	16	24
FN	36	32	27	32	34	43	38	38	30
FP	19	24	49	28	34	10	15	19	26
TN	123	118	93	114	108	132	127	123	116
ACC	0.71	0.71	0.61	0.69	0.65	0.72	0.72	0.70	0.71

Table 5.33: Test results and accuracy of predictions

Figure 5.28: Normalised confusion matrices of different prediction methods

	LR	MNB	KNN	MLP	DT	RF	SVM	LDA	QDA
TP	8	8	9	5	9	4	6	6	0
FN	16	16	15	19	15	20	18	18	24
FP	14	17	32	5	23	3	9	12	0
TN	145	142	127	154	136	156	150	147	159
ACC	0.83	0.81	0.74	0.86	0.79	0.87	0.85	0.83	0.86

Table 5.34: Test results and accuracy of predictions

5.6.4 Summary

Compared to the binary classification using stacked bar plots described in Section 5.4, using the predictive models earlier this section does not improve the performance of classification of the students. This may due to several problems involved in the predictive models.

- The population of the data set, especially the data for students from 2017, is quite small and does not enough for training the predictive models. When fitting Quadratic Discrimi-

nant Analysis model for students from 2017, we are warned by the `python` package that the variables are collinear.

- The categorical variables are converted to ordinal variables. For example, using natural ordering in variables fee status group and school may result in bad performance, because no ordinal relationship exists in these variables. Alternatively, they should be represented using binary vectors, i.e. using 'one-hot' (also known as 'one-of-k') encoding.

Additionally, the predictive models used in this section are based on the recommendation from Marbouti [12] and we do not have further exploration of how to improve the models because of time limitation in this summer project. However, as the number of students taking MDT increases, and the performance of the predictive models improves, using machine learning methods would have better prediction and classification of students based on their risk status than using visual representation from the stacked bar plots.

Bibliography

[1] D. J. Bartholomew et al. *Analysis of Multivariate Social Science Data*. Chapman and Hall/CRC, 2008.

[2] B. Bridgeman, N. Burton, and F. Cline. "A Note on Presenting What Predictive Validity Numbers Mean". In: *Applied Measurement in Education* 22.2 (2009), pp. 109–119. DOI: 10.1080/08957340902754577.

[3] T. Burgetova and J. R. Imanuel. *Maths Diagnostic Test Project*. 2017.

[4] R. P. Chalmers. "mirt: A Multidimensional Item Response Theory Package for the R Environment". In: *Journal of Statistical Software* 48.6 (2012), pp. 1–29. DOI: 10.18637/jss.v048.i06.

[5] E. Darlington. "Contrasts in mathematical challenges in A-level Mathematics and Further Mathematics, and undergraduate mathematics examinations". In: *Teaching Mathematics and Its Applications* 33 (2014), pp. 213–229. DOI: 10.1093/teamat/hru021.

[6] S. E. Embretson and S. P. Reise. *Item Response Theory for Psychologists*. Mahwah, New Jersey: Lawrence Erlbaum Associates, 2000.

[7] T. Fawcett. "An Introduction to ROC Analysis". In: *Pattern Recognition Letters* 27.8 (2006), pp. 861–874. DOI: 10.1016/j.patrec.2005.10.010.

[8] *Improving the mathematics diagnostic test*. The University of Edinburgh, Aug. 2017. URL: https://www.ed.ac.uk/institute-academic-development/learning-teaching/funding/funding/previous-projects/year/march-2017/mathematics-diagnostic-test.

[9] M. G. Jodoin, L. A. Keller, and H. Swaminathan. "A Comparison of Linear, Fixed Common Item, and Concurrent Parameter Estimation Equating Procedures in Capturing Academic Growth". In: *The Journal of Experimental Education* 71.3 (2003), pp. 229–250. DOI: 10.1080/00220970309602064.

[10] T. Kang and T. T. Chen. "Performance of the Generalized S-X2 Item-Fit Index for Polytomous IRT Models". In: *Journal of Educational Measurement* 45.4 (2008), pp. 391–406. DOI: 10.1111/j.1745-3984.2008.00071.x.

[11] R. King. *MATH08051: Statistics (Year 2) Course Notes: 2015/16*. School of Mathematics, The University of Edinburgh, 2015.

[12] F. Marbouti, H. A. Diefes-Dux, and K. Madhavan. "Models for early prediction of at-risk students in a course using standards-based grading". In: *Computer & Education* 103 (2016), pp. 1–15. DOI: 10.1016/j.compedu.2016.09.005.

[13] C. E. Metz. "Basic Principles of ROC Analysis". In: *Seminars in Nuclear Medicine* 8.4 (1978), pp. 283–298. DOI: 10.1016/S0001-2998(78)80014-2.

[14] E. Muraki. "Information Functions of the Generalized Partial Credit Model". In: *Applied Psychological Measurement* 17.4 (1993), pp. 351–363. DOI: 10.1177/014662169301700403.

[15] M. Orlando and D. Thissen. "Likelihood-Based Item-Fit Indices for Dichotomous Item Response Theory Models". In: *Applied Psychological Measurement* 48.1 (2000), pp. 50–64. DOI: 10.1177/01466216000241003.

[16] I. Partchev. *A visual guide to item response theory*. Feb. 2004. URL: https://www.metheval.uni-jena.de/irt/VisualIRT.pdf.

[17] Preetish. *Exploratory Factor Analysis in R*. Feb. 2017. URL: https://www.promptcloud.com/blog/exploratory-factor-analysis-in-r/.

[18] S. P. Reise and N. G. Waller. "Fitting the Two-Parameter Model to Personality Data". In: *Applied Psychological Measurement* 14 (1990), pp. 45–58. DOI: 10.1177/014662169001400105.

[19] D. Rizopoulos. "ltm: An R Package for Latent Variable Modeling and Item Response Theory Analyses". In: *Journal of Statistical Software* 17.5 (2006), pp. 1–25. DOI: 10.18637/jss.v017.i05.

[20] F. Samejima. "Normal Ogive Model on the Continuous Response Level in the Multidimensional Latent Space". In: *Psychometrika* 39.1 (1974), pp. 111–121. DOI: 10.1007/BF02291580.

[21] *STACK*. The University of Edinburgh. URL: http://www.stack.ed.ac.uk/stack.

[22] H. Zhang. *The Optimality of Naive Bayes*. Proceedings of the 17th International FLAIRS conference (FLAIRS2004). AAAI Press, 2004.