Student Performance Prediction based on Course Grade Correlation

by

Cheng Lei
B.Sc., Beijing University of Technology, 2008

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of

MASTER OF APPLIED SCIENCE

in the Department of Electrical and Computer Engineering

Student Performance Prediction based on Course Grade Correlation

by

Cheng Lei
B.Sc., Beijing University of Technology, 2008

Supervisory Committee

Dr. Kin Fun Li, Department of Electrical and Computer Engineering
Supervisor

Dr. Fayze Gebali, Department of Electrical and Computer Engineering
Departmental Member

ABSTRACT

Supervisory Committee

Dr. Kin Fun Li, Department of Electrical and Computer Engineering
Supervisor

Dr. Fayze Gebali, Department of Electrical and Computer Engineering
Departmental Member

This research explored the relationship between an earlier-year technical course and one later year technical course, for students who graduated between 2010 and 2015 with the degree of bachelor of engineering. The research only focuses on the courses in the program of Electrical Engineering at the University of Victoria. Three approaches based on the two major factors, coefficient and enrolment, were established to select the course grade predictor including Max(Pearson Coefficient), Max(Enrolment), and Max(P_i) which is a combination of the two factors. The prediction algorithm used is linear regression and the prediction results were evaluated by Mean Absolute Error and prediction precision. The results show that the predictions of most course pairs could not be reliably used for the student performance in one course based on another one. However, the fourth-year courses are specialization-related and have relatively small enrolments in general, some of the course pairs with fourth-year CourseYs and having acceptable MAE and prediction precision could be used as early references and advices for the students to select the specialization direction while they are in their first or second academic year.

Contents

List of Tables

List of Figures

Glossary

# of Student	The number of students
# of Technical Courses	The number of technical courses
CourseX	The earlier year course
CourseY	The later year course relative to CourseX
Enrolment	The number of students registered in both CourseX and CourseY
#points	The enrolment of one course pair
pValue	*p*-value of the t-test
coefficient	Pearson Coefficient
CrsXCode	The course code of CourseX
CrsXNum	The course number of CourseX
CrsYCode	The course code of CourseY
CrsYNum	The course number of CourseY
0~1.0	The number of testing enrolments that have prediction error in the range of from 0 to 1.0
%	Prediction precision

Acknowledgements

First, I would like to thank my supervisor, Dr. Kin Fun Li, for providing advice, support and encouragement in the research process. His experience and support have been invaluable to me during my graduate study.

Second, I would like to thank my supervisor committee member Dr. Fayze Gebali and the examiners Dr. Alex Thomo for spending time on reviewing my thesis.

Last, I would like to thank my family, especially my wife June, who supported me during my graduate study.

Chapter 1 Study Rationale and Literature Review

The academic performance assessment in institutions is popular and essential for both the students and instructors, even for the institutions themselves. Academic prediction in junior academic years does not only facilitate the students to adjust their study plans and study ways to avoid poor academic performance and failure in advance, but it also helps the instructors and the institutions to adjust the curricula to improve the teaching quality and decrease dropout rate. For the universities, they can optimize the resources to help their students based on the individual issues. At present, there are a number of papers that investigate student academic performance prediction by using various data related to both the students and the universities.

Academic performance prediction involves a wide range of knowledge, including the parameters used in the prediction models. There is a variety of predictors used to predict the academic performance, for instances, prerequisite academic performance, mathematical skills, and grade point average (GPA) in earlier studies. They can also be the student's demographic profiles, such as gender, race, nationality, family background, etc. The prediction models are mostly based on the field of machine learning and vary from clusters to decision tree, to Bayesian, to regression and so on.

The academic performance itself have different definitions in different aspects based on the specific goals of the research. For examples, it can be one student's performance in a course at a certain year, it can also be his or her program performance indicating if he or she will pass or fail the program.

This chapter focuses on this research background and also describes some referential researches done by other researchers. It presents the factors and models they used in their experiments and the academic performance domains representing on what aspect the researchers focused in their prediction. Also presented are the prediction results concluding what factors or predictors have crucial impacts on the academic performance, what factors have less, even with no influence, and also which models have better prediction performance. In addition, the prediction results, assessment methods, and tools in these researches are of significance and presented in Section 1.1 in this chapter.

Our research goals and objectives are discussed in Section 1.2. The structure of this thesis is presented in Section 1.3.

1.1 Literature Review and Related Research

There are several ongoing researches related to academic performance. The research datasets applied to the experiments mainly include two different types, namely, previous academic performance, such as GPA, mathematical skills, etc., and demographic profiles, for instances, gender, family background, and institution background. The evaluation methods also vary, for instance, the average prediction accuracy (APA) indicating how well the model predicts in average and percentage of accurate prediction (PAP) illustrating the percentage of accurate prediction over all predictions. Lei and Li presented a paper gathering the attributes used in the student performance prediction including academic attributes and student's profile [2].

There are some researches that only applied the previous academic performance as predictor variables. Huang and Fang applied four different prediction models: multiple linear regression (MLR), multilayer perceptron network (MLP), radial basis function network (RBFN) and support vector machine (SVM) on the student's cumulative GPA (CGPA), grades on these four prerequisite courses: Statics, Calculus One and Two, and Physics and three mid-term examinations' scores to assess the student's final exam performance of Engineering Dynamics [3]. These algorithms have their own advantages in different aspect of predictions. The four models outputted the accuracy of APA from 88% to 86% for MLR, MLP, RBFN and SVM, respectively, while it gave the accuracy of PAP from 61% to 64% for the four models, respectively. These results indicate that MLR prevails when predicting the average academic performance of Engineering Dynamics class as a whole while the SVM is better when predicting individual's academic performance of the Engineering Dynamics course.

Mussoab et al. proved that the natural science and mathematics (Physics, Chemistry and Mathematics) performance in high school are the strongest factors in the prediction of the first year GPA while the Fine Arts and gender factors are weak prediction factors [5]. Li et al. did an experiment to evaluate whether the students would dropout or fail the

program using the first-year engineering students' academic records by principal component analysis (PCA) and found that mathematical skills are highly relevant to their engineering work [6].

Asif et al. found the underlying relationship between the academic performance in early years and the degree completion via decision tree (DT), k-Nearest Neighbor (KNN), Naïve Bayesian (NB) and artificial neural network (ANN), which indicates it is possible to predict degree completion by applying only pre-university marks and the marks of first and second year courses [8]. Huang developed some mathematical models including MLR, MLP, RBFN and SVM, by using the course scores of Engineering Dynamics, prerequisite courses scores, scores of midterms and GPAs as predictors and found RBFN and SVM perform better in general in APA and PAP [10].

In addition to the academic records in early years, some researchers combined other potential factors, such as the student's gender, nationality, race, family financial conditions, etc. Ibrahim and Rushli applied information technology application knowledge, programming knowledge, previous school type (boarding or non-boarding) together with family financial status to predict the student's final CGPA on graduation by using ANN, DT and linear regression (LR), all of which produced over 80% accuracy [1].

Chen et al. utilized more estimators including gender, tertiary education entrance exam scores, high school graduation exam results, high school location, school type (public or private) and the time between from high school and university admission to predict a student's average academic performance of the first academic year by using ANN associated with cuckoo search (CS) and cuckoo optimization algorithm (COA) [7]. The cuckoo search and cuckoo optimization algorithm are inspired by the lifestyle of the birds called cuckoo by laying eggs in nests of other host birds [25] [26].

Oladokun et al. did not only used the students' subjects scores (Math, Physics, Chemistry, etc.) and university entrance examination scores, but also the students' demographic profile including gender, age, parent educational status, secondary school background (public or private, location) into the ANN to classify the students' CGPA into good, average or poor with 70% of precision [9].

Bhardwaj and S. Pal used naïve Bayesian (NB) to determine which group (First: Grades obtained in Bachelor of Computer Applications > 60%, Second: 45% < Grades < 60%, Third: 36% < Grades < 45% and Fail: Grades < 35%) a student would be in based on their demographic profiles [11]. These include gender, food habits (vegetarian or non-vegetarian), living location (village, town or city), family status (joint, individual), family size, etc., as well as grades from secondary schools. They also found that the grades from secondary school have the most importance in the prediction followed by the living location. Yadav and Pal applied similar predictors but employed different prediction algorithms such as the statistical classifier C4.5, iterative Dichotomiser 3 (ID3), and classification and regression tree (CART) to assess the students' final exam outcomes (fail or pass) with the highest precision of 68% from C4.5 [12].

Osmanbegović and Suljic tried NB, MLP and J48, a Java implemented C4.5 decision tree algorithm, to classify a student's grade level by using gender, resident's distance to university, family annual earnings, etc, as well as prior academic performance including GPA in high school and entrance exams. [14]. The research reveals that NB gained the highest accuracy of 77%. Similarly, Ramesh et al. also tried NB, MLP, sequential minimal optimization (SMO), J48, and REPTree from Weka to evaluate what level of grade the student will obtain in higher secondary school [16]. The student's characters and the family background including parents' occupations and also the student's primary school academic records were explored. It was found that the types of school (private or public) has least influence on the performance while the parents' occupations are of significance to the performance prediction.

Agrawal and Mavani categorized the students into four different levels, na,ely poor, average, good and excellent, by using ANN and NB with their secondary school performance, living places (town, village, city and etc.) and teaching languages [17]. The ANN outperformed the other algorithms with an accuracy of 70%. Both Cortez and Silva, and Berhanu and Abera analyzed the students' academic records in early year along with the demographic profiles such as parents' occupations, living place (urban or rural), gender, age and so on, to predict the students' final academic performance in later years [19] [20]. The former experimenters tried four different algorithms including DT, random forest,

neural Network and SVN. It was found that prior academic performances highly affect the student's achievement in later time. The latter experimenters just tried DT and gained the accuracy of 85%.

Some researchers also added more predictors such as leadership, time management, and study motivation, to predict the student's academic performance. The research done by Mussoab et al. [4] used ANN and four main factors: working memory capacity; attentional network test results; learning strategies such as attitude, motivation, time management, anxiety, and concentration; background variables, for instances, gender, parents' highest education level, parents' occupations, and secondary schools. Their goal is to predict a student's GPA of all courses at end of the academic years. They gained greater accuracy compared with traditional methods such as discriminant analysis with precision of 100% at identifying the top 33% and lowest 33% groups and precision from 87% to 100% at identifying low, mid and high performance levels.

Minaei-Bidgoli et al. introduced two different groups of classifiers: tree classifiers (C5.0, CART, QUEST, CRUISE) and non-tree classifiers (Bayes, 1-nearest neighbor (1NN), KNN, Parzen and MLP) to explore the students' academic performance of an introductory physics course for scientists and engineers and gained over 80% of precision [13]. The predictor variables used contain problem resolution ability including interactions with both other students and instructors, the time they spent, the attempt times they tried, and success rate of their first try and final success rate. To improve the accuracy, the genetic algorithm (GA) was applied, which achieved over 10% improvement of accuracy.

Al-Malaise et al. experimented on number of solved quizzes, number of submitted assignments, hours spent, etc., and used different algorithms: AdaBoost.M1 [22], LogitBoost [23], C4.5, and stage-wise additive modeling using a multi-class exponential loss function (SAMME) [24] to assess if the students will fail or pass the course [21]. SAMME and AdaBoost.M1 obtained the same prediction accuracy of 80% at the 5^{th} and 10^{th} iteration but the prediction accuracies of the two algorithms decrease as iteration numbers increase, but the prediction accuracy increases in LogitBoost.

Pleskac et al. used hierarchical regression to predict a student's GPA based on two types of predictors which include cognitive predictors, for instances, leadership and

responsibility, and noncognitive predictors such as high school scores and the demographic profiles [15]. The results proved that the students' previous academic performance played an essential role in further academic performance. Ahmed and Elaraby applied the student's academic performance in early year and detailed information including attendance, assignment, lab performance, midterms' marks and the institution background, to ID3 to predict the student's final marks in information system courses [18].

1.2 Research Goals

The present research aims at exploring academic performance in the program of Electrical Engineering at UVic in the four year program. In other words, the goal is to check the possibility of using one technical course's grade to predict another one's grade. It is important to study such underlying relationships so that the students can have further academic performance references and so that the instructors can adjust their curriculums and teaching strategies.

In details, the research started from finding the correlation of two different courses in different years, followed by picking the predictor for one of the courses using first their correlation, then enrolment, and followed by the combination of the two. Once one course's predictor is picked, linear regression is applied as the prediction technique, and the MAE and precision are employed to evaluate the prediction results. Finally, conclusions are made on the basis of the prediction results.

1.3 Thesis Structure

This thesis consists of five chapters, each of which focuses on one specific topic. The first chapter primarily talks about the background and literature review. It states the related work done by other researchers and the theories as well as methodologies applied in their researches.

Then, it is followed by Chapter 2 with the introduction of the research datasets and the preprocessing. The introduction section describes data presentation including the availability and limitations. The preprocessing section classifies one course into a technical course or a non-technical course, removes non-technical courses data, and adjusts courses

which have different names in different years. The privacy protection strategy is also designed and implemented in this chapter.

The third chapter introduces the preparation work of research datasets. It depicts the steps and strategies applied in the datasets preprocessed in chapter two. It prepares the datasets for the correlation exploration among technical courses in the last section of this chapter.

The fourth chapter is the predictor selection and prediction evaluation. It shows how to apply the theories and methodologies to the predicting ready datasets, which produces course predictors and the corresponding prediction results. Therefore, the evaluations of these computation results are followed after each methodology explanation. The last section in this chapter concludes the different methodologies in the prediction.

Chapter 5 summarizes the research and gives future work related to this research.

Chapter 2 Datasets and Preprocessing

The raw datasets obtained by queries from the metaserver at UVic were preprocessed including the grouping of technical courses and non-technical courses, removing redundant course records and aggregating courses renamed in different semesters. Moreover, the sensitive issues related to students' privacy and required by Human Research Ethics Board (HREB) [26] were also discussed and implemented in this chapter.

2.1 Data Description and Format

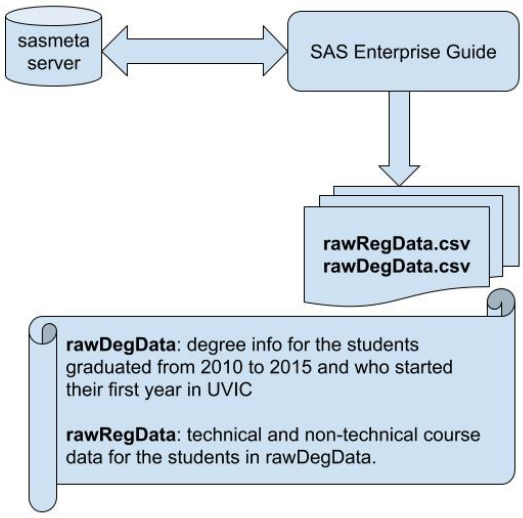

Figure 1 Data Query Workflow from SAS Meta Server

The datasets involved in the current research are stored in the metaserver at UVic and the research-aimed usage permission of the students' academic records was granted by the university. The tool used to query datasets from the server is SAS Enterprise Guide 5.0, a user-friendly graphic user interface (GUI) and a subclass of SAS [27] [28] [29] which is a third party software containing data mining algorithms as well as other simple functionalities such as data querying, sorting, import, export and so on. The workflow of raw data query between the server and the SAS software is illustrated in Figure 1.

The students' academic records are stored in the metaserver as tables associated with the student-related information such as name, student numbers, nationality, etc., and institution-related metadata, for instancea, faculty, department, program and so on. The academic records are from those who graduated from the program of Electrical Engineering (EE) at UVic with bachelor degree of engineering (BEng) from 2010 to 2015. In order to prevent the final research results being biased, the academic records from those who did not start their first year of Electrical Engineering program at UVic were excluded by checking if they have the records of Laboratory of Engineering Fundamentals (ELEC 199). The course of ELEC 199 is used as the criterion to judge if a student started his or her EE program at UVic because it is a fundamental engineering course and every electrical engineering student who started their program at UVic prior to 2014 has to enroll in it.

2.2 Strategy for Privacy Protection

The research presented in this thesis follows strict privacy guidelines. The students' private data must be protected so that their identities cannot be traced and revealed. The regulations and guidelines for privacy protection were defined by Human Research Ethics [27] at UVic. Therefore, the privacy protection policy is essential and vital to be set and implemented in order to complete the research.

In order to identify the students enrolled in the courses, the student numbers or V-Numbers or V#s were queried from the server. Moreover, the course records have certain row orders when they were exported from the database so that the student could be traced by comparing the row index of the course records. Therefore, to keep the student number safe and unidentifiable is the key to prevent violating the students' privacy in any possible way anywhere in the research.

There is one workflow, as shown in Figure 2, designed and implemented to achieve this goal. The workflow primarily consists of two sequential steps, namely, exported order shuffle of course record and V# encryption, respectively. The shuffle process used in the anonymization workflow was implemented by a randomization mechanism which dynamically used the time as seeds to generate a random number. In other words, the machine time was repeatedly picked as the seed when generating the random number,

which ensures the seed's uniqueness and non-traceability. Once the random number was generated, its uniqueness was checked in runtime to ensure the course record's integrity.

Figure 2 Students' Number Replacement Process

The course records were shuffled using a shuffle process and were saved to a csv file, each row of which represents one course record. Each course record has its own row index in the file when exported. Therefore, the shuffle process avoids the trace by comparing the course record row index. Secondly, the student numbers in the course record dataset were anonymized. This part contains two sub-steps, namely, student number substitution which used the random number generated by the shuffle mechanism to replace the student number, and substituted text encryption using the SHA256 (a cryptographic hash algorithm which generates an almost unique 256-bit signature for a text) [30] to convert the replaced student number into non-human readable texts.

2.3 Course Grouping

The raw course records exported from the metaserver contain both technical and non-technical courses taken by the students through their degree years. The non-technical courses are not a prerequisite to technical courses and it seems that they have no bearing on the outcome of subsequent technical courses. Therefore, they are identified and excluded from the research datasets. On the other hand, the technical courses need to be classified into different groups by academic year in order to meet the research needs. Therefore, the following section depicts how to group them.

2.3.1 Technical and Non-Technical Course Grouping

The concept of a technical course is that the course is directly related to the program of Electrical Engineering and is in the pool of core courses in the program academic schedules

or in the program requirements as stated on the homepages of Department of Electrical and Computer Engineering from 2005 to 2015 [32] - [42]. The technical courses in the present research datasets contain the core courses for the program of Electrical Engineering and technical electives as well. As there are several specializations in the program of Electrical Engineering such as Mechatronics and Embedded Systems, Physics, Computer Music, etc., the courses belong to these specializations are treated as technical courses as well. Therefore, the technical courses can be simply collected from the courses listed in the academic schedules or in the degree program requirements.

However, after careful inspection of these courses from the academic schedules or program requirements, to treat all courses in the academic schedule as technical courses is not accurate enough since there are some courses which cannot be branded as technical courses as they are not directly related to the technical aspects of Electrical Engineering. For instance, ENGR 280 (Engineering Economics), is a third-year course about the relationship between engineering and economics, and the fourth-year course ENGR 297 (Technology and Society) illustrates how the society is affected by technology.

The non-technical courses are the ones, such as English, that are not directly related to the technical knowledge of the program but are helpful for the students' development in other fields related to soft skills and professional development. These courses are offered by the program of Electrical Engineering, and can also be from other programs or faculties.

2.3.2 Courses Grouped by Calendar Year

The courses are listed differently in the program requirements or the academic schedules in Electrical Engineering. For instances, the courses in UVic Calendar 2005-2006 for BEng in Electrical Engineering were grouped by terms [32] while these courses in UVic Calendar 2014-2015 for BEng in Electrical Engineering were grouped by years as program requirements [42]. There are eight terms listed in the academic schedule of BEng in Electrical Engineering, namely, Term 1A, Term 1B, Term 2A, Term 2B, Term 3A, Term 3B, Term 4A and Term 4B. Likewise, there are four years of Year 1, Year 2, Year 3 and Year 4 listed in the program requirements of Electrical Engineering. By comparing the courses in each term and in each year, it can be concluded that Term 1A and Term 1B form

Year 1, Term 2A and Term 2B form Year 2, Term 3A and Term 3B form Year 3, and Term 4A and Term 4B form Year 4, respectively. The mapping of years and terms is shown in Table 1.

The courses listed in Year 1, Year 2 and Year 3 or Term 1A and Term 1B, Term 2A and Term 2B, and Term 3A and Term 3B have the course numbers starting with the academic year number or the term number (1, 2 or 3). The exception is ENGR 280 (Engineering Economics) in Term 3B or Year 3. There are three different leading digits in the course numbers: 2 (such as ENGR 297, Technology and Society), 3 (such as ELEC 395, Seminar) and 4 (such as ELEC 499, Design Project II) in Year 4 or Term 4A or Term 4B. In addition, both the technical electives and specialization courses are scheduled for Year 4 or Term 4A or Term 4B only. Although most of the technical electives have course numbers with first digit starting with 4, there are exceptions, for instance, SENG 330 (Object-Oriented Software Development). Therefore, except such mis-numbered courses, the year of a course can be identified by the first digit of its course number.

Table 1 Year and Term Mapping

Year 1	Year 2	Year 3	Year 4
Term 1A, Term 1B	Term 2A, Term 2B	Term 3A, Term 3B	Term 4A, Term 4B

2.4 Data Redundancy Removal and Adjustment

Our research focuses on the course records only from those who graduated between 2010 and 2015 with a bachelor degree of engineering in Electrical Engineering. Moreover, the non-technical courses need to be removed from the datasets as well, as discussed earlier.

As the metaserver at UVic backups the data at a certain time according to its backup policy, there are hundreds of thousands of course records having the same contents except the backup timestamp. Thus, these duplicate records were eliminated and only the latest time stamped ones were kept in the research datasets. Moreover, the students are able to re-register in the courses if they fail, or drop the courses in previous attempts. Therefore, the courses with multiple grade points were kept with the lowest grade to reflect course

grade of interest. The dropped courses were also eliminated from the research datasets since they are not assigned with final grades.

The grading criteria [43]- [53] at UVic academic calendars show that the grade given for a course could be a numeric value between 0 and 9, or a label indicating the course is failed or passed or at other status, such as COM (Complete), CTN (Continuing), F/X (Unsatisfactory Performance), INP (In Progress), N/X (Did not complete course requirements by the end of the term) or WDR (Withdrawal under extenuating circumstances). Since no useful information can be inferred from these courses, then, the courses that only have the text grading labels were excluded from the research data as well.

Another issue is that some technical courses' course number was changed in a subsequent academic year, for examples, MATH 133 was changed to MATH 110 and ENGR 110 was changed to ENGR 111. There are also cases that a course was completely renamed. For instance, MECH 141 was renamed to ENGR 141 in the academic year 2009. Therefore, for such courses, their names have to be unified and made unique for consistency in the research results.

One other interesting case is that a student with a degree in Electrical Engineering may be transferred from other institutions or other faculties or departments at some time point. As such, their academic records may be incomplete, which means that they have not registered in some of the prerequisite courses for the program. The research results may be skewed if such course data was applied. Therefore, the record of such students was excluded from the research data as well.

Courses with only one or two students enrolled were also excluded from the dataset. In the next chapter, the courses are paired to compute the Pearson correlation and the enrolments are the data points in the correlation computation. The courses with one enrolment are invalid for the computation and the courses with only two enrolments have insufficient information to explore the correlation. However, the one or two enrolments does not imply there were only one or two students in the class, instead, it tells that there are only one or two students whose course data meet the research requirements as described in Section 3.1.

2.5 Chapter Summary

With the preprocessing completed, student distribution in academic years is shown in Table 2 while technical course distribution by academic year is shown in Table 3. The technical courses including both compulsory technical courses and technical electives from the calendars [32] - [42] are presented in Appendix 1, while technical courses in the research data, with improper courses eliminated, are listed in Appendix 2.

The enrolment of a technical course varies dramatically from 1 to 120 as not all students started their degrees at UVic. Some students may be transferred from another universities or colleges. Moreover, the program of Electrical Engineering contains several specializations, for instances, Computer Music Option and Biomedical Engineering Option, which dilute the enrolment because students have different specializations. The enrolment table for every technical course extracted from the research data was listed in Appendix 3.

Table 2 BEng Student Distribution in Electrical Engineering in Calendar Year

Year	2010	2011	2012	2013	2014	2015
# of Student	15	27	30	24	17	7

Table 3 Technical Course Distribution in Program Year

Year	1ST Year	2ND Year	3RD Year	4TH Year
# of Technical Courses	12	12	11	45

There are almost one third of the fourth-year courses with enrolment less than 10, and another one third with enrolment between 10 and 20. The reason for the wide spread of enrolment of the fourth-year courses is that most of these courses are technical electives except a small number of courses which were listed as compulsory in the academic schedule, for example, ELEC 499. By contrast, the enrolment for the courses in the first three years was at much stable levels, especially for the third-year courses where all enrolments were 120. Most of the enrolments of the technical courses in Year 2 were over 100 and only three of the second-year technical courses having enrolment less than 100,

namely, 95 for ELEC 216, 57 for STAT 254, and 16 for CSC 230, respectively. Similarly, the enrolments of Year 1 technical courses were in the range from 47 to 120.

Therefore, the course pairs with a fourth-year course have small enrolment and the Pearson Correlation Coefficients of those course pairs would be in a wide range. On the other hand, the correlation coefficients of course pairs with second- and third-year courses would be more clustered within a certain range. At this point, the raw data has been polished and ready for the correlation computation.

Chapter 3 Course Grade Correlation

This chapter introduces the approach used to explore the correlation between grades obtained in two technical courses. The Pearson Coefficient is used to examine the correlation. The way that the dataset is partitioned into the training sets and testing sets is described. Also, the correlation results of the course pairs are presented in this chapter.

3.1 Pearson Correlation and its Strength Determination

This part of the research is trying to find how two courses are correlated. In other words, it tries to find how important one course's performance is to another course's performance. Various correlation approaches can be used. In addition to the Pearson Correlation, which is introduced to compute how strong the course grades are correlated with each other, other correlation approaches were also explored. For instance, Spearman rank-order correlation [59] examines the monotonic relationship between two continual or ordinal variables. However, it uses the ranked value instead of unprocessed data. Kendall rank correlation [60] is also based on ranks, which is not appropriate for our data. On the other hand, Pearson Correlation is widely used to evaluate the linear relationship between two variables and can use unprocessed data. Therefore, Pearson Correlation is chosen for the course grade correlation analysis.

The Pearson Coefficient [54] [55] [56], also known as Pearson product-moment correlation coefficient, is used to represent the correlation coefficient between the grades of two technical courses. It is defined as the result of the covariance of two variables X and Y, divided by the product of their standard deviations, where the two variables X and Y have the same dimension, with n points and noted as $X = (x_1, x_2, \dots, x_n)$, and $Y = (y_1, y_2, \dots, y_n)$. It is a common metric to measure the linear strength between two variables X and Y. It gives the best linear fit for all data points of the two variables and measures the distances of the points to the best fit line. The formula to compute Pearson Correlation Coefficient, r, for a sample dataset of $X = (x_1, x_2, x_{3,\dots,} x_n)$ and $Y = (y_1, y_2, y_{3,\dots,} y_n)$ is shown in (1):

$$r = r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

where:

- n is the sample size
- x_i, y_i are the ith indexed samples
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, is the mean of X; similarly for $\bar{y}$

The coefficient of Pearson Correlation, r, has an inclusive continuous value, ranging from negative 1.0 to positive 1.0, [-1.0, +1.0]. The sign of the coefficient indicates whether the linear correlation is positive or negative. In the positive quadrant, the bigger the coefficient r is, the stronger the Pearson Correlation of the two variables X and Y is, which means that variable Y changes as X changes in the same direction. Conversely, in the negative quadrant, the correlation of the two variables X and Y, is inversely proportional to the coefficient, r, which means that if the variable Y increases, the variable X decreases.

The two extreme values of the coefficient at the two ends, -1.0 and +1.0, show the perfect linear correlation between the two variables. The value of +1.0 means the two variables have perfect positive linear correlation while the value of -1.0 indicates the two variables have perfect negative linear correlation. The middle point, 0, of r, indicates that the two variables do not have linear correlation.

In addition, there are some pre-conditions and assumptions made before one can use this metric. Before the Pearson Correlation is applied, the data must meet these three constraints. The first condition is that the data is not categorical and has a known interval. The second assumption is that the data points scattered in the plot must be linearly related. Finally, the points of each variable are assumed to be normally distributed.

To determine how strong the Pearson Correlation of two variables is, the Null Hypothesis, H_0 [57] is used. H_0 indicates that there is no linear correlation between the two variables against the alternative hypothesis, H_1, which shows the two variable has linear correlation. Then, the p-value of the two-tailed test [58] is applied as an indicator to determine the correlation significance. The magnitude of the p-value gives the strength of

rejecting the null hypothesis. Therefore, a p-value cut-off has to be set so that it can be used to compare against the p-values from the datasets. Usually, 0.05 is selected as the p-value, which means there are 95% possibility to reject the null hypothesis, H_0 [61]. In other words, if the p-value of one correlation coefficient is less than or equal to the cut-off value, 0.05, it is confident to accept the correlation is strong. Otherwise, the linear correlation is deemed unacceptable.

3.2 Data Partition

The research goal of this project is to investigate whether one can estimate a student's course grade based on the performance in another course taken earlier. If the theory is proven to be valid, then, it is straightforward to use course performance in early years to predict course performance in later years. In other words, grades of Year-1 courses are used to predict grades of Year-2, Year-3 or Year-4 courses; grades of Year-2 courses are used to predict grades of Year-3 and Year-4 courses. In order to do so, course data from the students in earlier years of the program was applied to train the prediction models while course data from later years was used to test the trained models.

Table 4 Training Sets and Testing Sets

Training Sets	Testing Sets
Train2010-2011	Test2012, Test2013, Test2014, Test2015
Train2010-2012	Test2013, Test2014, Test2015
Train2010-2013	Test2014, Test2015
Train2010-2014	Test2015

The entire data set was split into the training datasets and testing datasets. The training sets consist of course data of at least two years while the test sets contain course data of one year. Therefore, based on the graduation year bin, there are four training datasets with the increment of one-year of course data, namely, Train2010-2011, Train2010-2012, Train2010-2013 and Train2010-2014. The corresponding testing sets for each training dataset are shown in Table 4.

3.3 Course Correlation Results and Analysis

Since it stated in Section 3.1, Pearson Correlation is computed between two variables which have the same dimension, which means that the courses without common students cannot be paired even if they are provided in different years. The two variables in the current research are formed from course pairs. A course pair is defined as two courses enrolled by the same students with grades given. The number of students in one course pair is defined as the enrolment of the course pair. CourseX is defined as the early year course and CourseY is the later year course in the course pair. For instance, CourseX in the course pair of ELEC 199 and MATH 200 is ELEC 199 while MATH 200 is CourseY.

As stated that the courses offered in different years were paired, then one course in one lower year can be paired with multiple courses in upper years. Thus, each course pair has one Pearson Coefficient, and therefore, for one specific course, it has multiple paired courses and the corresponding coefficients. Meanwhile, as the coefficients and the enrolments of these course pairs vary significantly, the p-value stated in Section 3.1 is applied to determine the coefficient strength with its typical value 0.05 [61]. Therefore, the course pairs having p-value less than or equal to 0.05 are chosen as strongly correlated course pairs which are one of the subsets of all the possible course pairs. In other words, for the course pairs with common enrolments, the selected course pairs are the ones with p-value <= 0.05 while the unselected course pairs are the ones with higher p-value (> 0.05).

All strongly correlated course pairs computed in the training set of Train2010-2011 are listed in the table in Appendix 4. As expected, there are multiple CourseYs for one CourseX with strong correlation and vice versa. The frequency of CourseYs strongly correlated with CourseX is plotted in Figure 3. The bar chart clearly shows that for each CourseX, it has at least one strongly correlated CourseY, ranging from 1 (CSC 115, out of 33 total paired courses from Year 2, Year 3 and Year 4) to 22 (MECH 141, out of 54 total paired courses from Year 2, Year 3 and Year 4).

Similarly, the frequency of CourseXs strongly correlated with CourseY is shown in Figure 4. The figure also shows that each CourseY has at least one strongly correlated CourseX with a minimum of 1 out of 11 (CENG 255) and out of 21 (MECH 410), and a maximum of 17 out of 23 (ELEC 340).

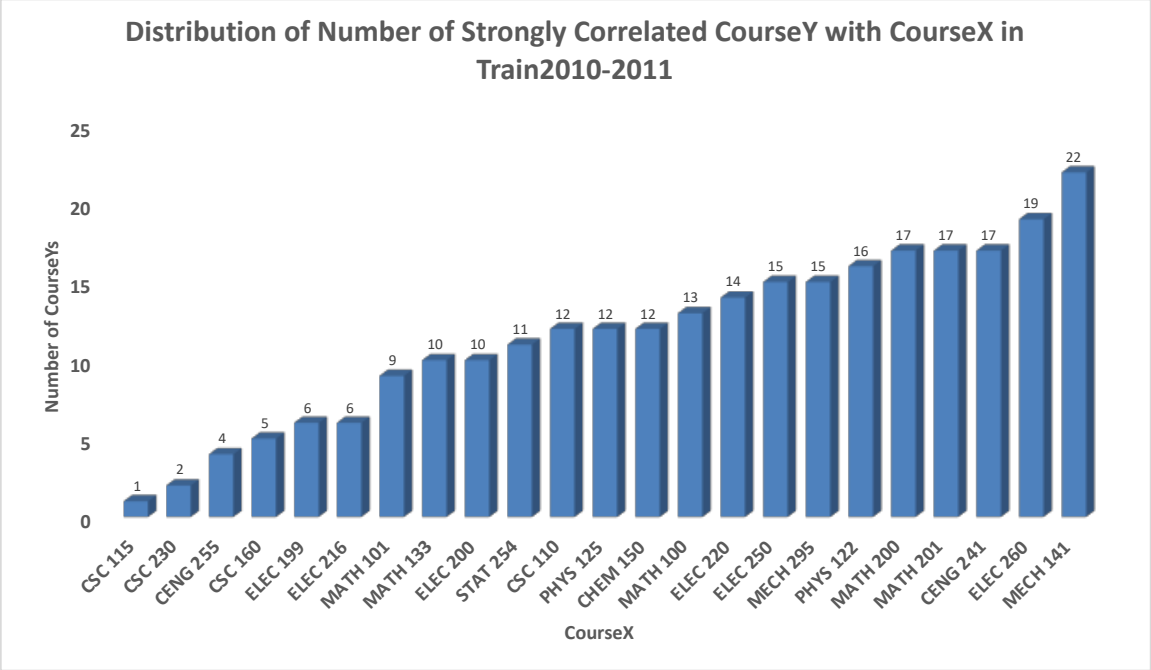

Figure 3 Distribution of Strongly Correlated CourseYs with CourseX in Train2010-2011

Comparing the two figures, the frequency of strongly correlated courses for CourseX is greater than the one for CourseY in general. As stated in Section 3.2, a CourseX in the training set is a first-year or second-year courses while a CourseY is a second-year, third-year or fourth-year courses. It can be concluded from this histogram that the first-year and second-year courses are more fundamental courses and may have significant impact on third-year and fourth-year courses which are more specialized.

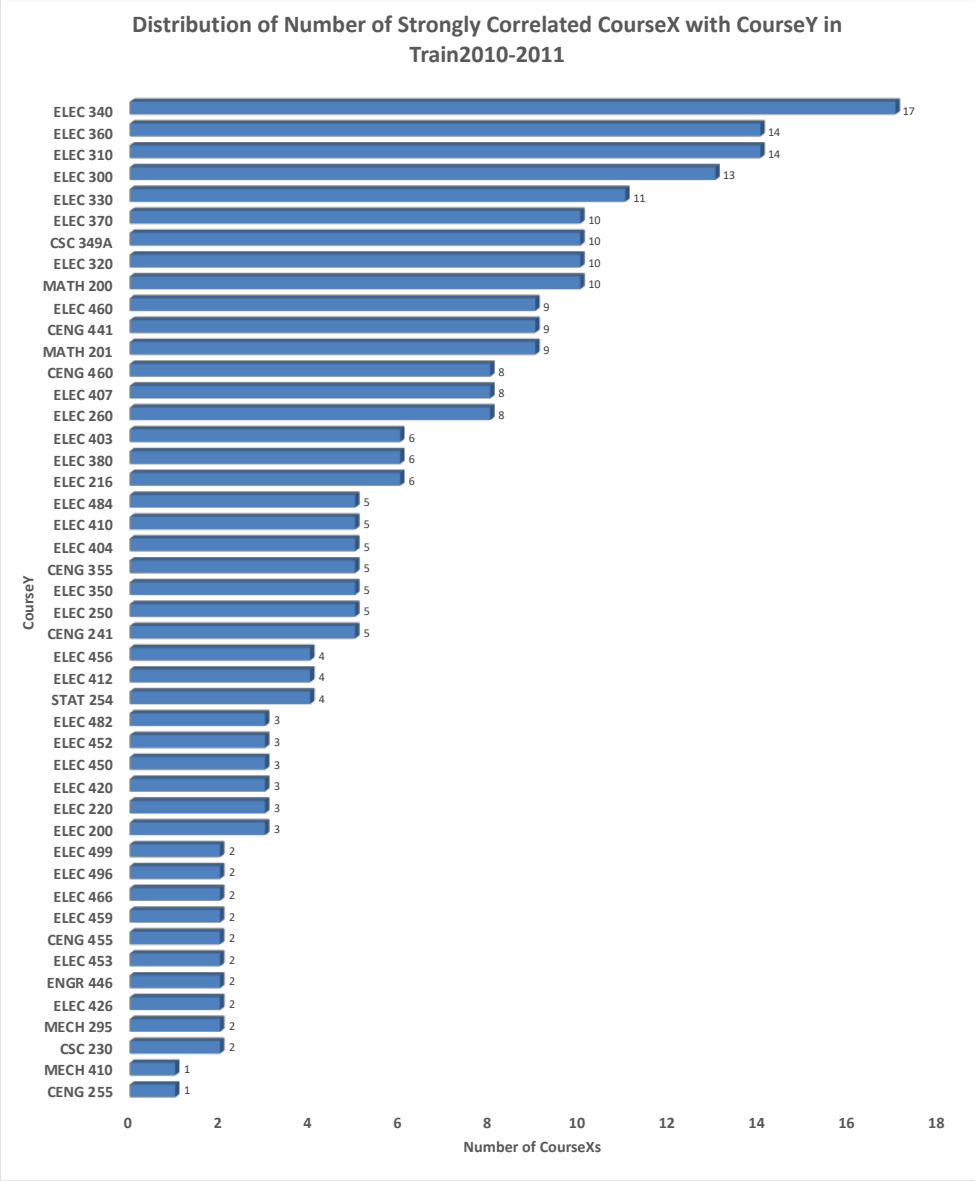

Figure 4 Distribution of Strongly Correlated CourseXs with CourseY in Train2010-2011

Similarly, the strongly correlated course pairs in the other three training sets, Train2010-2012, Train2010-2013 and Train2010-2014, were analyzed in the same way and similar results were obtained. The course pairs distributions for both CourseX and CourseY are shown in Appendix 5.

The distribution figures from the four training sets also show that a technical course, A, can be strongly correlated with several other technical courses, B_i $(i = 1, 2, , 3, \dots, n)$. Therefore, how to select one of B_i as the predictor, or predicting course that is best for

course A will be discussed in Chapter 4. As the ranges of strongly correlated courses span from 1 to 22 in the X to Y case, and 1 to 17 in the Y to X case in Train2010-2011, there should be sufficient courses, B_i's, to choose from.

The coefficient distribution in Train2010-2011 is shown in the histogram in Figure 5, which shows that most of the coefficients are around 0.5 and the coefficients in this training set fall mostly into the range of 0.3 to 1. There are 14 course pairs with coefficients between 0.9 and 1 shown in Figure 5. Therefore, it appears that these courses can be predicted perfectly if just based on their coefficient. However, as explained, these courses are paired with 4-year courses and have a small enrolment and that is why they have seemingly perfect coefficients. Also, that is why there are three different ways to select predictors discussed in Chapter 4.

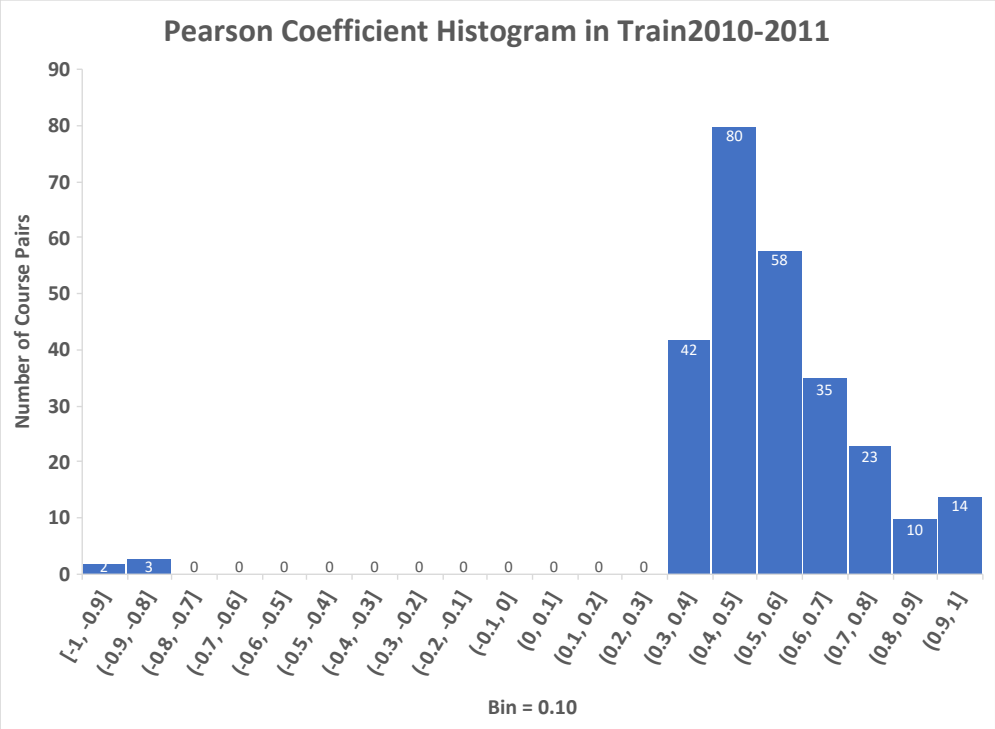

Figure 5 Histogram of Pearson Coefficients in Train2010-2011

The coefficient histograms in the other three training sets, Train2010-2012, Train2010-2013 and Train2010-2014, have similar characteristics as shown in Appendix 6. These three histograms show the coefficients clustered around 0.5 in most cases and also

have extreme values at the two ends of the histograms. Comparing the distributions of the four histograms, their similar trends ascertain that the data in the four years are consistent.

The enrolment and coefficient computed in the training set of Train2010-2011 are shown as a scatter plot in Figure 6. The figure shows the same trend of the coefficients presented in the histogram of Figure 5 that most coefficients are in the range from 0.3 to 1.0 and a small number of the coefficients are in the interval of from -1.0 to -0.8. Although the coefficients were filtered by the p-value and deemed as strong coefficients, they still have coefficients close to the extreme value of 1.0 or -1.0 in each training set. In particular, some course pairs have small enrolments, such as 3, 4, or 5, which indicates that they just have 3, 4, or 5 course marks. Meanwhile, some marks have the same values, which causes one grade in the coordinate system to present several marks. For example, the course pair of ELEC 250 and CENG 412 from Train2010-2011 has three points, (3, 7), (4, 6) and (3, 7) two of which have the same value of (3, 7). Therefore, the Pearson Coefficient of this course pair is -1.0 because the two points coincided. In most cases, the course pairs with seemingly perfect coefficients are paired with fourth-year courses because the fourth-year courses have small enrolments as concluded from Figure 3 and Figure 4.

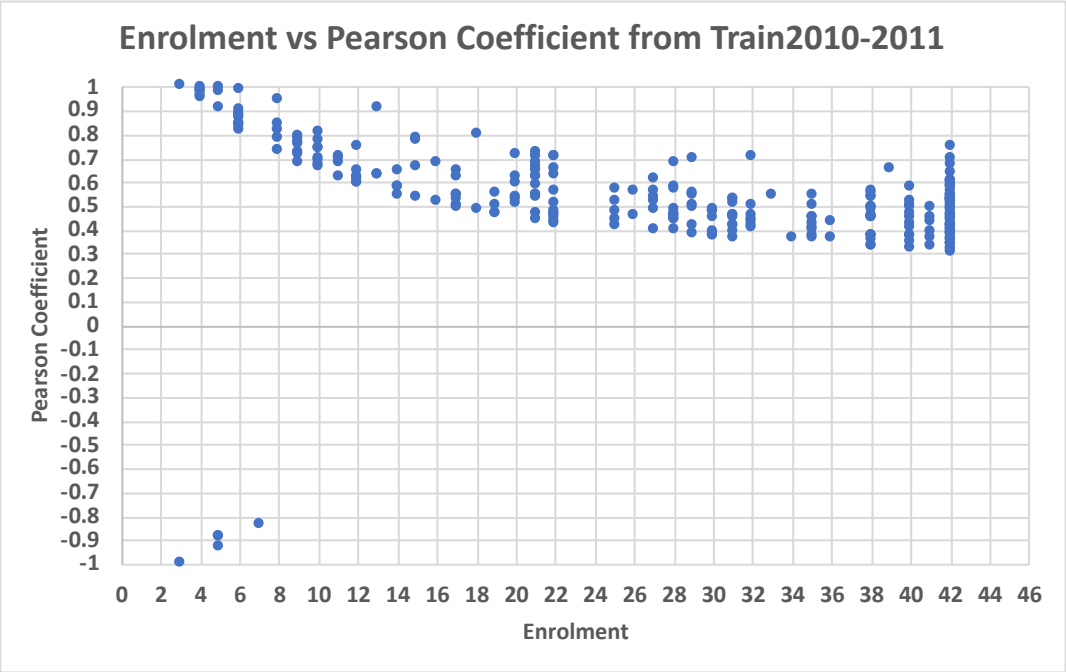

Figure 6 Enrolment and Pearson Coefficient from Train2010-2011

Also, the enrolment that produced the strong coefficients varies from 3 to 113 in Train2010-2011 to Train2010-2014. The other three training sets have similar distributions of enrolment and Pearson Coefficients as shown in Appendix 7. Figure 6 also shows that the coefficient decreases as the enrolment increases. In other words, more samples used in the computation of Pearson Coefficient produce more reliable coefficient.

From the bar charts of strongly correlated course pairs shown in Figure 3 and Figure 4, it can be seen that one technical course may have multiple strongly correlated courses. From the histogram of the coefficients, it shows that the coefficients have a wide range with different strongly correlated courses because of the enrolment. The scatter plot of the enrolment and coefficient shows that bigger enrolment generates a more reliable coefficient. Therefore, in the next chapter, the enrolment and the coefficient are the two major factors used to select the predictor for a technical course which have multiple strongly correlated predicting courses.

Chapter 4 Course Pair Prediction Analysis

The course pairs described in Chapter 3 were employed to train the prediction models to predict course grades as shown in this chapter. The course pairs were filtered by the strength of their Pearson Correlation which forms the best linear models for the points in the course pairs. Linear regression is introduced in this chapter. Also, for each course of the course pairs, there are multiple predictor-candidate courses which have strong correlation with it. Therefore, in order to reduce complexity and effort, three predictor-selection methods were designed and implemented. Once the predictor for each technical course is selected, the prediction starts with the model trained by the selected predictor course and the predicted course. The accuracy of the prediction results is measured by Mean Absolute Error and prediction precision.

4.1 Prediction Algorithm

Polynomial regression [62] demonstrates the relationship in statistics between the dependent variable y and the single independent variable x with an *n*th order polynomial. It is mathematically represented as below.

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_n x^n + \epsilon \quad (2)$$

where a_0, a_1, a_2,…, a_n are unknown estimators; ε is an unobserved random error with mean zero, conditioned on a scalar variable x; x is the independent variable and y is the dependent variable.

It can be seen from equation (2) that linear regression [63] is a special case of the polynomial regression, where the degree of the independent variable x equals to 1. Therefore, the linear regression model is a straight line as shown in equation (3).

$$y = a_0 + a_1 x \quad (3)$$

As stated in Section 3.3, the Pearson Correlation measures the linear strength between two variables and gives the best linear fit for the data points, which performs the same as linear regression. Therefore, it is simple to use linear regression as the prediction algorithm.

Mean Absolute Error [64], MAE, is the average of absolute errors, which measures the closeness of the predictions to the real outcomes. Its formula is defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i| \quad (4)$$

where n is the total number of instances; f_i is the ith prediction and y_i is the ith real outcome; correspondingly, $|e_i| = |f_i - y_i|$ is the error between f_i and y_i, that is, the difference between the real value and its estimate. The MAE is the metric used to assess the performance of the predictor in this research. The acceptable MAEs are the ones with the value less than or equal to 1.0 in the research, which means that the predicted grade of a course has ±1.0 average error margin.

Before applying this algorithm to predict one technical course's grade based on its paired strongly correlated course's grade, the strongly correlated course has to be selected first. As each technical course has several strongly correlated courses in the training sets as mentioned in Section 3.3, the approaches to select the strongly correlated course for one technical course is described in the next section, Section 4.2.

4.2 Predictor-Selection Approaches

As stated in section 3.3 in Chapter 3, each CourseX or CourseY has strong correlation with multiple courses and these strongly correlated courses have different characteristics in enrolment, correlation coefficient, etc. For example, there exist two course pairs, namely, Pair AB and Pair AC, where each capital letter represents a technical course. Pair AB has more enrolments than Pair AC but its coefficient is smaller than that of Pair AC. Meanwhile, the two course pairs' coefficients are treated as strong coefficients according to the p-value cut-off criterion. It can be seen from this example that enrolment and coefficient are the two major factors to select a predicting course.

In order to balance the influence from different factors, three different methods were explored to select predictor or predicting course for predicted courses in later years. The first two ways are simple, just by using the corresponding extreme values, the maximum enrolment or maximum Pearson Coefficient, as the predictor-picking criterion. In other words, for one technical course, the strongly correlated course with the maximum enrolment or maximum coefficient among all of its strongly correlated courses is chosen as the predictor, or predicting course. The third one is developed from the combination of the two weighed factors in coefficient and enrolment. The predictors are from the course pairs filtered using the p-value cutoff in previous chapter.

4.2.1 Predictor-Selection by Pearson Coefficient

This straightforward predictor-selection way means the predicting course having the maximum coefficient with the predicted course is selected as the predictor. If there are multiple predicting courses which have the same coefficients with the predicted courses, then the predicting course with the smallest course number is selected as predictor because a lower course number implies the course is offered in lower academic years. If the earlier year course helps identify potential problems earlier, then the earlier course selected is better. Meanwhile, earlier courses are from first or second year so the enrolment is expected to be higher and thus the result is more reliable.

The course pairs shown in Table 5 are selected from the strongly correlated course pairs in Train2010-2011 based on the maximum Pearson Coefficient with CourseX. It can be seen from the table that most of the CourseYs are fourth-year courses except the course pair of CSC 115 and ELEC 300.

The bar chart Figure 7 shows the coefficient distribution of the course pairs selected based on CourseX in the training set of Train2010-2011. As stated in Section 3.3 the coefficient decreases when enrolment increases, the course pairs with fourth-year courses have relative bigger coefficient than others because the fourth-year courses are more specialization-related, thus having small enrolment. The enrolments shown in the fourth column in Table 5 support this conclusion. The negative coefficient is also due to the small number of samples for the course pair of MATH 101 and ELEC 452.

Table 5 Course Pairs Picked by Max(r) by CourseX from Train2010-2011

CourseX	CourseY	coefficient	#points	pValue
MATH 101	ELEC 452	-0.9325	5	0.0208
CSC 160	ELEC 499	0.4743	30	0.0081
CENG 255	ELEC 482	0.648	12	0.0227
ELEC 200	ELEC 453	0.7016	11	0.0161
ELEC 199	ELEC 484	0.7109	9	0.0318
PHYS 125	ELEC 407	0.7459	12	0.0053
STAT 254	CENG 441	0.7514	9	0.0196
MATH 100	ELEC 403	0.762	9	0.017
ELEC 260	ELEC 407	0.7756	15	0.0007
ELEC 250	CENG 455	0.7819	8	0.0219
ELEC 216	CENG 460	0.8109	8	0.0146
MECH 295	ELEC 459	0.8363	6	0.038
CENG 241	ELEC 452	0.8771	6	0.0217
CSC 115	ELEC 300	0.8807	6	0.0205
MECH 141	ELEC 420	0.9068	5	0.0336
ELEC 220	ELEC 456	0.9081	13	0
PHYS 122	ELEC 459	0.9623	4	0.0377
MATH 200	ELEC 420	0.9733	4	0.0267
CSC 230	ELEC 482	0.9744	4	0.0256
CSC 110	CENG 460	0.9865	6	0.0003
CHEM 150	ELEC 466	0.9869	4	0.0131
MATH 133	CENG 455	0.9898	5	0.0012
MATH 201	ELEC 420	0.9969	4	0.0031

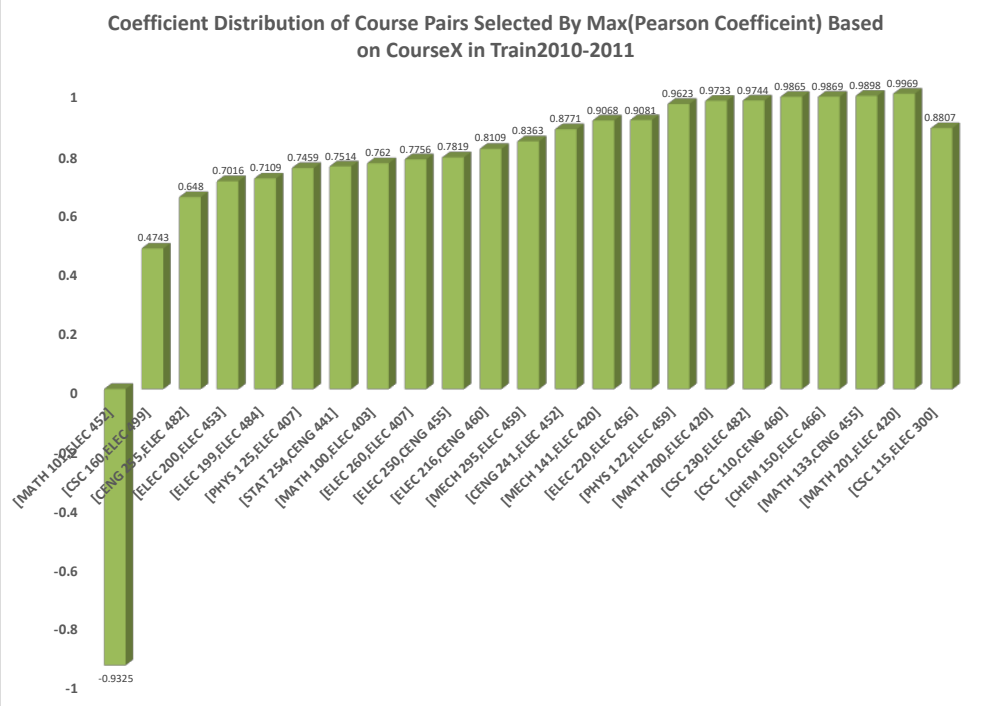

Figure 7 Pearson Correlation Distribution of Course Pairs Selected By Max(Pearson Coefficient) Based on CourseX in Train2010-2011

Table 6 Course Pairs Picked by Max(r) by CourseY from Train2010-2011

CourseX	CourseY	coefficient	#points	pValue	CourseX	CourseY	coefficient	#points	pValue
MATH 100	MATH 200	0.7054	22	0.0002	MATH 100	ELEC 403	0.762	9	0.017
MECH 141	ELEC 200	0.6553	39	0	ELEC 200	ELEC 404	0.5698	25	0.0029
MATH 100	MATH 201	0.7257	21	0.0002	CENG 241	ELEC 407	0.7777	15	0.0006
MECH 141	ELEC 216	0.7036	32	0	CSC 110	MECH 410	0.6909	10	0.0269
MATH 133	ELEC 220	0.4545	28	0.0151	MATH 201	ELEC 410	0.797	18	0.0001
CHEM 150	CSC 230	0.902	6	0.0139	ELEC 220	ELEC 412	0.8037	10	0.0051
CHEM 150	CENG 241	0.5659	28	0.0017	MATH 201	ELEC 420	0.9969	4	0.0031
MATH 100	ELEC 250	0.6261	22	0.0018	CHEM 150	ELEC 426	-0.8922	5	0.0418
MATH 101	STAT 254	0.6407	17	0.0056	CSC 230	CENG 441	0.9457	8	0.0004
MECH 141	CENG 255	0.4774	30	0.0076	PHYS 125	ENGR 446	0.421	35	0.0118
PHYS 122	ELEC 260	0.5223	31	0.0026	PHYS 125	ELEC 450	0.5159	16	0.0408
MECH 141	MECH 295	0.573	40	0.0001	MATH 101	ELEC 452	-0.9325	5	0.0208
CSC 115	ELEC 300	0.8807	6	0.0205	CSC 110	ELEC 453	0.8365	8	0.0096
ELEC 260	ELEC 310	0.7506	42	0	MATH 133	CENG 455	0.9898	5	0.0012
ELEC 250	ELEC 320	0.6942	42	0	ELEC 220	ELEC 456	0.9081	13	0
ELEC 260	ELEC 330	0.6715	42	0	PHYS 122	ELEC 459	0.9623	4	0.0377
STAT 254	ELEC 340	0.7036	21	0.0004	CSC 110	CENG 460	0.9865	6	0.0003
STAT 254	CSC 349A	0.619	21	0.0028	ELEC 260	ELEC 460	0.701	22	0.0003
ELEC 260	ELEC 350	0.5915	42	0	CHEM 150	ELEC 466	0.9869	4	0.0131
MECH 141	CENG 355	0.4252	40	0.0062	CSC 230	ELEC 482	0.9744	4	0.0256
STAT 254	ELEC 360	0.6484	21	0.0015	MATH 133	ELEC 484	0.8641	6	0.0265
STAT 254	ELEC 370	0.5444	21	0.0107	MATH 200	ELEC 496	0.733	8	0.0386
ELEC 260	ELEC 380	0.5253	42	0.0004	CSC 160	ELEC 499	0.4743	30	0.0081

Similarly, the strongly correlated courses with CourseY in Train2010-2011 selected by the maximum Pearson Correlation are shown in Table 6. There are more course pairs in the table than the ones selected for CourseX because the CourseXs are only the first- and second-year courses while the CourseYs are the second-, third- and fourth-year courses as stated in Chapter 2. In other words, CourseXs only contain the first- and second-year courses while CourseYs consist of second-, third- and fourth-year courses.

The coefficient distribution of the course pairs selected by the maximum of Pearson Coefficient based on CourseY is shown in Figure 8. As explained above for the course pairs selected based on CourseX, due to the small number of samples, these coefficients are close to 1.0 or -1.0, especially for the course pairs with fourth-year courses. The two negative coefficients are from the course pairs with fourth-year courses because of the small sample size for these two course pairs in the training set. The graph also shows that the coefficients of course pairs with second- and third-year courses in general are relatively

smaller than the ones of course pairs with fourth-year courses, although some of the course pairs with second- or third-year courses have coefficient close to 1.0.

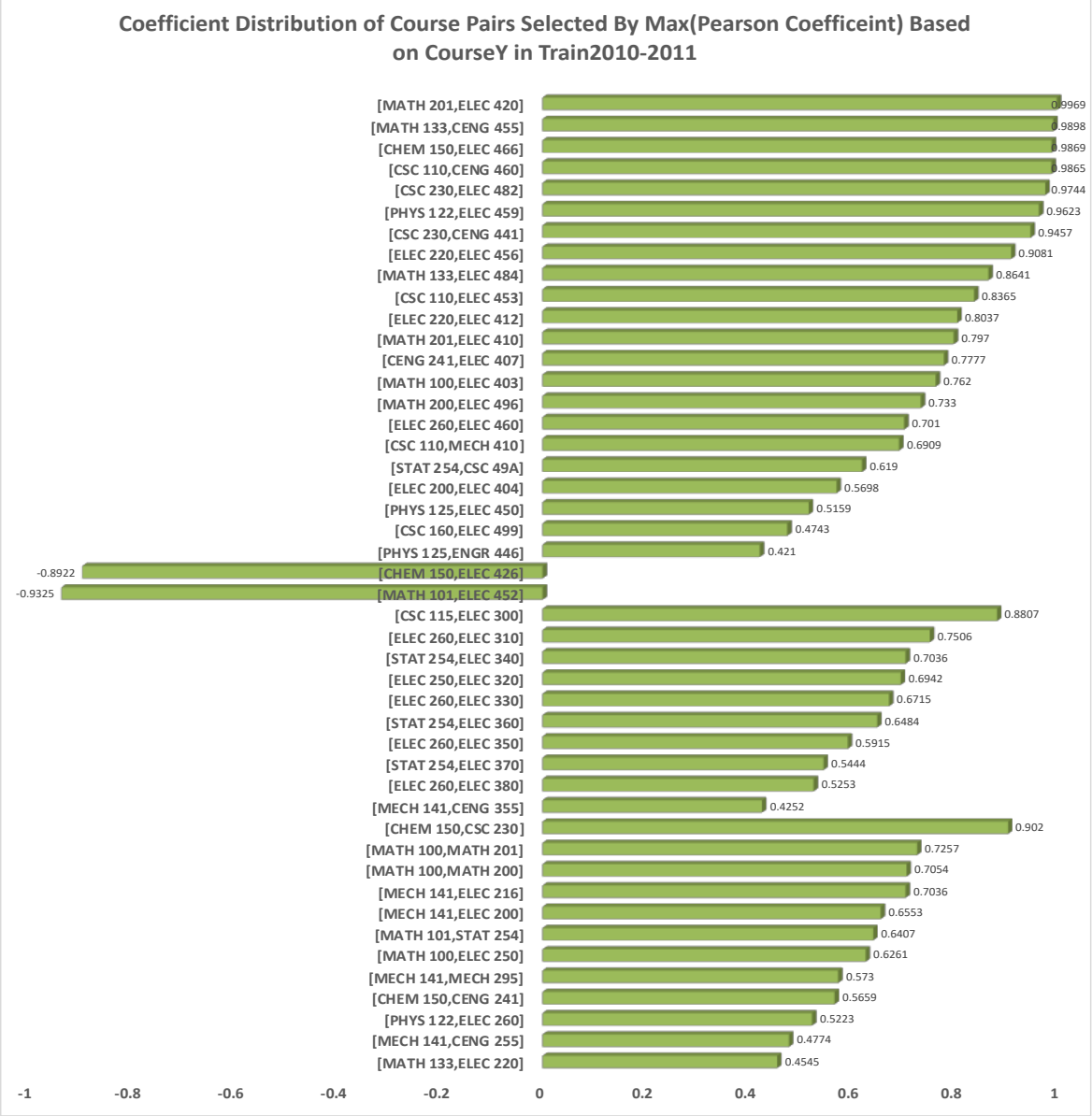

Figure 8 Pearson Correlation Distribution of Course Pairs Selected By Max(Pearson Coefficient) Based on CourseY in Train2010-2011

The course pairs were also selected using the criterion of maximum coefficient in the other three training sets of Train2010-2012, Train2010-2013 and Train2010-2014. Their

coefficient distributions for the selected course pairs have similar trend as the one in Train2010-2011 and they are shown in Appendix 8.

After comparing the course pairs in the four different training sets from the coefficient graphs, it can be concluded that the course pairs selected based on CourseX are prone to be paired with the fourth-year courses due to the small enrolment. Also, the course pairs with second- or third-year courses produce relatively smaller coefficients. Meanwhile, since the smaller enrolments produce bigger Pearson Coefficients close to 1.0 or -1.0, the prediction results may be skewed. New approaches will be investigated and applied in the following sections to balance this discrepancy.

Since the predictors were selected using the maximum Pearson Coefficient in all of the training sets, the next step is to use these selected predictors as the first attempt for the course grade predictions discussed in next section.

4.2.1.1 Testing Enrolment Distribution of Course Pairs Selected by Max(Pearson Coefficient)

As stated in Section 3.2, the research data excluding the training data was partitioned into subsets, each of which contains one-year course data and is one of the testing sets. The testing enrolments in each testing set for the course pairs selected by Max(Pearson Coefficient) based on CourseX from Train2010-2011 are shown in Figure 9. The horizontal axis shows the selected strongly correlated course pair while the vertical axis represents the testing enrolment.

The arrows in Figure 9 point to the course pairs discussed in the section. It can be seen from Figure 9 that the testing enrolments for the course pairs in each testing set for Train2010-2011 have a wide range from 0 to 22. There are only three course pairs in Test2012 with a relatively large testing enrolment, two course pairs in Test2013 and in Test2014, and none in Test2015.

Because of the small testing enrolment in the one-year testing data, the prediction precisions of these course pairs may not be reliable. Moreover, some course pairs do not have testing enrolments in certain testing sets. For instance, the course pair of CSC 110 and CENG 460 (indicated by the orange arrow in Figure 9) does not have testing enrolment

in Test2014 and Test2015, so there is no prediction. Therefore, the testing sets were merged to avoid the above scenarios and all of the testing sets with one-year course data were merged into one and denoted as Test2012-2015X as the new testing set. The testing enrolment distribution of the merged testing set, Test2012-2015X, for these course pairs are shown in Figure 10 and the course pairs discussed in this section are marked by its green testing enrolment bars.

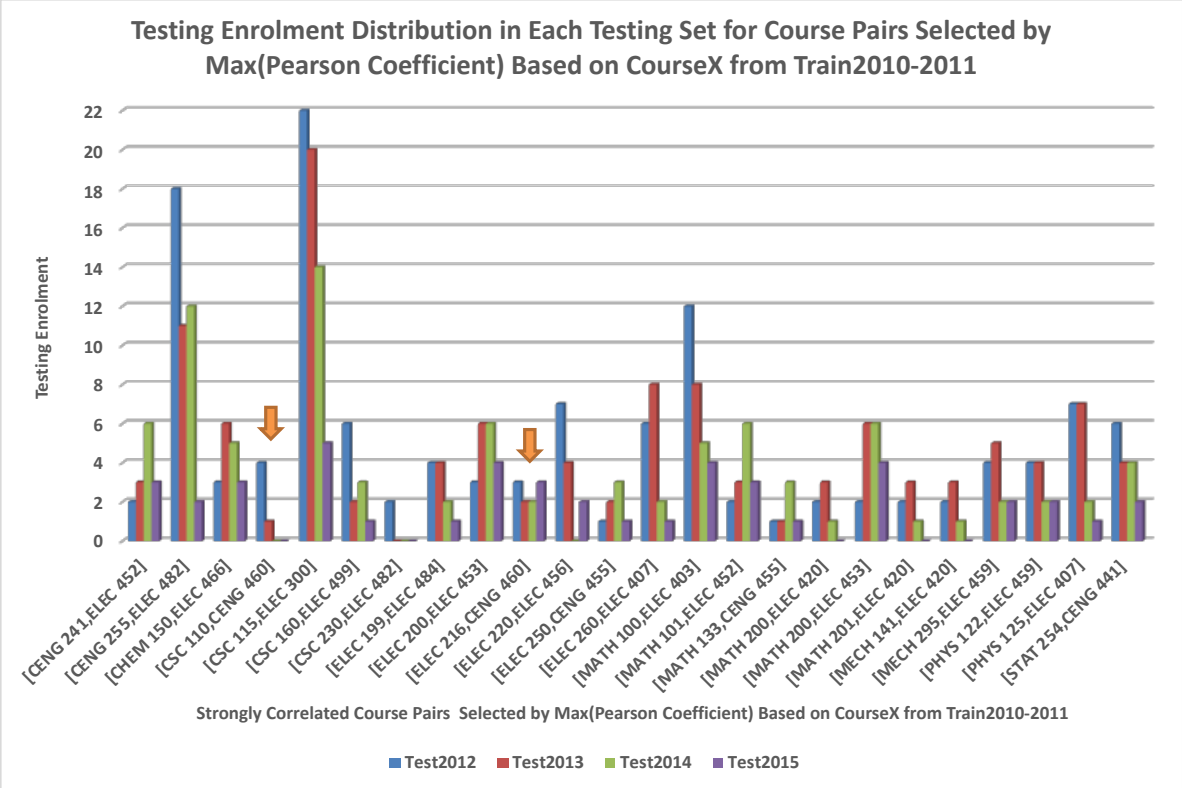

Figure 9 Testing Enrolment of Course Pairs Selected by Max(Pearson Coefficient) Based on CourseX in Each Testing Set for Train2010-2011

The enrolment for each course pair in Figure 9 and Figure 10 increases dramatically for most of the course pairs in the merged testing set of Test2012-2015X, although there still have course pairs with small testing enrolments. For example, prior to the testing set merge, the course pair of ELEC 216 and CENG 460 has 3, 2, 2, 3 testing enrolments in the four different testing sets of Test2012, Test2013, Test2014 and Test2015, respectively (highlighted by an orange arrow in Figure 9). Afterwards, its testing enrolment increased

to 10 after the merge (displaying in green bar in Figure 10) so that it can be used for prediction more reliably than using only 2 or 3 testing enrolments. Moreover, there is no course pairs having 0 testing enrolments in Figure 10 so that all course pairs could be used for prediction.

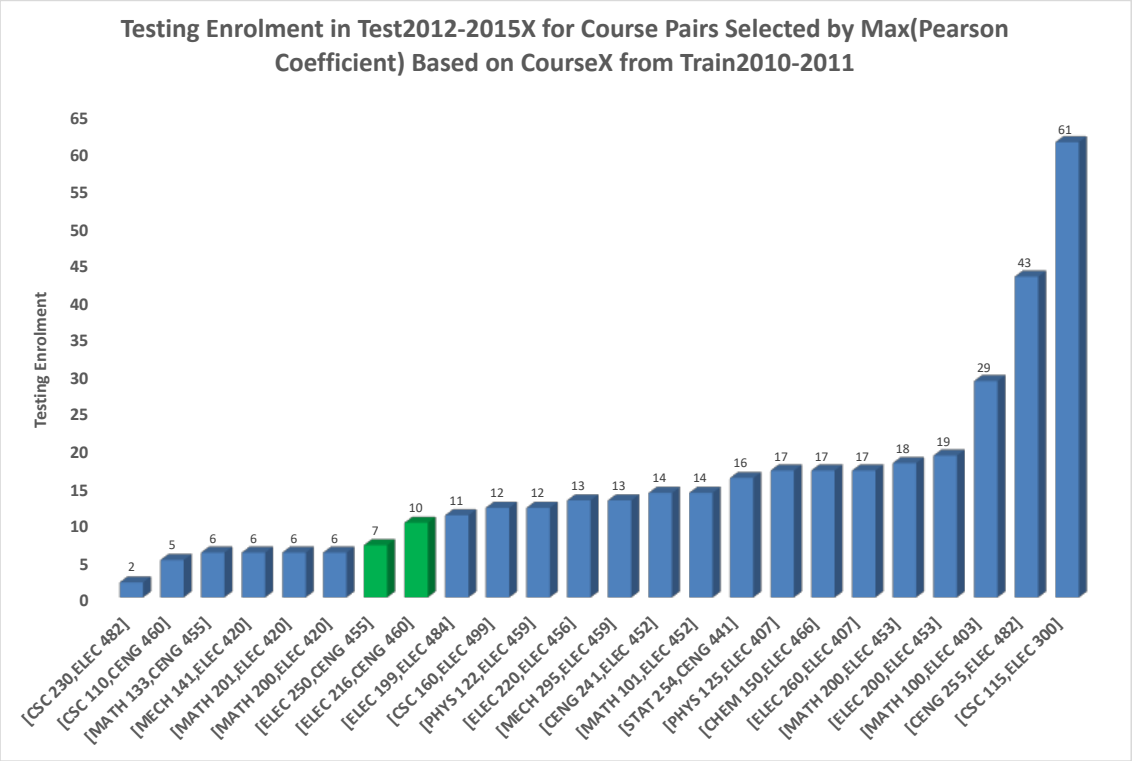

Figure 10 Testing Enrolment of course Pairs Selected by Max(Pearson Coefficient) Based on CourseX from Train2010-2011

The testing enrolment with course pairs whose CourseY is one of the courses in second- or third-year, for instance, the pair of ELEC 250 and ELEC 320 having 78 testing enrolments shown in Figure 12 (presented in green bar), the testing enrolments for the course pairs whose CourseY is one of the courses in fourth-year are relatively small. Due to the small testing enrolment, the prediction precisions of such course pairs may be less reliable even if the precisions are considerably high. For example, the course pair of ELEC 250 and CENG 455 in Test2012-2015X shown in green bar in Figure 10 only has 7 enrolments and therefore even with the relatively high precision of 71%, using this course pair for prediction may not be reliable.

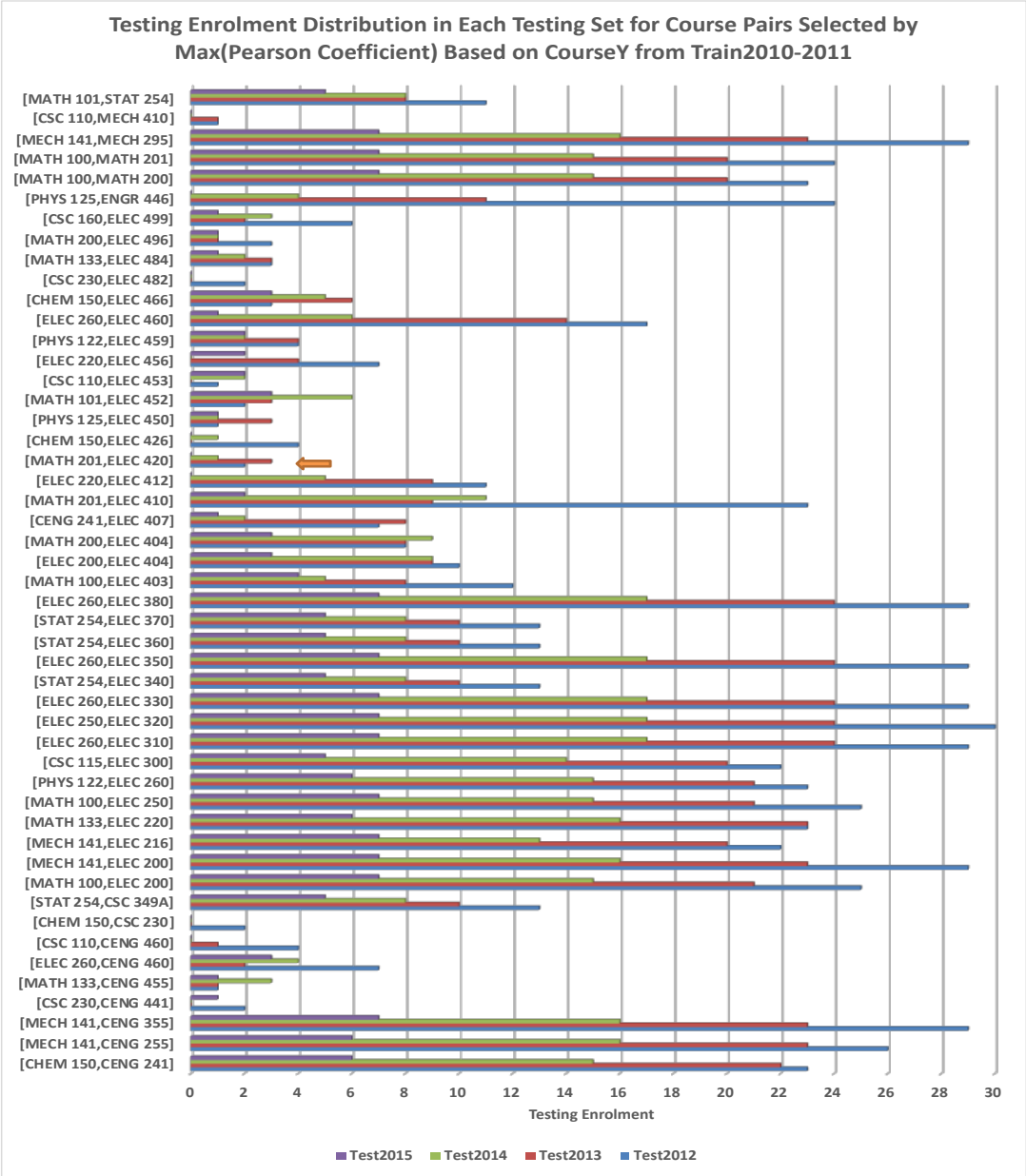

Figure 11 Testing Enrolment of Course Pairs Selected by Max(Pearson Coefficient) Based on CourseY in Each Testing Set for Train2010-2011

The testing enrolment distribution for course pairs selected by maximum of Pearson Coefficient based on CourseY is presented in Figure 11. It can be seen from Figure 11 that the testing enrolments have a wide range from 0 to 30, similar to the distribution in Figure 9. In addition, there are some course pairs that do not have testing enrolments in certain testing sets, for instance, the course pair of MATH 201 and ELEC 420 have no testing

enrolment in Test2015 (highlighted by orange arrow in Figure 11), and therefore this course pair has no basis to be used for prediction.

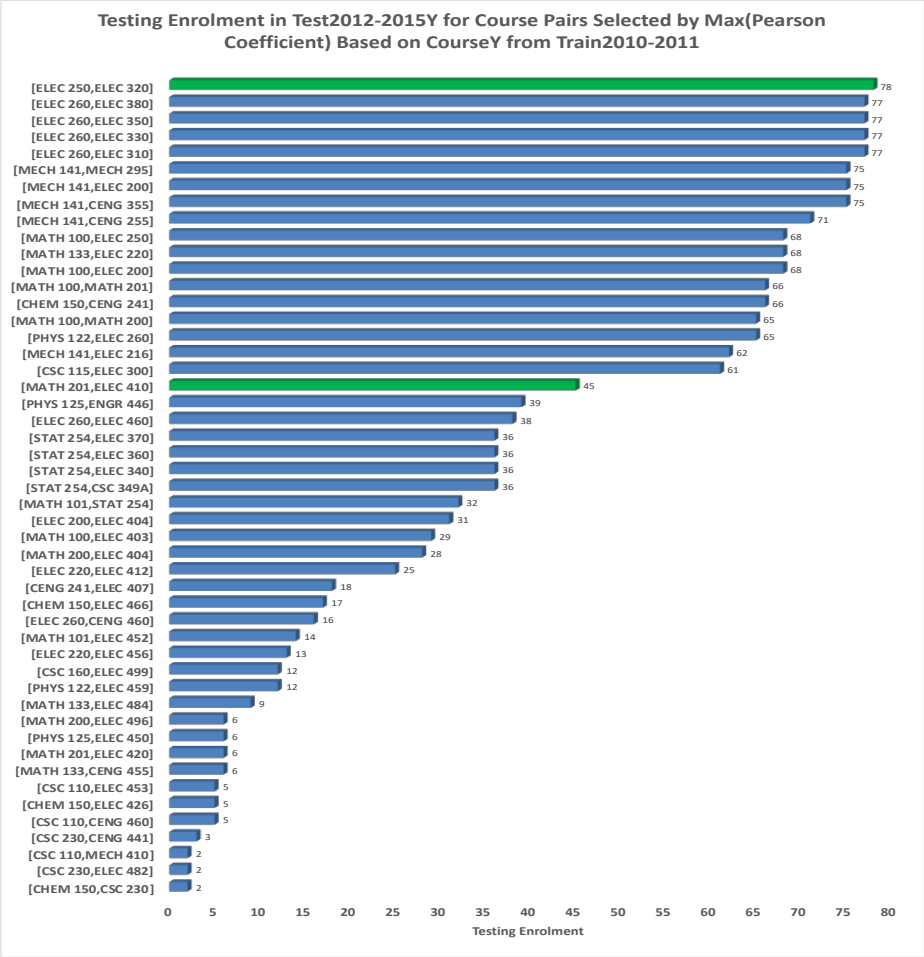

Figure 12 Testing Enrolment of course Pairs Selected by Max(Pearson Coefficient) Based on CourseY from Train2010-2011

Similarly, the course pairs with CourseY in second-year or third-year have more testing samples because the second-year and third-year CourseYs are fundamental courses while the course pairs with CourseY in fourth-year have fewer testing enrolments because of the specialization-related attribute.

In order to keep prediction results consistent with the ones selected based on CourseX and avoid the case of no testing enrolment for some course pairs in a testing set, the four different testing sets, Test2012, Test2013, Test2014 and Test2015, were also merged into

one and denoted as Test2012-2015Y. The testing enrolment distribution of the merged testing set for course pairs selected based on CourseY is shown in Figure 12.

The figure shows similar enrolment distribution trend as the course pairs selected based on CourseX shown in Figure 10 , namely that the testing enrolments increased after the merge, especially for most of the course pairs with second- or third-year courses. For instance, the testing enrolment of the pair of ELEC 250 and ELEC 320 increased from 30 in Test2012 to 78 in the merged testing set of Test2012-2015Y. The testing enrolments of the course pairs with fourth-year courses also increased, such as the testing enrolment of the course pair of MATH 201 and ELEC 410 which increased from 23 in Test2012 to 45 in Test2012-2015Y (shown in green bar in Figure 12).

For the two training sets of Train2010-2012 and Train2010-2013, the testing enrolments have similar distributions in the corresponding testing sets for the course pairs selected based on both CourseX and CourseY. Therefore, these testing sets were merged into one for each training set in the same way for Train2010-2011, namely, Test2013-2015X and Test2013-2015Y for Train2010-2012, and Test2014-2015X and Test2014-2015Y for Train2010-2013.

However, for the training set of Train2010-2014, there is just one testing set of Test2015 which contains only one-year course data. The testing enrolment distributions for the course pairs selected based on CourseX and CourseY are shown in Figure 13 and Figure 14. It can be seen that the testing enrolment for the course pairs selected based on CourseX has a maximum of 5 for the pair of CSC 115 and CENG 241 in Figure 13 (presented in a green bar) and the maximum of testing enrolments of the course pairs selected based on CourseY is 7 as shown in Figure 14. Compared with the testing enrolments shown in Figure 10 and Figure 12, the testing enrolments for the training set of Train2010-2014 shown in Figure 13 and Figure 14 are much smaller for most course pairs. Therefore, the training set Train2010-2014 was deemed invalid and not used in further analysis because of insufficient testing data.

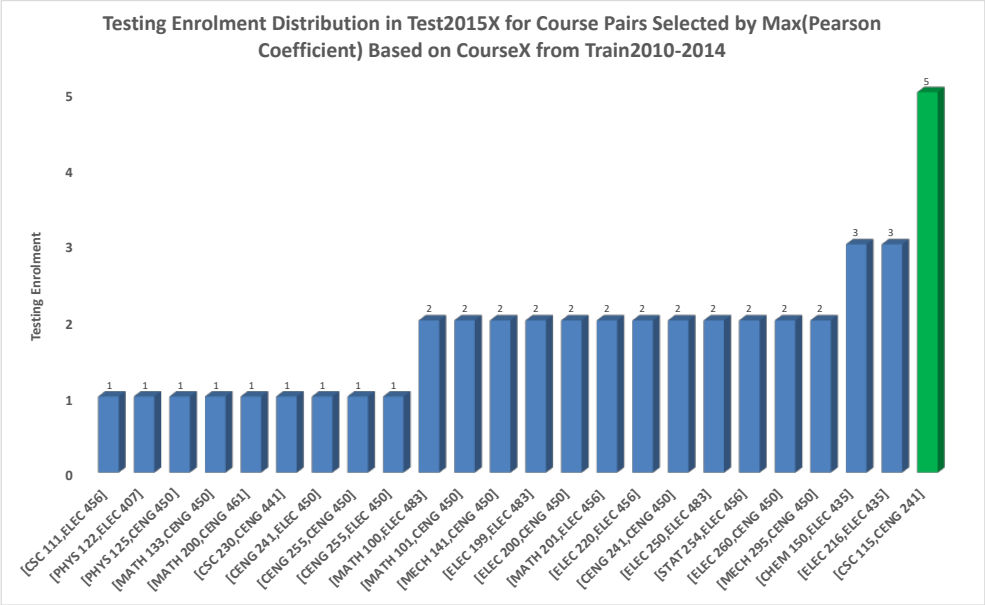

Figure 13 Testing Enrolment of course Pairs Selected by Max(Pearson Coefficient) Based on CourseX from Train2010-2014

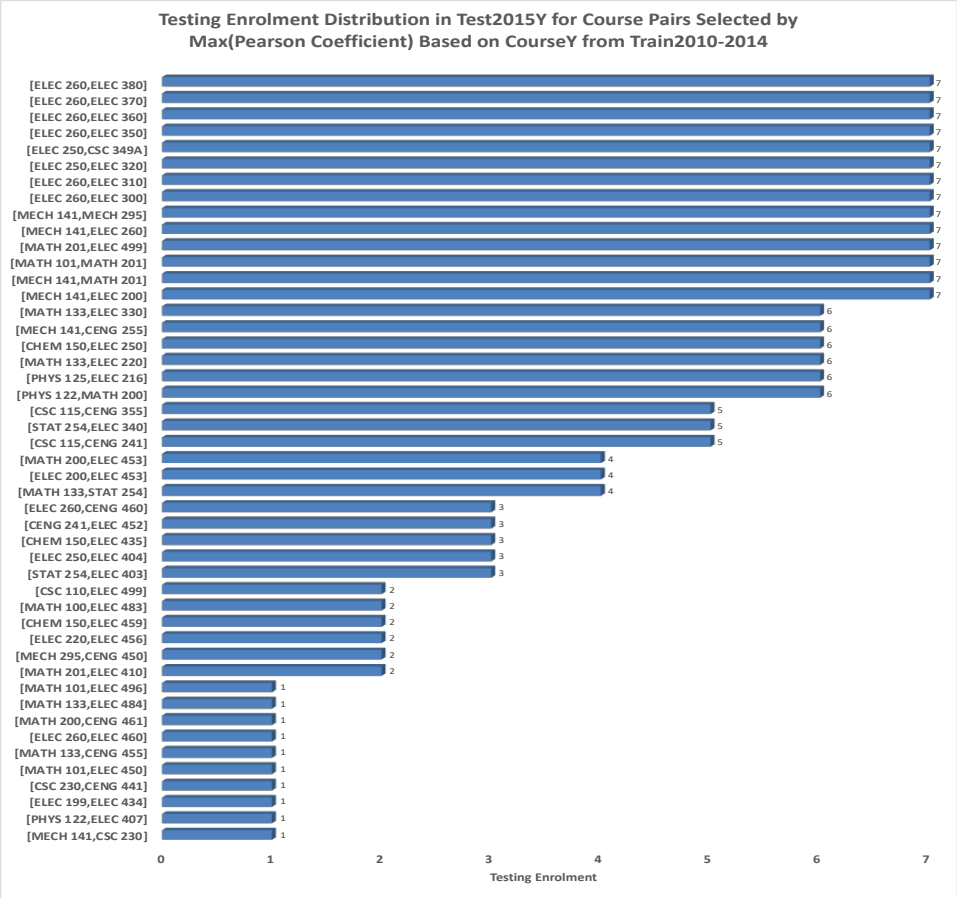

Figure 14 Testing Enrolment of course Pairs Selected by Max(Pearson Coefficient) Based on CourseY from Train2010-2014

The datasets have been pre-processed and can now be used to carry out further analysis, which includes computing the MAE and prediction precision as discussed in the next two sections.

4.2.1.2 MAE Analysis of Selection with Max(Pearson Coefficient)

As stated in Section 4.1, the MAE is used as the metric to assess the prediction results. The MAEs computed in Test2012-2015X for the course pairs selected by using maximum of Pearson Coefficient based on CourseX from Train2010-2011 is presented in Figure 15.

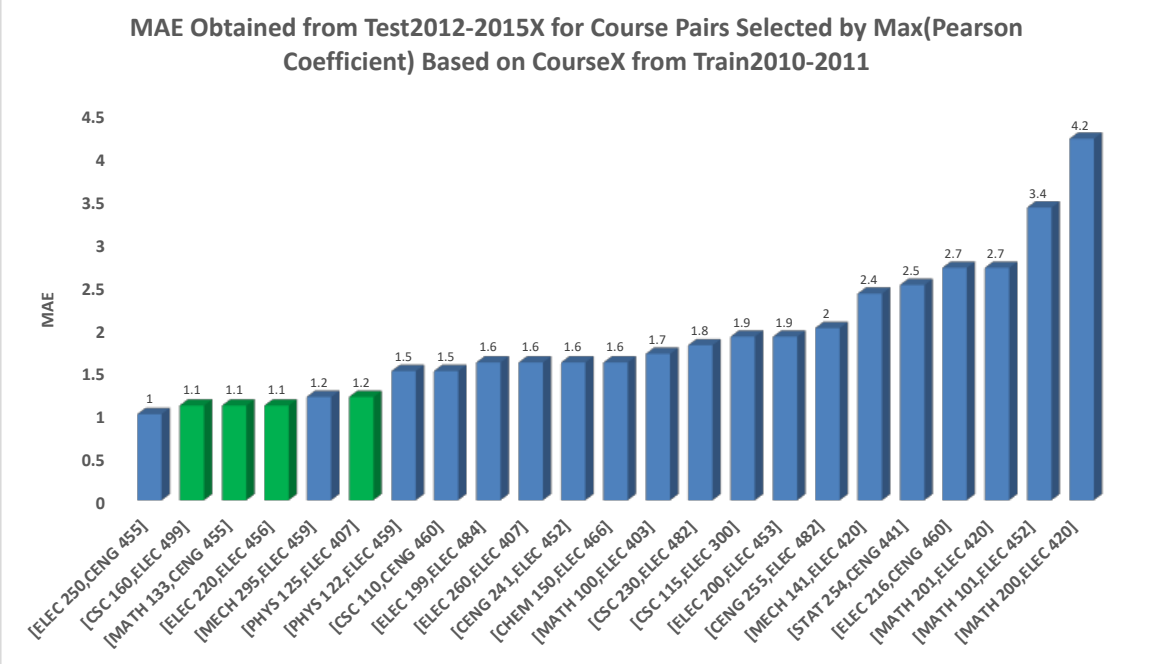

Figure 15 Prediction MAEs for Course Pairs Selected by Max(Pearson Coefficient) Based on CourseX from Train2010-2011 in Test2012-2015X

The MAEs were computed from the testing set of Test2012-2015X, resulting in a range of 1.0 to 4.2. According to the MAE acceptance criteria defined in Section 4.1, the MAEs within 1.0 are deemed acceptable since ±1.0 is a small error in our context. In other words, the predicted grade is just one grade away from the actual grade.

Table 7 Course Pairs Selected Based on CourseX with MAE<=1.2 in Test2012-2015X

CrsXCode	CrsXNum	CrsYCode	CrsYNum	Enrolment	0~1.0	%	MAE
ELEC	250	CENG	455	7	5	71	1
MATH	133	CENG	455	6	3	50	1.1
CSC	160	ELEC	499	12	4	33	1.1
ELEC	220	ELEC	456	13	7	54	1.1
PHYS	125	ELEC	407	17	12	71	1.2

- CrsXCode: the course code of CourseX
- CrsXNum: the course number of CourseX
- CrsYCode: the course code of CourseY
- CrsYNum: the course number of CourseY
- Enrolment: the number of testing enrolments in Test2012-2015X
- 0~1.0: the number of testing enrolments that have prediction error in the range of from 0 to 1.0
- %: the percentage of testing enrolments that have prediction errors in the range of from 0 to 1.0

However, according to its definition, the MAE just shows the average error margin and its results may be skewed if some errors are very big or some are very small, which may skew the average. Meanwhile, it cannot show how the predicted errors are distributed. The five course pairs selected from Train2010-2011 have MAE equal to 1.0 or very close to 1.0 as shown in Table 7 in Test2012-2015X. Although the MAE of 1.1 is out of the acceptable MAE range, it is just 0.1 away from 1.0. Therefore, it can be treated as acceptable.

On the other hand, the three course pairs with MAE equal to 1.1 in Table 7 have small number of enrolments with errors in the acceptable range (3 out of 6 for the course pair of MATH 133 and CENG 455, 4 out of 12 for course pair of CSC 160 and ELEC 499, and 7 out of 13 for course pair of ELEC 220 and ELEC 456). However, the course pair of PHYS 125 and ELEC 407 has 12 out of 17 enrolments with error within 1.0, but its MAE is 1.2 which is 20% higher than the threshold of acceptable range.

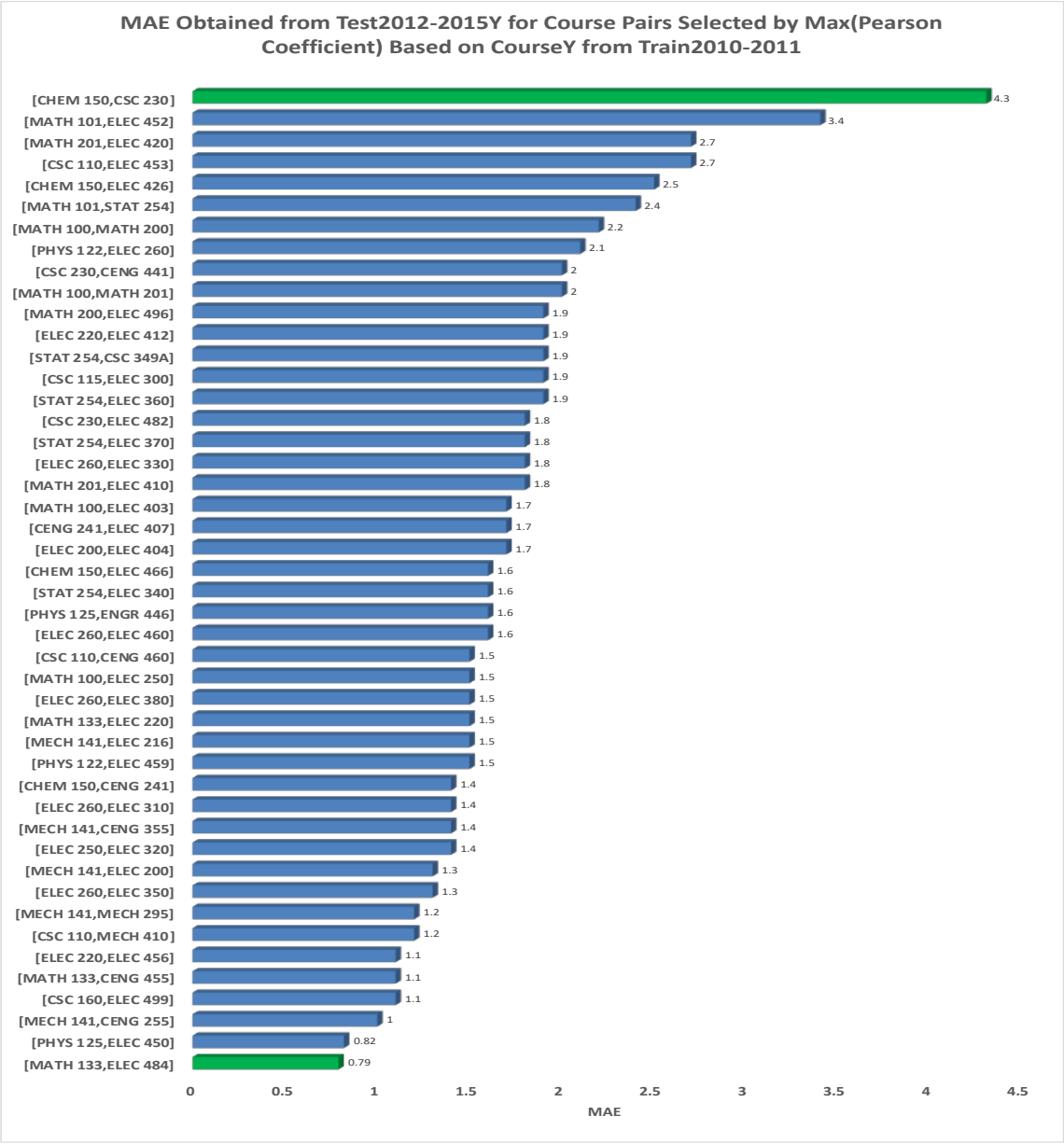

Figure 16 Prediction MAEs for Course Pairs Selected by Max(Pearson Coefficient) Based on CourseY from Train2010-2011 in Test2012-2015Y

Similarly, the course pairs selected by maximum of Pearson Coefficient based on CourseY from the training set of Train2010-2011 were predicted and their MAEs computed in the testing set of Test2012-2015Y are shown in the bar chart of Figure 16 with maximum MAE of 4.3 from the course pair of CHEM 150 and CSC 230 and minimum MAE of 0.79 from the course pair of MATH 133 and ELEC 484. Although there are several

course pairs with MAE within or a bit over 1.0, they have the same issue as mentioned above for the course pairs selected based on CourseX that some course pairs have small number of enrolments with errors within 1.0 shown in Table 8. It can be seen from Table 8 that there is only one course pair, MATH 133 and ELEC 484, that has 78% of tested enrolments with predicted error in the acceptable range and the other 5 pairs have less than 60% of tested enrolments with predicted error in the acceptable range. Therefore, MAE outliers can bias the prediction and a new approach is required for the course pairs selected based on CourseY as well.

Table 8 Course Pairs Selected Based on CourseY with MAE<=1.1 in Test2012-2015Y

CrsXCode	CrsXNum	CrsYCode	CrsYNum	Enrolment	0~1.0	%	MAE
MATH	133	ELEC	484	9	7	78	0.79
PHYS	125	ELEC	450	6	3	50	0.82
MECH	141	CENG	255	71	42	59	1
MATH	133	CENG	455	6	3	50	1.1
CSC	160	ELEC	499	12	4	33	1.1
ELEC	220	ELEC	456	13	7	54	1.1

The MAEs for course pairs selected by the maximum of Pearson Coefficient based on both CourseX and CourseY for the remaining training sets of Train2010-2012 and Train2010-2013 are shown in Appendix 9. These 4 figures demonstrate similar trends as the figures for Train2010-2011 that the course pairs with acceptable MAE have small number of predicted errors in the acceptable range (See Table 7 and Table 8).

Although the predictions of course pairs with MAE less than or equal to 1.0 are more acceptable based on the MAE acceptance criteria compared with the ones with MAE greater than 1.0 shown in Figure 15 and Figure 16, some of such course pairs, for instance, the pair of CSC 160 and ELEC 499 shown in Table 7 only has half and the pair of PHYS 125 and ELEC 450 shown in Table 8 has only one-third of predicted errors less than or equal to 1.0. The MAE only presents the average error, but it does not show how the errors are distributed. Then, the results may be skewed due to outliers if some errors are either too big or too small. Therefore, another measure, prediction precision, is introduced to depict how the errors distribute and discussed in next section.

4.2.1.3 Prediction Precision Analysis with Max(Pearson Coefficient)

The course pairs for both CourseXs and CourseYs are picked from the computations of the training sets based on their Pearson Coefficients' strength. As stated in Section 4.2.1, the predictor-selection approach is to select the course having maximum Pearson Coefficient with the candidate predicted course as predictor, or predicting course by using Pearson Coefficient for CourseXs and CourseYs.

The Pearson Correlation Coefficient measures the linear relationship between two variables, which forms a best line for the points of two variables. In other words, the distance of the points of the two variables to that line is the shortest. Meanwhile, the linear regression also models the linear relationship between two variables. Therefore, the linear regression is applied as the model to start the prediction. The prediction error of one course grade, e, the difference between the actual course grade and the predicted course grade, is recognized as the indicator to determine how well the linear regression model performs. An accurate prediction is defined as a prediction with error less than or equal to 1.0. In other words, when the difference between prediction grade and the actual grade of one course is less than or equal to 1.0, the prediction is accepted and recognized as an accurate prediction for the predicted course. The prediction precision of one course is defined as the percentage of accurate predictions.

The prediction precisions obtained from the testing set of Test2012-2015X for the course pairs selected by using maximum of Pearson Coefficient based on CourseX from Train2010-2011 are shown in Figure 17. It can be seen from Figure 17 that there are only two course pairs with precision over 70%, 71% from the course pair of PHYS 125 and ELEC 407 with 17 testing enrolments, and the course pair of ELEC 250 and CENG 455 with 7 testing enrolments. The remaining precisions vary from 0% to 60% with testing enrolment ranging from 2 to 61.

The two course pairs with precisions over 70%, one from the course pair of ELEC 250 and CENG 455 just has 7 testing enrolments and the other one is from the course pair of PHYS 125 and ELEC 407 which has 17 testing enrolments (See Figure 17). Compared with the quantity of the testing enrolments of the two course pairs, the 71% precision of the latter course pair is relatively more reliable than the one of the former course pair

because the more testing enrolments would give narrower margin of errors with high confidence than the small testing enrolments.

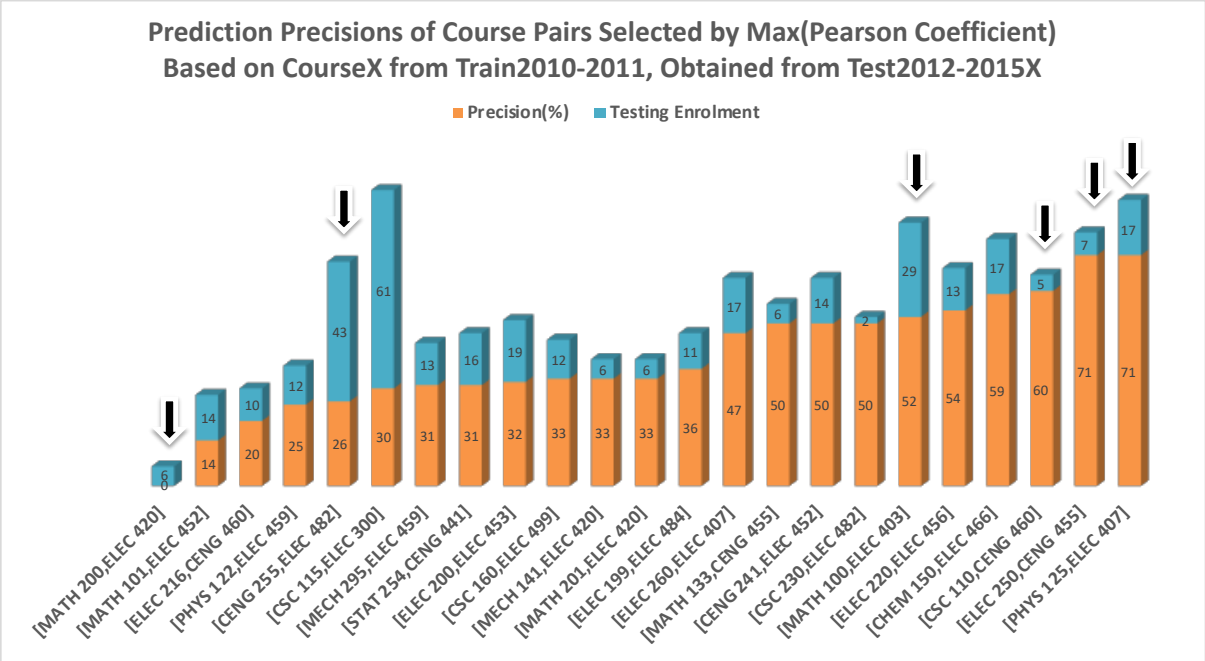

Figure 17 Prediction Precisions of Course Pairs in Test2012-2015X, Trained by Train2010-2011

Relatively speaking, a big testing enrolment can produce both high and low precision, for instance, the pair of MATH 100 and ELEC 403 having precision of 52% with 29 testing enrolments and the pair of CENG 255 and ELEC 482 having precision of 26% with 43 testing enrolments. The small testing enrolment can generate both high and low precision as well: the pair of CSC 110 and CENG 460 having precision of 60% with 5 testing enrolments and the pair of MATH 200 and ELEC 420 having precision of 0% with 6 testing enrolments.

The prediction precisions shown in Figure 17 indicate that the course of PHYS 125 is the only course that produces more acceptable result when trying to select strongly correlated courses based on CourseX by using the Pearson Correlation Coefficient only.

The strongly correlated course pairs were also selected based on CourseY. The prediction precisions in Test2012-2015Y for these course pairs from Train2010-2011 are plotted in Figure 18. Obviously, there are more course pairs as the CourseY can be one

course in second-, third- or fourth-year course while the CourseX is one course in first- or second-year declared in Chapter 2.

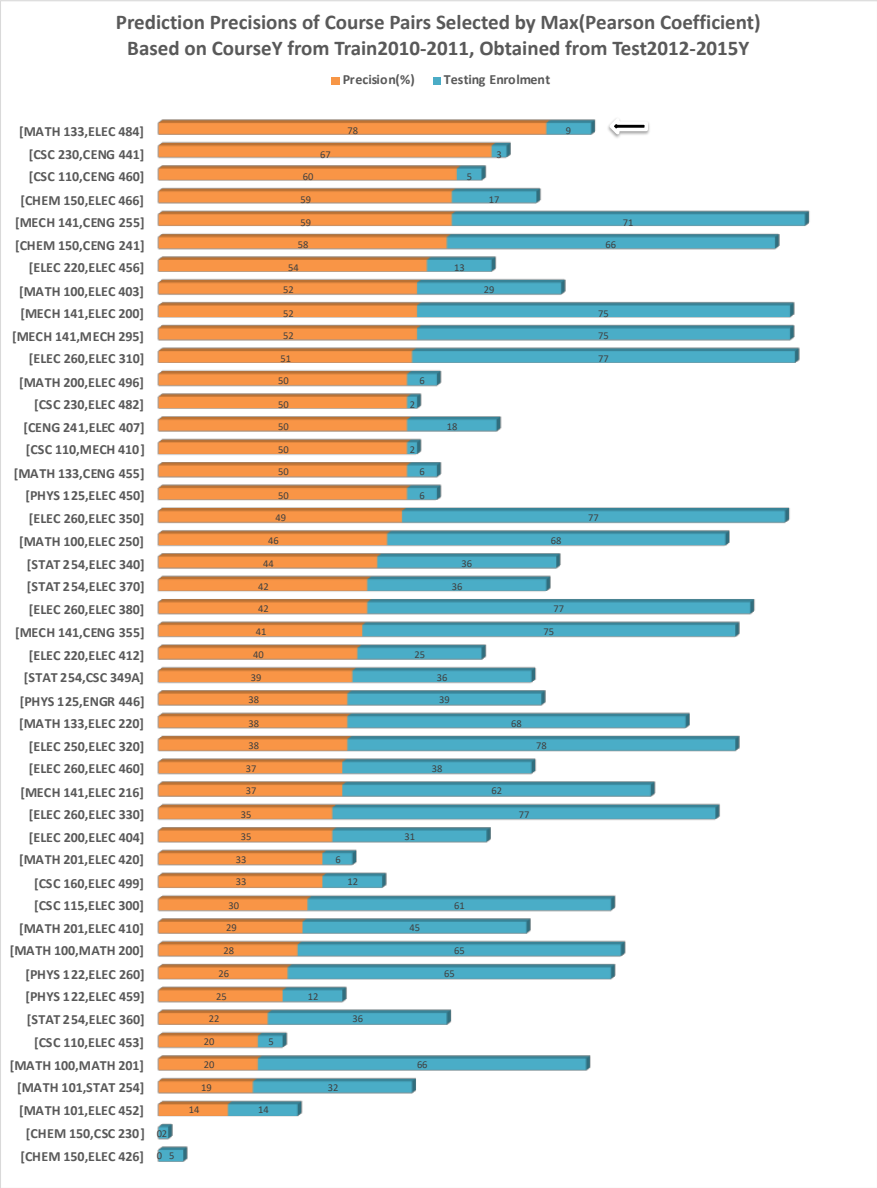

Figure 18 Prediction Precisions of Course Pairs in Test2012-2015Y, Trained by Train2010-2011

In Figure 18, there is only one course pair with precision of over 70%, the pair of MATH 133 and ELEC 484 having precision of 78% with 9 testing enrolments while the rest have a wide range of precisions from 0 to 67% with testing enrolment from 2 to 77.

The precisions and testing enrolments distributions shown in Figure 18 show similar trends as the ones produced by course pairs selected based on CourseX shown in Figure 17 that big testing enrolment can produce either high or low precision and small testing enrolment can also produce high and low precisions.

As stated in Section 3.2, the training size of the training dataset just contains two-year course grade data. The prediction precision may have different observations if the training dataset increases. Therefore, the training dataset was increased three times with one-year course grade data increment at each time. In other words, three different training datasets, namely, Train2010-2012, Train2010-2013 and Train2010-2014 were partitioned. However, the training set of Train2010-2014 was deemed as invalid because there are insufficient testing enrolments for the course pairs selected in this training set discussed and confirmed in Section 4.2.1.1.

The prediction precisions obtained from the four testing sets of Test2013-2015X, Test2013-2015Y, Test2014-2015X and Test2014-2015Y for the selected strongly correlated course pairs are shown in Appendix 10. These 4 bar charts present similar trends as the ones for Train2010-2011 that most of the course pairs have prediction precisions lower than 70% and a small part of the course pairs have precisions equal to or greater than 70% in general.

For example, it can be seen from Figure 63 that 21 out of 24 course pairs have precisions under 70% with maximum of testing enrolment of 39 from the course pair of CSC 115 and CENG 241, and only 3 of 24 of them have precisions greater than 70% with maximum of testing enrolment of 5 from the course pair of MATH 100 and ELEC 483. Meanwhile, comparing the distributions of the precision and testing enrolment in the four figures listed in Appendix 10 with the ones in Figure 17 and Figure 18 shows that relatively bigger enrolment can generate either high or low precision while comparatively smaller enrolment can also produce high and low precision.

There are some course pairs having prediction precisions of 100% with small testing enrolments as shown in Appendix 10. For examples, (CSC 110 and CENG 460, and CSC 230 and CENG 441, both with 1 testing enrolment in Figure 63; CSC 230 and CENG 441 within 1 testing enrolment in Figure 63, Figure 65 and Figure 66; MATH 101 and ELEC

450 with 5 testing enrolments in Figure 64 and 2 testing enrolment in Figure 66; MATH 201 and ELEC 481 with 1 testing enrolment in Figure 66; ELEC 260 and CENG 461 with 2 testing enrolments in Figure 66). However, such course pairs contain fourth-year CourseYs and the corresponding testing enrolments are very small. Therefore, special attention is required when using such high precisions with small testing enrolments. The class size of such courses may have an impact on the prediction.

It can be seen from Figure 17, Figure 63 and Figure 65 that that the courses in the course pairs selected by the maximum of Pearson Coefficient based on CourseX are usually the fourth-year courses. Because the fourth-year courses are more specialization-related and have few enrolments in the dataset, the prediction precisions of such course pairs could be less trustful due to the small training and testing enrolments.

Similarly, the course pairs with fourth-year course selected using this approach based on CourseY have the same issue that the testing enrolments are small (See Figure 18, Figure 64 and Figure 66). Therefore, the approach of selecting predictor by only using the maximum Pearson Coefficient is not reliable for the course pairs with small enrolments, that is, using a statistically small sample only.

On the other hand, the course pairs with second- or third-year courses selected based on CourseY have enough training and testing enrolments in the datasets, but the prediction precisions are small, under 70% for most of the course pairs. Therefore, one technical course's performance cannot be simply assessed by another technical course's performance just based on their correlations.

In other words, the predictor selection only based on the Pearson Correlation produce poor prediction results for most of the course pairs regardless of whether they were selected based on CourseX or CourseY. As stated in Chapter 3 there are two major factors, coefficient and enrolment, which have impact on the correlation of two courses. The training enrolment was employed to select the predictor and assess the predictions in next section.

4.2.2 Predictor Selection by Enrolment

As stated in Chapter 3 the enrolment of one course pair is one of the two major factors, which have an impact on the correlation of the course pair. This section talks about how to select predictors by using enrolment.

Every technical course A can pair with n other technical courses. Specifically, a predictor course of A can be selected, among the n courses, as the one that has the most common enrolment with A. If there are multiple courses having the same enrolment, the course with the smallest first digit of its course number is chosen as the predictor as the smaller course number indicates that course was typically taken by a student in an earlier term in the academic schedules. If the predictor candidates have same first digit of course numbers, all of them are selected as predictors and used for prediction.

Similarly to the approach of selecting predictors using the maximum of Pearson Coefficient in Section 4.2.1, this approach also was utilized to select predictors based on enrolments for both CourseX and CourseY. The selected course pairs for both CourseX and CourseY in Train2010-2011 are shown in Figure 19 and Figure 20, respectively.

It can be seen from Figure 19 that the CourseXs from the course are mostly paired with second- or third-year courses except the pair of CSC 230 and CENG 441 (shown in green). As stated in Chapter 3 the course pairs with fourth-year courses have relatively small enrolments due to the specialization-related attribute, compared to the ones from the course pairs with second- or third-year courses. Therefore, the course pairs with fourth-year courses are less likely to be selected as the predictor when using enrolment as the criterion. Also, most of the enrolments of the selected course pairs shown in Figure 19 are relatively big because the second- and third-year courses are fundamental courses which are required to be taken by the student.

The course pairs in Figure 19 show that most of the CourseX from first-year (10 out of 11) are paired with second-year courses and most of CourseX from second-year (11 out of 12) are paired with third-year courses. Therefore, it seems that when trying to estimate the course performance, it could be helpful to use the early year courses to estimate the next year courses performance.

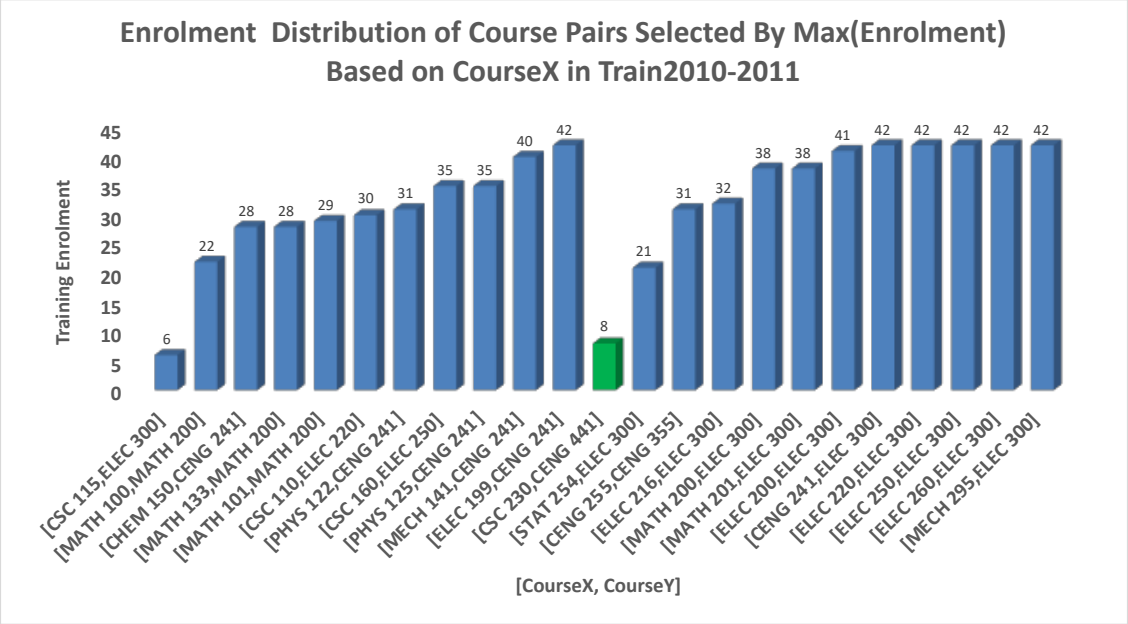

Figure 19 Enrolment Distribution of Course Pairs Selected by Max(Enrolment) Based on CourseX in Train2010-2011

The training enrolment distribution for the course pairs selected based on CourseY in Train2010-2011 is shown in Figure 20. The enrolments of course pairs with CourseY from fourth year have a wide range from 4 from the pair of PHYS 122 and ELEC 466 to 35 from the pair of PHYS 125 and ENGR 446.

The enrolments of course pairs with CourseY from second year also have a wide range from 11 from the pair of MECH 141 and CSC 230 to 42 from the two pairs of ELEC 199 and CENG 241, and ELEC 199 and ELEC 260. But the enrolment variation of the course pairs with CourseY from second year is relatively smaller than the one from the course pairs with CourseY from fourth year because the second-year courses are compulsory courses. All of the enrolments of course pairs with CourseY from third year are 42 because they are compulsory courses that students must take and the research is considering one single year of a student in the Electrical Engineering program.

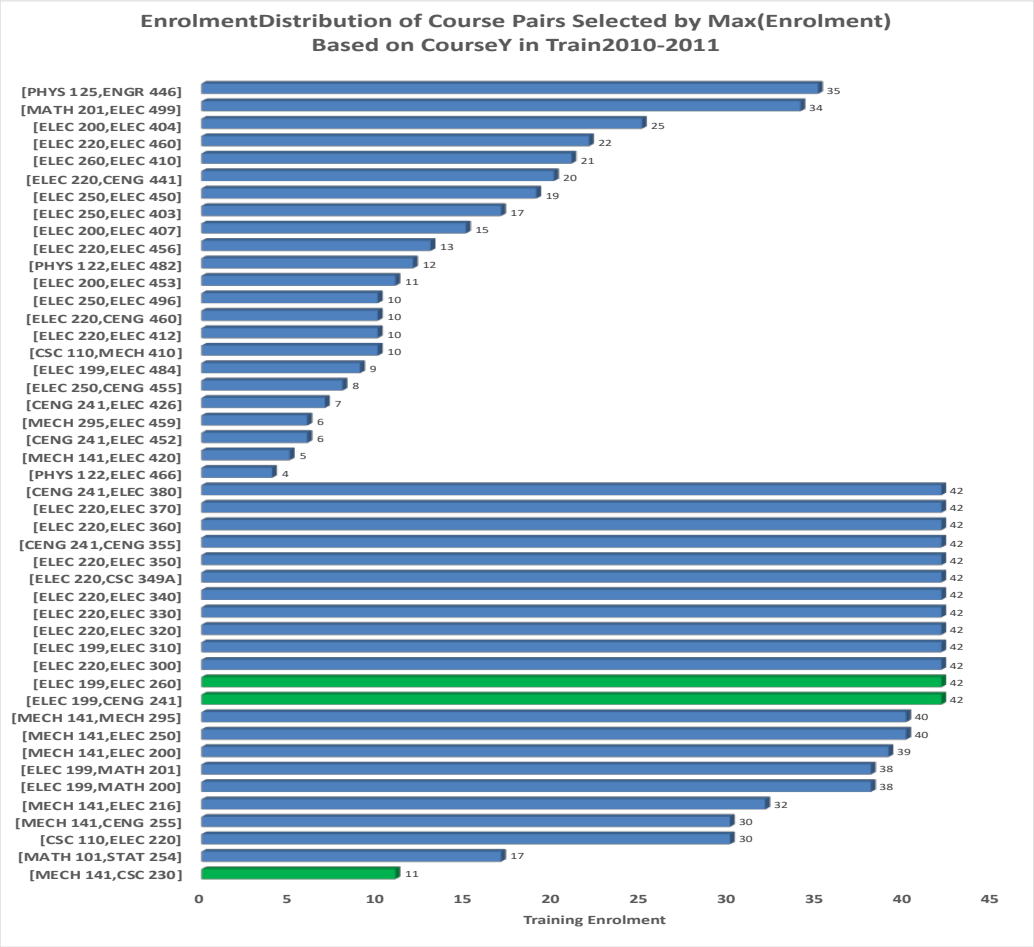

Figure 20 Enrolment Distribution of Course Pairs Selected by Max(Enrolment) Based on CourseY in Train2010-2011

The course pairs in the training sets of Train2010-2012, Train2010-2013 and Train2010-2014 were also selected by maximum of training enrolment based on both CourseX and CourseY shown in Appendix 11.

The prediction course pairs selected based on CourseX are listed in Figure 67, Figure 69 and Figure 71 from the three training sets, respectively. Similarly to the course pairs shown in Figure 19, the course pairs shown in the three figures present that the first-year course is paired with second-year course and second-year course is paired with third-year course. Also, the training enrolments of most of the course pairs are comparatively big as CourseYs in these course pairs are from second- or third-year compulsory courses.

The course pairs selected based on CourseY in the three training sets are shown in Figure 68, Figure 70 and Figure 72. The figures show similar trends as the one shown in Figure 20 that the course pairs with CourseY from fourth year have a wide range of training enrolment. For example, the training enrolment of the course pairs with CourseY from fourth year in Train2010-2012 is from 5 from the pair of CENG 241 and ELEC 435 to 72 from the pair of ELEC 250 and ENGR 466 in Figure 68. Moreover, the training enrolment distribution of the course pairs with CourseY from second or third year behave the same as the ones in Figure 20 in that they are the same as compulsory courses from the students in the same program.

In this section the prediction course pairs have been selected using maximum enrolment from the four training sets, the following sections will discuss the prediction results produced by these prediction course pairs.

4.2.2.1 Testing Enrolment Distribution of Course Pairs Selected by Max(Enrolment)

As the predictors selected by maximum of enrolment are different from the ones selected by maximum of correlation coefficient discussed in Section 4.2.1, the testing enrolments for the strongly correlated course pairs selected are thus different. As shown in Figure 1, the testing enrolment distribution using Max(Enrolment) based on CourseX is different than the distribution using Max(Pearson Coefficient).

It can be seen from Figure 21 that the testing set of Test2012 has the most testing enrolments for each course pair with the maximum of 30 and it has relatively big testing enrolments for most course pairs except the pair of CSC 160 and ELEC 250 with 6 testing enrolments and the pair of CSC 230 and CENG 441 with 2 testing enrolments (indicated by the orange arrows in Figure 21). Therefore, the prediction results of these two course pairs obtained in Test2012 may not be reliable due to their small testing enrolments in the testing set.

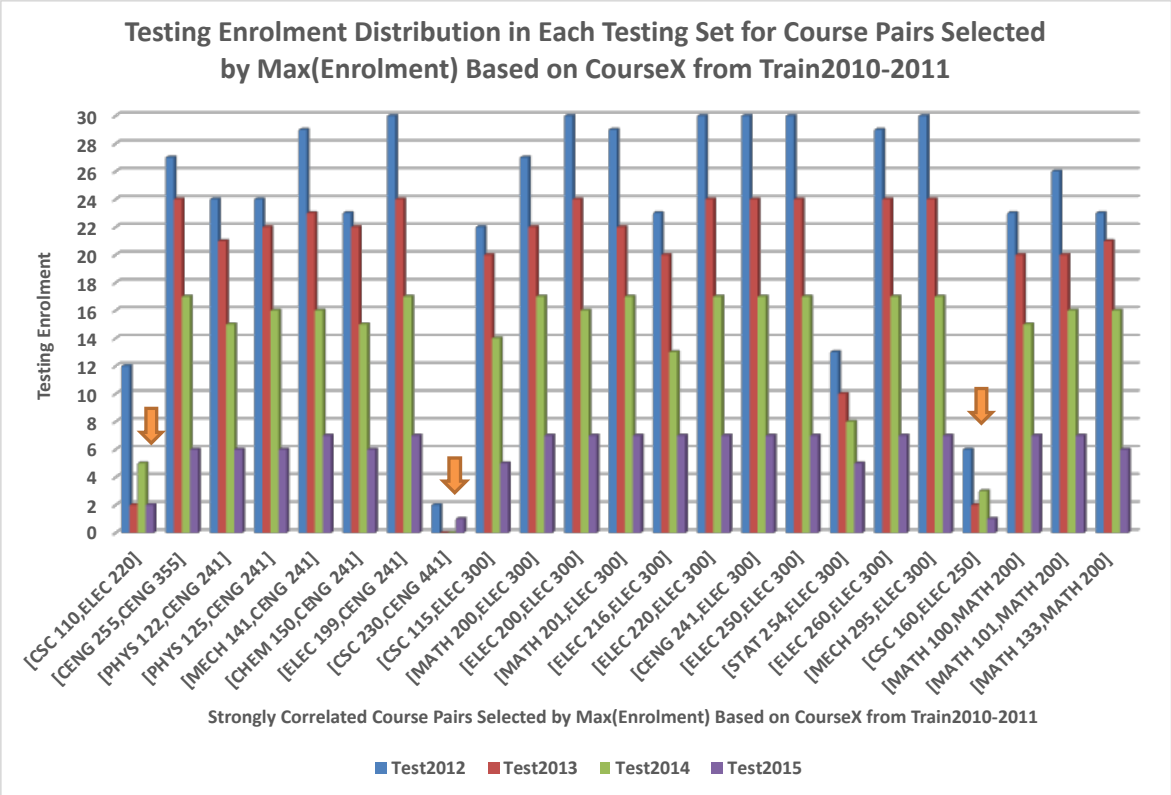

Figure 21 Testing Enrolment of Course Pairs Selected by Max(Enrolment) Based on CourseX in Each Testing Set for Train2010-2011

The testing enrolments in Test2013 and Test2014 have similar trend that their testing enrolments are relatively big compared to the ones in Test2015 for most course pairs with the maximum of 24 in Test2013 and maximum of 17 in Test2014, respectively. Moreover, there is no testing enrolments in these two testing sets for the pair of CSC 230 and CENG 441. Thus, there is no prediction for this course pair. Some of the course pairs in the testing sets also have the same issue as the one in Test2012 that the prediction results of the course pairs with relatively small testing enrolments may be not reliable. For instance, the pair of CSC 110 and ELEC 220 with 2 and 5 testing enrolments in Test2013 and Test2014, respectively.

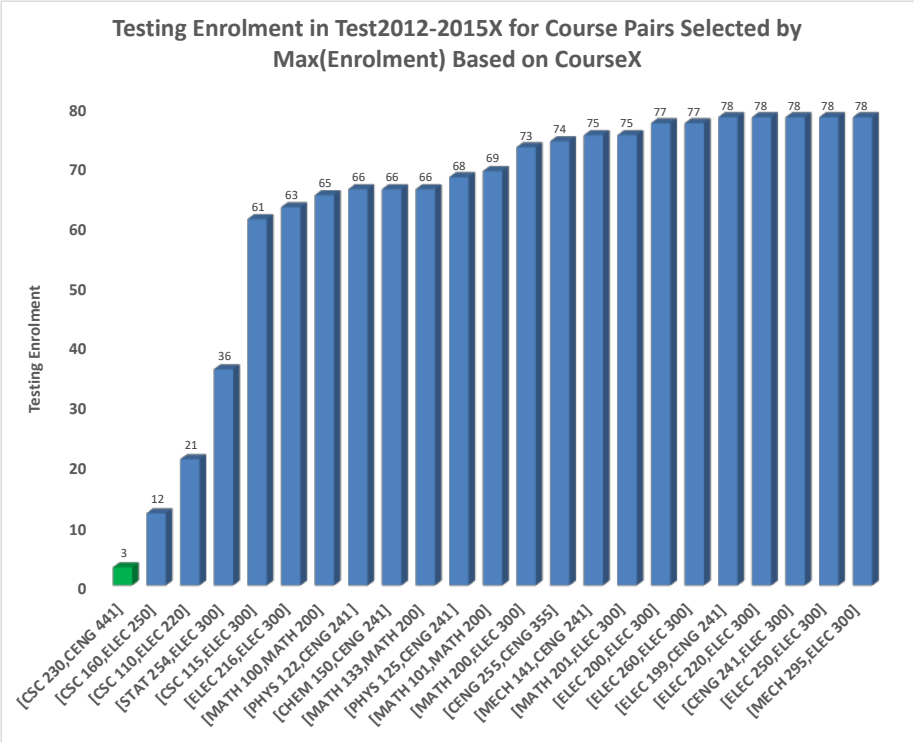

Figure 22 Testing Enrolment Distribution in Test2012-2015X for course Pairs Selected by Max(Enrolment) Based on CourseX from Train2010-2011

The testing set of Test2015 has the least testing enrolments among these four testing sets for most course pairs. The maximum of testing enrolments in this testing set is 7 and the minimum is 1. Compared with the testing enrolments in the other three testing sets, the prediction results obtained in Test2015 are not reliable because of insufficient testing enrolment.

Therefore, in order to prevent the above issues including small testing enrolments for some course pairs (i.e.: the pair of CSC 230 and CENG 441 with 2 testing enrolments in Test2012 and 1 in Test2015) and no prediction (the pair of CSC 230 and CENG 441 with 0 in Test2013 and Test2014), the four testing sets were merged and denoted as Test2012-2015X for prediction. The testing enrolment distribution in the merged testing set of Test2012-2015X is shown in Figure 22.

It can be seen from Figure 22 that most course pairs have considerable testing enrolments in the merged testing set of Test2012-2015X with maximum of 78 and the

predictions obtained from this testing set for those course pairs are more reliable compared to the ones obtained in each of the four unmerged testing sets. However, there is still one course pair, CSC 230 and CENG 441 (represented by the green bar in Figure 22), which has just 3 testing enrolments. Then, the prediction of this course pair is more doubtful due to the small testing enrolments.

The testing enrolments for the course pairs selected by maximum of enrolment based on CourseY in the four testing sets of Test2012, Test2013, Test2014 and Test2015 shown in Figure 23 have similar distributions as the ones for the course pairs selected based on CourseX in the four testing sets that some course pairs have relatively small testing enrolments, especially the ones in Test2015 and some course pairs do not have testing enrolments, for instance, the course pair of CSC 110 and MECH 410 in Test2014 and Test2015.

In other words, some course pairs have relatively smaller testing enrolments, which may produce skewed prediction results and some course pairs do not have predictions due to the testing enrolment absence in some testing set. Therefore, the four testing sets also were merged as one Test2012-2015Y for these strongly correlated course pairs selected based on CourseY. The testing enrolment distribution of the merged testing set is shown in Figure 24.

Because the second- or third-year courses are fundamental courses while the fourth-year courses are specialization-related, it can be seen from Figure 24 that most of the course pairs with CourseY from second or third year have relatively big testing enrolments while most of course pairs with CourseY from fourth year have relatively small testing enrolments. Although the testing sets were merged, there are still some course pairs having relatively smaller testing enrolments, such as the pair of CSC 110 and MECH 410 with 2 testing enrolments (the top green bar in Figure 24). Therefore, the prediction results of such course pairs are suspicious. On the other hand, the predictions of the course pairs with relatively big testing enrolments, for instance, the pair of ELEC 220 and CSC 349A with 78 testing enrolments (the top green bars in Figure 24), are more reliable.

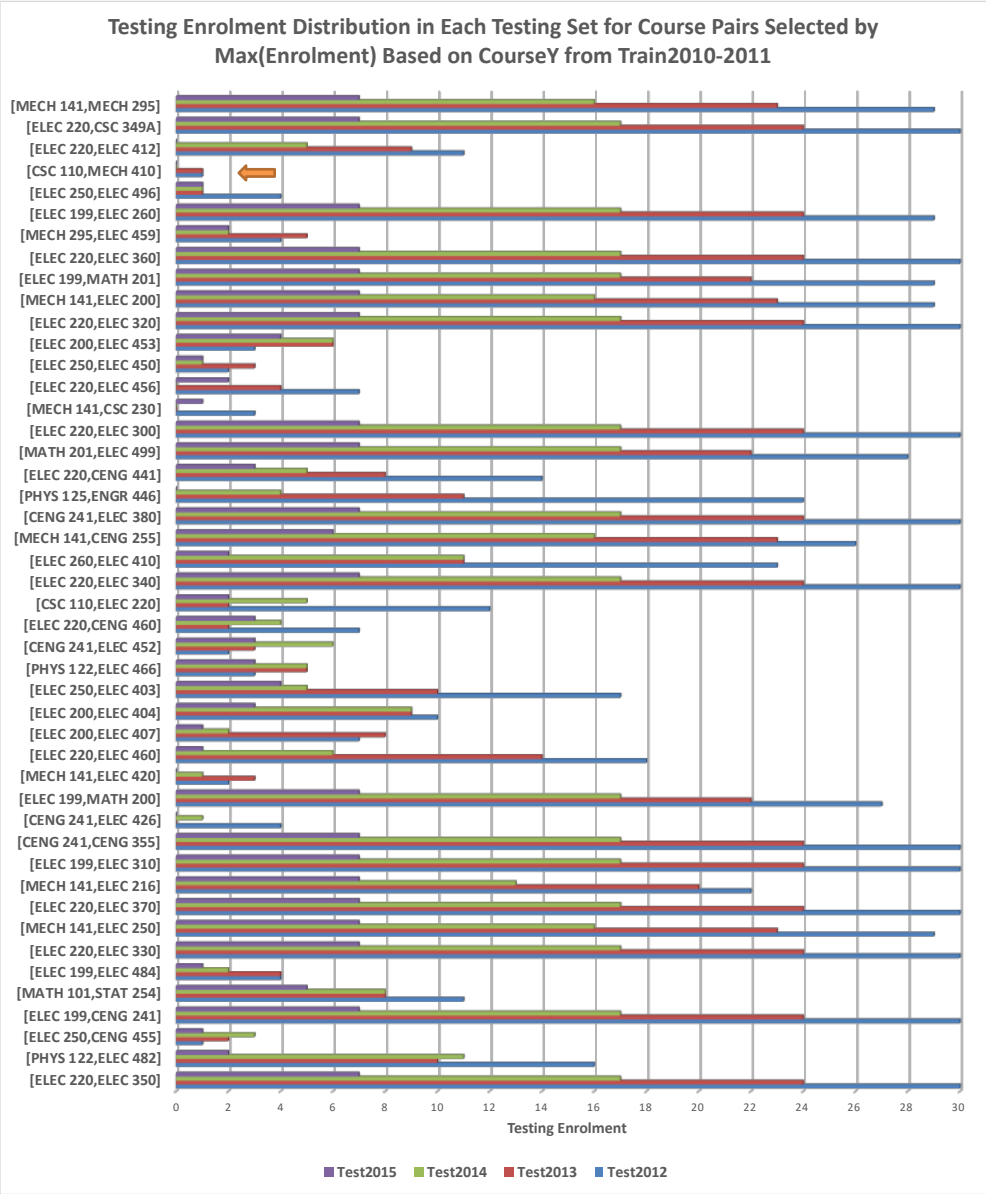

Figure 23 Testing Enrolment of Course Pairs Selected by Max(Enrolment) Based on CourseY in Each Testing Set for Train2010-2011

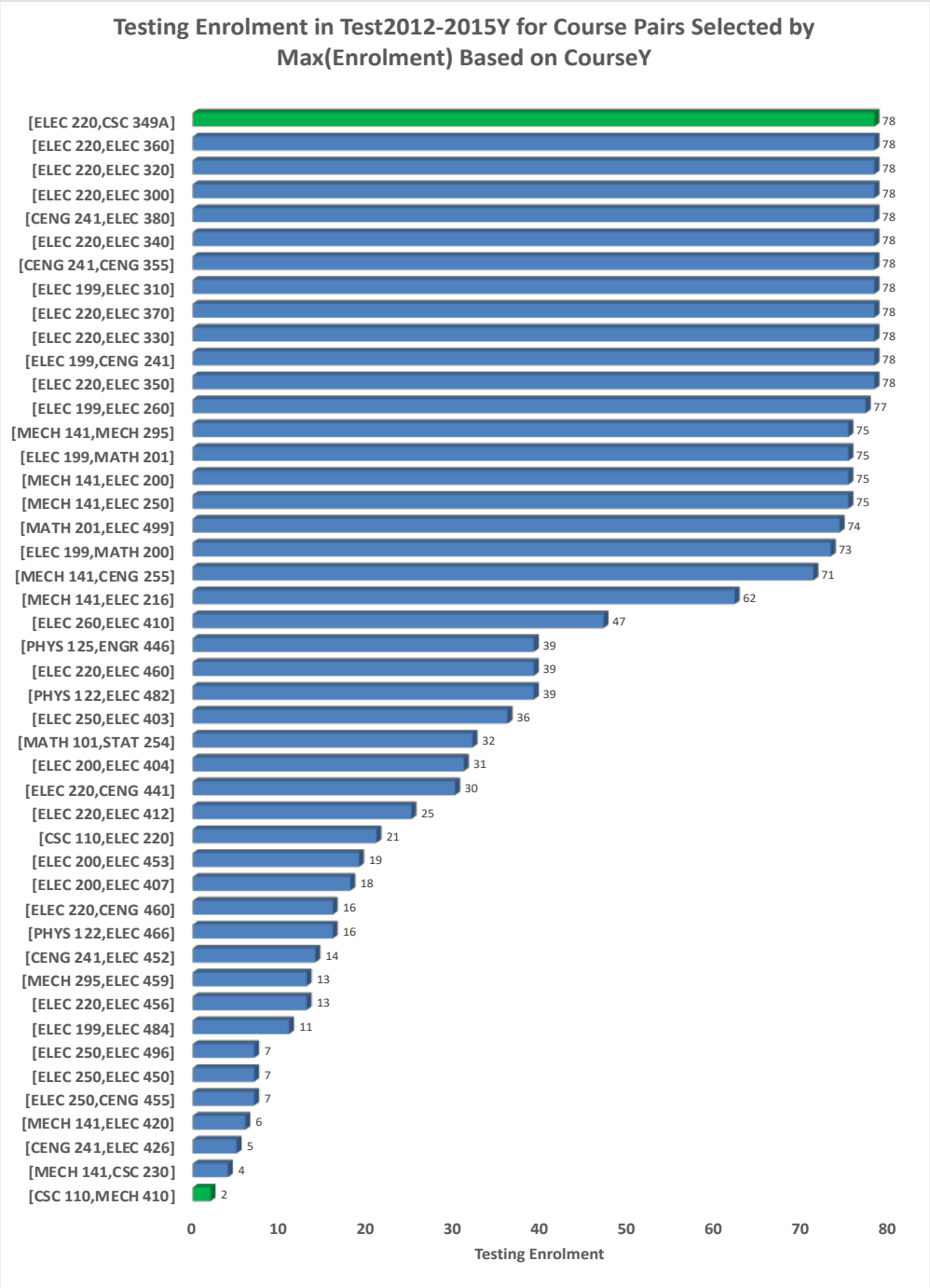

Figure 24 Testing Enrolment Distribution in Test2012-2015Y for course Pairs Selected by Max(Enrolment) Based on CourseY from Train2010-2011

Similarly, the testing sets for the course pairs selected based on CourseX and CourseY in Train2010-2012 and Train2010-2013 were merged. The merged testing sets were

denoted as Test2013-2015X and Test2013-2015Y for Train2010-2012, and Test2014-2015X and Test2014-2015Y for Train2010-2013, respectively.

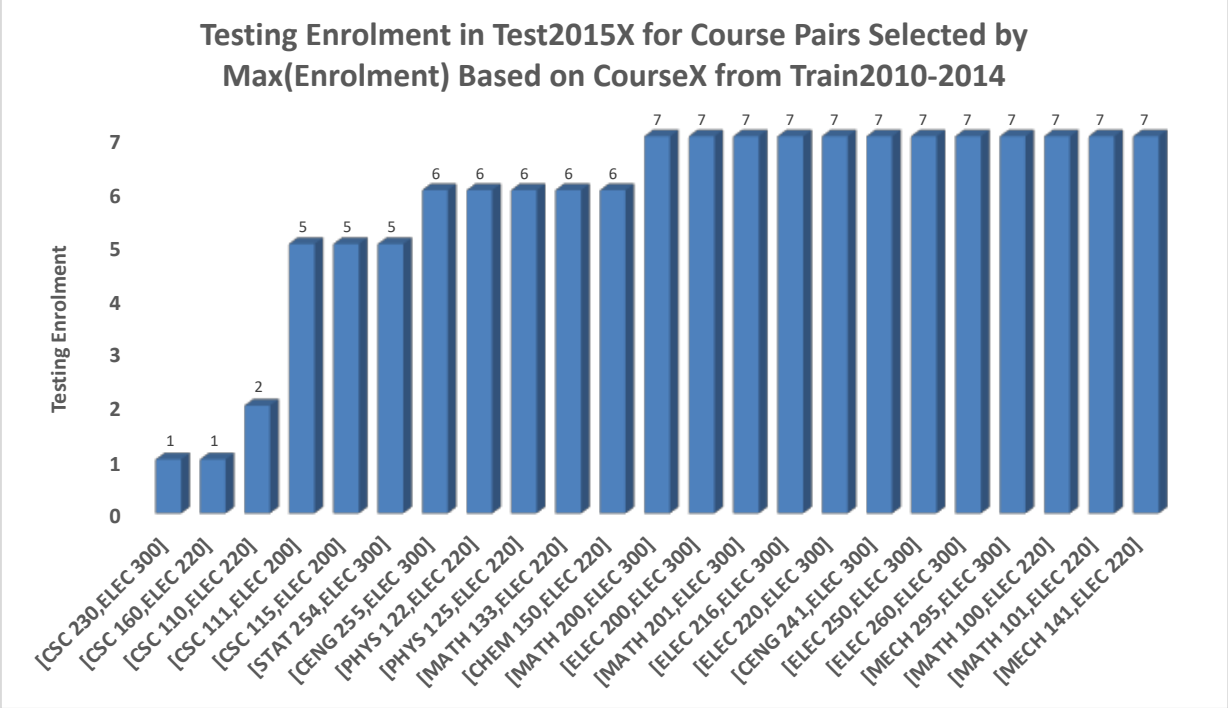

Figure 25 Testing Enrolment Distribution in Test2015X for course Pairs Selected by Max(Enrolment) Based on CourseY from Train2010-2014

Because of the same issues discussed above, the training set of Train2010-2014 is deemed invalid due to the small testing enrolments in its only testing set of Test2015, which was discussed in Section 4.2.1.1. The testing enrolments of Test2015X and Test2015Y for Train2010-2014 are shown in Figure 25 and Figure 26 with a maximum of 7.

As the testing datasets were preprocessed, they are used in the MAE and prediction precision analysis in the next two sections.

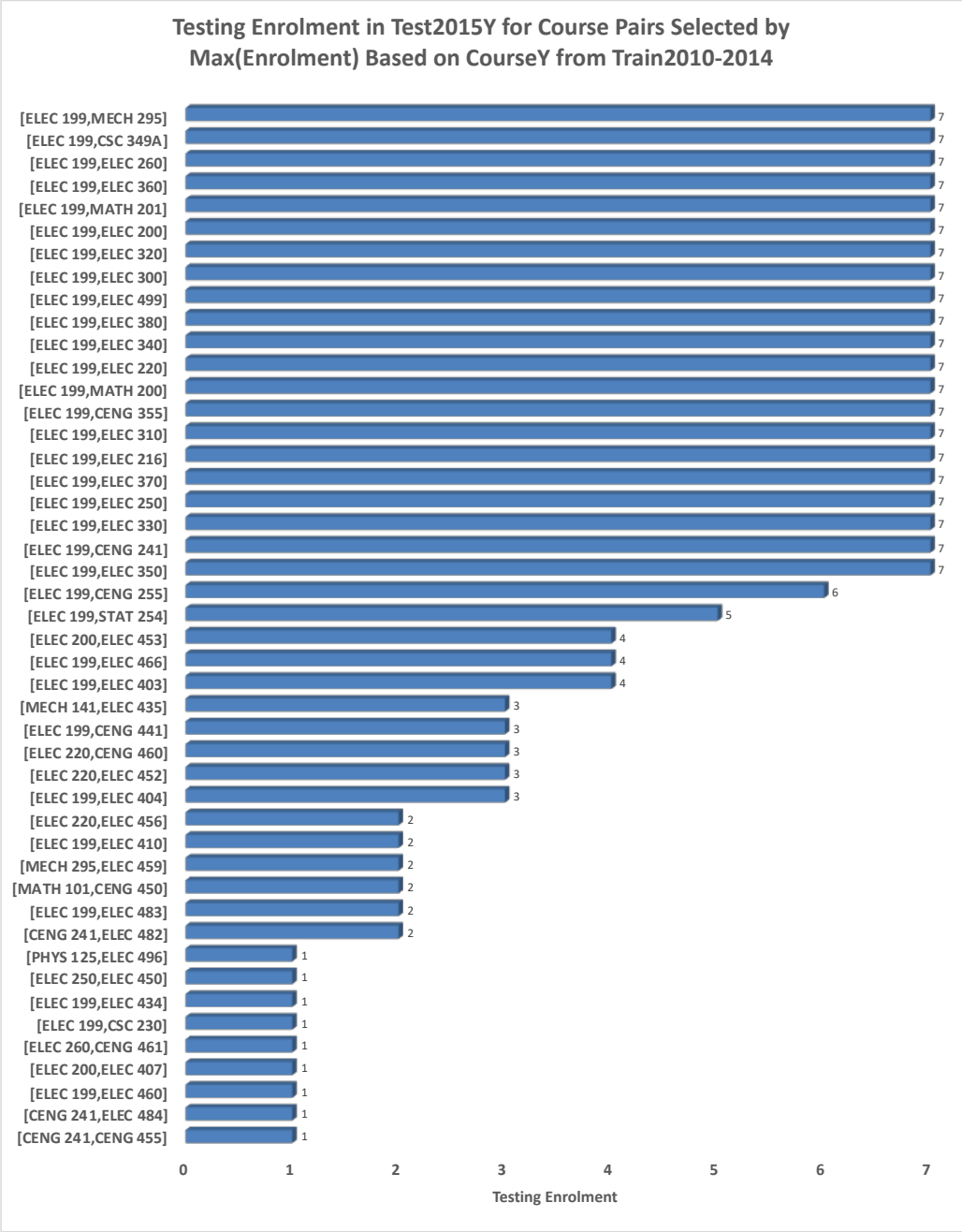

Figure 26 Testing Enrolment Distribution in Test2015X for course Pairs Selected by Max(Enrolment) Based on CourseY from Train2010-2014

4.2.2.2 MAE Analysis of Selection with Max(Enrolment)

The MAE obtained from the testing set of Test2012-2015X for the course pairs selected by Max(Enrolment) based on CourseX are shown in Figure 27.

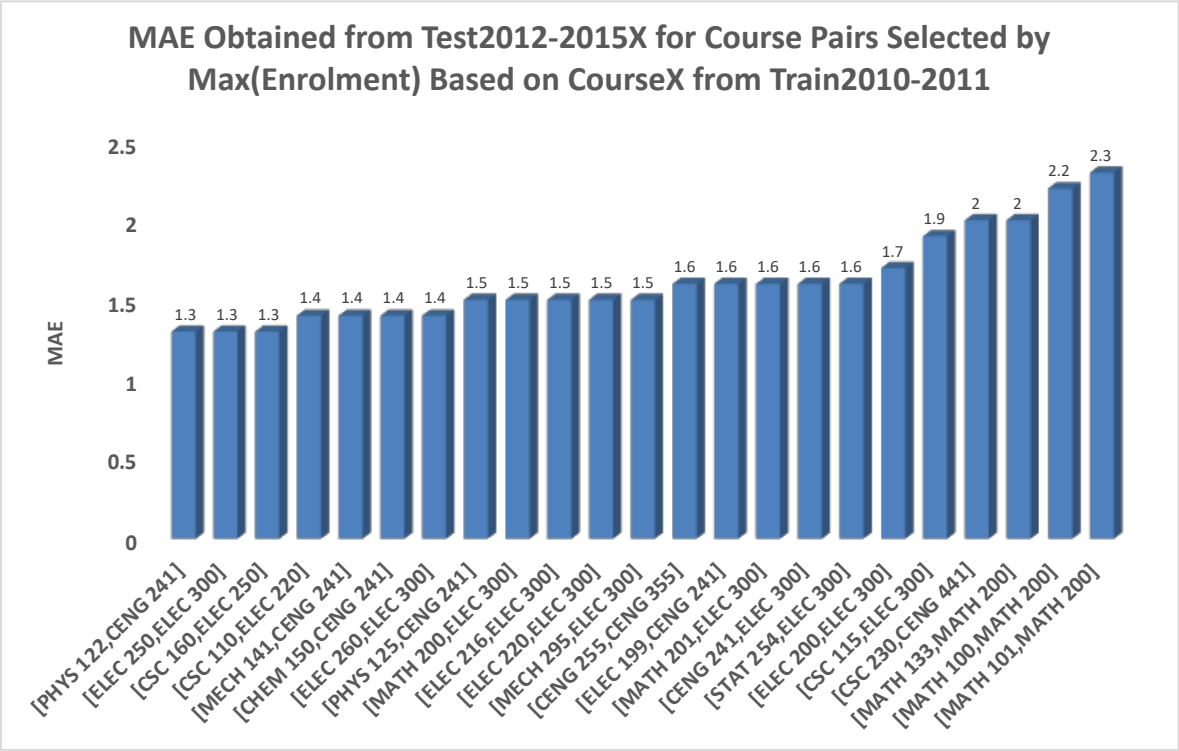

Figure 27 Prediction MAEs for Course Pairs Selected by Max(Enrolment) Based on CourseX from Train2010-2011 in Test2012-2015X

It can be seen from Figure 27 that none of the MAEs meet the acceptance criterion defined in Section 4.1. All of the MAEs are greater than 1.0 in the range of 1.3 to 2.3. Table 9 lists, for each course pairs, the testing enrolment, amount of predicted error in the range of 0 to 1.0, percentage of predicted errors in the range of from 0 to 1.0, and the MAE.

It can be seen from Table 9 that the testing enrolment has a wide range from 3 to 78 and none of the percentages of predicted errors in the range of 0 to 1.0 are above 70%. Instead, the percentages are in the range from 26% to 67%. Therefore, together with the unacceptable MAEs of the selected course pairs which were rejected, the predictions of

these course pairs are also not acceptable. In other words, the CourseXs in the course pairs are not suitable to be used as the predictors for the performance of CourseYs.

Table 9 Course Pairs Selected Based on CourseX in Test2012-2015X

CrsXCode	CrsXNum	CrsYCode	CrsYNum	Enrolment	0~1.0	%	MAE
CSC	160	ELEC	250	12	4	33	1.3
ELEC	250	ELEC	300	78	36	46	1.3
PHYS	122	CENG	241	66	38	58	1.3
ELEC	260	ELEC	300	77	31	40	1.4
CSC	110	ELEC	220	21	10	48	1.4
MECH	141	CENG	241	75	38	51	1.4
CHEM	150	CENG	241	66	38	58	1.4
ELEC	216	ELEC	300	63	23	37	1.5
ELEC	220	ELEC	300	78	32	41	1.5
MECH	295	ELEC	300	78	33	42	1.5
PHYS	125	CENG	241	68	30	44	1.5
MATH	200	ELEC	300	73	32	44	1.5
CENG	241	ELEC	300	78	26	33	1.6
CENG	255	CENG	355	74	26	35	1.6
MATH	201	ELEC	300	75	27	36	1.6
STAT	254	ELEC	300	36	13	36	1.6
ELEC	199	CENG	241	78	31	40	1.6
ELEC	200	ELEC	300	77	29	38	1.7
CSC	115	ELEC	300	61	18	30	1.9
MATH	133	MATH	200	66	18	27	2
CSC	230	CENG	441	3	2	67	2
MATH	100	MATH	200	65	18	28	2.2
MATH	101	MATH	200	69	18	26	2.3

Similarly, the MAEs for course pairs selected based on CourseY were computed from Test2012-2015Y and shown in Figure 28. It can be seen from Figure 28 that there are four course pairs with acceptable MAE (shown in green). Two of the course pairs have MAE equal to 1.0 from the pair of MECH 141 and CENG 255 with 71 testing enrolments and 59% of errors within 1.0, and the pair of ELEC 250 and CENG 455 with 7 testing enrolments and 71% of errors within 1.0 shown in Table 10.

The two course pairs with MAE <= 1.0 only have about 60% of predicted errors in the acceptable range with 74 testing enrolments for the pair of MATH 201 and ELEC 499 and 71 testing enrolments for the pair of MECH 141 and CENG 255, respectively. Based on the MAEs, the two CourseXs are reliable predictors for the corresponding CourseYs. But based on the percentage of predicted errors in the range of from 0 to 1.0, their prediction performance is poor. Therefore, this predictor-selection approach does not work well for the course pairs and some other approach to select the appropriate predictor is required.The

two course pairs with 71% of predicted errors <= 1.0 shown in Table 10 only have 7 testing enrolments in total. Therefore, although they have a low percentage of errors less than or equal to 1.0, the prediction may be skewed due to the relatively small testing enrolment.

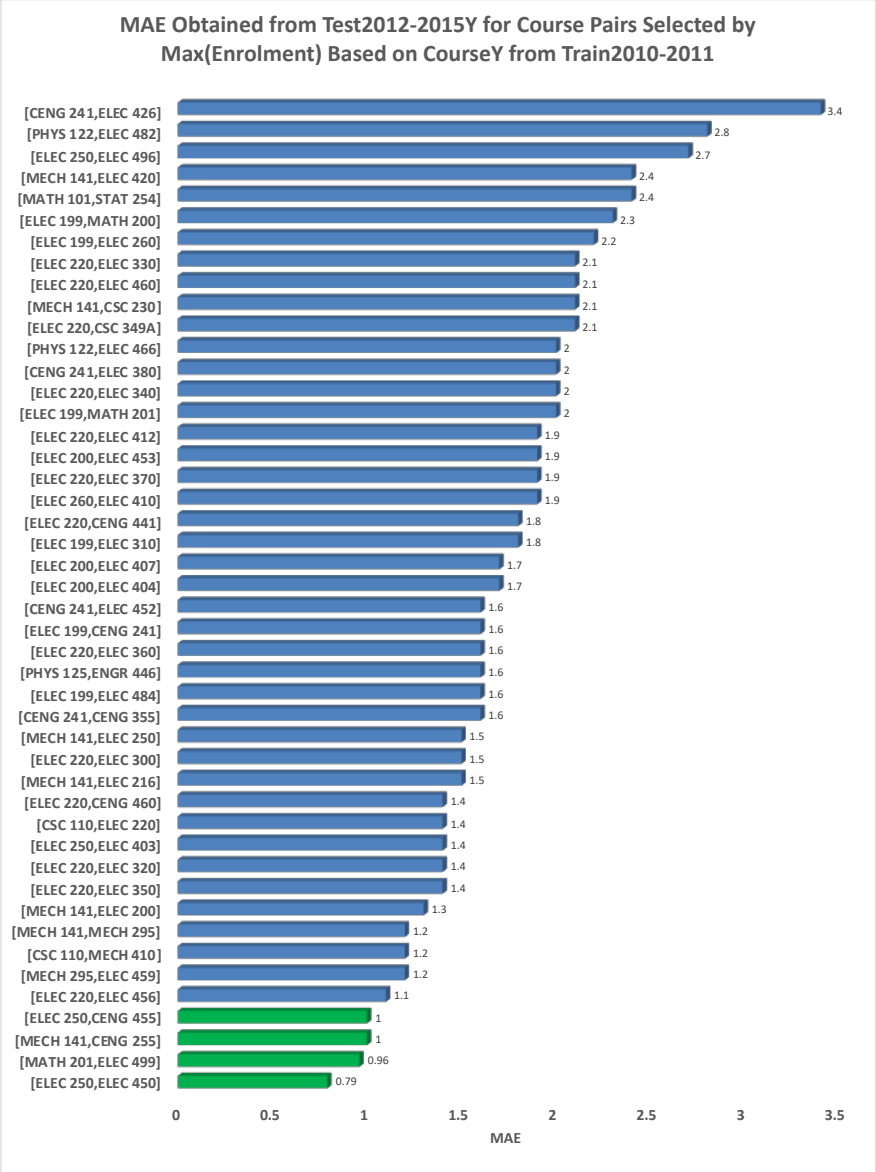

Figure 28 Prediction MAEs for Course Pairs Selected by Max(Enrolment) Based on CourseY from Train2010-2011 in Test2012-2015Y

Table 10 Course Pairs Selected Based on CourseY with MAE<=1.0 in Test2012-2015Y

CrsXCode	CrsXNum	CrsYCode	CrsYNum	Enrolment	0~1.0	%	MAE
ELEC	250	ELEC	450	7	5	71	0.79
MATH	201	ELEC	499	74	44	59	0.96
MECH	141	CENG	255	71	42	59	1
ELEC	250	CENG	455	7	5	71	1

The MAEs of course pairs selected based on CourseX obtained from Test2013-2015X and Test2014-2015X are shown in Figure 73 and Figure 75 in Appendix 12, respectively. The two figures have the same trend as the one shown in Figure 27 that all of the MAEs are out of the acceptable range. Meanwhile, the predicted errors less than or equal to 1.0 still have some course pairs with less than 70% upon further examination.

Similarly, the MAEs of course pairs selected based on CourseY obtained from Test2013-2015Y and Test2014-2015Y are shown in Figure 74 and Figure 76 in Appendix 12. There are a few course pairs with MAE less than or equal to 1.0 similar to the ones in Figure 28. For such course pairs, there are two scenarios. The first scenario is that the course pair has a relatively big testing enrolment but a relatively small portion of errors less than or equal to 1.0, for instance, the course pair of ELEC 199 and ELEC 499 with MAE of 0.87, 31 out of 48 errors less than or equal to 1.0 (65%) from Test2013-2015Y. Then, for such course pairs, although their MAE are in acceptable range, the MAE cannot tell how much predicted errors are acceptable. In other words, the MAE cannot provide a full picture of the prediction and other means is required to enhance the prediction.

The second scenario is that the course pair have a relatively high portion of errors less than or equal to 1.0 but have a relatively small number of testing enrolments, for example, the pair of ELEC 250 and ELEC 450 with MAE of 0.63, 4 out of 5 errors less than or equal to 1.0 (80%) from Test2013-2015Y. For such course pairs, both of their MAEs and precisions are acceptable, however, the prediction also may be skewed due to the small testing enrolment.

Therefore, it can be concluded in general that the CourseXs selected by Max(Enrolment) from the course pairs based on CourseX are not good predictors because their MAEs are greater than 1.0 and their prediction precisions are relatively low (less than

70%), regardless of whether the testing enrolment is relative big or relatively small. Moreover, the MAEs less than or equal to 1.0 from the course pairs selected based on CourseY cannot really indicate if the prediction is reliable or not because of two scenarios mentioned above. Therefore, another prediction result evaluation approach, prediction precision, is carried out for further analysis in next section.

4.2.2.3 Prediction Precision Analysis with Max(Enrolment)

Using the same definitions as in previous sections, prediction precision analysis are presented in this section. The prediction precisions of the course pairs selected based on CourseX and obtained from the testing set of Test2012-2015X are shown in Figure 29 together with the corresponding testing enrolments. It can be seen from Figure 29 that the testing enrolments produce widespread prediction precisions, with a maximum of 67% from the course pair of CSC 230 and CENG 441 with just 3 testing enrolments and a minimum of 26% from the pair of MATH 101 and MATH 200 with 69 testing enrolments.

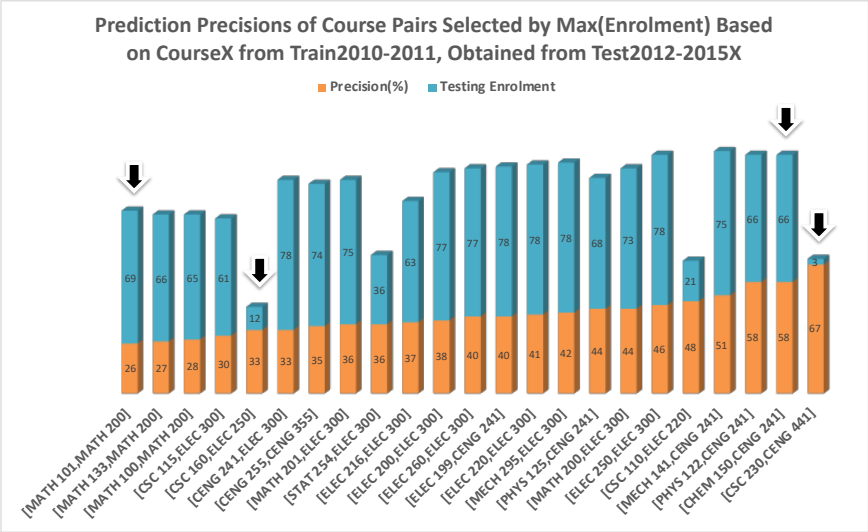

Figure 29 Prediction Precisions of Course Pairs Tested in Test2012-2015, Trained by Train2010-2011

Moreover, there are course pairs with relatively high prediction precisions with big testing enrolments, for instance, the pair of CHEM 150 and CENG 241 has prediction

precision of 58% with 66 testing enrolments. This implies that a big testing enrolment can produce either a high or low prediction precision. The course pair of CSC 160 and ELEC 250 has a relatively small testing enrolment, 12 with prediction precision of 33%. In other words, the relatively small testing enrolment also can generate both low and high prediction precision. Therefore, these course pairs perform poorly in the prediction and some other approach needs to be investigated to explore new course pairs to produce better prediction results.

The prediction precisions of course pairs selected based on CourseY from the testing set of Test2012-2015Y are presented in Figure 30. The prediction precisions and testing enrolments of each course pair show a similar trend as the one in Figure 29 that no matter what size the testing enrolment is, it can produce either high or low prediction precisions. For instance, the course pair of ELEC 250 and ELEC 450 has 71% precision with 7 testing enrolments while the pair of CENG 241 and ELEC 426 has 20% precision with 5 testing enrolments. Similarly, the course pair of MECH 141 and CENG 255 has 59% precision with 71 testing enrolments while the pair of ELEC 220 and CSC 349A has 18% precision with 78 testing enrolments. Therefore, the course pairs selected based on CourseY are not reliable to be used for further analysis.

Although the top two bars in Figure 30 have precisions of 71%, the pairs of ELEC 250 and CENG 455, and ELEC 250 and ELEC 450, they are not deemed reliable because of the relatively small testing enrolment of 7.

Similarly to the prediction precision that was analyzed for the course pairs selected from Train2010-2011, it was also explored for the other two valid training sets of Train2010-2012 and Train2010-2013. The prediction precisions with the corresponding testing enrolment of the course pairs selected from these training sets are shown in Appendix 13.

None of the prediction precisions obtained from Test2013-2015X and Test2014-2015X are above 70%. Meanwhile, the testing enrolment distribution behaves the same as the ones in Figure 29 namely it shows that small testing enrolment have low precision and also high precision. Also, the big testing enrolment has the same trend as the small testing

enrolment, for example, MATH 133 and MATH 200 has 26% with 43 testing enrolments and CSC 111 and CENG 241 has 63% with 38 testing enrolments.

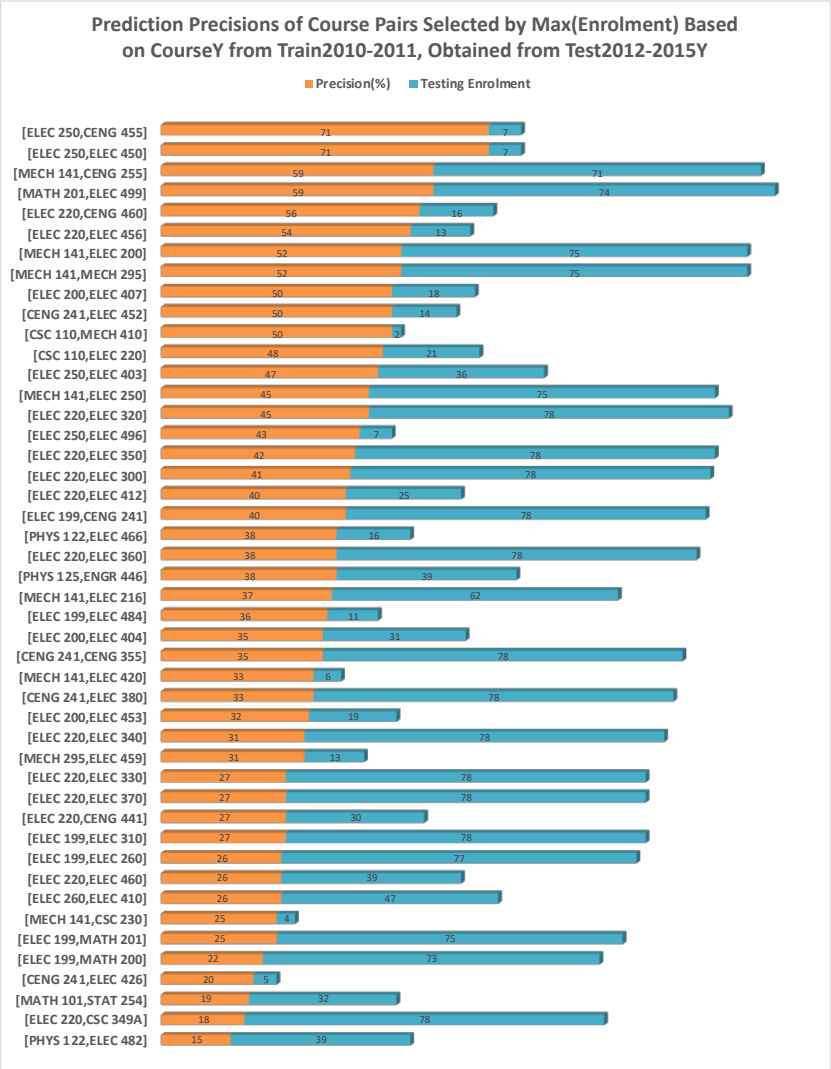

Figure 30 Prediction Precisions of Course Pairs Tested in Test2012-2015, Trained by Train2010-2011

The testing enrolments from Test2013-2015Y and Test2014-2015Y for the course pairs selected based on CourseY shown in Figure 78 and Figure 80 in Appendix 13 have the same trend as the testing enrolments discussed above.

There are 7 course pairs with prediction precisions over 70%, 3 from Test2013-2015Y shown in Table 11 and 4 from Test2014-2015Y shown in Table 12. It can be seen from

Table 11 that the testing enrolment are relatively small compared to the ones shown in Figure 78 in Appendix 13 with the maximum of 12 from the pair of CENG 241 and ELEC 452, followed by 5 and 1 from the pair of ELEC 250 and. ELEC 450, and ELEC 220 and ELEC 426, respectively. There are just 1 or 2 testing enrolments for the course pairs with precision over 70% from Test2014-2015shown in Table 12. Therefore, although the precisions are relatively high and acceptable, they may be not reliable due to the small testing enrolment.

Table 11 Course Pairs with Precision over 70% from Test2013-2015Y

CrsXCode	CrsXNum	CrsYCode	CrsYNum	Enrolment	0~1.0	%
ELEC	220	ELEC	426	1	1	100
ELEC	250	ELEC	450	5	4	80
CENG	241	ELEC	452	12	9	75

Table 12 Course Pairs with Precision over 70% from Test2014-2015Y

CrsXCode	CrsXNum	CrsYCode	CrsYNum	Enrolment	0~1.0	%
ELEC	220	ELEC	426	1	1	100
ELEC	250	ELEC	450	2	2	100
ELEC	260	CENG	461	2	2	100
MATH	201	ELEC	481	1	1	100

Based on the prediction precision and testing enrolment analyzed above, the course pairs selected by maximum enrolment perform poorly in the prediction. Therefore, a new predictor selection approach is explored in the next section.

4.2.3 Predictor Selection by Coefficient and Enrolment Combined

As concluded from the previous two sections, using either the coefficient or enrolment alone to select predictor may not give reliable results, and therefore another approach needs to be investigated. The third predictor selection method is to use the combination of the two factors, coefficient and enrolment.

Table 13 Coefficients and Enrolments of Course Pairs of Course$i$ and Course$S$

Course$X$	Course$Y$	Coefficient$(X-Y)$	Enrolment$(X-Y)$
Course1	Course$S$	r_1	e_1
Course2	Course$S$	r_2	e_2
…	…	…	…
Course$i$	Course$S$	r_i	e_i
…	…	…	…
Course$n$	Course$S$	r_n	e_n

As shown in Table 13, there are n course pairs each of which is paired with the specific course$S$, with the corresponding coefficients and enrolments. As discussed in the previous two predictor selection approaches, either the maximum of coefficient or maximum of enrolment is the indicator to select the predictors. Therefore, a simple combination is to add the two factors together, denoted as P_i for the course pair of Course$i$ and Course S shown in equation (5). The predictor selection criterion is to use the maximum of the combination, $Max(P_i)$.

$$P_i = r_i + e_i \tag{5}$$

As seen in the Pearson Coefficient definition, the coefficient has a range of from -1.0 to 1.0. However, the enrolment is the number of students who enrolled in the same courses Course$i$ and Course$S$. Typically, the enrolment is a number larger than the coefficient. Then, the enrolment would dominate in the equation (5) and the coefficient could be ignored. Therefore, the equation is likely to be the equation (6) and the predictor selection criterion, $Max(P_i)$, is the same as $Max(Enrolment)$ which was already discussed in the second predictor selection approach in Section 4.2.2.

$$P_i = e_i \tag{6}$$

Therefore, the two factors must be normalized [65], as shown in equations (7) and (8), to eliminate the enrolment bias.

- Normalized i^{th} coefficient r_i:

$$r_{normi} = \frac{r_i - r_{min}}{r_{max} - r_{min}}$$

where $r_{min} = \min(r_1, r_2, r_3, \dots, r_n)$

$$r_{max} = \max(r_1, r_2, r_3, \dots, r_n) \quad (7)$$

- Normalized i^{th} enrolment e_i:

$$e_{normi} = \frac{e_i - e_{min}}{e_{max} - e_{min}}$$

where $e_{min} = \min(e_1, e_2, e_3, \dots, e_n)$

$$\mathrm{e_{max}} = \max(e_1, e_2, \dots, e_n) \quad (8)$$

Accordingly, the combination for the course pair of Course i and Course S, P_i, becomes the sum of the two normalized factors shown in equation (9).

$$P_i = r_{normi} + e_{normi} \quad (9)$$

In order to eliminate the domination of either of two factors, weights are assigned to each normalized factor, w_r for normalized coefficient and w_e for normalized enrolment. In addition, one can emphasize the importance of one factor over the other by adjusting the weights. Therefore, P_i becomes (10).

$$P_i = w_r * r_{normi} + w_e * e_{normi} \quad (10)$$

Various values of the two weights are used to compute P_i, in order to find the optimal results. The two weights, w_r and w_e, are initialized to 1.0 for one course pair when computing its P_i. The enrolment weight, w_e, is expected to remain at the initial value of 1.0 as the enrolment is much larger compared to the coefficient. Indeed, experimentation with the current dataset using different values of w_e confirms this conjecture. However, the enrolment weight may have to be adjusted accordingly when using a different dataset. The coefficient weight, w_r, is incremented by 0.1 in each iteration in our trial computations of P_i. From our experimentation, when w_r reaches 2.5, the maximum P_i is always

generated from the same course pair among the n course pairs. For each course pair, there are 16 iterations in computing P_i with 16 weight pairs shown in Table 14.

Table 14 Weight Pairs of Coefficient and Enrolment for P_i Computation of One Course Pair

w_1	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0	2.1	2.2	2.3	2.4	2.5
w_2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Two examples of using the coefficient-enrolment approach to select the predictors are shown in Table 15 and Table 16. The course pairs in the examples are from the training set of Train2010-2011. The columns showing the results for P_i have the headers $w_r\_w_e$, for coefficient weight and enrolment weights, respectively.

All of the maximum of P_i of each weight pair for all of the predicted courses were highlighted as shown in Table 15. The CourseX in the predicted course pair is the predicting course. If there are more than one predicting courses with the same P_i, the one with the smallest first digit of the course number is used as predictor for the predicted course as the smaller course number indicates that course was typically taken by a student in an earlier term in the academic schedules. If multiple predicting courses have same first digit of course numbers, all of them are selected as predictors and use for prediction.

The maximum P_i for each weight pair for the course pairs in Table 16 is highlighted. The maximum P_i is generated from the course pair MATH 100 and MATH 201when w_r reaches 1.4. Therefore, MATH 100 is picked as the predictor for course MATH 201.

Based on the P_i computed by using the weights associated with the Pearson Coefficient and enrolment, the predictors are picked from each training set. Similar to the other two approaches discussed in Section 4.2.1 and 4.2.2, the predictors are selected for both CourseX and CourseY.

Table 15 Predictor-Selection for Course STAT 254 by using P_i

xsubCode	xnum	ysubCode	ynum	coefficient	#points	1_1	1.1_1	1.2_1	1.3_1	1.4_1	1.5_1	1.6_1	1.7_1	1.8_1	1.9_1	2.0_1	2.1_1	2.2_1	2.3_1	2.4_1	2.5_1
MATH	101	STAT	254	0.6407	17	1.8095	1.9095	2.0095	2.1095	2.2095	2.3095	2.4095	2.5095	2.6095	2.7095	2.8095	2.9095	3.0095	3.1095	3.2095	3.3095
PHYS	122	STAT	254	0.5392	17	1.6511	1.7353	1.8194	1.9036	1.9877	2.0719	2.1561	2.2402	2.3244	2.4085	2.4927	2.5768	2.661	2.7452	2.8293	2.9135
MATH	133	STAT	254	0.5268	17	1.6317	1.714	1.7962	1.8784	1.9606	2.0429	2.1251	2.2073	2.2895	2.3718	2.454	2.5362	2.6184	2.7006	2.7829	2.8651
CSC	110	STAT	254	0.3953	18	1.4741	1.5358	1.5975	1.6592	1.7209	1.7826	1.8443	1.906	1.9677	2.0294	2.0911	2.1528	2.2145	2.2762	2.3379	2.3996
CHEM	150	STAT	254	0.3727	16	1.3436	1.4018	1.46	1.5181	1.5763	1.6345	1.6926	1.7508	1.809	1.8671	1.9253	1.9835	2.0417	2.0998	2.158	2.2162
MECH	141	STAT	254	0.3366	21	1.5254	1.5779	1.6304	1.683	1.7355	1.788	1.8406	1.8931	1.9457	1.9982	2.0507	2.1033	2.1558	2.2083	2.2609	2.3134
CSC	160	STAT	254	0.322	18	1.3597	1.41	1.4602	1.5105	1.5607	1.611	1.6613	1.7115	1.7618	1.812	1.8623	1.9126	1.9628	2.0131	2.0633	2.1136
PHYS	125	STAT	254	0.2006	20	1.2655	1.2968	1.3281	1.3594	1.3907	1.422	1.4533	1.4846	1.516	1.5473	1.5786	1.6099	1.6412	1.6725	1.7038	1.7351
ELEC	199	STAT	254	0.1533	21	1.2393	1.2632	1.2871	1.3111	1.335	1.3589	1.3828	1.4068	1.4307	1.4546	1.4785	1.5025	1.5264	1.5503	1.5742	1.5982

Table 16 Predictor-Selection for Course MATH 201 by using P_i

xsubCode	xnum	ysubCode	ynum	coefficient	#points	1_1	1.1_1	1.2_1	1.3_1	1.4_1	1.5_1	1.6_1	1.7_1	1.8_1	1.9_1	2.0_1	2.1_1	2.2_1	2.3_1	2.4_1	2.5_1
MATH	100	MATH	201	0.7257	21	1.5526	1.6526	1.7526	1.8526	1.9526	2.0526	2.1526	2.2526	2.3526	2.4526	2.5526	2.6526	2.7526	2.8526	2.9526	3.0526
MATH	101	MATH	201	0.6077	27	1.5479	1.6317	1.7154	1.7991	1.8829	1.9666	2.0504	2.1341	2.2178	2.3016	2.3853	2.4691	2.5528	2.6365	2.7203	2.804
MATH	133	MATH	201	0.5563	26	1.4508	1.5274	1.6041	1.6808	1.7574	1.8341	1.9107	1.9874	2.064	2.1407	2.2174	2.294	2.3707	2.4473	2.524	2.6006
CHEM	150	MATH	201	0.5193	27	1.4261	1.4977	1.5692	1.6408	1.7123	1.7839	1.8555	1.927	1.9986	2.0701	2.1417	2.2133	2.2848	2.3564	2.4279	2.4995
PHYS	122	MATH	201	0.4973	29	1.4484	1.517	1.5855	1.654	1.7225	1.7911	1.8596	1.9281	1.9966	2.0652	2.1337	2.2022	2.2708	2.3393	2.4078	2.4763
ELEC	199	MATH	201	0.4867	38	1.6707	1.7377	1.8048	1.8719	1.9389	2.006	2.0731	2.1401	2.2072	2.2743	2.3413	2.4084	2.4755	2.5425	2.6096	2.6767
CSC	160	MATH	201	0.4547	31	1.4424	1.505	1.5677	1.6303	1.693	1.7556	1.8183	1.881	1.9436	2.0063	2.0689	2.1316	2.1942	2.2569	2.3196	2.3822
MECH	141	MATH	201	0.4289	36	1.5384	1.5975	1.6566	1.7157	1.7748	1.8339	1.893	1.9521	2.0112	2.0703	2.1294	2.1885	2.2476	2.3067	2.3658	2.4249
PHYS	125	MATH	201	0.4276	32	1.4313	1.4903	1.5492	1.6081	1.667	1.7259	1.7849	1.8438	1.9027	1.9616	2.0206	2.0795	2.1384	2.1973	2.2562	2.3152
CSC	110	MATH	201	0.1572	28	0.9535	0.9751	0.9968	1.0184	1.0401	1.0618	1.0834	1.1051	1.1268	1.1484	1.1701	1.1917	1.2134	1.2351	1.2567	1.2784

The course pairs are selected from the four different training sets of Tain2010-2011, Train2010-2012, Train2010-2013 and Train2010-2014 by using P_i based on both CourseX and CourseY. It can be seen from the course pairs selected based on CourseX shown in Table 19, Table 21, Table 23 and Table 25 in Appendix 14 that there are some course pairs with CourseY from fourth year, which is similar as the course pairs selected by maximum coefficient based on CoruseX discussed in Section 4.2.1. Therefore, it prevents the enrolment's domination in the predictor selection discussed in Section 4.2.2, where the fourth-year courses are removed from the selected course pairs because the fourth-year courses are specialization-related and have relatively small enrolments.

As the course pairs were selected from the training sets, they will be used for prediction in the following sections, including testing set preprocessing, MAE analysis and precision analysis.

4.2.3.1 Testing Enrolment Distribution of Course Pairs Selected by Max(P_i)

The testing enrolments of course pairs selected by Max(P_i) based on CourseX in the four testing sets of Test2012, Test2013, Test2014 and Test2015 are shown in Figure 31. It can be seen from Figure 31 that, compared to other years, there are more testing enrolments for most course pairs in Test2012 with a maximum of 30, followed by the testing enrolments in Test2013 with 24. Test2015 has the least testing enrolments for most course pairs compared to the other three testing enrolments.

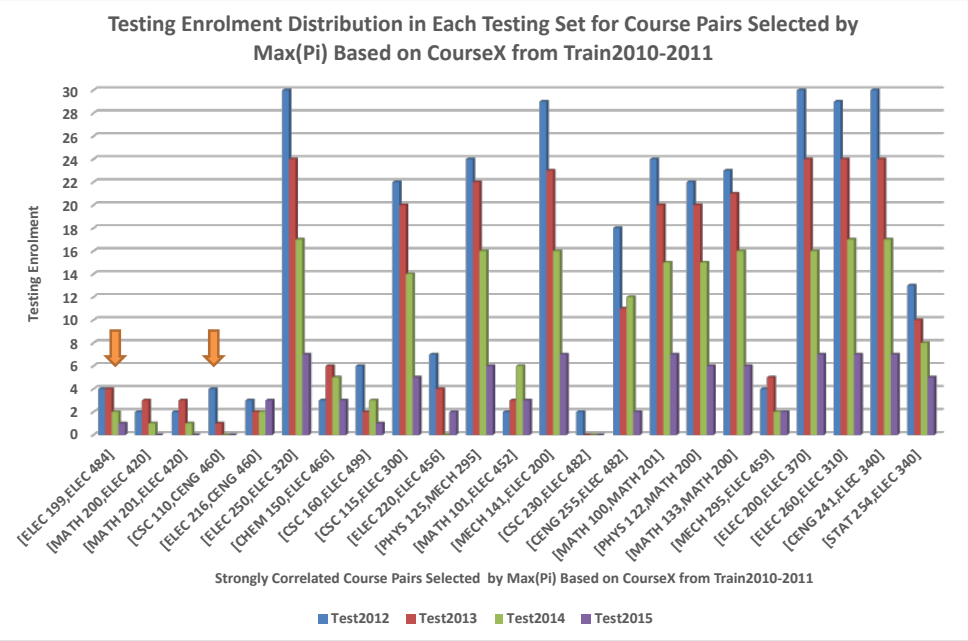

Figure 31 Testing Enrolments of Course Pairs Selected by Max(P_i) Based on CourseX in Each Testing Set for Train2010-2011

There are course pairs in each testing set that have relatively small testing enrolments, for instance, the course pair of ELEC 199 and ELEC 484 with 4 testing enrolments in Test2012 and Test2013, 2 and 1 testing enrolments in Test2014 and Test2015, respectively. Then, for such course pairs, the prediction results may be biased due to the small testing enrolment.

Moreover, there are some course pairs that do not have testing enrolments in some testing sets, for instance, the pair of CSC 110 and CENG 460 in Test2014 and Test2015. Therefore, no prediction is possible for such course pairs.

Because of the above issues of testing enrolment for the course pairs, the four different testing sets were merged as one and denoted as Test2012-2015X.

The testing enrolment distribution in the merged testing set of Test2012-2015X is shown in Figure 32. It can be seen that every course pair has testing enrolments, although some of the course pairs still have relatively small ones. The course pairs with CourseY from second or third year have comparatively bigger testing enrolments with the maximum of 78 and minimum of 36. By contrast, the course pairs with CourseY from the fourth year have relatively smaller testing enrolments with maximum of 17 from the course pair of CHEM 150 and ELEC 466. The reason is because fourth-year courses are more specialization-related and have small enrolments but the second- and third-year courses are fundamental courses.

The testing enrolment distribution of course pairs selected based on CourseY is shown in Figure 33. The testing enrolments in each testing set show similar trend that Test2012 has the most testing enrolment for many course pairs with maximum of 30 while Test2015 has the least testing enrolments for most course pairs with the maximum of 7. There still have course pairs having relatively small testing enrolment, for instance, the pair of MATH 133 and CENG 455 with 1 testing enrolment in Test2012. Also, there are some course pairs not having testing enrolment, for instance, the pair of CHEM 150 and ELEC 426 in Test2013 and Test2015. Therefore, the four testing sets also were merged as Test2012-2015Y for the same reason as the one for Test2012-2015X discussed above.

The testing enrolment in Test2012-2015Y for the course pairs selected based on CourseY is shown in Figure 34. The testing enrolments in Figure 34 have a similar trend as the one exhibited by CourseX in Figure 32 that every course pair has a testing enrolment with a maximum of 78, although there are some course pairs with relatively small testing enrolment, for instance, the pair of CSC 230 and ELEC 482 with 2 testing enrolments.

The testing sets for Train2010-2012 and Train2010-2013 were also merged due to the same issues as the ones discussed in this section for Train2010-2011. The merged testing

sets of Test2013-2015X and Test2013-2015Y are for Train2010-2012, and Test2014-2015X and Test2014-2015Y is for Train2010-2013.

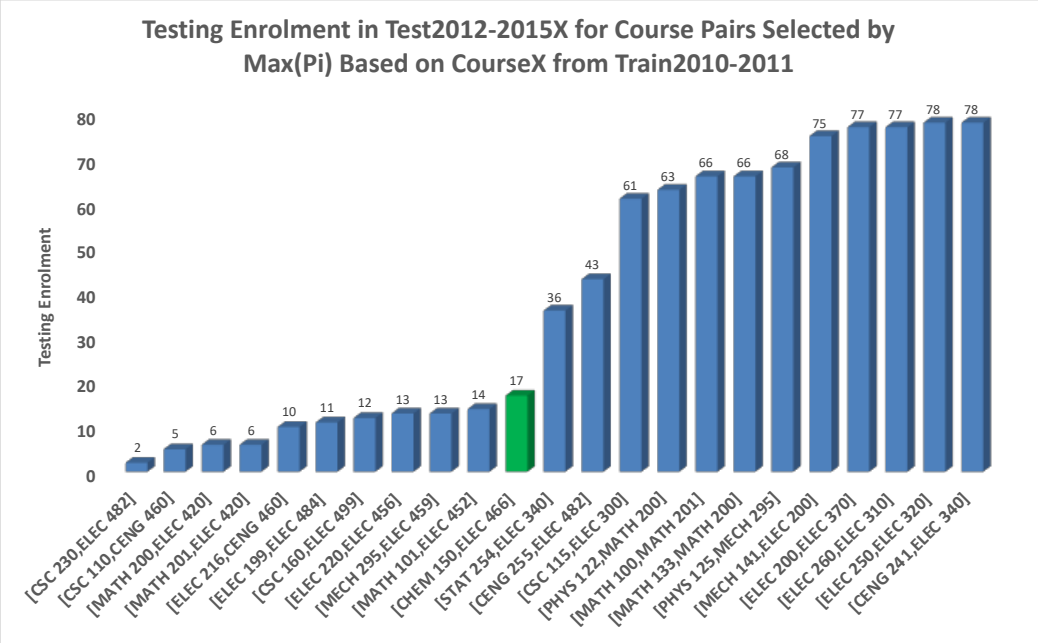

Figure 32 Testing Enrolments of course Pairs Selected by Max(p_i) Based on CourseX from Train2010-2011

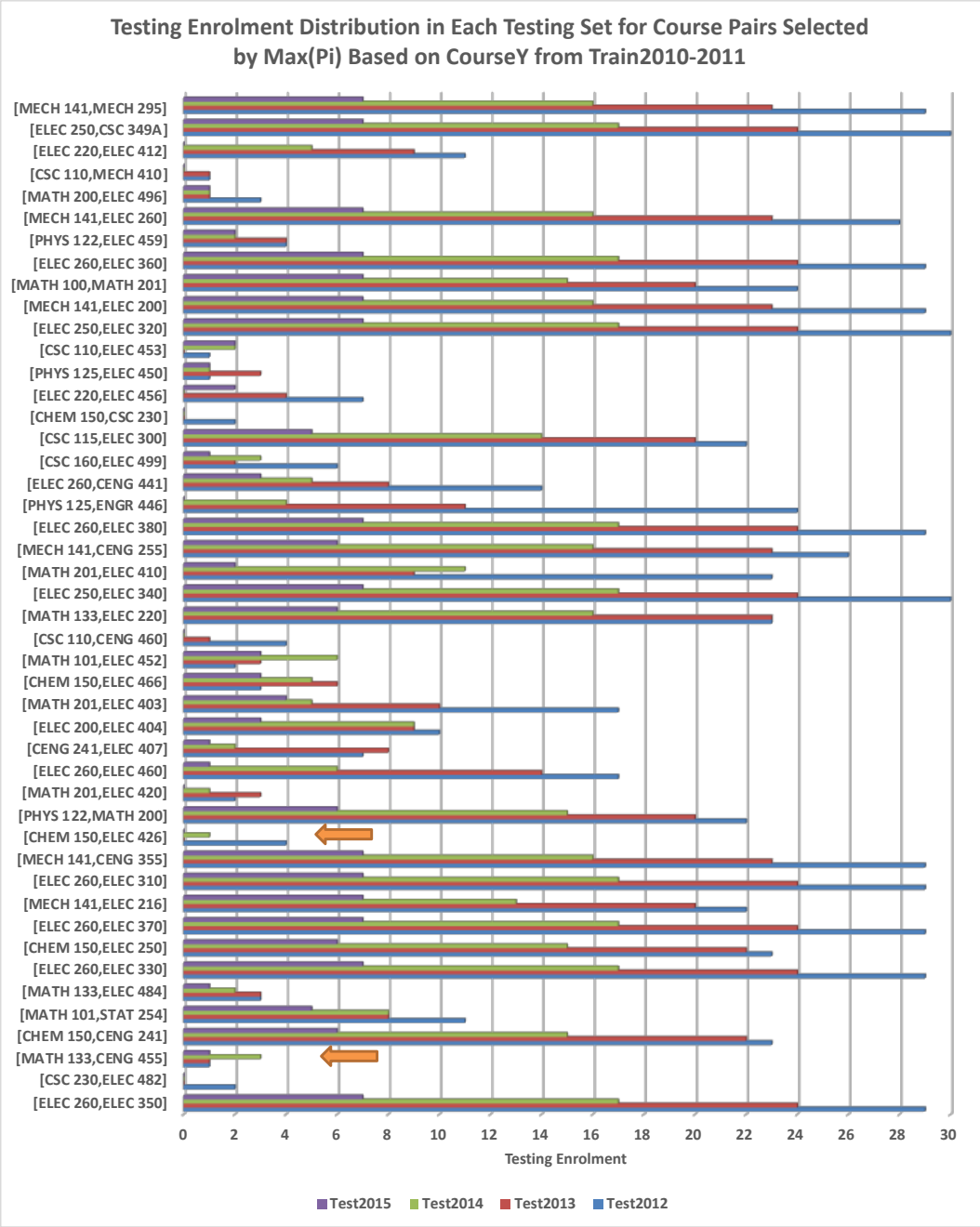

Figure 33 Testing Enrolment of Course Pairs Selected by Max(P_i) Based on CourseY in Each Testing Set for Train2010-2011

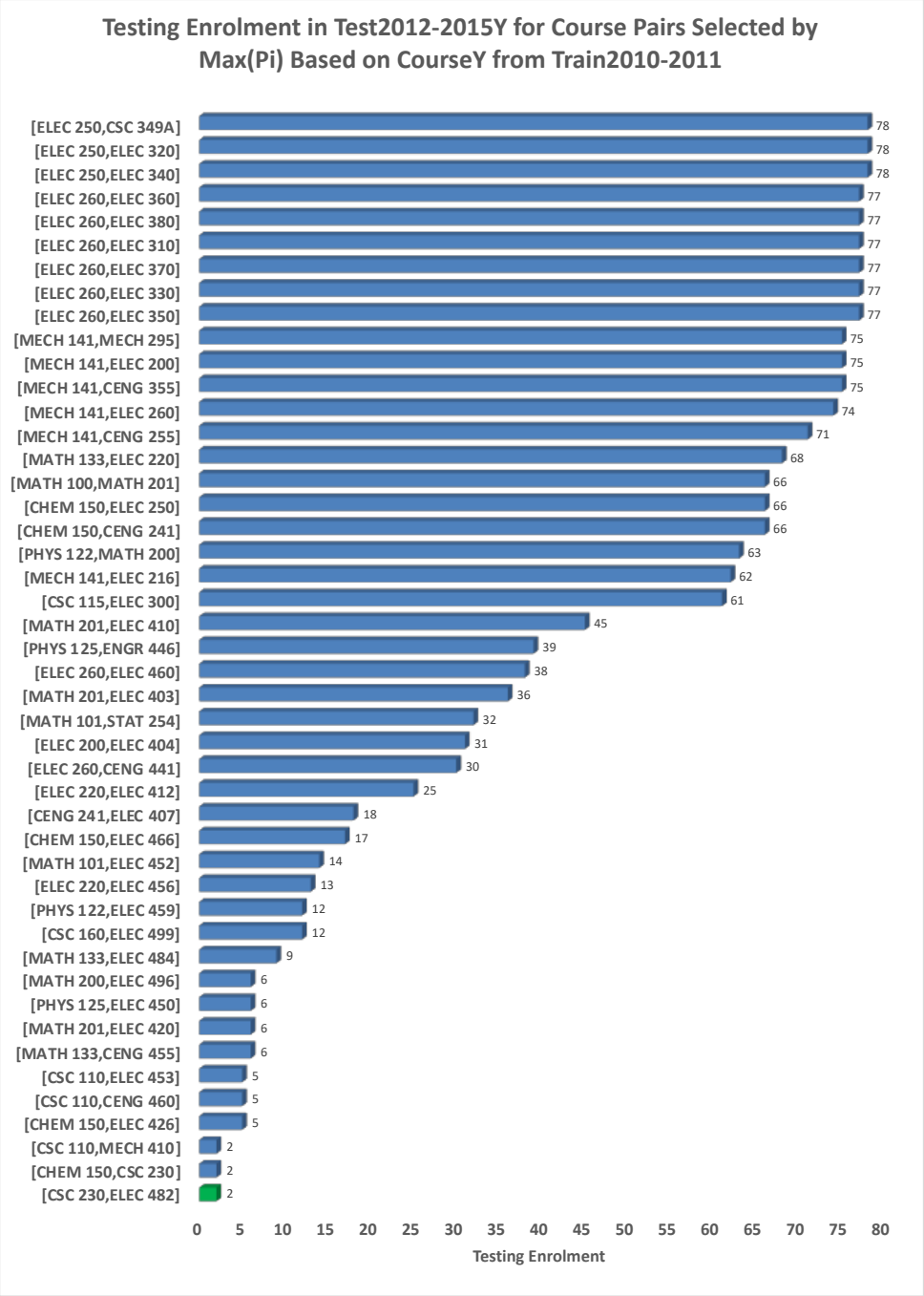

Figure 34 Testing Enrolments of course Pairs Selected by Max(p_i) Based on CourseY from Train2010-2011

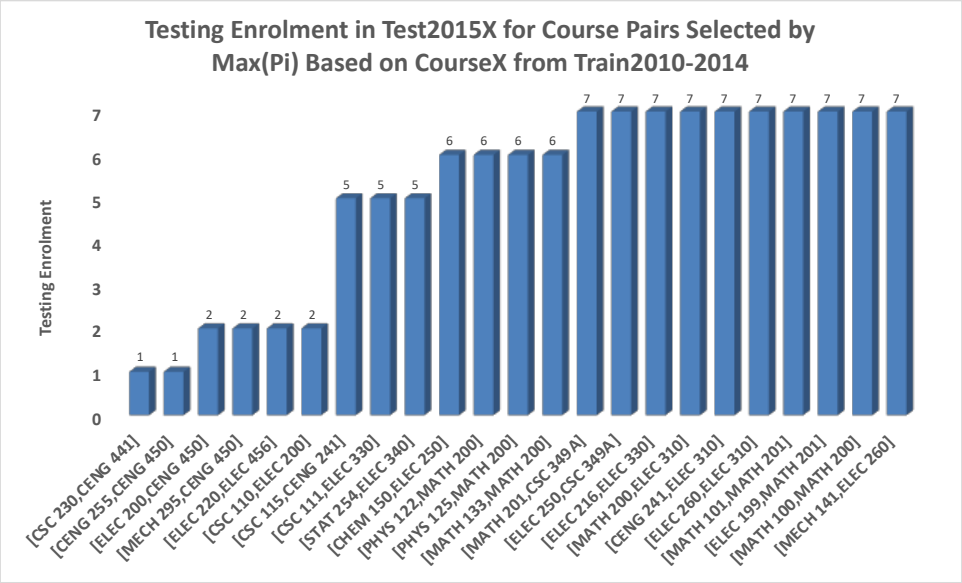

Figure 35 Testing Enrolments of course Pairs Selected by Max(P_i) Based on CourseX from Train2010-2014

The testing set for Train2010-2014 is Test2015. As there is just one testing set for Train2010-2014, there is no need to be merged. However, the testing enrolments in Test2015 is much smaller than the ones in the merged testing sets. Due to the relatively small testing enrolment, the prediction results would be biased even if the MAE or prediction precision looks considerably reliable. Therefore, the Train2010-2014 is deemed invalid and not used in latter prediction analysis. The testing enrolment distributions of course pairs selected based on both CourseX and CourseY from Train010-2014 are shown in Figure 35 and Figure 36 with a maximum of 7.

As the testing sets for all the training sets were preprocessed, they are used in the prediction analysis in the following sections including the MAE analysis and prediction precision analysis.

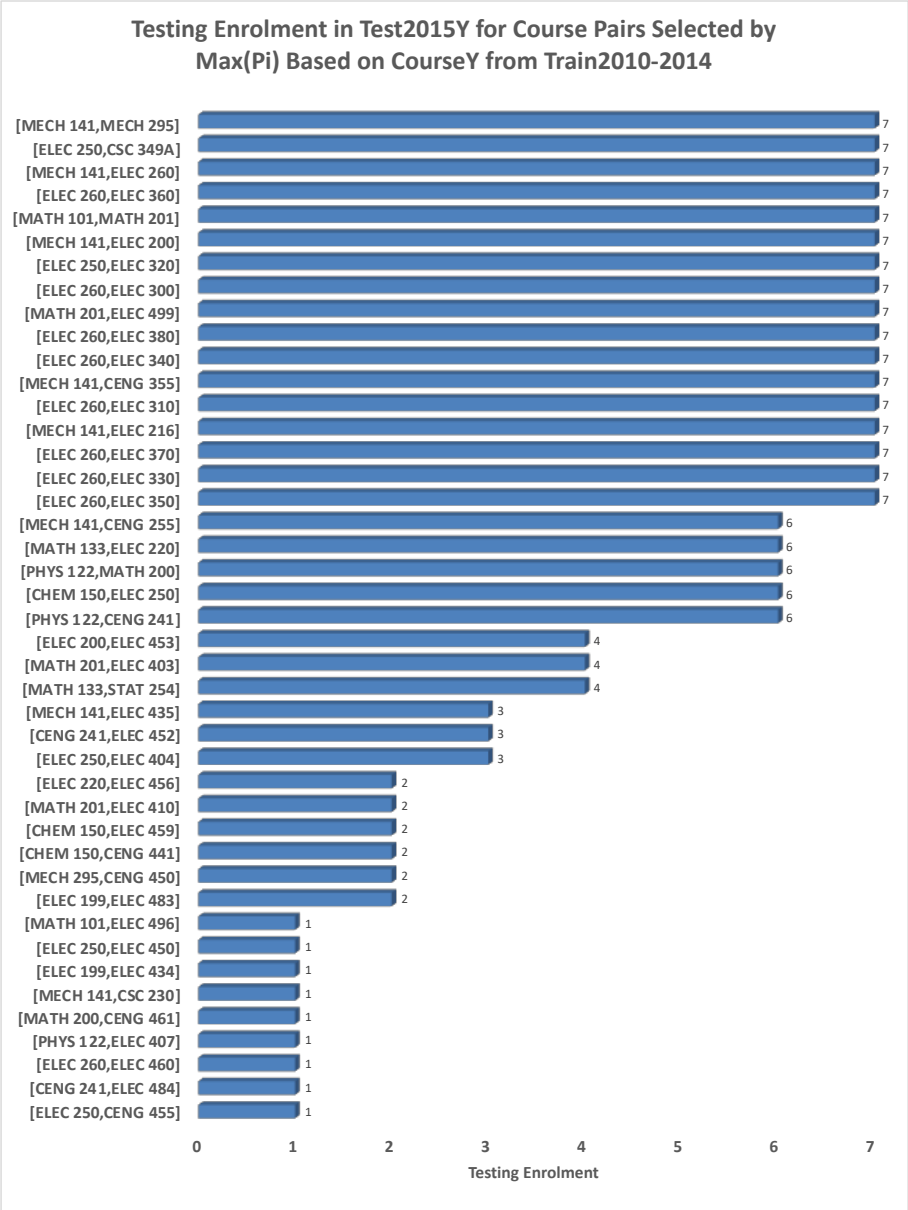

Figure 36 Testing Enrolments of course Pairs Selected by Max(P_i) Based on CourseY from Train2010-2014

4.2.3.2 MAE Analysis of Selection with Max(P_i)

The MAEs of course pairs selected by maximum of P_i based on CourseX and obtained from the testing set of Test2012-2015X are shown in Figure 37. According to the MAE acceptance criterion discussed in Section 4.1, none of them are acceptable with the

maximum of 4.2 from the course pair of MATH 200 and ELEC 420 and minimum of 1.1 from the two pairs of CSC 160 and ELEC 499, and ELEC 220 and ELEC 456.

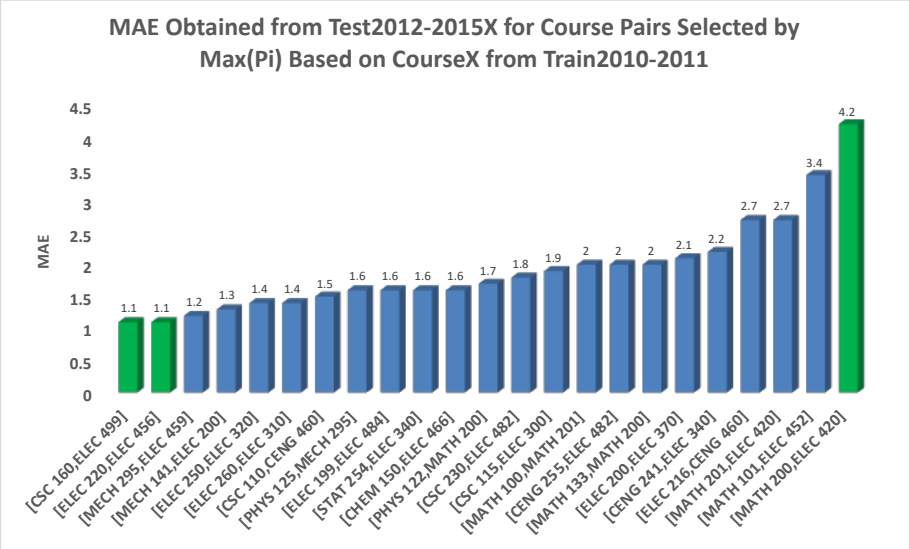

Figure 37 Prediction MAEs for Course Pairs Selected by Max(P_i) Based on CourseX from Train2010-2011 in Test2012-2015X

The statistics of the course pairs selected based on CourseX listed in Table 17 show that the testing enrolments have a wide range from 2 to 78 and none of the precisions is over 70% independent of the size of the test enrolments. That is, the relatively small testing enrolment can produce relatively small MAE, for example, the course pair of CSC 160 and ELEC 499 having MAE of 1.1 with 12 testing enrolments, and relatively big MAE, for instance, the course pair of MATH 101 and ELEC 452 having MAE of 3.4 with 14 testing enrolments (See orange rows in Table 17). Similar trends can be observed for relatively big testing enrolment, for instance, the course pair of MECH 141 and ELEC 200 having MAE of 1.3 with 75 testing enrolments, and the course pair of CENG 241 and ELEC 340 having MAE of 2.2 with 78 testing enrolments (See green rows in Table 17). Therefore, these course pairs are not reliable course pairs for use in prediction.

Table 17 Course Pairs Selected Based on CourseX in Test2012-2015X

CrsXCode	CrsXNum	CrsYCode	CrsYNum	Enrolment	0~1.0	%	MAE
CSC	160	ELEC	499	12	4	33	1.1
ELEC	220	ELEC	456	13	7	54	1.1
MECH	295	ELEC	459	13	4	31	1.2
MECH	141	ELEC	200	75	39	52	1.3
ELEC	250	ELEC	320	78	30	38	1.4
ELEC	260	ELEC	310	77	39	51	1.4
CSC	110	CENG	460	5	3	60	1.5
PHYS	125	MECH	295	68	23	34	1.6
ELEC	199	ELEC	484	11	4	36	1.6
STAT	254	ELEC	340	36	16	44	1.6
CHEM	150	ELEC	466	17	10	59	1.6
PHYS	122	MATH	200	63	25	40	1.7
CSC	230	ELEC	482	2	1	50	1.8
CSC	115	ELEC	300	61	18	30	1.9
MATH	100	MATH	201	66	13	20	2
CENG	255	ELEC	482	43	11	26	2
MATH	133	MATH	200	66	18	27	2
ELEC	200	ELEC	370	77	24	31	2.1
CENG	241	ELEC	340	78	23	29	2.2
ELEC	216	CENG	460	10	2	20	2.7
MATH	201	ELEC	420	6	2	33	2.7
MATH	101	ELEC	452	14	2	14	3.4
MATH	200	ELEC	420	6	0	0	4.2

The MAEs obtained from Test2012-2015Y for course pairs selected by Max(P_i) based on CourseY from the training set of Train2010-2011 are shown in Figure 38. There are six qualifying pairs: two course pairs with MAE < 1.0: PHYS 125 and ELEC 450 with MAE of 0.82, and MATH 133 and ELEC 484 with MAE of 0.79. MECH 141 and CENG 255 has MAE = 1.0. Three course pairs have MAE = 1.1: ELEC 220 and ELEC 456, MATH 133 and CENG 455, and CSC 160 and ELEC 499. The statistics of the six course pairs are listed in Table 18.

The course pair of MATH 133 and ELEC 484 has MAE of 0.79 and precision of 78% with 9 testing enrolments (yellow row in Table 18). Although both MAE and precision of this course pair are acceptable, they may not be used reliably for prediction because the testing enrolment is relatively small (9 testing enrolments). The MAEs of the two course pairs of PHYS 125 and ELEC 450, and MECH 141 and CENG 255, are in the acceptable range, but their precisions are 50% and 59% (green rows in Table 18). Therefore, the predictions using these two course pairs may be skewed.

The MAEs of the other three course pairs in Table 18 are 1.1 which can be treated as acceptable MAE because it is just 0.1 away from the acceptable MAE. However, there is only one-third of prediction errors in the acceptable range for CSC 160 and ELEC 499, half for MATH 133 and CENG 455, and 54% for ELEC 220 and ELEC

456 (un-highlighted rows in Table 18). The percentages for these course pairs are too low to use them in prediction.

Table 18 Course Pairs Selected Based on CourseX with MAE<= 1.1 in Test2012-2015Y

CrsXCode	CrsXNum	CrsYCode	CrsYNum	Enrolment	0~1.0	%	MAE
MATH	133	ELEC	484	9	7	78	0.79
PHYS	125	ELEC	450	6	3	50	0.82
MECH	141	CENG	255	71	42	59	1
CSC	160	ELEC	499	12	4	33	1.1
MATH	133	CENG	455	6	3	50	1.1
ELEC	220	ELEC	456	13	7	54	1.1

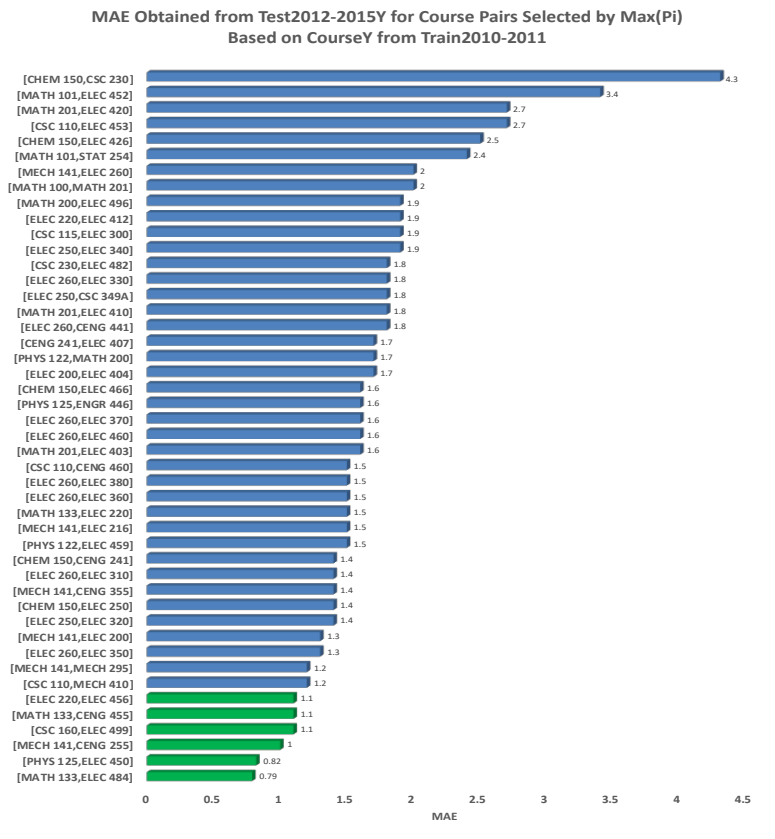

Figure 38 Prediction MAEs for Course Pairs Selected by Max(P_i) Based on CourseY from Train2010-2011 in Test2012-2015Y

The MAEs of the course pairs selected based on both CourseX and CourseY from Train2010-2012 and train2010-2013 are computed and shown in Appendix 15. The course

pairs with MAE less than or equal to 1.0 in all six figures have similar trends as the course pairs shown in Figure 37 and Figure 38 that although the MAEs of the course pairs are acceptable, their prediction precisions are relatively small, or even if both the MAE and the precision are acceptable, their testing enrolment is relatively small.

Based on the MAE assessment, the predictions of the course pairs selected from the three different training sets are not acceptable. Therefore, prediction precision is presented in the next section for prediction evaluation.

4.2.3.3 Prediction Precision Analysis of Selection with Max(P_i)

The prediction precisions for CourseX tested in the merged testing set of Test2012-2015X are shown in Figure 39. The highest precision is 60% from the course pair of CSC 110 and CENG 460 with 5 testing enrolments and the lowest is 0% from the course pair of MATH 200 and ELEC 420 with 6 testing enrolments. Also, the small testing enrolment produced both high and low precisions, for instances, precision of 60% is obtained from course pair of CSC 110 and CENG 460 with 5 testing enrolments, and precision of 0% from course pair of MATH 200 and ELEC 420 with 6 testing enrolments. Bigger testing enrolments have similar trend that it can produce both high and low precisions, for instances, 52% from MECH 141 and ELEC 200 with 75 testing enrolments and 29% from CENG 241 and ELEC 340 with 78 testing enrolments. Therefore, the predictions using the course pairs in Figure 39 are not acceptable.

The prediction precisions of the course pairs selected based on CourseY are presented in the Figure 40. The maximum precision is 78% from the course pair of MATH 133 and ELEC 484 with 9 testing enrolments and it is the only precision that is over 70%. However, because of the relatively small testing enrolment of 9 as compared to other course pairs, this prediction may not be reliable.

It can also be seen from Figure 40 that the testing enrolment has the same trend as the testing enrolment shown in Figure 39 that small testing enrolment can produce high and low precisions: 78% from MATH 133 and ELEC 484 with 9 testing enrolments, 0% from CHEM 150 and CSC 230 with 2 testing enrolments, and CHEM 150 and ELEC 426 with 5 testing enrolments. The biggers testing enrolments produce high and low precisions: 59%

from MECH 141 and CENG 255 with 71 testing enrolments, 20% from MATH 100 and MATH 201 with 66 testing enrolments. Therefore, the predictions of the course pairs selected based on CourseY are not reliable, either. In other words, the predictor selection of using Max(P_i) does not work well.

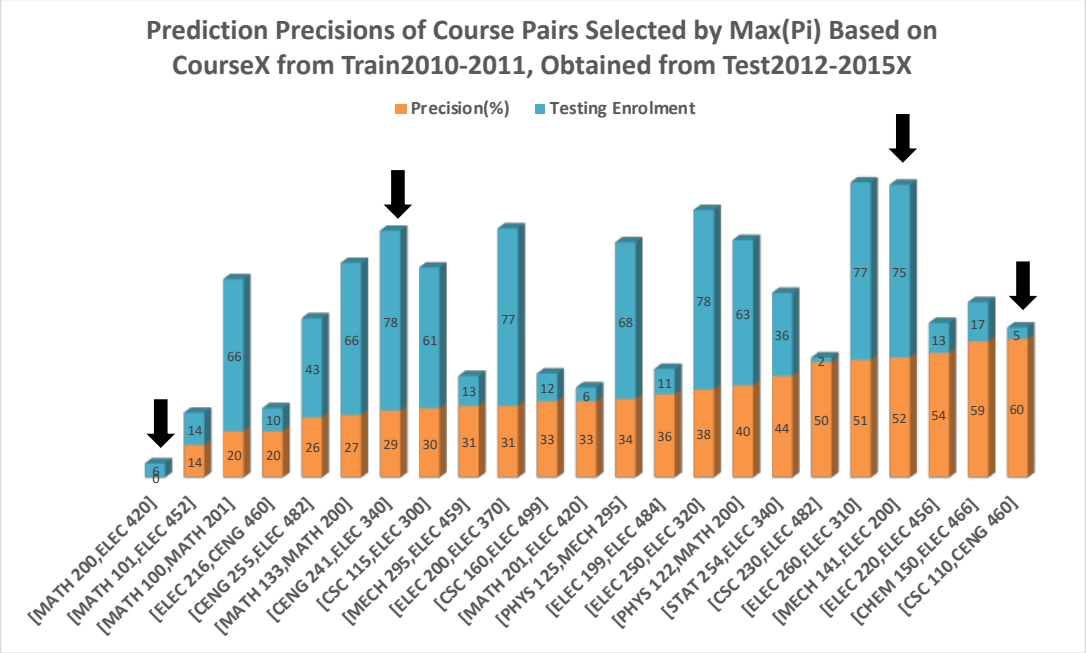

Figure 39 Prediction Precisions of Course Pairs in Test2012-2015X, Trained by Train2010-2011

Similar analysis was done for course pairs selected based on CourseX and CourseY from the other two training sets of Train2010-2012 and Train2010-2013. The precisions of these course pairs selected from the two training sets are shown in Appendix 16. These figures show similar results as the ones from the testing set of Test2012-2015X and Test2012-2015Y shown in Figure 39 and Figure 40. The precisions of course pairs over 70% have relatively small testing enrolments and a small testing enrolment can produce both high and low precisions, and bigger testing enrolment can also produce high and low precisions.

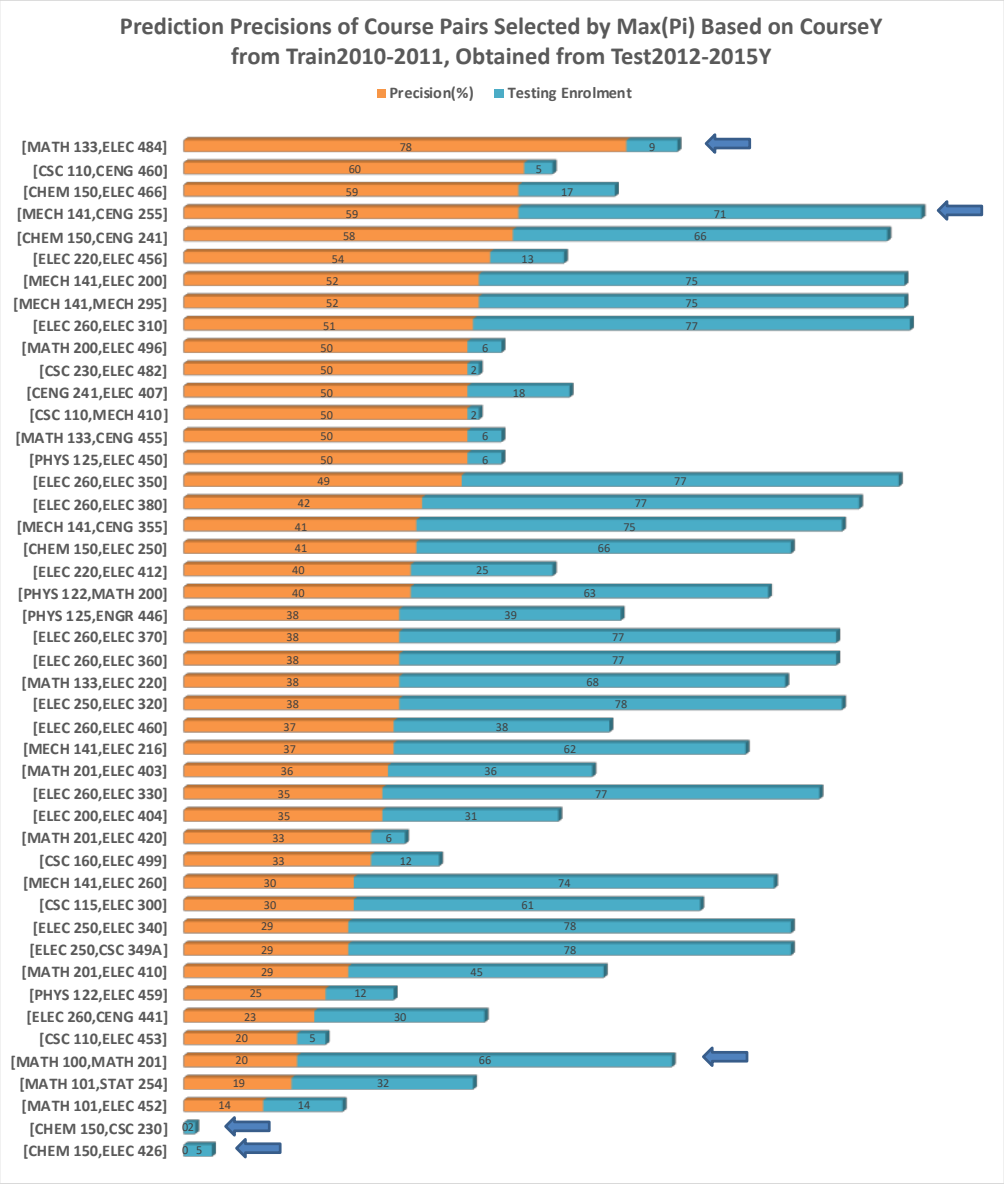

Figure 40 Prediction Precisions of Course Pairs in Test2012-2015Y, Trained by Train2010-2011

Although this predictor selection approach considers both Pearson Coefficient and enrolment to balance the influence of these factors, it does not work well either for CourseX or CourseY to assess one course's performance based on another course. The poor prediction results may be due to other course- or student-related factors which are are not

considered, such as family financial background, parents' education background, and prior institutional information.

4.3 Comparing the Three Predictor Selection Approaches

This section compares the results produced by the three predictor selection methods. The metrics used include the selected courses, the MAE trends and prediction precisions.

4.3.1 Selected Course Pairs Comparison

The predictors selected by Max(Pearson Coefficient) discussed in Section 4.2.1 show that the predicted courses selected based on an earlier year CourseX are prone to be the fourth-year courses because of the fact that the coefficients from the course pairs with fourth-year courses are more likely to have relatively bigger coefficients in the research training datasets. In other words, the coefficient dominates in the predictor selection for the course pairs with fourth-year courses. Therefore, the selected course pairs may be biased, which may cause later analysis to be not reliable.

The second predictor selection approach presented in Section 4.2.2 is to use the enrolment as the only indicator to select the predictor course. The predicted courses selected by this approach based on CourseX show completely different properties as compared with the ones selected by Max(Pearson Coefficient). In this case, the fourth-year courses are removed from predicted course candidates because of the relatively small enrolments of the course pairs with fourth-year courses. Because of the enrolment domination in the predictor selection, the course pair selection would be skewed, too.

The last approach studied in Section 4.2.3 eliminates the coefficient domination (Section 4.2.1) and enrolment domination (Section 4.2.2) in the predictor selection because both factors are employed to balance each other in the predictor selection. The predictors selected by using this approach based on CourseX show that some of the fourth-year courses selected were removed in the second approach, Max(Enrolment) in Section 4.2.2. Though, the number of removed fourth-year courses is less than the ones selected using

the first approach, Max(Pearson Coefficient) in Section 4.2.1. Then, the selected course pairs are relatively balanced as compared to the ones generated by the other two approaches.

4.3.2 MAE Trends Comparison

The MAEs of most course pairs produced by the three approaches are greater than the acceptable threshold, 1.0. Course pairs with MAE <= 1.0 have relatively low precisions, such as the course pair of MECH 141 and CENG 255 having precision of 59% (42 out of 71 testing enrolments) with MAE = 1.0 as discussed in Section 4.2.1.2. Such course pairs prediction accuracy is not acceptable. As stated earlier, the MAE only represents the average prediction error, but not how the prediction errors are distributed. Therefore, a prediction may be skewed if MAE is the only assessment used.

There exist course pairs having precision over 70% with MAE <= 1.0, for instance, the pair of MATH 133 and ELEC 484 having precision of 78% (7 out of 9 testing enrolments) with MAE = 0.79 in Section 4.2.1.2. However, such course pairs have relatively small testing enrolments. Therefore, these course pairs may not be used in prediction reliably, either.

4.3.3 Prediction Precisions Comparison

No matter which one of the threes predictor-selection method is used, the prediction precisions for course pairs with CourseY in Year 2 or Year 3 are under 70% in most cases. Although there are some courses from these two years with prediction precisions above 70%. For instance, the course pair of CSC 160 and ELEC 360 has precision of 75% with 4 testing enrolments as presented in Section 4.2.3.3, but their testing enrolments are relatively small. For the course pairs with CourseY in Year 4, there are more prediction accuracies over 70%, such as the pair of ELEC 250 and CENG 455 having precision of 71% with 7 testing enrolments, as shown in Section 4.2.1.3. But, again, the testing enrolments are relatively small. Therefore, the prediction precisions are not acceptable and such course pairs also cannot be used reliably in prediction.

There are three valid training sets with incremental training size and the predictions are analyzed. As the training size increased, the prediction results, measured using MAE

and prediction precision, did not improve. Therefore, the training size does not have much influence on the prediction as typically would be expected.

In conclusion, although the coefficient domination and enrolment domination are balanced in the third prediction selection approach and, also, the training size is increased, it is hardly helpful in using use a course's grade to predict another course grade by linear regression. It is better to consider additional factors, such as a student's performance in mathematical subjects at high school. Also, to predict performance in a course, other algorithm should be investigated, for instance, Bayesian, classification, or association.

Chapter 5 Conclusions and Future Work

This chapter concludes the research done including the Pearson Coefficient introduced and results obtained in Chapter 3, the three predictor selection methods, and their prediction results evaluated by MAE and prediction precision. Also, based on the research conclusions, possible future work is discussed.

In order to protect a student's privacy, critical information, such as student number, is hidden from the analysis by using a well-defined encryption algorithm as discussed in Section 2.2. First, the course record order exported from the metaserver is shuffled to avoid student identification by the ordering of student records. Then, the student numbers in each course record are replaced by a random number which was encrypted by SHA256. The last step was to shuffle the reordered and V# encrypted course records. This way the course records are free of sensitive information and are ready for further analysis.

All technical courses were extracted from the training sets. Each course was paired with another course from a different year in Chapter 3. Then, the Pearson Coefficients of these course pairs were computed and the strength of each coefficient was determined by a p-value of 0.05. The strongly correlated course pairs from each training set were selected using the criterion of p-value <= 0.05. The statistics of the selected strongly correlated course pairs from each training set show that each earlier year CourseX has at least one strongly correlated later year CourseY in the course pair, and each CourseY has at least one strongly correlated CourseX.

The histograms of the selected coefficients with p-value <= 0.05 from the four training sets show that the coefficients clustered around 0.5 in most cases and minor portions of the coefficients have extreme values at the two ends of the histograms. The relationship between the coefficient and enrolment shown in Chapter 3 shows that the coefficient decreases as the enrolment increases, which indicates that more enrolments generate a more reliable coefficient.

As shown in Chapter 3, the enrolment has influence on the coefficient. Thus, three predictor selection approaches were investigated based on the two factors, coefficient and enrolment, to select predictor or predicting course as discussed in Chapter 4.

The predictor selection results generated using only maximum coefficient as the criterion show that CourseX is prone to be paired with a course from the fourth year. As shown in Chapter 3, a course pair with a fourth-year course typically has a relatively bigger coefficient close to +1.0 or –1.0.

The approach using maximum enrolment as the predictor selection criterion eliminates the fourth-year courses from the predictor candidates as these courses are specialization-related and have relatively small enrolments.

The third approach uses a combination of coefficient and enrolment as an attempt to balance the influence of the two factors on the predictor selection, thus eliminating a heavier domination by either one of the two factors.

As stated in the testing enrolment distribution section in Chapter 4, the testing sets with one-year course data were merged as one because some course pairs have relatively small testing enrolments and some have no testing enrolment. To avoid the issues that may cause problems in prediction later on, they were merged together as one.

The MAE statistics from the three predictor selection methods show that the MAEs are greater than 1.0 in most cases and a small portion of MAEs are in the acceptable range of [0, 1.0]. However, although some course pairs have MAE <= 1.0, their prediction precisions are relatively low. There are still a few course pairs with MAE <= 1.0 and precision >= 70%. But the testing enrolments of such course pairs are relatively small, which indicates these course pairs may not be used reliably in prediction.

The prediction precisions generated by the three predictor selection methods have similar trends as the MAEs'. The precision statistics show that in most cases it is under 70%, and a few are above 70% but with relatively small testing enrolments. Because of the relatively small enrolments, the course pair with precision equal to or greater than 70% may not be used reliably in prediction, either.

As stated in Chapter 3, the research dataset was split into four different training sets with one-year course data increment, the three predictor selection methods were applied to all of the training sets except Train2010-2014 which was shown to be invalid in Chapter 4. The MAE and prediction precision analysis were also done in the testing sets of the three valid training sets. However, the statistics of both MAEs and prediction precisions did not

improve when the training size increased. Therefore, the training size does not have much influence in the prediction.

In conclusion, the predictions of most course pairs cannot provide helpful information in the student performance prediction. However, as the fourth-year courses are specialization-related and have relatively small enrolments in general, the course pairs with CourseYs from the fourth-year and having acceptable MAEs and prediction precisions could be used as references and advice for the students to select the study directions while in their first or second academic year.

In this research, only two attributes, the enrolment and the Pearson Correlation Coefficient, were used in prediction. The prediction statistics show that none of the predictions is acceptable or reliable. It implies that additional attributes should be explored in grade prediction.

Moreover, linear regression is the only prediction algorithm applied in the research to estimate one course's grade based on its predicting course's grade. Other algorithms should be investigated in future research.

Bibliography

[1] Z. Ibrahim, D. Rushli, Predicting Students' Academic Performance: Comparing Artificial Neural Network, Decision Tree and Linear Regression, *21st Annual SAS Malaysia Forum, 5th September 2007, Shangri-La Hotel, Kuala Lumpur*

[2] C. Lei, K.F. Li, "Academic performance predictors", *29th International Conference on Advanced Information Networking and Applications Workshops*, 2015

[3] S. B. Huang, N. Fang, Predicting Student Academic Performance in An Engineering Dynamics Course: A Comparison of Four Types of Predictive Mathematical Models, *Computer & Education 61 (2013)* 133-145

[4] M. F. Mussoab, E. Kyndtac, E. C. Cascallarad, F. Dochya, Predicting General Academic Performance and Identifying the Differential Contribution of Participating Variables Using Artificial Neural Networks, *Frontline Learning Research 1 (2013)* 42 - 71 ISSN 2295-3159

[5] J. C. F. DE WINTER and D. DODOU, Predicting Academic Performance in Engineering Using High School Exam Scores, *International Journal of Engineering Education* Vol. 27, No. 6, pp. 1343–1351, 2011

[6] K. F. Li, D. Rusk and F. Song, Predicting Student Academic Performance, DOI 10.1109/CISIS.2013.15

[7] J. F. Chen, H. N. Hsieh and Q. H. Do, Predicting Student Academic Performance: A Comparison of Two Meta-Heuristic Algorithms Inspired by Cuckoo Birds for Training Neural Networks, *Algorithms 2014*, 7, 538-553; doi:10.3390/a7040538

[8] R. Asif, A. Merceron and M. K. Pathan, Predicting Student Academic Performance at Degree Level: A Case Study, DOI: 10.5815/ijisa.2015.01.05

[9] V.O. Oladokun, A.T. Adebanjo, and O.E. Charles-Owaba, Predicting Students' Academic Performance using Artificial Neural Network: A Case Study of an Engineering Course, Volume 9. Number 1. May-June 2008 (Spring)

[10] S. B. Huang, "Predictive Modeling and Analysis of Student Academic Performance in an Engineering Dynamics Course" (2011). All Graduate Theses and Dissertations. Paper 1086.

[11] B. K. Bhardwaj, S. Pal, Data Mining: A prediction for performance improvement using classification, *(IJCSIS) International Journal of Computer Science and Information Security*, Vol. 9, No. 4, April 2011

[12] S. K. Yadav, S. Pal, Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification, *World of Computer Science and Information Technology Journal (WCSIT)* ISSN: 2221-0741 Vol. 2, No. 2, 51-56, 2012

[13] B. Minaei-Bidgoli, D. A. Kashy, G. Kortemeyer, W. F. Punch, Predicting Student Performance: An Application of Data Mining Methods with The Educational Web-Based System Lon-Capa, *33rd ASEE/IEEE Frontiers in Education Conference*

[14] E. Osmanbegović, M. Suljic, Data Mining Approach for Predicting Student Performance, *Economic Review – Journal of Economics and Business*, Vol. X, Issue 1, May 2012

[15] T. J. Pleskac, A. Q. Billington, R. Sinha, M. Zorzie, N. Schmitt, Jessica Keeney and F. L. Oswald, Prediction of 4-Year College Student Performance Using Cognitive and Noncognitive Predictors and the Impact on Demographic Status of Admitted Students, *Journal of Applied Psychology, 2009*, Vol. 94, No. 6, 1479 –1497

[16] V. Ramesh, P. Parkavi, K. Ramar, Predicting Student Performance: A Statistical and Data Mining Approach, *International Journal of Computer Applications (0975 – 8887)* Volume 63– No.8, February 2013

[17] H. Agrawal, H. Mavani, Student Performance Prediction using Machine Learning, *International Journal of Engineering Research & Technology (IJERT)* ISSN: 2278-0181 Vol. 4 Issue 03, March-2015

[18] A. B. E. D. Ahmed, I. S. Elaraby, Data Mining: A prediction for Student's Performance Using Classification Method, *World Journal of Computer Application and Technology* 2(2): 43-47, 2014, DOI: 10.13189/wjcat.2014.020203

[19] P. Cortez, A. Silva, Using Data Mining To Predict Secondary School Student Performance, BRITO, A.; TEIXEIRA, J., eds. lit. – "*Proceedings of 5th Annual Future*

Business Technology Conference, Porto, 2008". [S.l. : EUROSIS, 2008]. ISBN 978-9077381-39-7. p. 5-12.

[20] F. Berhanu, A. Abera, Students' Performance Prediction based on their Academic Record, *International Journal of Computer Applications (0975 – 8887)* Volume 131 – No.5, December2015

[21] A. AL-Malaise, A. Malibari and M. Alkhozae, Students' Performance Prediction System Using Multi Agent Data Mining Technique, *International Journal of Data Mining & Knowledge Management Process (IJDKP)* Vol.4, No.5, September 2014

[22] Y. Freund and R. Schapire, 'A decision-theoretic generalization of on-line learning and an application to boosting', *Journal of Computer and System Sciences*, vol 55, iss 1, pp.119--139, 1997.

[23] J. Friedman, T. Hastie, and R. Tibshirani, 'Additive logistic regression: a statistical view of boosting', *The Annals of Statistics*, vol 28, iss 2, pp. 337-- 407, 2000.

[24] J. Zhu, H. Zou, S. Rosset and T. Hastie, 'Multi-class adaboost', *Statistics and Its Interface*, vol 2, pp. 349--360, 2009.

[25] Cuckoo Search: https://en.wikipedia.org/wiki/Cuckoo_search. Access Date: 2019-03-08.

[26] S. Jafari, O. Bozorg-Haddad, X. Chu, (2018) Cuckoo Optimization Algorithm (COA). In: Bozorg-Haddad O. (eds) *Advanced Optimization by Nature-Inspired Algorithms. Studies in Computational Intelligence*, vol 720.

[27] Human Research Ethics at UVic: http://www.uvic.ca/research/conduct/home/regapproval/humanethics/. Access Date: 2019-03-08.

[28] SAS homepage: http://www.sas.com/en_ca/home.html. Access Date: 2019-03-08.

[29] SAS software: https://en.wikipedia.org/wiki/SAS_(software) . Access Date: 2019-03-08.

[30] Introduction to SAS Enterprise Guide: http://support.sas.com/publishing/pubcat/chaps/61625.pdf. Access Date: 2019-03-08.

[31] SHA-2: https://en.wikipedia.org/wiki/SHA-2. Access Date: 2019-03-08.

[32] UVic Calendar 2005-2006 for BEng in Electrical Engineering: http://web.uvic.ca/calendar2005/FACS/FoEn/DoElaCE/ASBEiEE.html. Access Date: 2019-03-08.

[33] UVic Calendar 2006-2007 for BEng in Electrical Engineering: http://web.uvic.ca/calendar2006/FACS/FoEn/DoElaCE/ASBEiEE.html. Access Date: 2019-03-08.

[34] UVic Calendar 2007-2008 for BEng in Electrical Engineering: http://web.uvic.ca/calendar2007/FACS/FoEn/DoElaCE/ASBEiEE.html. Access Date: 2019-03-08.

[35] UVic Calendar 2008-2009 for BEng in Electrical Engineering: http://web.uvic.ca/calendar2008/FACS/FoEn/DoElaCE/ASBEiEE.html. Access Date: 2019-03-08.

[36] UVic Calendar 2009-2010 for BEng in Electrical Engineering: http://web.uvic.ca/calendar2009/FACS/FoEn/DoElaCE/ASBEiEE.html. Access Date: 2019-03-08.

[37] UVic Calendar 2010-2011 for BEng in Electrical Engineering: http://web.uvic.ca/calendar2010/FACS/FoEn/DoElaCE/ASBEiEE.html. Access Date: 2019-03-08.

[38] UVic Calendar 2011-2012 for BEng in Electrical Engineering: http://web.uvic.ca/calendar2011/FACS/FoEn/DoElaCE/ASBEiEE.html. Access Date: 2019-03-08.

[39] UVic Calendar 2012-2013 for BEng in Electrical Engineering: http://web.uvic.ca/calendar2012/FACS/FoEn/DoElaCE/ASBEiEE.html. Access Date: 2019-03-08.

[40] UVic Calendar 2013-2014 for BEng in Electrical Engineering: http://web.uvic.ca/calendar2013/FACS/FoEn/ElEn/AcSc.html. Access Date: 2019-03-08.

[41] UVic Calendar 2014-2015 for BEng in Electrical Engineering: http://web.uvic.ca/calendar2014/FACS/FoEn/ElEn/index.html. Access Date: 2019-03-08.

[42] UVic Calendar 2015-2016 for BEng in Electrical Engineering: https://web.uvic.ca/calendar2016-01/undergrad/engineering/elec.html#. Access Date: 2019-03-08.

[43] Grading in UVic Calendar 2005-2006: http://web.uvic.ca/calendar2005/FACS/UnIn/UARe/Grad.html. Access Date: 2019-03-08.

[44] Grading in UVic Calendar 2006-2007: http://web.uvic.ca/calendar2006/FACS/UnIn/UARe/Grad.html. Access Date: 2019-03-08.

[45] Grading in UVic Calendar 2007-2008: http://web.uvic.ca/calendar2007/FACS/UnIn/UARe/Grad.html. Access Date: 2019-03-08.

[46] Grading in UVic Calendar 2008-2009: http://web.uvic.ca/calendar2008/FACS/UnIn/UARe/Grad.html. Access Date: 2019-03-08.

[47] Grading in UVic Calendar 2009-2010: http://web.uvic.ca/calendar2009/FACS/UnIn/UARe/Grad.html. Access Date: 2019-03-08.

[48] Grading in UVic Calendar 2010-2011: http://web.uvic.ca/calendar2010/FACS/UnIn/UARe/Grad.html. Access Date: 2019-03-08.

[49] Grading in UVic Calendar 2011-2012: http://web.uvic.ca/calendar2011/FACS/UnIn/UARe/Grad.html. Access Date: 2019-03-08.

[50] Grading in UVic Calendar 2012-2013: http://web.uvic.ca/calendar2012/FACS/UnIn/UARe/Grad.html. Access Date: 2019-03-08.

[51] Grading in UVic Calendar 2013-2014: http://web.uvic.ca/calendar2013/FACS/UnIn/UARe/Grad.html. Access Date: 2019-03-08.

[52] Grading in UVic Calendar 2014-2015: http://web.uvic.ca/calendar2014/FACS/UnIn/UARe/Grad.html. Access Date: 2019-03-08.

[53] Grading in UVic Calendar 2015-2016: https://web.uvic.ca/calendar2016-01/undergrad/info/regulations/grading.html#. Access Date: 2019-03-08.

[54] Pearson correlation coefficient: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient. Access Date: 2019-03-08.

[55] Pearson Correlation: https://libguides.library.kent.edu/SPSS/PearsonCorr. Access Date: 2019-03-08.

[56] Pearson Product-Moment Correlation: https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php. Access Date: 2019-03-08.

[57] Null hypothesis: http://psc.dss.ucdavis.edu/sommerb/sommerdemo/stat_inf/null.htm. Access Date: 2019-03-08.

[58] One- and two tailed tests: https://en.wikipedia.org/wiki/One-_and_two-tailed_tests. Access Date: 2019-03-08.

[59] Spearman's rank correlation coefficient: https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient. Access Date: 2019-03-08.

[60] Kendall rank correlation coefficient: https://en.wikipedia.org/wiki/Kendall_rank_correlation_coefficient. Access Date: 2019-03-08.

[61] *p*-value, https://en.wikipedia.org/wiki/P-value. Access Date: 2019-03-08.

[62] Polynomial regression: https://en.wikipedia.org/wiki/Polynomial_regression. Access Date: 2019-03-08.

[63] Linear Regression: https://en.wikipedia.org/wiki/Linear_regression. Access Date: 2019-03-08.

[64] Mean Absolute Error: https://en.wikipedia.org/wiki/Mean_absolute_error. Access Date: 2019-03-08.

[65] Normalization: http://www.statisticshowto.com/normalized/. Access Date: 2019-03-08.

Appendix 1 Technical Courses in Calendar

All compulsory technical courses and technical electives in the UVic calendars from 2005-2006 to 2009-2010 of BEng in Electrical Engineering:

Year	Course List	# of Courses
Year 1	CHEM150 CSC110, CSC111, CSC115, CSC160 ELEC199 ENGR110(111)[1] MATH100, MATH101, MATH133(110) MECH141(ENGR141) PHYS122, PHYS125	13
Year 2	CENG241, CENG255, CENG290 CSC230 ELEC200, ELEC216, ELEC220, ELEC250, ELEC260 MATH200, MATH201 MECH295 STAT254	13
Year 3	CENG355 CSC349A ELEC300, ELEC310, ELEC320, ELEC330, ELEC340, ELEC350 ELEC360, ELEC370, ELEC380 ENGR280	12
Year 4	ENGR446, ENGR297 CENG412, CENG420, CENG421, CENG440, CENG441, CENG442	71

[1] Course number in the round brackets was changed or renamed as shown in later academic schedule, though they are still the same course with very similar materials.

	CENG450, CENG453, CENG455, CENG 460, CENG461, CENG465 CENG496, CENG499, CENG499A, CENG499B CSC349B, CSC405, CSC450, CSC454 ELEC400, ELEC403, ELEC404, ELEC405, ELEC407, ELEC408 ELEC410, ELEC412, ELEC420, ELEC426, ELEC434, ELEC435 ELEC450, ELEC452, ELEC453, ELEC454, ELEC456, ELEC459 ELEC460, ELEC466, ELEC481, ELEC482, ELEC483, ELEC484 ELEC485, ELEC486, ELEC496, ELEC499A, ELEC499B, ELEC499 ELEC395 MECH410, MECH460, MECH486 SENG330, SENG365, SENG410, SENG412, SENG422, SENG426 SENG440, SENG454, SENG460, SENG461, SENG462, SENG466 SENG474, SENG480, SENG490	
Total		109

Appendix 2 Technical Courses in Dataset

The table below show all the technical courses in research data grouped by academic year:

Year	Course List	# of Courses
Year 1	CHEM150 CSC110, CSC111, CSC115, CSC160 ELEC199 MATH100, MATH101, MATH133 MECH141(ENGR141) PHYS122, PHYS125	12
Year 2	CENG241, CENG255 CSC230 ELEC200, ELEC216, ELEC220, ELEC250, ELEC260 MATH200, MATH201 MECH295 STAT254	12
Year 3	CENG355 CSC349A ELEC300, ELEC310, ELEC320, ELEC330, ELEC340, ELEC350 ELEC360, ELEC370, ELEC380	11
Year 4	CENG412, CENG420, CENG421, CENG441, CENG450 CENG453, CENG455, CENG460, CENG461 ELEC403, ELEC404, ELEC405, ELEC407, ELEC410, ELEC412 ELEC420, ELEC426, ELEC434, ELEC435, ELEC450, ELEC452	45

	ELEC453, ELEC454, ELEC456, ELEC459, ELEC460, ELEC466 ELEC481, ELEC482, ELEC483, ELEC484, ELEC485, ELEC486 ELEC496, ELEC499 ENGR446 MECH410, MECH460 SENG330, SENG422, SENG426, SENG440, SENG460, SENG461 SENG466	
Total		80

Appendix 3 Enrolments of Technical Courses in Dataset

All technical courses in the research dataset with the corresponding enrolments (the number of students who registered in and completed the course with a given grade) and years in research data:

Subject	Number	Frequency	Year		Subject	Number	Frequency	Year
ELEC	199	120	1		ELEC	460	61	4
MECH	141	115	1		ELEC	404	56	4
PHYS	125	103	1		ELEC	403	53	4
MATH	101	99	1		CENG	441	50	4
PHYS	122	97	1		ELEC	412	35	4
MATH	133	96	1		ELEC	407	33	4
CHEM	150	94	1		ELEC	453	30	4
MATH	100	90	1		ELEC	456	26	4
CSC	115	67	1		ELEC	450	26	4
CSC	111	55	1		ELEC	466	26	4
CSC	110	51	1		CENG	460	26	4
CSC	160	47	1		ELEC	452	20	4
ELEC	250	120	2		ELEC	484	20	4
CENG	241	120	2		ELEC	459	19	4
ELEC	220	120	2		ELEC	481	19	4
MECH	295	120	2		MECH	410	18	4
ELEC	260	119	2		ELEC	496	17	4
ELEC	200	118	2		ELEC	434	15	4
MATH	201	113	2		CENG	455	15	4
MATH	200	111	2		ELEC	435	14	4
CENG	255	105	2		ELEC	486	13	4
ELEC	216	95	2		CENG	421	13	4
STAT	254	57	2		ELEC	426	12	4
CSC	230	16	2		ELEC	483	11	4
ELEC	370	120	3		ELEC	420	11	4
ELEC	310	120	3		CENG	461	9	4
ELEC	350	120	3		SENG	440	9	4
ELEC	380	120	3		CENG	450	8	4
ELEC	320	120	3		SENG	466	6	4
ELEC	300	120	3		CENG	420	4	4
ELEC	340	120	3		CENG	412	3	4
ELEC	360	120	3		ELEC	405	3	4
CSC	349A	120	3		MECH	460	3	4
CENG	355	120	3		ELEC	485	2	4
ELEC	330	120	3		CENG	453	2	4
ELEC	499	114	4		SENG	461	2	4
ENGR	446	89	4		SENG	460	2	4
ELEC	410	69	4		SENG	422	1	4
ELEC	482	61	4		SENG	426	1	4

Appendix 4 Strongly Correlated Course Pairs in Train2010-2011

In the following table, it presents all possible strongly correlated course pairs selected based on the p-value. The p-value is set to be 0.05 which is widely used in statistics. Ther are 265 course pairs in total with p-value under 0.05.

xsubCode	xnum	ysubCode	ynum	coefficient	#points	pValue
MATH	100	MATH	200	0.7054	22	0.0002
MATH	100	MATH	201	0.7257	21	0.0002
MATH	100	ELEC	216	0.4669	21	0.0329
MATH	100	ELEC	220	0.427	22	0.0475
MATH	100	ELEC	250	0.6261	22	0.0018
MATH	100	STAT	254	0.6066	12	0.0365
MATH	100	ELEC	260	0.4348	22	0.0432
MATH	100	ELEC	300	0.4639	22	0.0297
MATH	100	ELEC	310	0.474	22	0.0259
MATH	100	ELEC	320	0.509	22	0.0155
MATH	100	ELEC	340	0.5627	22	0.0064
MATH	100	ELEC	360	0.4448	22	0.0381
MATH	100	ELEC	403	0.762	9	0.017
MATH	101	MATH	200	0.5411	29	0.0024
MATH	101	MATH	201	0.6077	27	0.0008
MATH	101	STAT	254	0.6407	17	0.0056
MATH	101	ELEC	340	0.4875	29	0.0073
MATH	101	CSC	349A	0.4159	29	0.0248
MATH	101	ELEC	370	0.3843	29	0.0395
MATH	101	ELEC	403	0.7348	10	0.0155

MATH	101	ELEC	410	0.6303	13	0.0209
MATH	101	ELEC	452	-0.9325	5	0.0208
CSC	110	MATH	200	0.3972	28	0.0363
CSC	110	ELEC	200	0.5524	29	0.0019
CSC	110	ELEC	216	0.3983	27	0.0396
CSC	110	ELEC	220	0.372	30	0.043
CSC	110	ELEC	260	0.484	30	0.0067
CSC	110	ELEC	330	0.3776	30	0.0397
CSC	110	ELEC	340	0.3915	30	0.0324
CSC	110	CENG	355	0.376	30	0.0406
CSC	110	MECH	410	0.6909	10	0.0269
CSC	110	ELEC	453	0.8365	8	0.0096
CSC	110	CENG	460	0.9865	6	0.0003
CSC	110	ELEC	484	0.813	6	0.0492
CSC	115	ELEC	300	0.8807	6	0.0205
PHYS	122	MATH	200	0.6994	29	0
PHYS	122	ELEC	200	0.449	30	0.0128
PHYS	122	MATH	201	0.4973	29	0.0061
PHYS	122	ELEC	216	0.5576	27	0.0025
PHYS	122	CENG	241	0.5273	31	0.0023
PHYS	122	STAT	254	0.5392	17	0.0255
PHYS	122	ELEC	260	0.5223	31	0.0026
PHYS	122	ELEC	310	0.4593	31	0.0093
PHYS	122	ELEC	330	0.4172	31	0.0195
PHYS	122	ELEC	340	0.5086	31	0.0035
PHYS	122	CSC	349A	0.3608	31	0.0462
PHYS	122	ELEC	407	0.623	11	0.0406
PHYS	122	ELEC	459	0.9623	4	0.0377
PHYS	122	ELEC	466	0.9535	4	0.0465

PHYS	122	ELEC	482	0.5924	12	0.0424
PHYS	122	ELEC	484	0.8427	6	0.0352
PHYS	125	MATH	200	0.5417	33	0.0011
PHYS	125	MATH	201	0.4276	32	0.0146
PHYS	125	ELEC	216	0.5306	27	0.0044
PHYS	125	CENG	241	0.4991	35	0.0023
PHYS	125	ELEC	250	0.4505	35	0.0066
PHYS	125	ELEC	260	0.4073	35	0.0152
PHYS	125	MECH	295	0.5446	35	0.0007
PHYS	125	ELEC	340	0.4523	35	0.0064
PHYS	125	ELEC	360	0.3952	35	0.0188
PHYS	125	ELEC	407	0.7459	12	0.0053
PHYS	125	ENGR	446	0.421	35	0.0118
PHYS	125	ELEC	450	0.5159	16	0.0408
MATH	133	MATH	200	0.6748	28	0.0001
MATH	133	MATH	201	0.5563	26	0.0032
MATH	133	ELEC	216	0.456	26	0.0192
MATH	133	ELEC	220	0.4545	28	0.0151
MATH	133	STAT	254	0.5268	17	0.0298
MATH	133	ELEC	260	0.4576	28	0.0144
MATH	133	ELEC	330	0.4425	28	0.0184
MATH	133	CENG	455	0.9898	5	0.0012
MATH	133	CENG	460	0.9788	5	0.0037
MATH	133	ELEC	484	0.8641	6	0.0265
MECH	141	MATH	200	0.3654	36	0.0284
MECH	141	ELEC	200	0.6553	39	0
MECH	141	MATH	201	0.4289	36	0.009
MECH	141	ELEC	216	0.7036	32	0
MECH	141	CSC	230	0.6831	11	0.0205

MECH	141	CENG	241	0.4449	40	0.004
MECH	141	ELEC	250	0.323	40	0.042
MECH	141	CENG	255	0.4774	30	0.0076
MECH	141	ELEC	260	0.4948	40	0.0012
MECH	141	MECH	295	0.573	40	0.0001
MECH	141	ELEC	300	0.3746	40	0.0172
MECH	141	ELEC	310	0.4039	40	0.0098
MECH	141	ELEC	320	0.3753	40	0.017
MECH	141	ELEC	330	0.3473	40	0.0281
MECH	141	ELEC	340	0.4971	40	0.0011
MECH	141	CENG	355	0.4252	40	0.0062
MECH	141	ELEC	360	0.5203	40	0.0006
MECH	141	ELEC	370	0.4699	40	0.0022
MECH	141	ELEC	407	0.641	14	0.0135
MECH	141	ELEC	420	0.9068	5	0.0336
MECH	141	CENG	441	0.4632	19	0.0458
MECH	141	ELEC	460	0.6212	21	0.0026
CHEM	150	MATH	200	0.4859	27	0.0102
CHEM	150	MATH	201	0.5193	27	0.0055
CHEM	150	CSC	230	0.902	6	0.0139
CHEM	150	CENG	241	0.5659	28	0.0017
CHEM	150	ELEC	250	0.5785	28	0.0013
CHEM	150	ELEC	260	0.4463	28	0.0173
CHEM	150	ELEC	320	0.4847	28	0.0089
CHEM	150	ELEC	340	0.466	28	0.0124
CHEM	150	CSC	349A	0.4372	28	0.02
CHEM	150	ELEC	426	-0.8922	5	0.0418
CHEM	150	CENG	460	0.8429	6	0.0351
CHEM	150	ELEC	466	0.9869	4	0.0131

CSC	160	MATH	200	0.4524	31	0.0106
CSC	160	MATH	201	0.4547	31	0.0102
CSC	160	ELEC	250	0.3605	35	0.0334
CSC	160	ELEC	360	0.3721	35	0.0277
CSC	160	ELEC	499	0.4743	30	0.0081
ELEC	199	MATH	200	0.3269	38	0.0452
ELEC	199	MATH	201	0.4867	38	0.0019
ELEC	199	CENG	241	0.3819	42	0.0126
ELEC	199	ELEC	260	0.3814	42	0.0127
ELEC	199	ELEC	310	0.3071	42	0.0479
ELEC	199	ELEC	484	0.7109	9	0.0318
MATH	200	ELEC	300	0.4929	38	0.0017
ELEC	200	ELEC	300	0.3875	41	0.0123
MATH	200	ELEC	310	0.5504	38	0.0003
ELEC	200	ELEC	310	0.3663	41	0.0185
ELEC	200	ELEC	320	0.328	41	0.0363
MATH	200	ELEC	330	0.5606	38	0.0003
MATH	200	ELEC	340	0.5379	38	0.0005
ELEC	200	ELEC	340	0.4527	41	0.003
MATH	200	CSC	349A	0.4516	38	0.0044
MATH	200	ELEC	360	0.4953	38	0.0016
ELEC	200	ELEC	360	0.4299	41	0.005
ELEC	200	ELEC	370	0.4902	41	0.0011
MATH	200	ELEC	380	0.356	38	0.0283
MATH	200	ELEC	403	0.5739	14	0.0319
ELEC	200	ELEC	404	0.5698	25	0.0029
MATH	200	ELEC	407	0.5387	14	0.0469
ELEC	200	ELEC	407	0.6626	15	0.0071
MATH	200	ELEC	410	0.6229	17	0.0076

MATH	200	ELEC	420	0.9733	4	0.0267
MATH	200	ELEC	450	0.4901	17	0.0458
MATH	200	ELEC	452	-0.8847	5	0.0462
ELEC	200	ELEC	453	0.7016	11	0.0161
MATH	200	ELEC	456	0.6224	12	0.0307
MATH	200	ELEC	460	0.527	20	0.017
MATH	200	CENG	460	0.7922	9	0.0109
ELEC	200	ELEC	460	0.5879	21	0.0051
MATH	200	ELEC	496	0.733	8	0.0386
MATH	201	ELEC	300	0.3286	38	0.044
MATH	201	ELEC	310	0.5318	38	0.0006
MATH	201	ELEC	330	0.4578	38	0.0039
MATH	201	ELEC	340	0.4516	38	0.0044
MATH	201	CSC	349A	0.3758	38	0.0201
MATH	201	ELEC	360	0.4496	38	0.0046
MATH	201	ELEC	370	0.3688	38	0.0227
MATH	201	ELEC	380	0.3705	38	0.022
MATH	201	ELEC	403	0.6794	16	0.0038
MATH	201	ELEC	410	0.797	18	0.0001
MATH	201	ELEC	412	0.7224	9	0.0279
MATH	201	ELEC	420	0.9969	4	0.0031
MATH	201	CENG	441	0.4672	19	0.0437
MATH	201	ELEC	456	0.5941	12	0.0416
MATH	201	ELEC	460	0.5532	19	0.014
MATH	201	CENG	460	0.7773	9	0.0137
MATH	201	ELEC	499	0.3648	34	0.0339
ELEC	216	ELEC	300	0.4086	32	0.0202
ELEC	216	ELEC	310	0.4126	32	0.0189
ELEC	216	ELEC	330	0.5005	32	0.0035

ELEC	216	ELEC	340	0.4555	32	0.0088
ELEC	216	ELEC	360	0.4373	32	0.0123
ELEC	216	CENG	460	0.8109	8	0.0146
ELEC	220	ELEC	300	0.441	42	0.0035
ELEC	220	ELEC	310	0.5817	42	0.0001
ELEC	220	ELEC	320	0.317	42	0.0408
ELEC	220	ELEC	330	0.5203	42	0.0004
ELEC	220	ELEC	340	0.4857	42	0.0011
ELEC	220	CSC	349A	0.3597	42	0.0193
ELEC	220	ELEC	350	0.4173	42	0.006
ELEC	220	ELEC	360	0.5075	42	0.0006
ELEC	220	ELEC	370	0.4561	42	0.0024
ELEC	220	ELEC	412	0.8037	10	0.0051
ELEC	220	CENG	441	0.6199	20	0.0036
ELEC	220	ELEC	456	0.9081	13	0
ELEC	220	ELEC	460	0.4555	22	0.0332
ELEC	220	CENG	460	0.7367	10	0.0151
CSC	230	CENG	441	0.9457	8	0.0004
CSC	230	ELEC	482	0.9744	4	0.0256
CENG	241	ELEC	300	0.5331	42	0.0003
CENG	241	ELEC	310	0.4805	42	0.0013
CENG	241	ELEC	320	0.558	42	0.0001
CENG	241	ELEC	330	0.4089	42	0.0072
CENG	241	ELEC	340	0.6023	42	0
CENG	241	CSC	349A	0.3959	42	0.0095
CENG	241	ELEC	350	0.4622	42	0.0021
CENG	241	CENG	355	0.3376	42	0.0288
CENG	241	ELEC	360	0.5251	42	0.0004
CENG	241	ELEC	370	0.3579	42	0.0199

CENG	241	ELEC	380	0.5129	42	0.0005
CENG	241	ELEC	404	0.4742	25	0.0166
CENG	241	ELEC	407	0.7777	15	0.0006
CENG	241	ELEC	426	-0.8346	7	0.0195
CENG	241	CENG	441	0.5338	20	0.0153
CENG	241	ELEC	452	0.8771	6	0.0217
CENG	241	ELEC	484	0.6818	9	0.0431
ELEC	250	ELEC	300	0.5223	42	0.0004
ELEC	250	ELEC	310	0.576	42	0.0001
ELEC	250	ELEC	320	0.6942	42	0
ELEC	250	ELEC	340	0.6063	42	0
ELEC	250	CSC	349A	0.5401	42	0.0002
ELEC	250	ELEC	350	0.3855	42	0.0117
ELEC	250	ELEC	360	0.5277	42	0.0003
ELEC	250	ELEC	370	0.4113	42	0.0068
ELEC	250	ELEC	380	0.3523	42	0.0221
ELEC	250	ELEC	403	0.5419	17	0.0246
ELEC	250	ELEC	404	0.5171	25	0.0081
ELEC	250	CENG	441	0.5977	20	0.0054
ELEC	250	ELEC	450	0.4992	19	0.0295
ELEC	250	CENG	455	0.7819	8	0.0219
ELEC	250	ELEC	496	0.6607	10	0.0375
STAT	254	ELEC	300	0.5389	21	0.0117
STAT	254	ELEC	310	0.6779	21	0.0007
STAT	254	ELEC	320	0.4429	21	0.0444
STAT	254	ELEC	330	0.6615	21	0.0011
STAT	254	ELEC	340	0.7036	21	0.0004
STAT	254	CSC	349A	0.619	21	0.0028
STAT	254	ELEC	360	0.6484	21	0.0015

STAT	254	ELEC	370	0.5444	21	0.0107
STAT	254	ELEC	410	0.6957	11	0.0174
STAT	254	CENG	441	0.7514	9	0.0196
STAT	254	ELEC	460	0.5747	14	0.0316
CENG	255	CENG	355	0.3878	31	0.0311
CENG	255	ENGR	446	0.4147	31	0.0204
CENG	255	ELEC	460	0.4808	18	0.0434
CENG	255	ELEC	482	0.648	12	0.0227
ELEC	260	ELEC	300	0.4978	42	0.0008
ELEC	260	ELEC	310	0.7506	42	0
ELEC	260	ELEC	320	0.4384	42	0.0037
ELEC	260	ELEC	330	0.6715	42	0
ELEC	260	ELEC	340	0.5763	42	0.0001
ELEC	260	CSC	349A	0.4465	42	0.003
ELEC	260	ELEC	350	0.5915	42	0
ELEC	260	ELEC	360	0.6346	42	0
ELEC	260	ELEC	370	0.5434	42	0.0002
ELEC	260	ELEC	380	0.5253	42	0.0004
ELEC	260	ELEC	403	0.5	17	0.041
ELEC	260	ELEC	404	0.4427	25	0.0267
ELEC	260	ELEC	407	0.7756	15	0.0007
ELEC	260	ELEC	410	0.5346	21	0.0125
ELEC	260	ELEC	412	0.7724	10	0.0088
ELEC	260	CENG	441	0.7114	20	0.0004
ELEC	260	ELEC	456	0.6295	13	0.0211
ELEC	260	ELEC	460	0.701	22	0.0003
ELEC	260	CENG	460	0.674	10	0.0326
MECH	295	ELEC	300	0.3392	42	0.028
MECH	295	ELEC	310	0.3104	42	0.0454

MECH	295	ELEC	320	0.4945	42	0.0009
MECH	295	ELEC	340	0.4147	42	0.0063
MECH	295	ELEC	350	0.4259	42	0.0049
MECH	295	CENG	355	0.3957	42	0.0095
MECH	295	ELEC	360	0.4579	42	0.0023
MECH	295	ELEC	370	0.3818	42	0.0126
MECH	295	ELEC	380	0.3348	42	0.0302
MECH	295	ELEC	404	0.413	25	0.0402
MECH	295	ELEC	407	0.5365	15	0.0392
MECH	295	ELEC	412	0.6963	10	0.0253
MECH	295	CENG	441	0.5051	20	0.0231
MECH	295	ELEC	459	0.8363	6	0.038
MECH	295	ELEC	460	0.6567	22	0.0009

Appendix 5 Strongly Correlated Course Distributions

The distribution of strongly correlated courses including CourseYs with CourseX and CourseXs with CourseY from Train2010-2012, Train2010-2013 and Train2010-2014 are shown below in from Figure 41 to Figure 46.

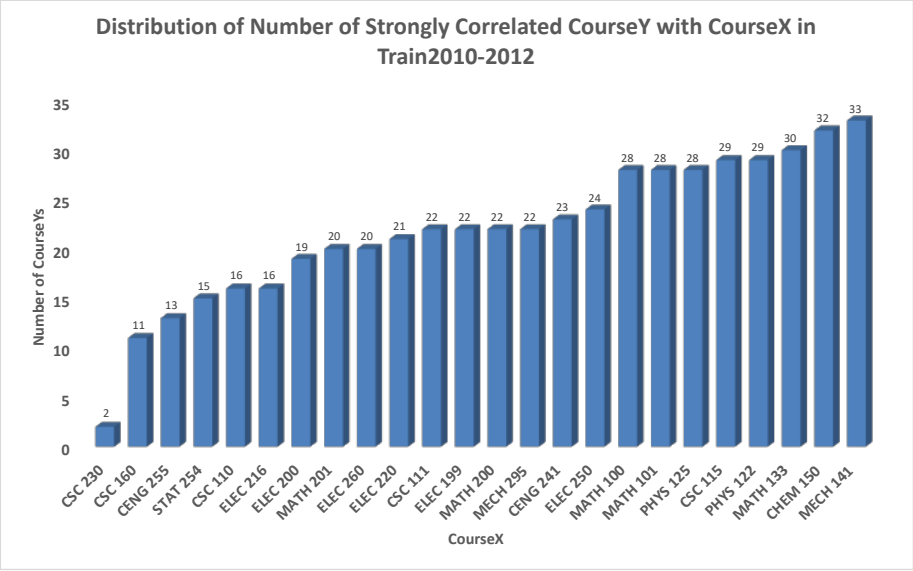

Figure 41 Distribution of Strongly Correlated CourseYs with CourseX in Train2010-2012

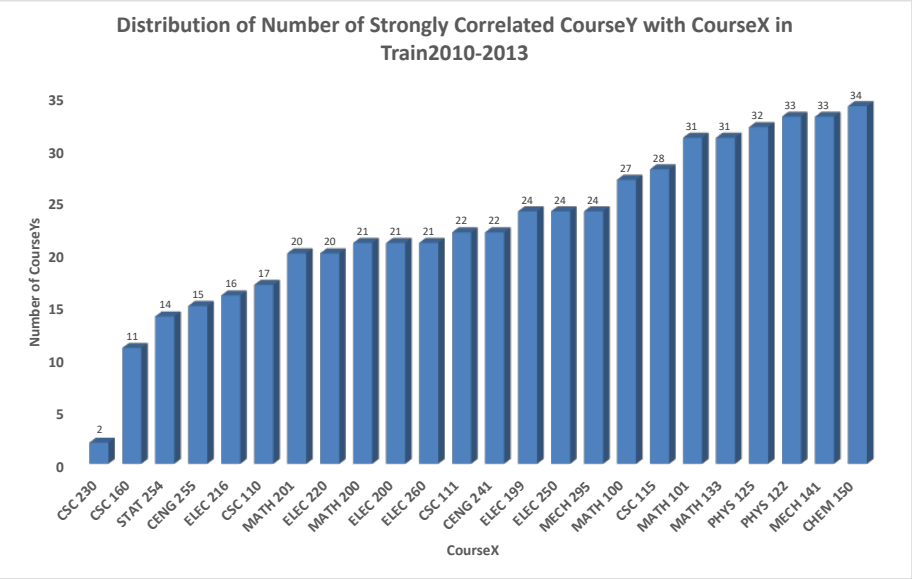

Figure 42 Distribution of Strongly Correlated CourseYs with CourseX in Train2010-2013

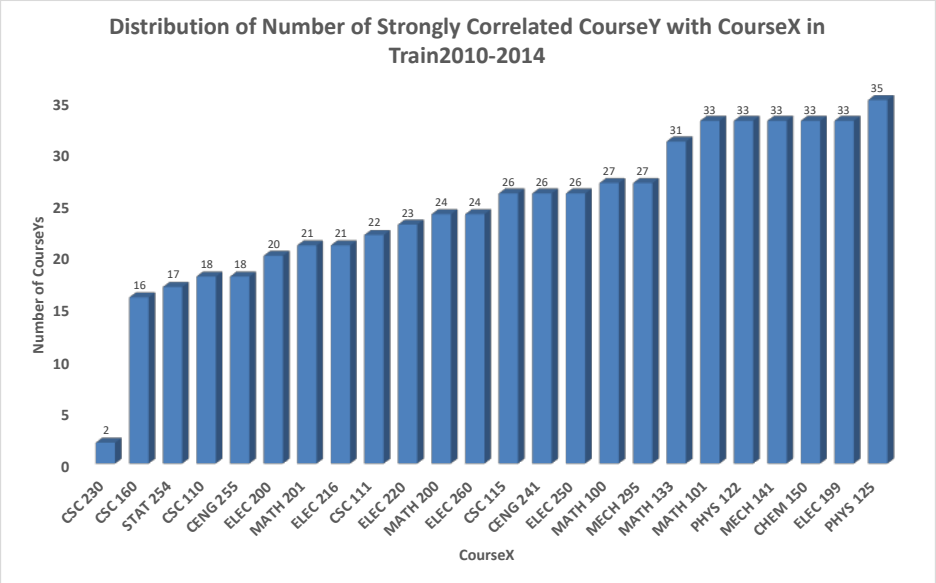

Figure 43 Distribution of Strongly Correlated CourseYs with CourseX in Train2010-2014

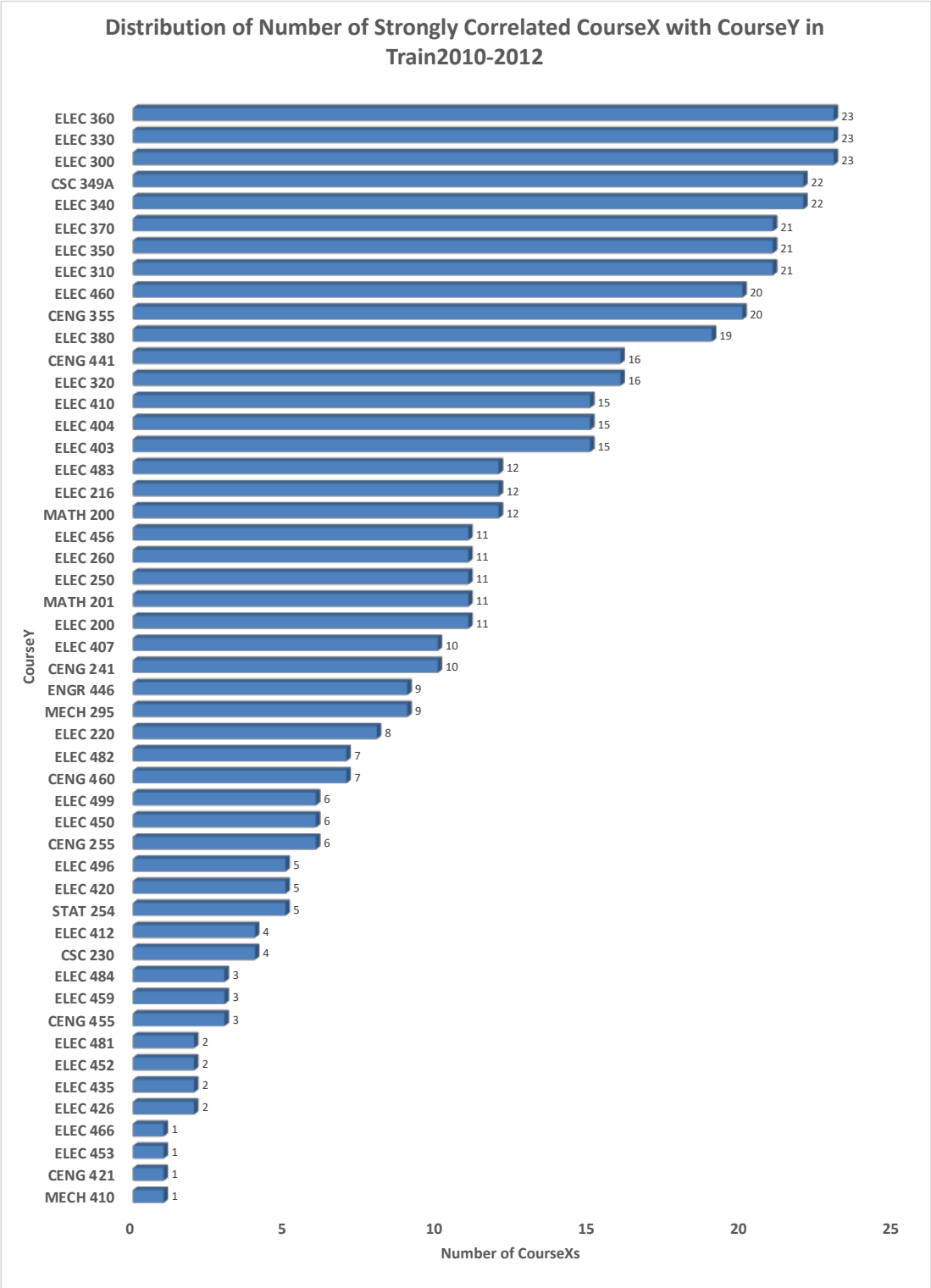

Figure 44 Distribution of Strongly Correlated CourseXs with CourseY in Train2010-2012

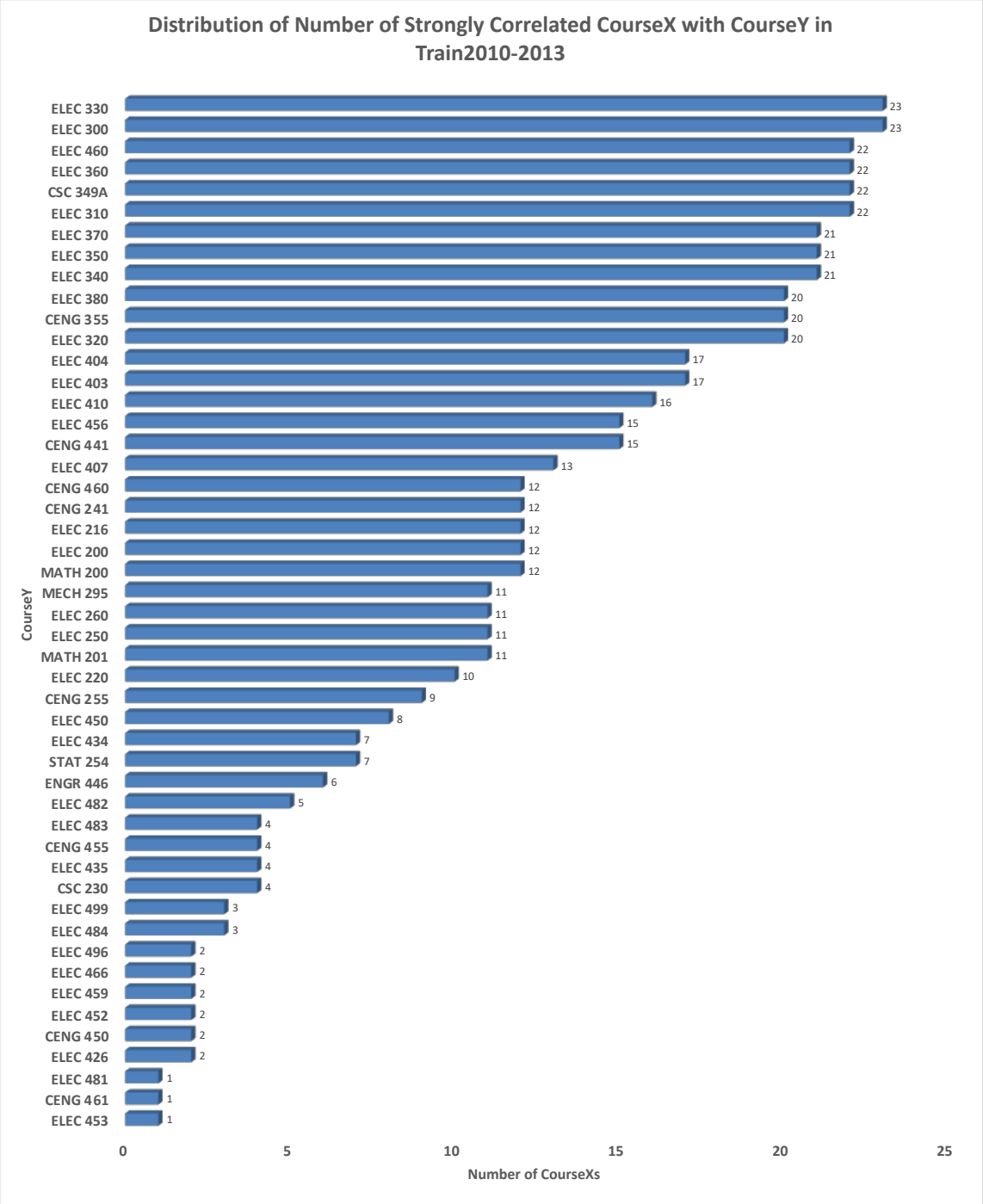

Figure 45 Distribution of Strongly Correlated CourseXs with CourseY in Train2010-2013

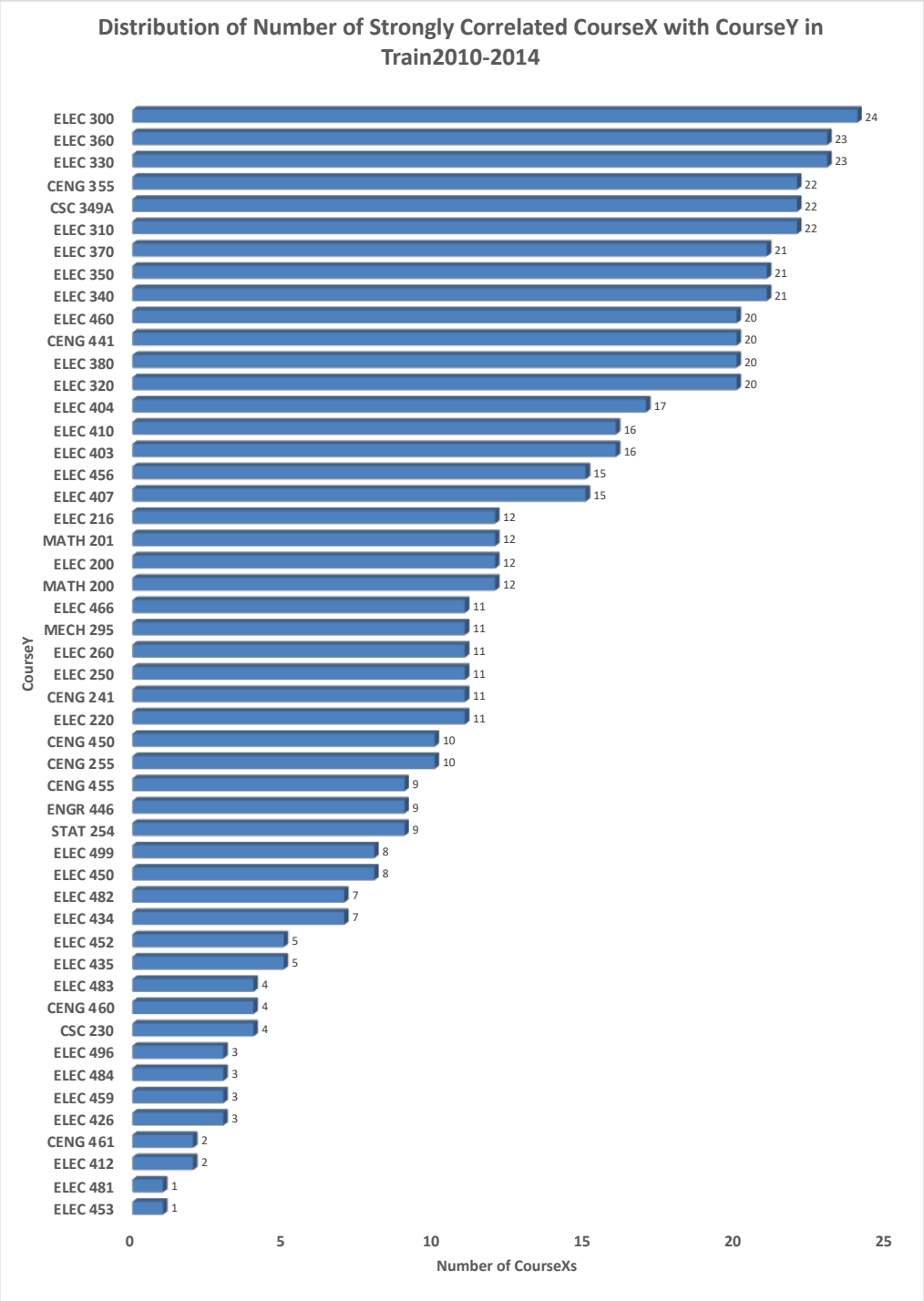

Figure 46 Distribution of Strongly Correlated CourseXs with CourseY in Train2010-2014

Appendix 6 Pearson Coefficient Histograms

The Pearson Coefficient histograms from Train2010-2012, Train2010-2013 and Train2010-2014 are shown in below from Figure 47 to Figure 49 and they show similar trend that most of the coefficients are around 0.5 and the coefficients in this training set fall mostly into the range of 0.3 to 1.

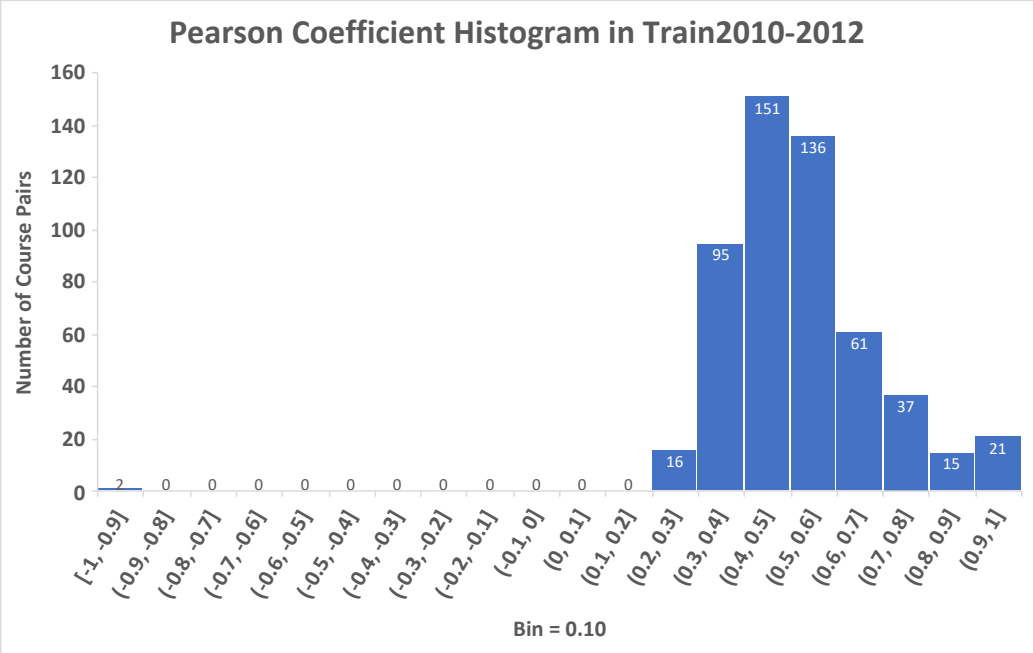

Figure 47 Histogram of Pearson Coefficients in Train2010-2012

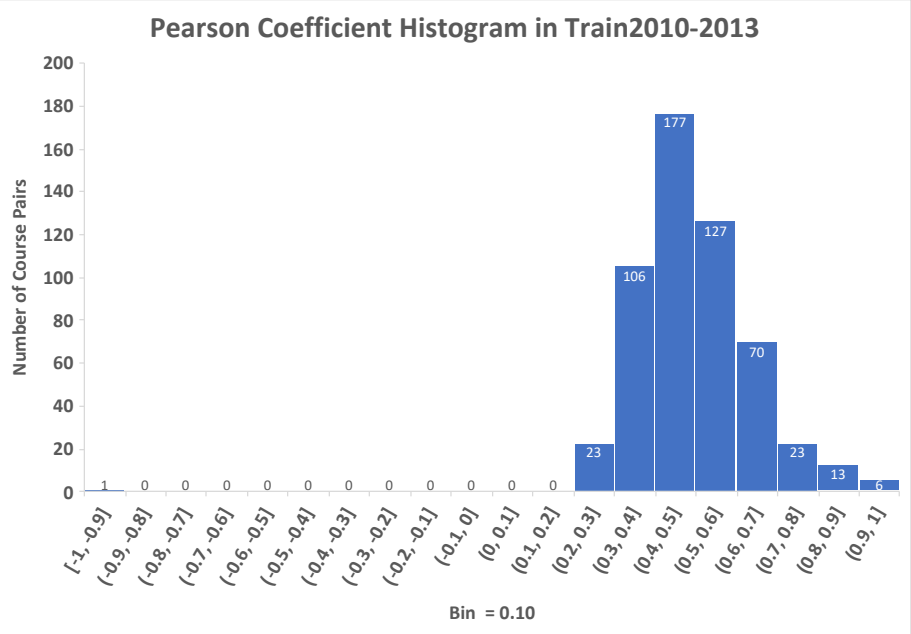

Figure 48 Histogram of Pearson Coefficients in Train2010-2013

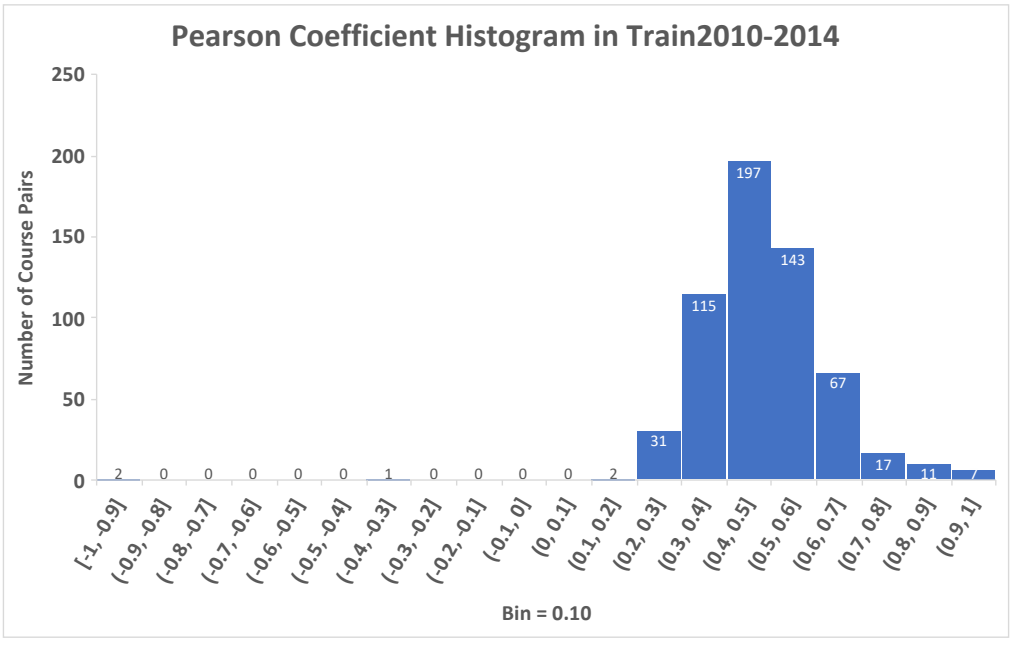

Figure 49 Histogram of Pearson Coefficients in Train2010-2014

Appendix 7 Distributions of Pearson Coefficient and Enrolment

The Pearson Coefficient and enrolment in the other three training sets of Train2010-2012, Train2010-2013 and Train2010-2014 also were scattered and shown below. These scatterplots show the same trend as the one in Train2010-2011 that most coefficients are in the range of from 0.3 to 1.0 and a small number of the coefficients are in the interval of from -1.0 to -0.8.

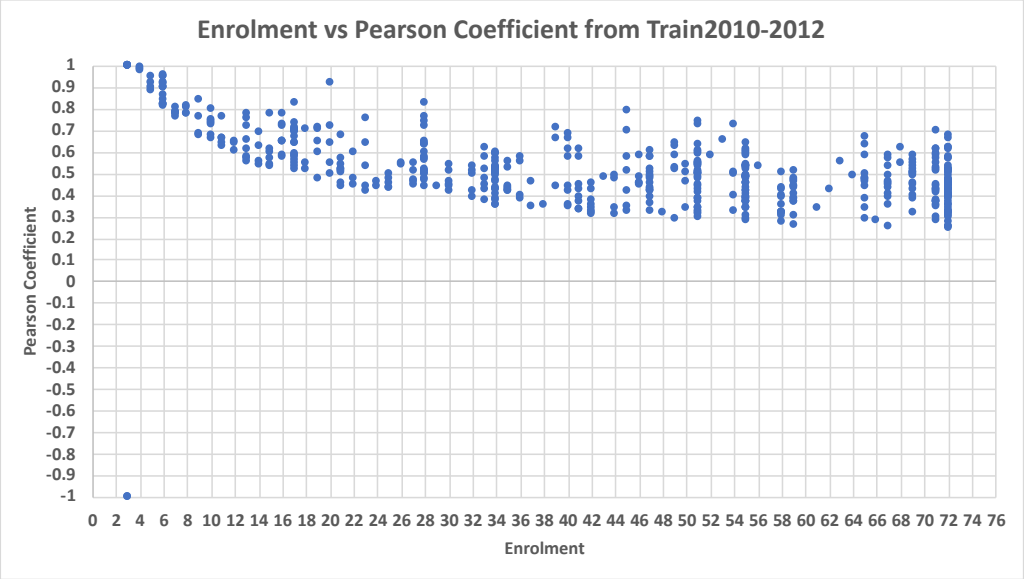

Figure 50 Enrolment and Pearson Coefficient from Train2010-2012

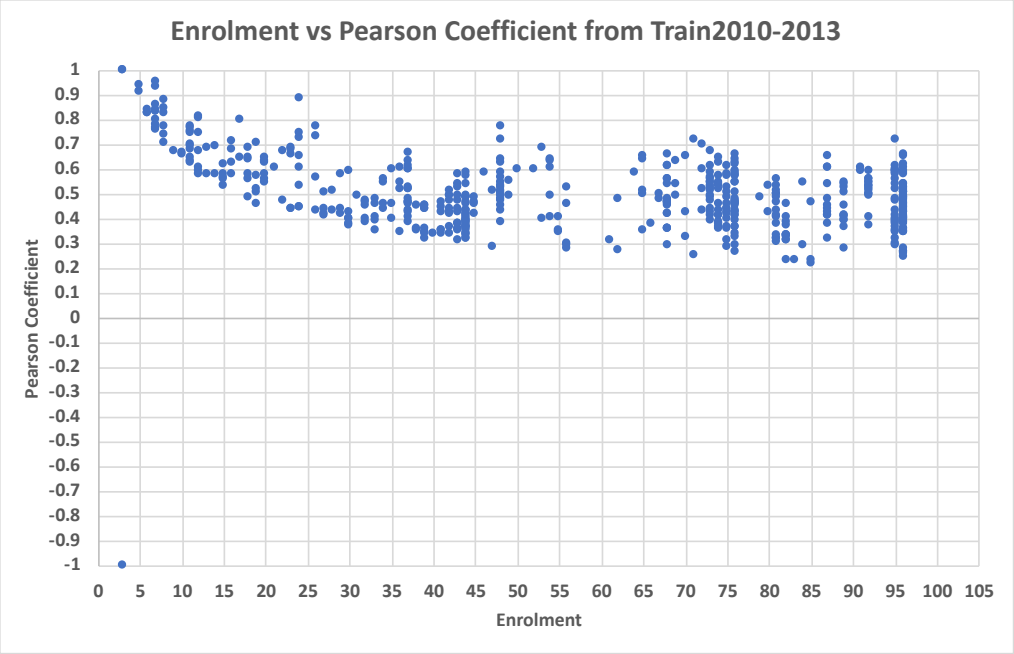

Figure 51 Enrolment and Pearson Coefficient from Train2010-2013

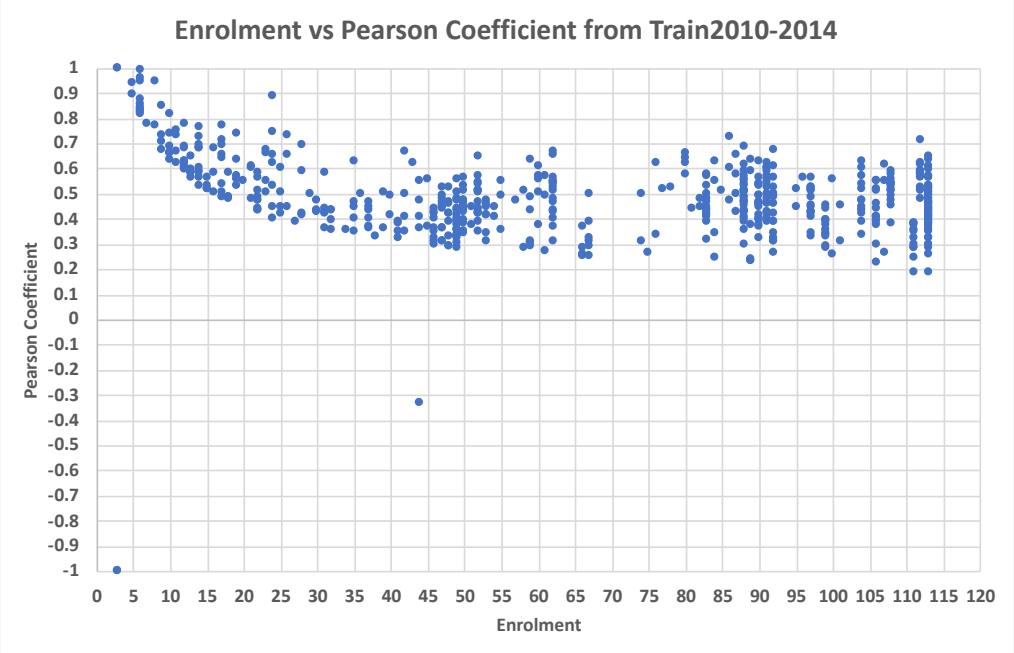

Figure 52 Enrolment and Pearson Coefficient from Train2010-2014

Appendix 8 Coefficient Distributions of Course Pairs Selected by Max(Pearson Coefficient)

The graphs below show the coefficient distributions of course pairs selected by the maximum of coefficient in Train2010-2012, Train2010-2013 and Train2010-2014. It can be from the figures below that the course pairs with fourth-year courses have relative bigger coefficient than others because the fourth-year courses are more specialization-related, thus having small enrolment.

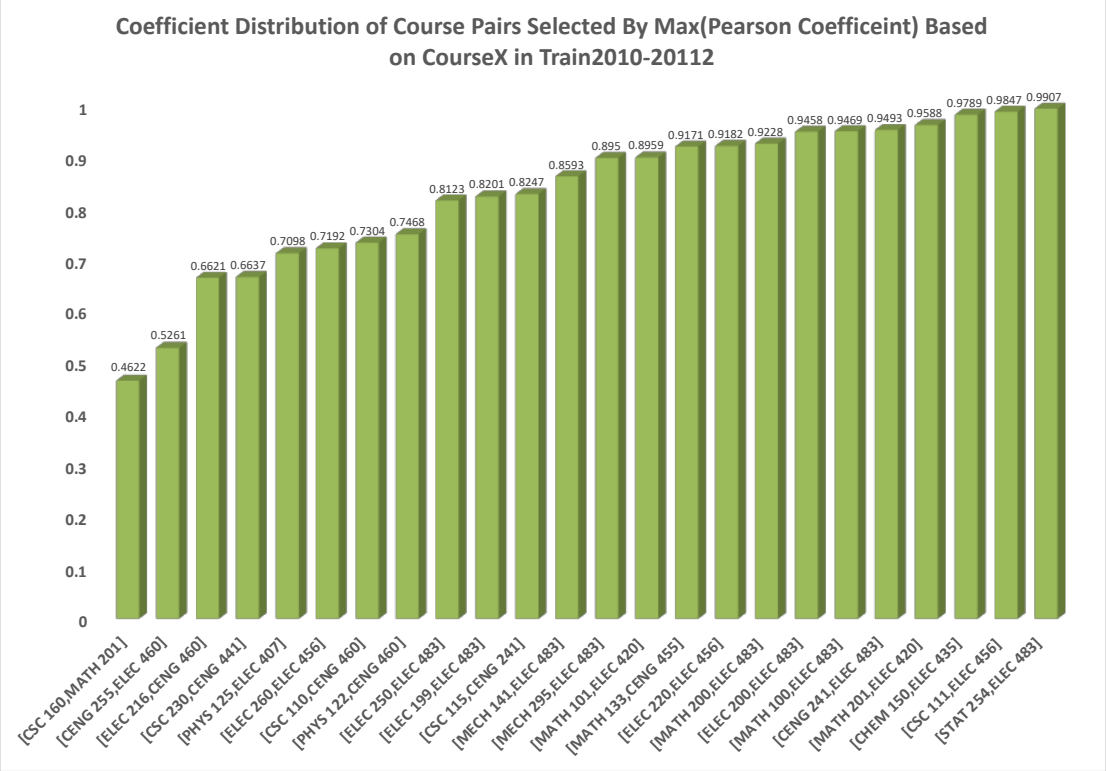

Figure 53 Pearson Correlation Distribution of Course Pairs Selected By Max(Pearson Coefficient) Based on CourseX in Train2010-20112

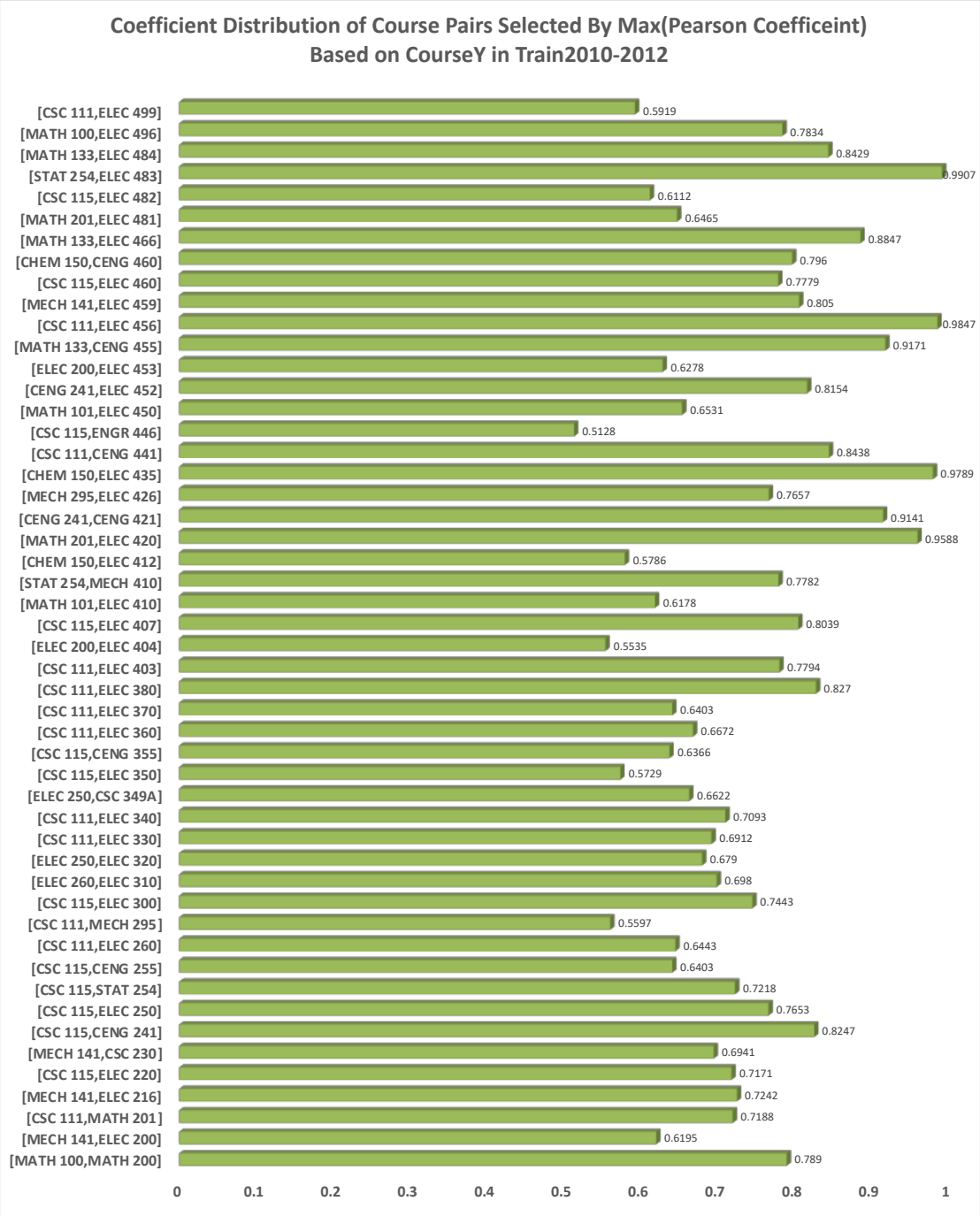

Figure 54 Pearson Correlation Distribution of Course Pairs Selected By Max(Pearson Coefficient) Based on CourseY in Train2010-2012

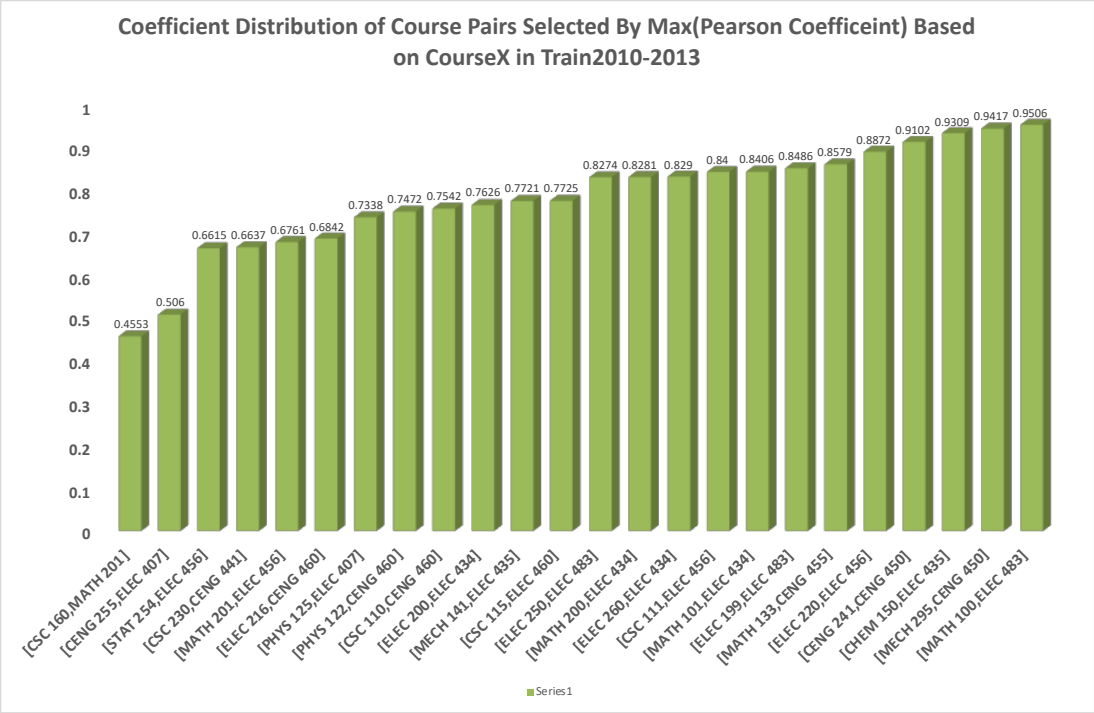

Figure 55 Pearson Correlation Distribution of Course Pairs Selected By Max(Pearson Coefficient) Based on CourseX in Train2010-2013

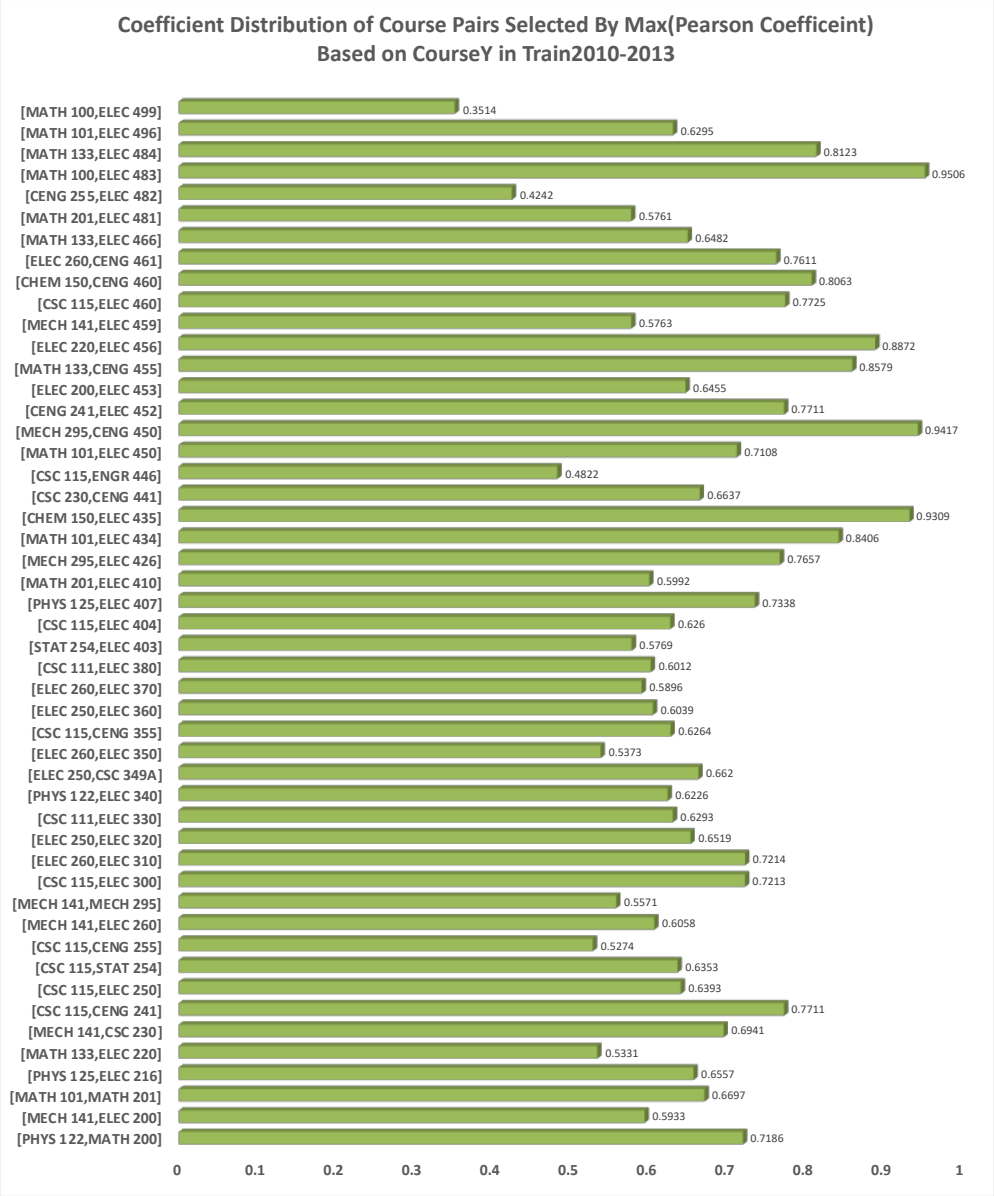

Figure 56 Pearson Correlation Distribution of Course Pairs Selected By Max(Pearson Coefficient) Based on CourseY in Train2010-2013

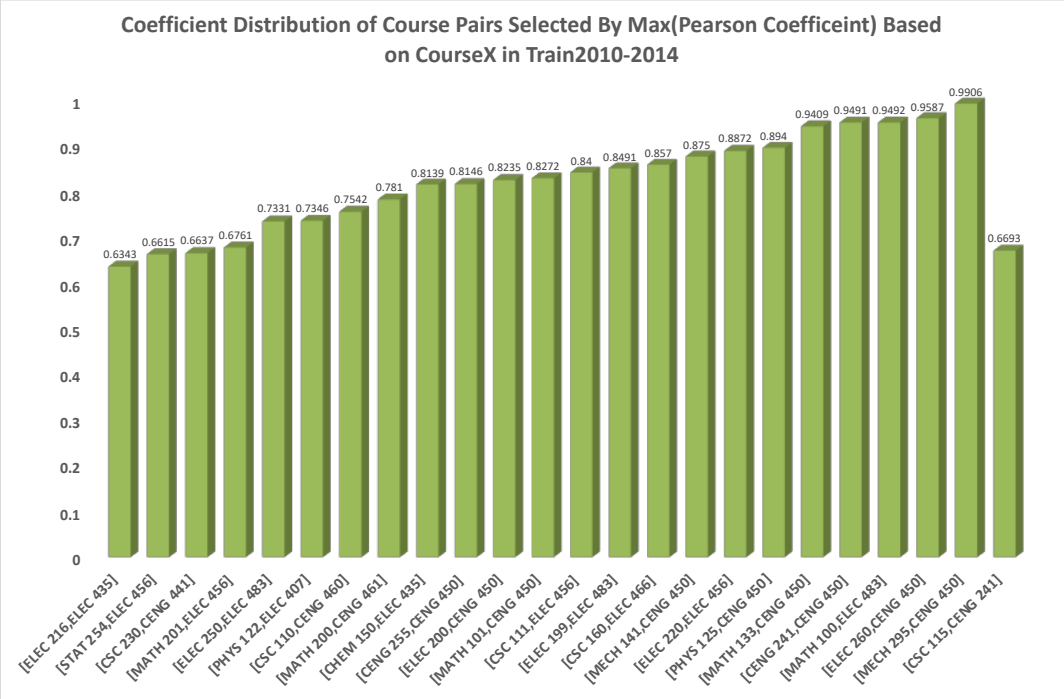

Figure 57 Pearson Correlation Distribution of Course Pairs Selected By Max(Pearson Coefficient) Based on CourseX in Train2010-2014

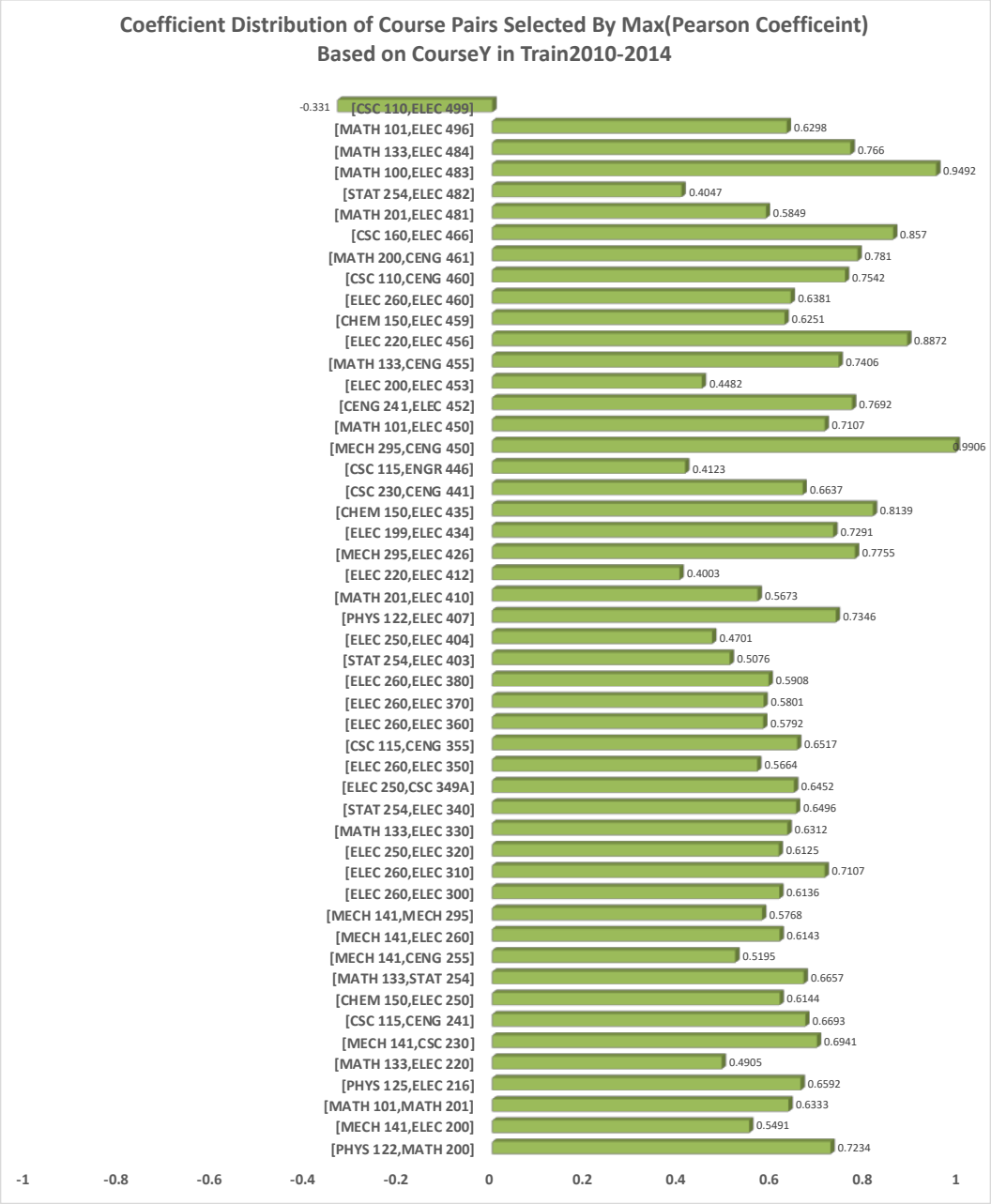

Figure 58 Pearson Correlation Distribution of Course Pairs Selected By Max(Pearson Coefficient) Based on CourseY in Train2010-2014

Appendix 9 MAEs of Course Pairs Selected by Max(Pearson Coefficient)

This appendix lists the MAE distributions of course pairs selected by maximum of coefficient based on both CourseX and CourseY from the three training sets of Train2010-2012 and Train2010-2013. It can be seen from the figures show that majority of the MAEs are greater than 1.0, regardless that they were computed from course pairs selected based on CourseX or CourseY.

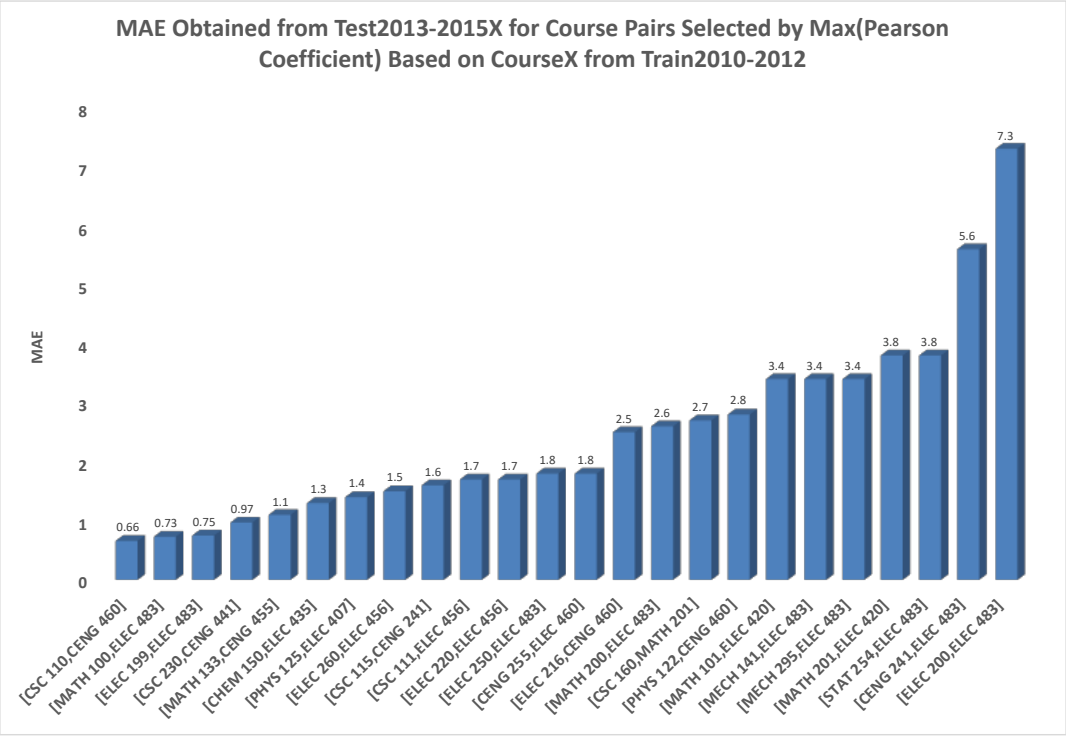

Figure 59 Prediction MAEs for Course Pairs Selected by Max(Pearson Coefficient) Based on CourseX from Train2010-2012 in Test2013-2015X

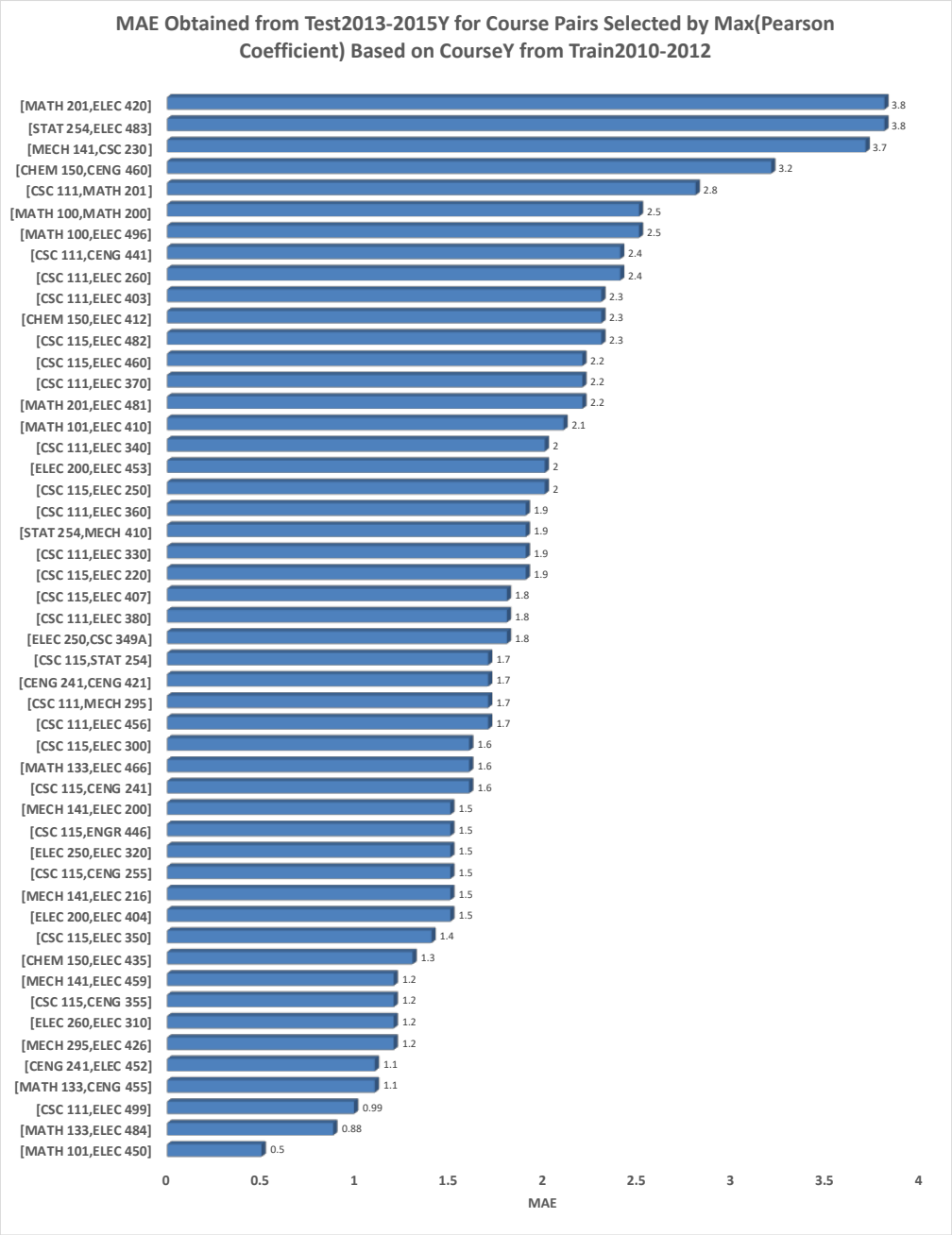

Figure 60 Prediction MAEs for Course Pairs Selected by Max(Pearson Coefficient) Based on CourseY from Train2010-2012 in Test2013-2015Y

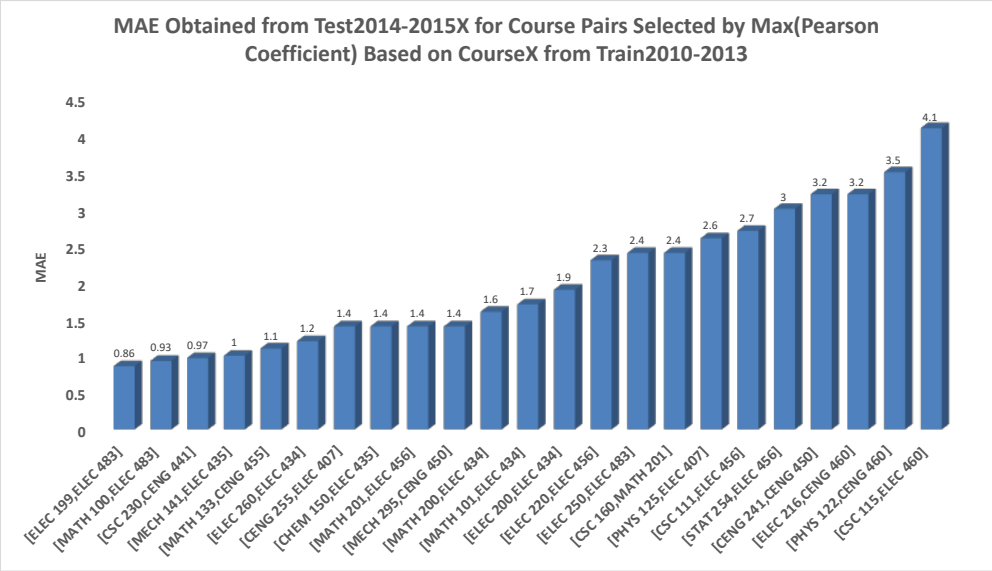

Figure 61 Prediction MAEs for Course Pairs Selected by Max(Pearson Coefficient) Based on CourseX from Train2010-2013 in Test2012-2015X

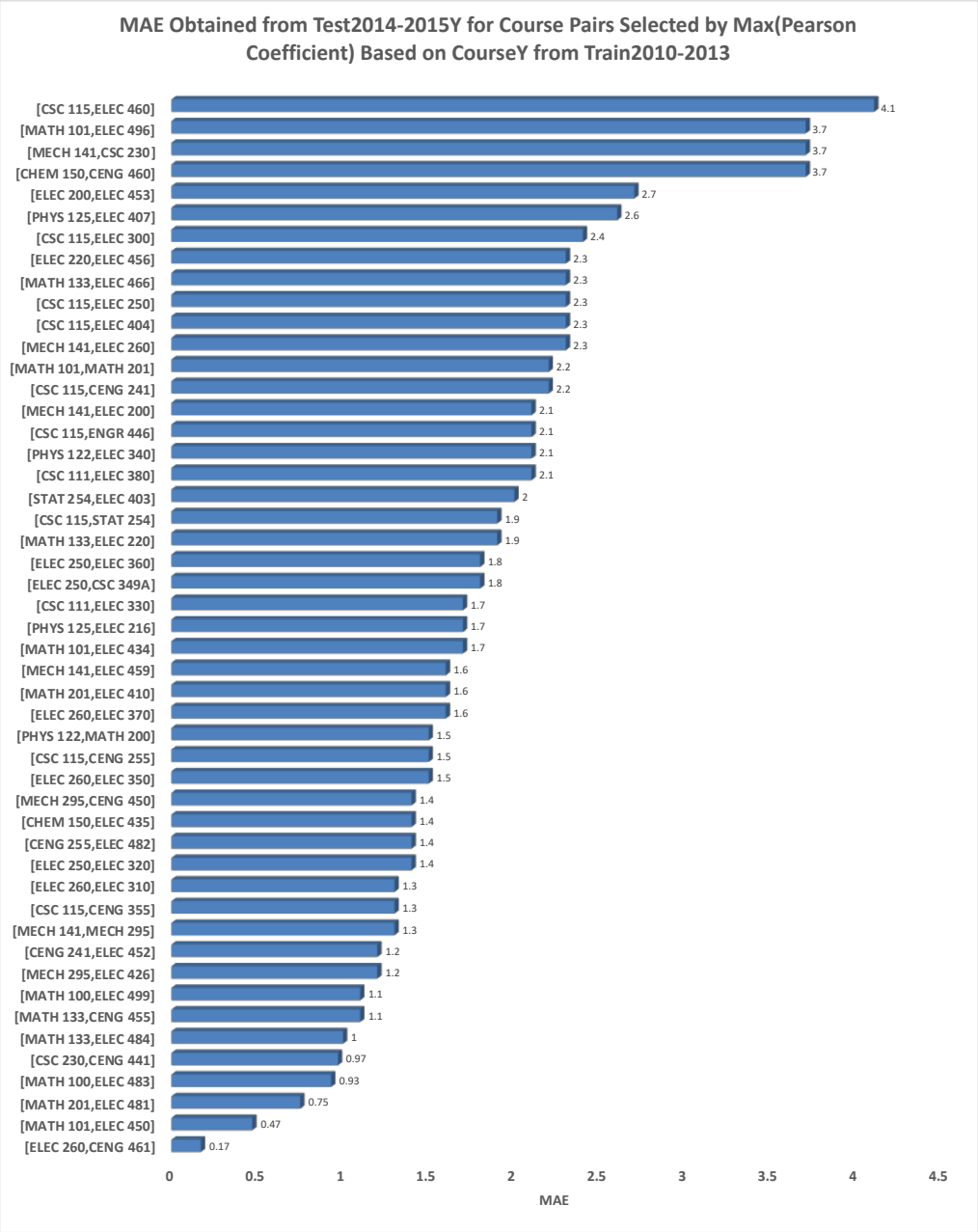

Figure 62 Prediction MAEs for Course Pairs Selected by Max(Pearson Coefficient) Based on CourseY from Train2010-2013 in Test2014-2015Y

Appendix 10 Precisions of Course Pairs Selected by Max(Pearson Coefficient)

The figures show the prediction precisions of course pairs selected by maximum of coefficient in Train2010-2012 and Train201-2013. The arrows in figures indicate the course pairs discussed in the content.

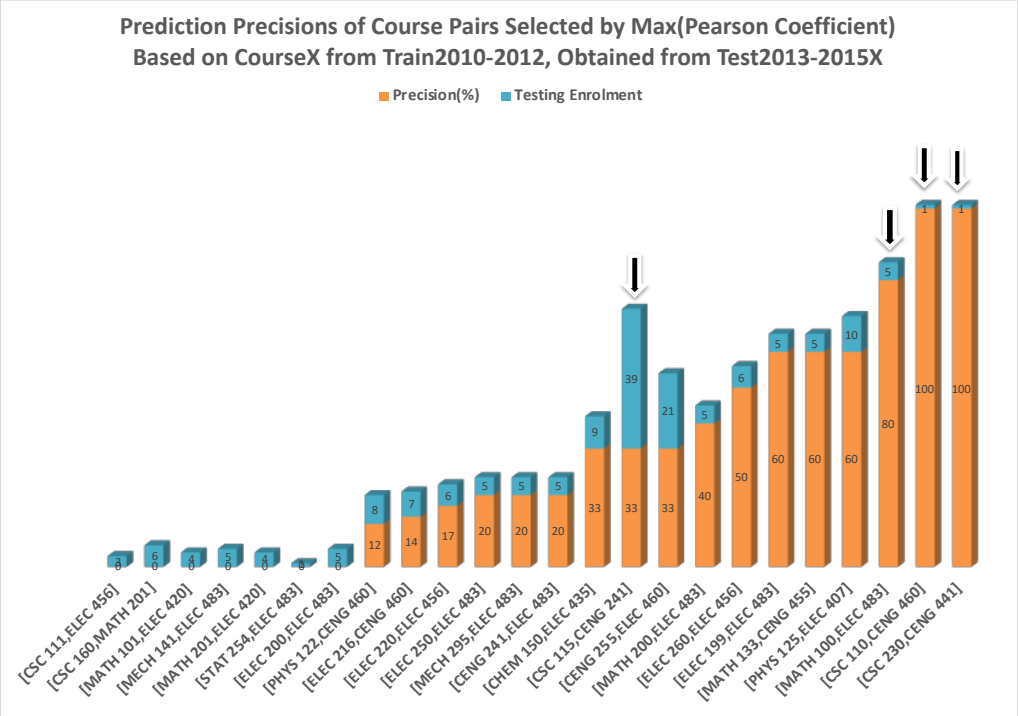

Figure 63 Prediction Precisions of Course Pairs in Test2013-2015X, Trained by Train2010-2012

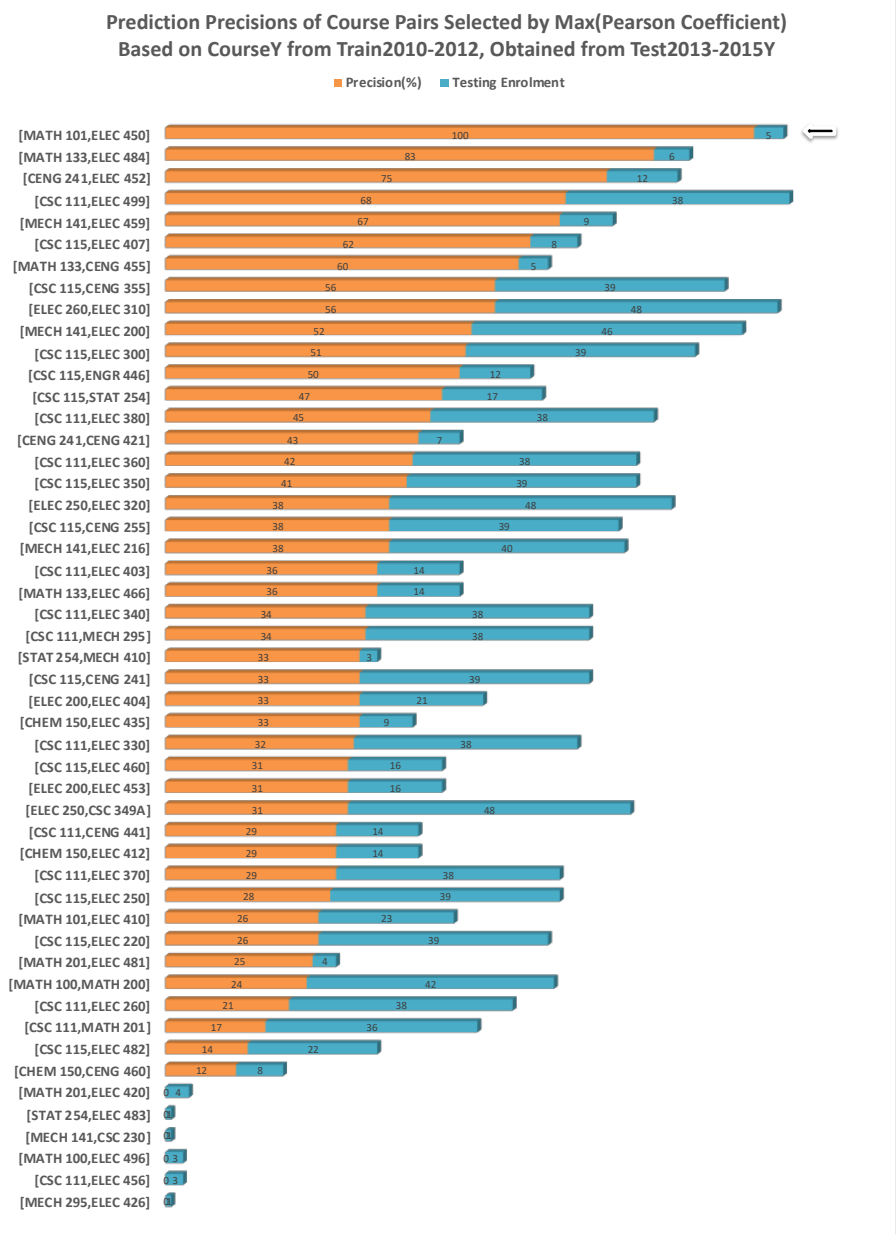

Figure 64 Prediction Precisions of Course Pairs in Test2013-2015Y, Trained by Train2010-2012

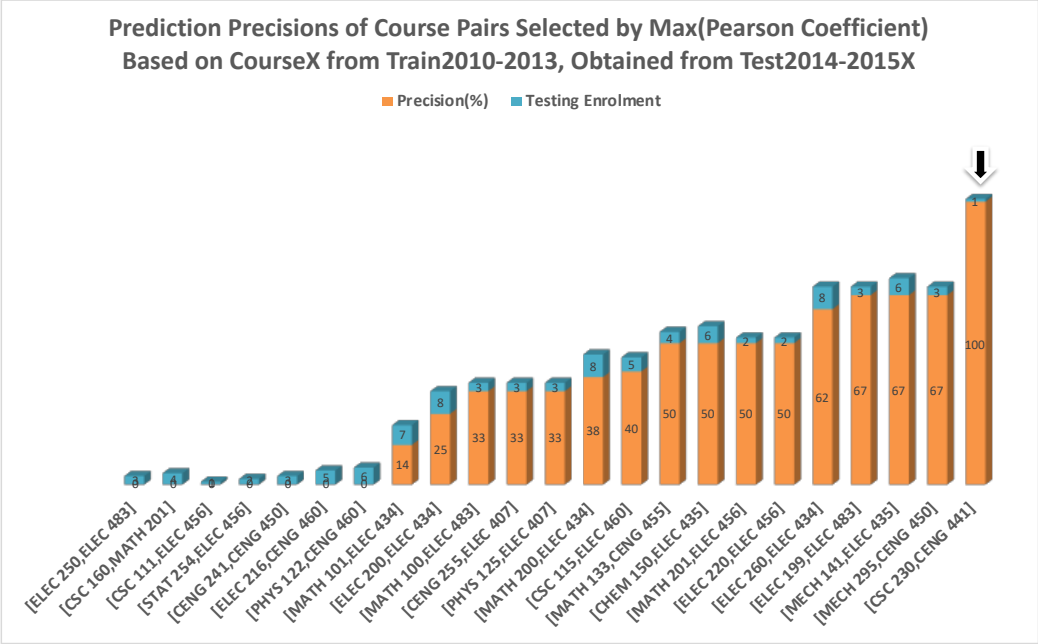

Figure 65 Prediction Precisions of Course Pairs in Test2014-2015X, Trained by Train2010-2013

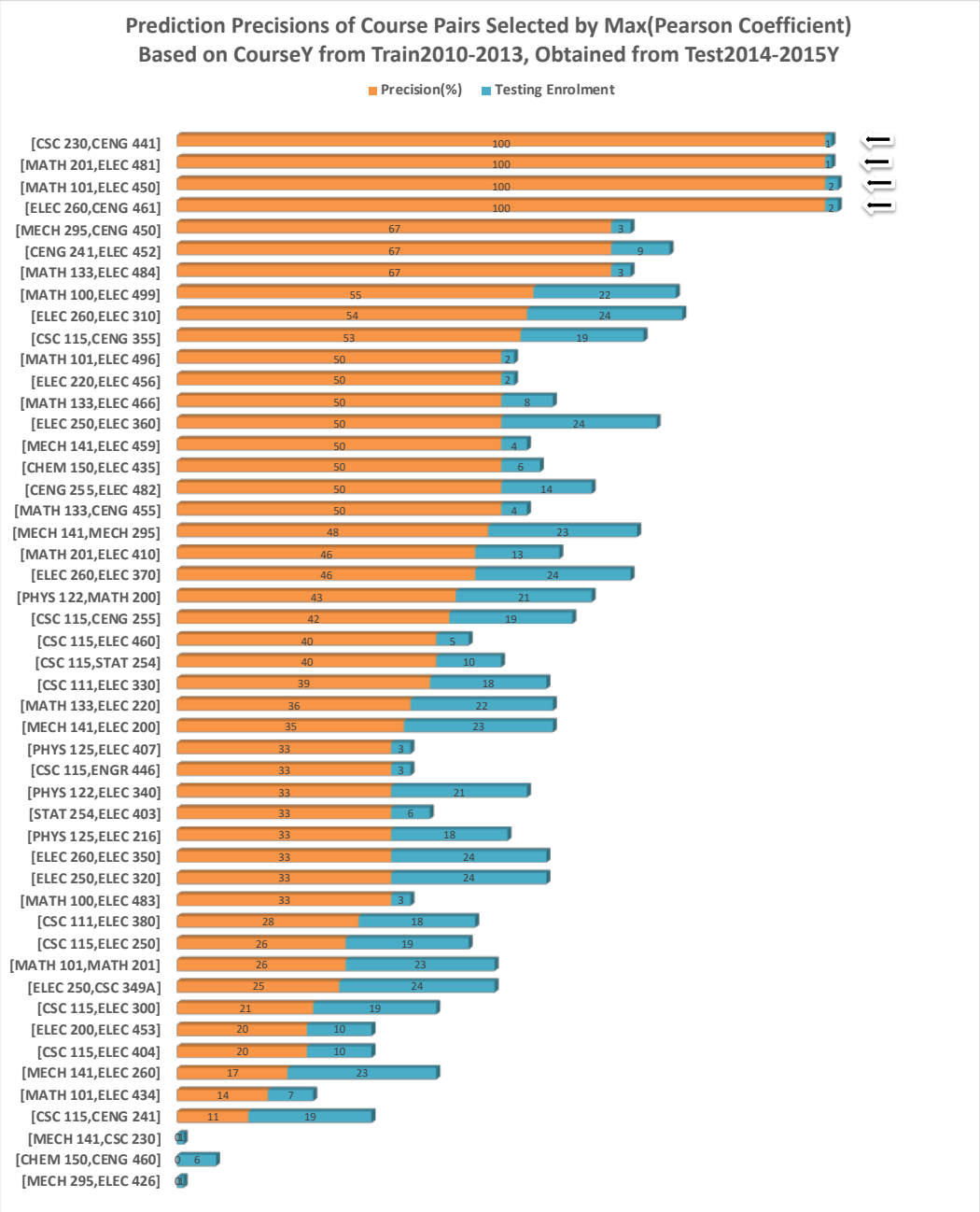

Figure 66 Prediction Precisions of Course Pairs in Test2014-2015Y, Trained by Train2010-2013

Appendix 11 Enrolment Distributions of Course Pairs Selected by Max(Enrolment)

The enrolments of course pairs selected by maximum of enrolment in the training set of Train2010-2012, Train2010-2013 and Train2010-2014 were plotted in bar charts shown below.

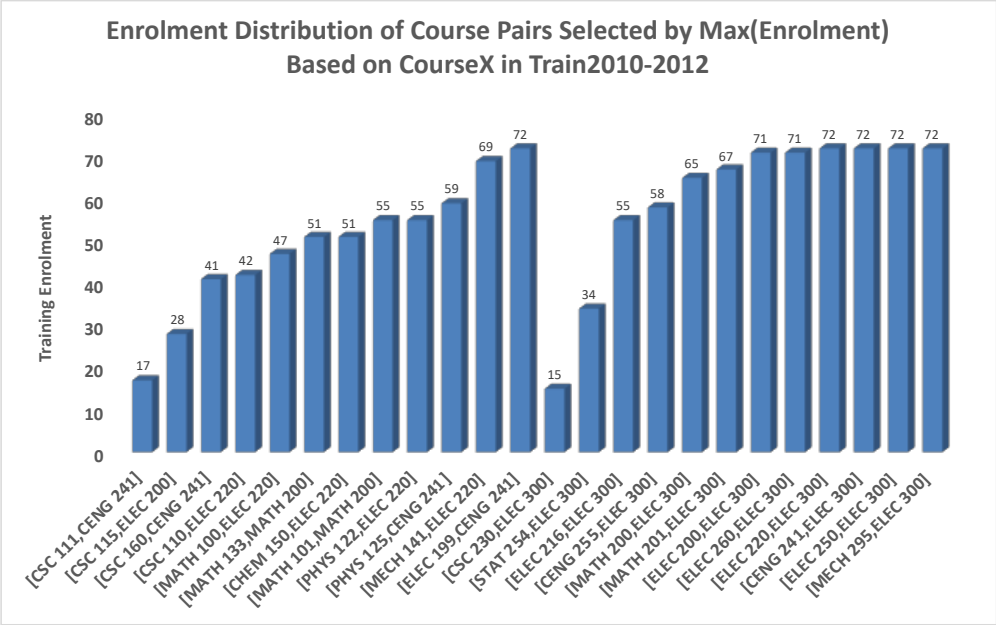

Figure 67 Enrolment Distribution of Course Pairs Selected by Max(Enrolment) Based on CourseX in Train2010-2012

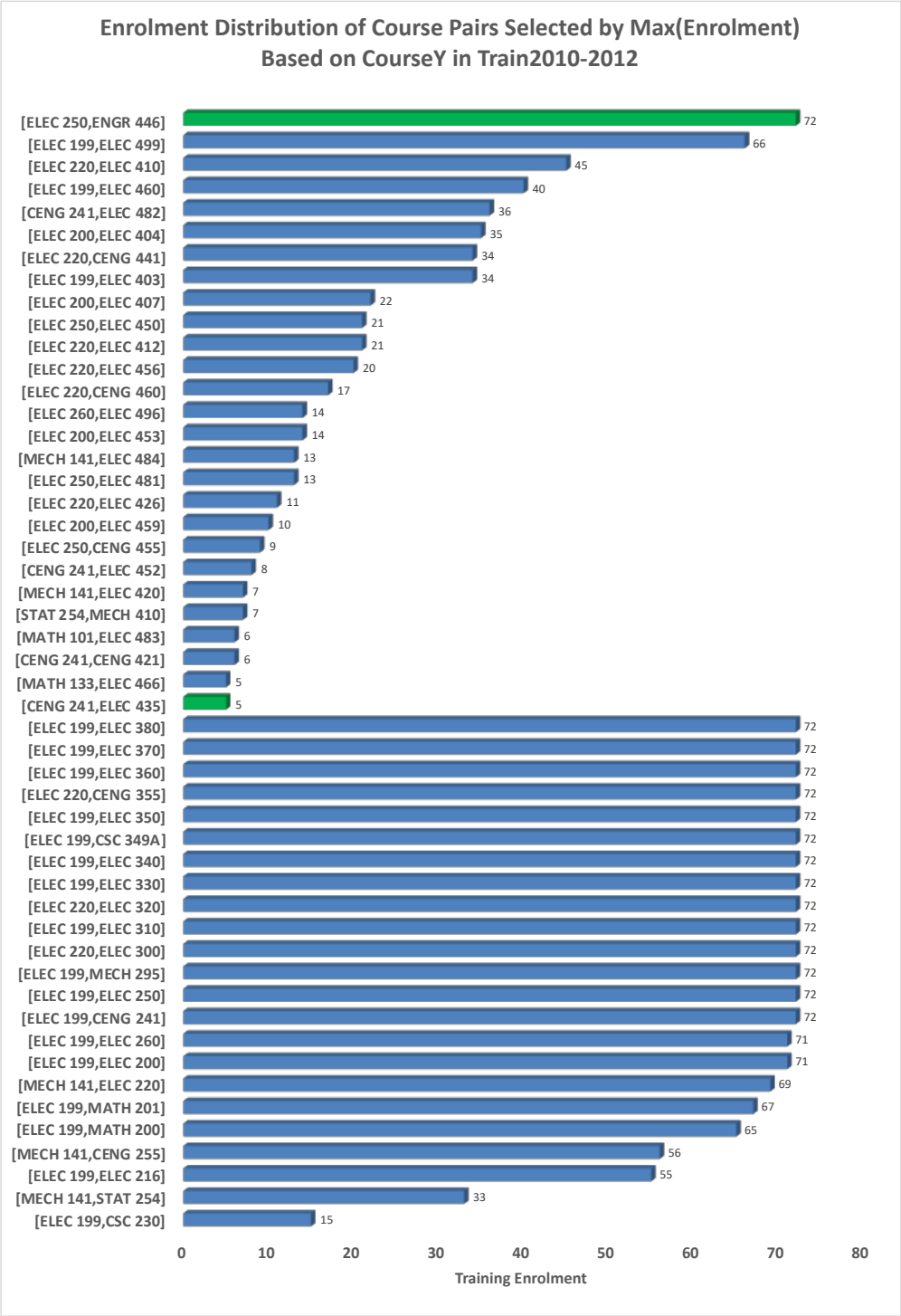

Figure 68 Enrolment Distribution of Course Pairs Selected by Max(Enrolment) Based on CourseY in Train2010-2012

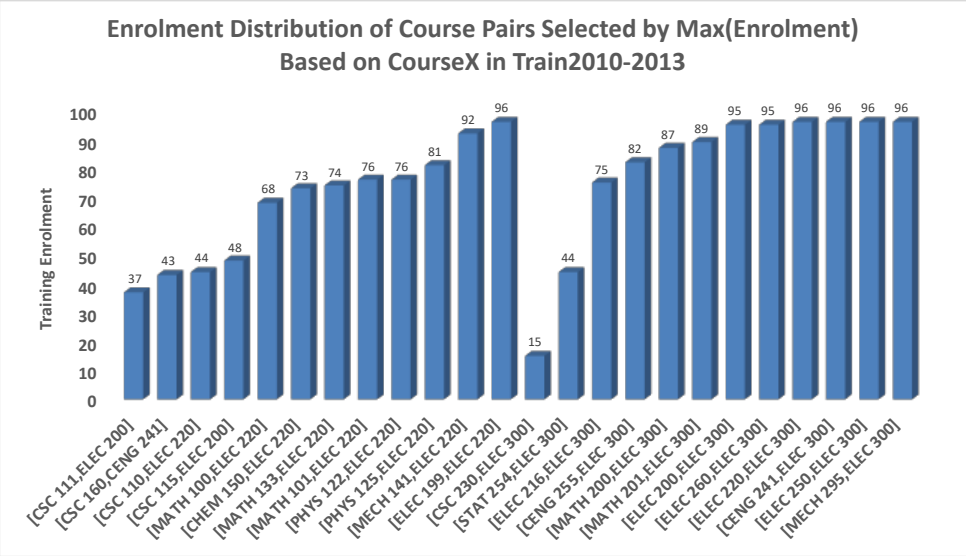

Figure 69 Enrolment Distribution of Course Pairs Selected by Max(Enrolment) Based on CourseX in Train2010-2013

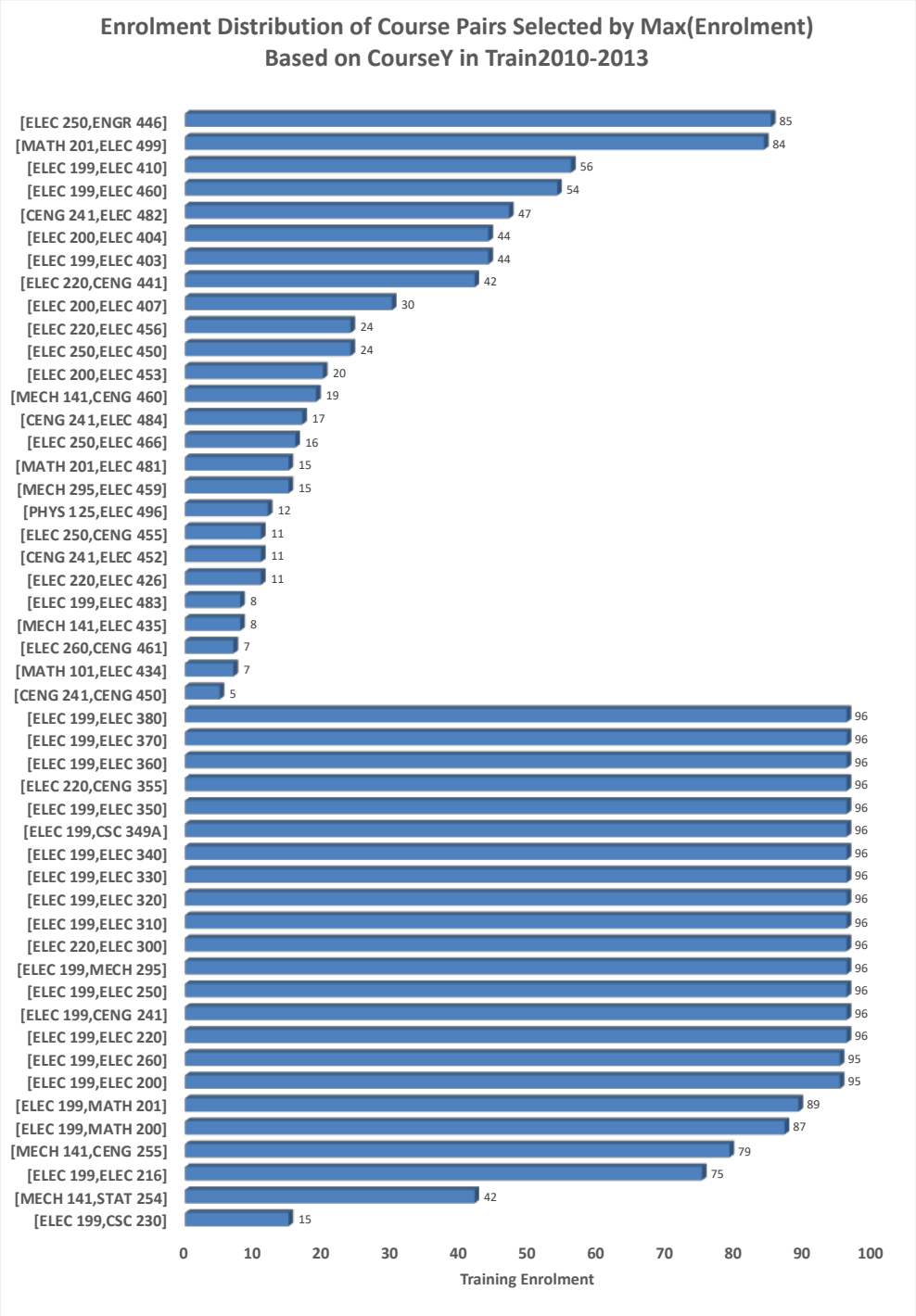

Figure 70 Enrolment Distribution of Course Pairs Selected by Max(Enrolment) Based on CourseY in Train2010-2013

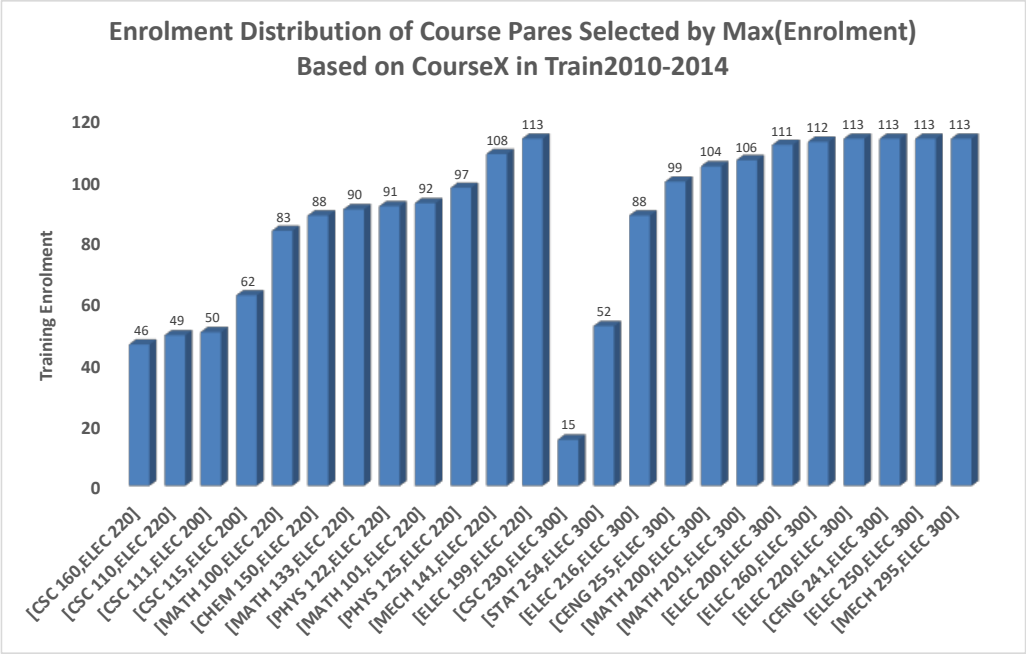

Figure 71 Enrolment Distribution of Course Pairs Selected by Max(Enrolment) Based on CourseX in Train2010-2014

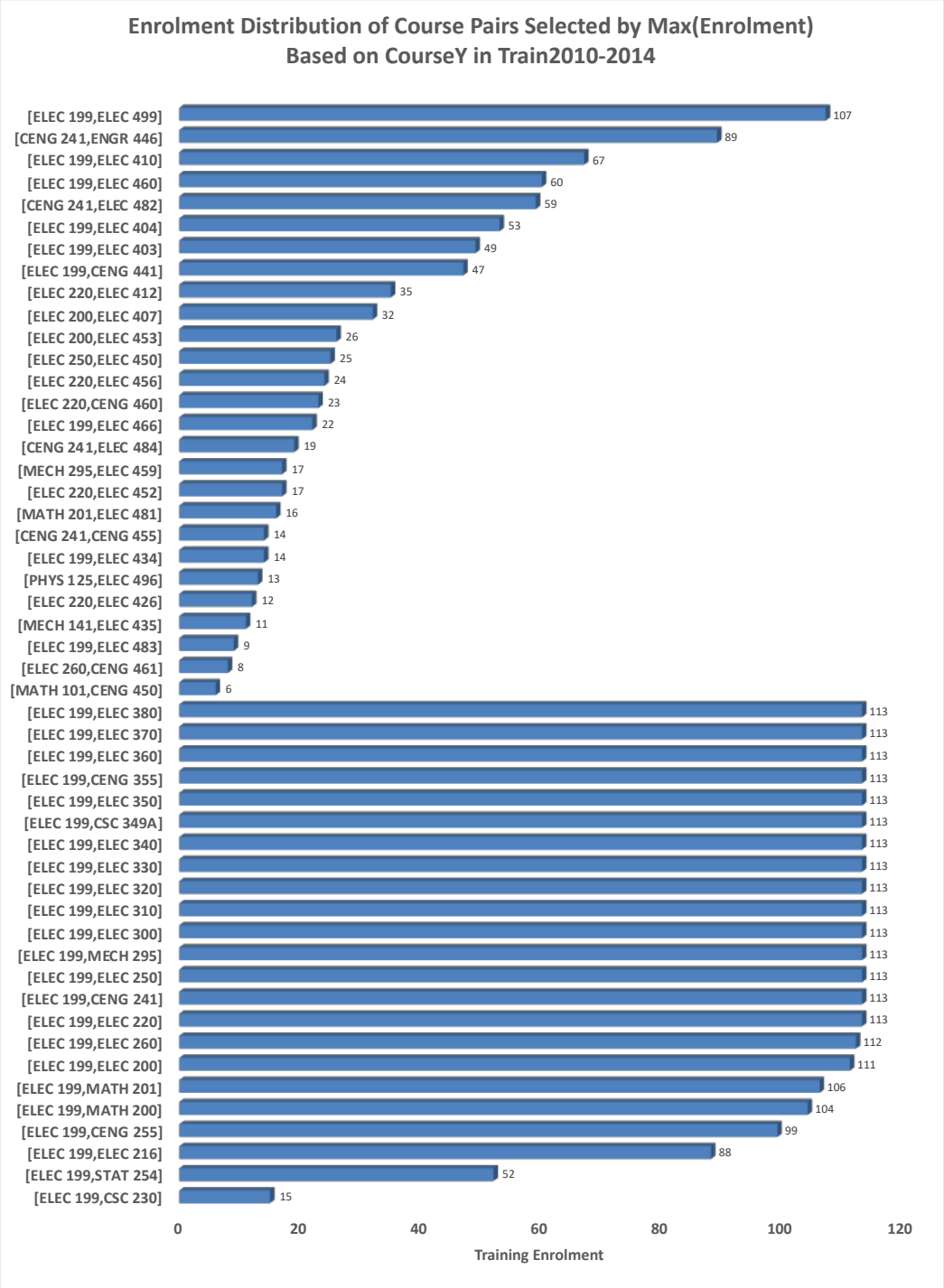

Figure 72 Enrolment Distribution of Course Pairs Selected by Max(Enrolment) Based on CourseY in Train2010-2014

Appendix 12 MAEs of Course Pairs Selected by Max(Enrolment)

The prediction MAEs of course pairs selected by the maximum of enrolment from Train2010-2012, Train2010-2013 are shown in figures below.

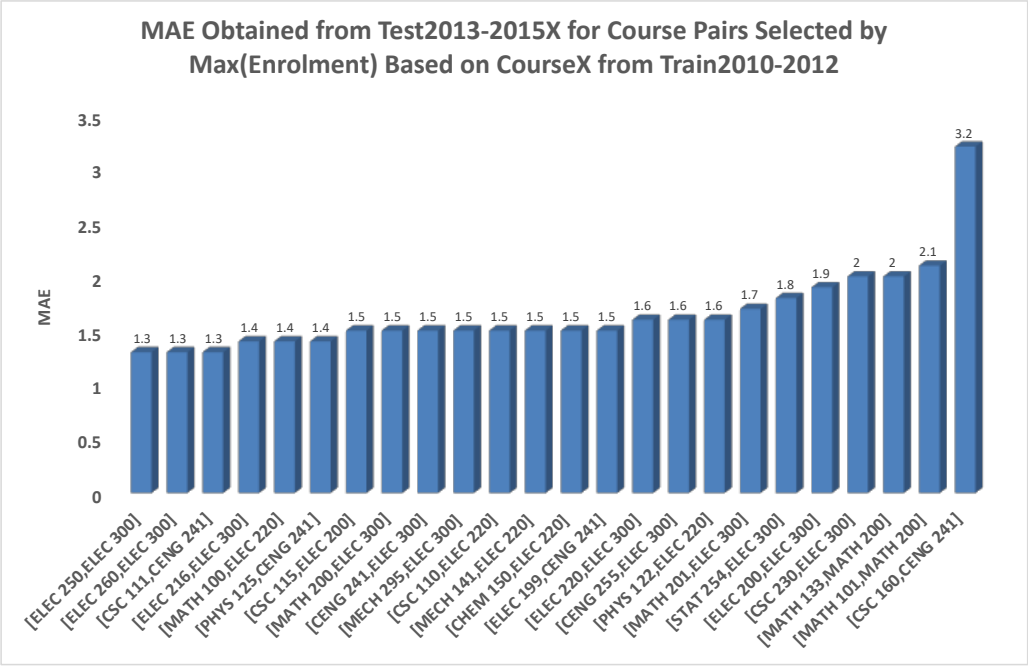

Figure 73 Prediction MAEs for Course Pairs Selected by Max(Enrolment) Based on CourseX from Train2010-2012 in Test2013-2015X

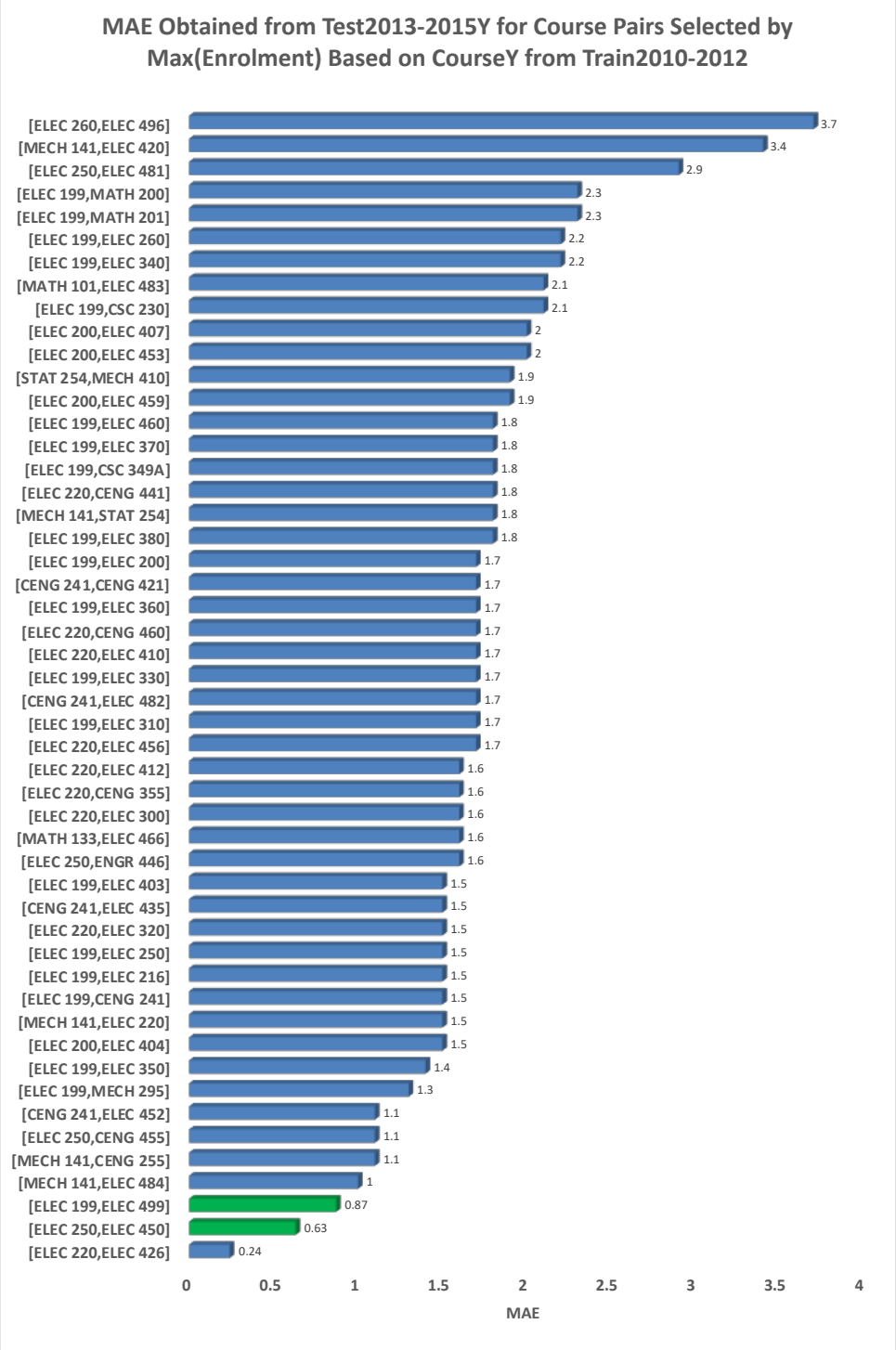

Figure 74 Prediction MAEs for Course Pairs Selected by Max(Enrolment) Based on CourseY from Train2010-2012 in Test2013-2015Y

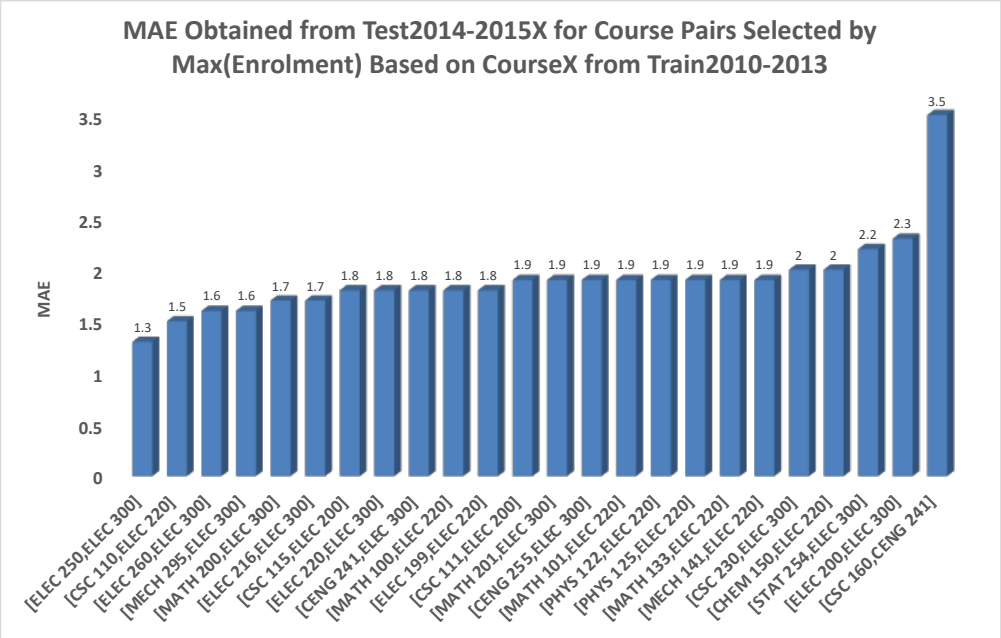

Figure 75 Prediction MAEs for Course Pairs Selected by Max(Enrolment) Based on CourseX from Train2010-2013 in Test2014-2015X

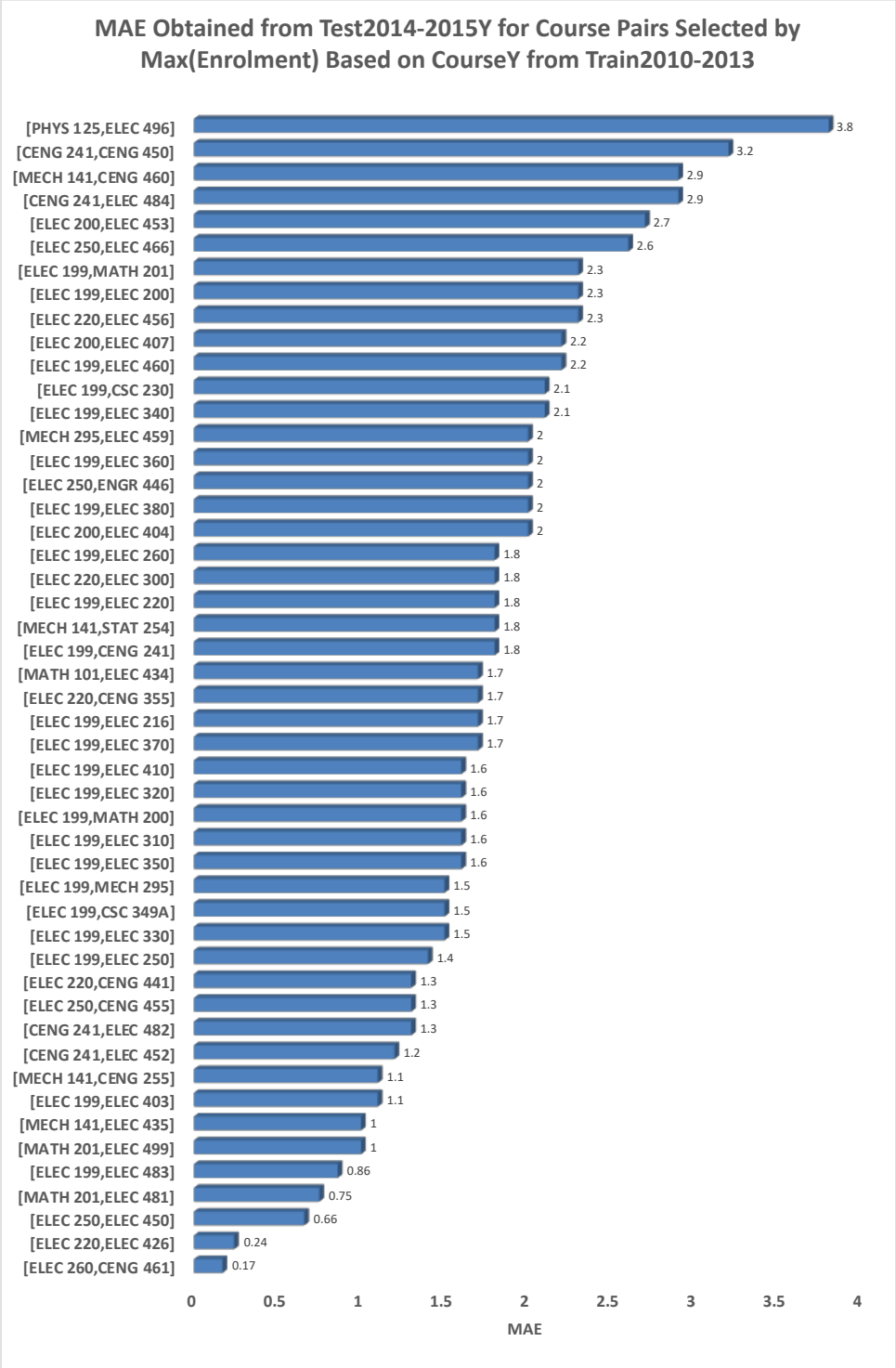

Figure 76 Prediction MAEs for Course Pairs Selected by Max(Enrolment) Based on CourseY from Train2010-2013 in Test2014-2015Y

Appendix 13 Precisions of Course Pairs Selected by Max(Enrolment)

The prediction precisions of course pairs selected by the maximum of the enrolment from Train2010-2012, Train2010-2013 are shown in figures below.

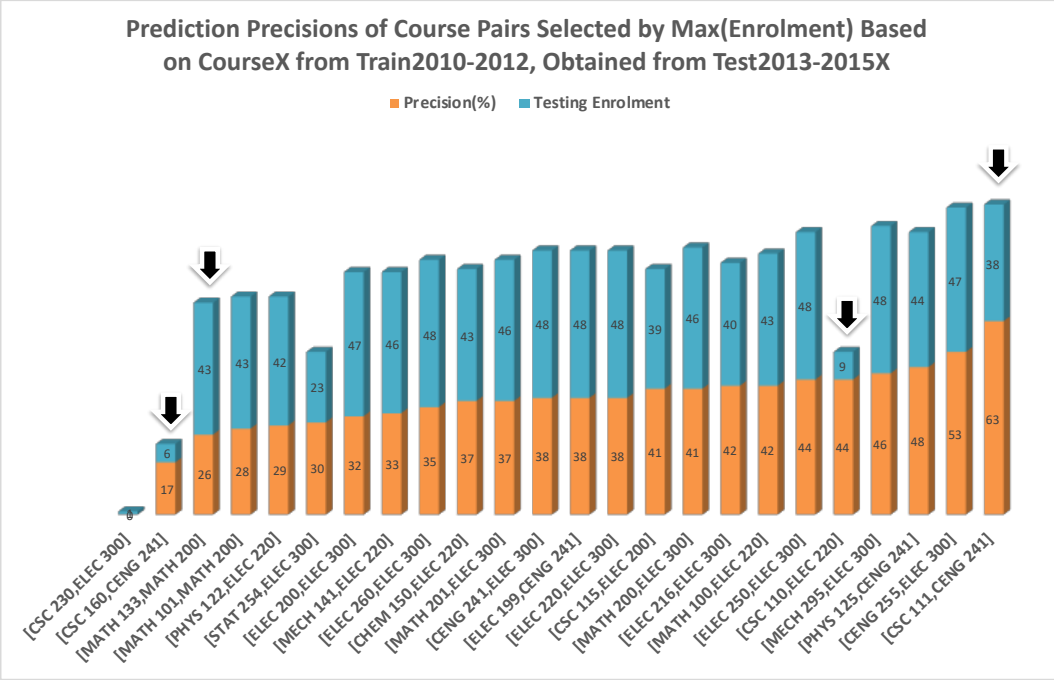

Figure 77 Prediction Precisions of Course Pairs Tested in Test2013-2015, Trained by Train2010-2012

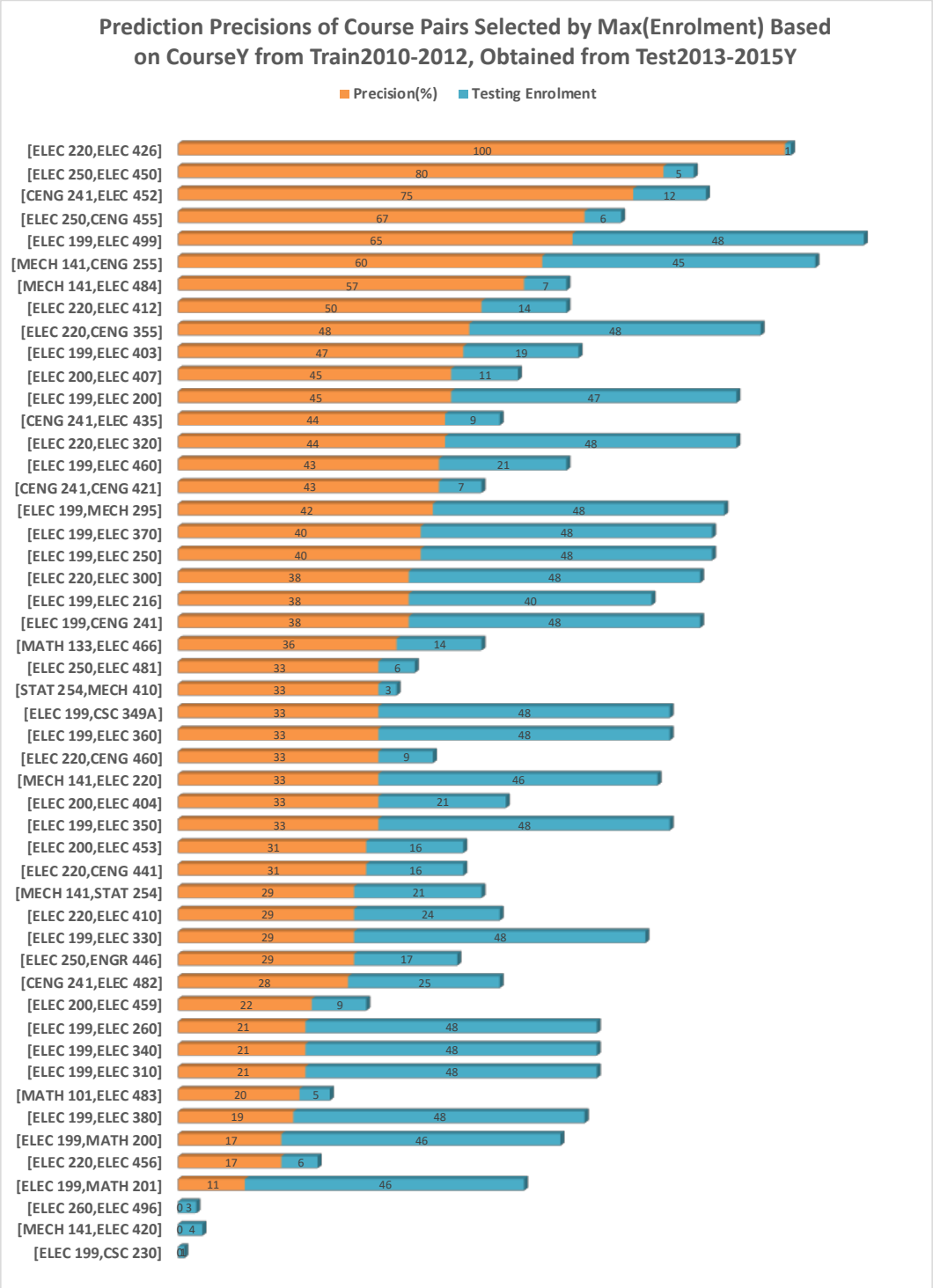

Figure 78 Prediction Precisions of Course Pairs Tested in Test2013-2015, Trained by Train2010-2012

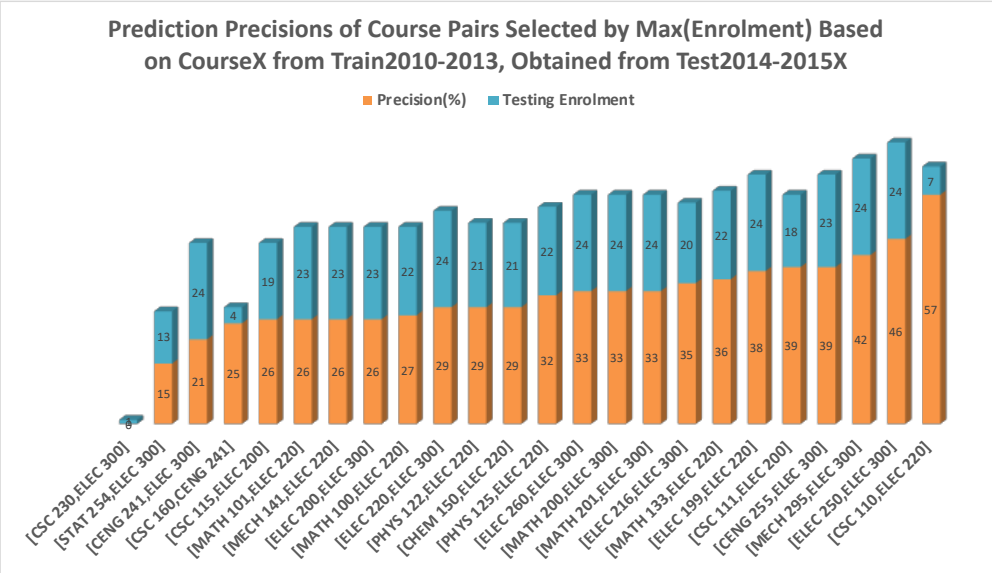

Figure 79 Prediction Precisions of Course Pairs Tested in Test2014-2015, Trained by Train2010-2013

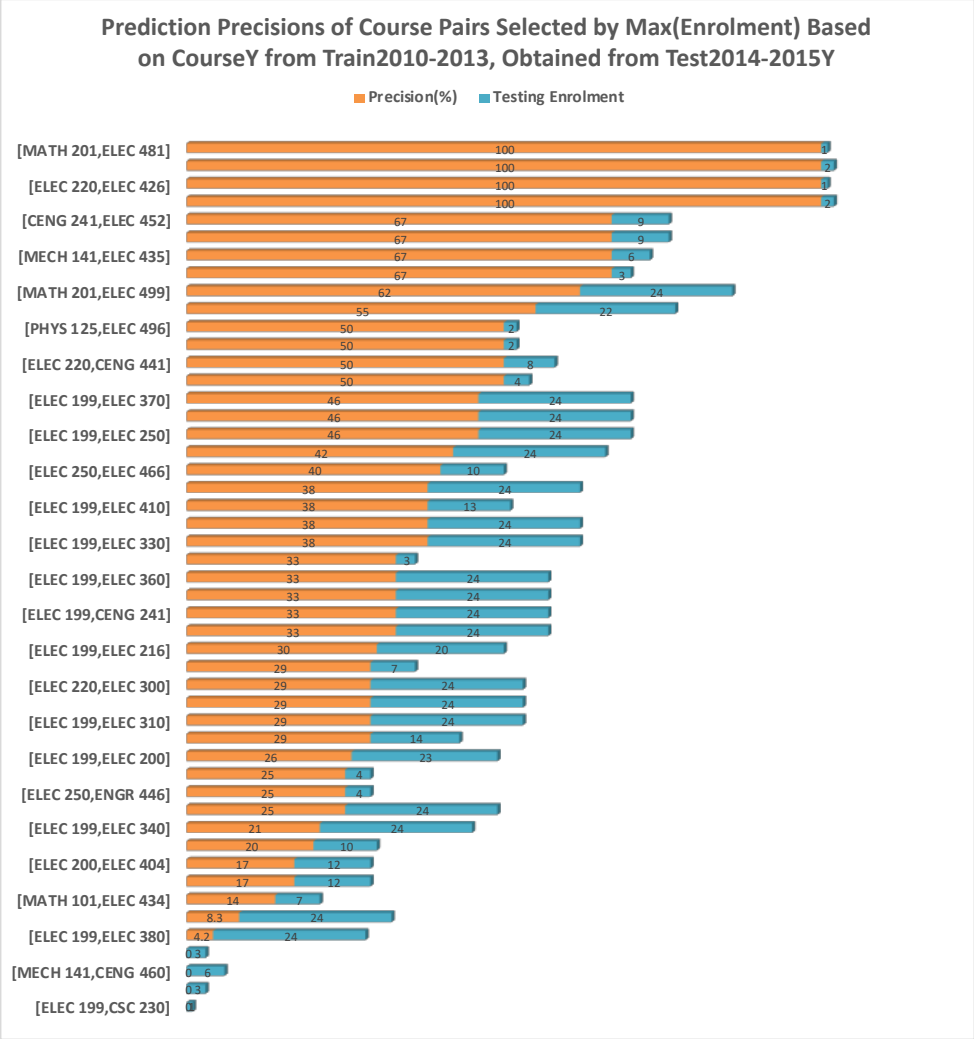

Figure 80 Prediction Precisions of Course Pairs Tested in Test2014-2015, Trained by Train2010-2013

Appendix 14 Course Pairs Selected by Max(P_i)

The course pairs selected by the maximum of the combination of coefficient and enrolment from train2010-2011, Train2010-2012, Train2010-2013 are shown in tables below.

Table 19 Selected Course Pairs for CourseX by Using P_i from Train2010-2011

CourseX	CourseY	CourseX	CourseY	CourseX	CourseY
MATH 100	MATH 201	CHEM 150	ELEC 466	CSC 230	ELEC 482
MATH 101	ELEC 452	CSC 160	ELEC 499	CENG 241	ELEC 340
CSC 110	CENG 460	ELEC 199	ELEC 484	ELEC 250	ELEC 320
CSC 115	ELEC 300	MATH 200	ELEC 420	STAT 254	ELEC 340
PHYS 122	MATH 200	ELEC 200	ELEC 370	CENG 255	ELEC 482
PHYS 125	MECH 295	MATH 201	ELEC 420	ELEC 260	ELEC 310
MATH 133	MATH 200	ELEC 216	CENG 460	MECH 295	ELEC 459
MECH 141	ELEC 200	ELEC 220	ELEC 456		

Table 20 Selected Course Pairs for CourseY by Using P_i from Train2010-2011

CourseX	CourseY	CourseX	CourseY	CourseX	CourseY
PHYS 122	MATH 200	ELEC 260	ELEC 330	CHEM 150	ELEC 426
MECH 141	ELEC 200	ELEC 250	ELEC 340	ELEC 260	CENG 441
MATH 100	MATH 201	ELEC 250	CSC 349A	PHYS 125	ENGR 446
MECH 141	ELEC 216	ELEC 260	ELEC 350	PHYS 125	ELEC 450
MATH 133	ELEC 220	MECH 141	CENG 355	MATH 101	ELEC 452
CHEM 150	CSC 230	ELEC 260	ELEC 360	CSC 110	ELEC 453
CHEM 150	CENG 241	ELEC 260	ELEC 370	MATH 133	CENG 455
CHEM 150	ELEC 250	ELEC 260	ELEC 380	ELEC 220	ELEC 456
MATH 101	STAT 254	MATH 201	ELEC 403	PHYS 122	ELEC 459
MECH 141	CENG 255	ELEC 200	ELEC 404	CSC 110	CENG 460

MECH 141	ELEC 260	CENG 241	ELEC 407	ELEC 260	ELEC 460
MECH 141	MECH 295	CSC 110	MECH 410	CHEM 150	ELEC 466
CSC 115	ELEC 300	MATH 201	ELEC 410	CSC 230	ELEC 482
ELEC 260	ELEC 310	ELEC 220	ELEC 412	MATH 133	ELEC 484
ELEC 250	ELEC 320	MATH 201	ELEC 420	MATH 200	ELEC 496
				CSC 160	ELEC 499

Table 21 Selected Course Pairs for CourseX by Using P_i from Train2010-2012

CourseX	CourseY	CourseX	CourseY	CourseX	CourseY
MATH 100	MATH 200	MECH 141	ELEC 216	ELEC 220	ELEC 456
MATH 101	MATH 201	CHEM 150	ELEC 250	CSC 230	CENG 441
CSC 110	ELEC 200	CSC 160	ELEC 360	CENG 241	ELEC 483
CSC 111	ELEC 380	ELEC 199	MATH 201	ELEC 250	ELEC 320
CSC 115	CENG 241	MATH 200	ELEC 310	STAT 254	ELEC 483
PHYS 122	MATH 200	ELEC 200	ELEC 483	CENG 255	ELEC 340
PHYS 125	MATH 200	MATH 201	ELEC 310	ELEC 260	ELEC 310
MATH 133	MATH 200	ELEC 216	ELEC 330	MECH 295	ELEC 360

Table 22 Selected Course Pairs for CourseY by Using P_i from Train2010-2012

CourseX	CourseY	CourseX	CourseY	CourseX	CourseY
MATH 100	MATH 200	ELEC 250	CSC 349A	CSC 115	ENGR 446
MECH 141	ELEC 200	ELEC 260	ELEC 350	MATH 101	ELEC 450
MATH 100	MATH 201	CHEM 150	CENG 355	CENG 241	ELEC 452
MECH 141	ELEC 216	ELEC 250	ELEC 360	ELEC 200	ELEC 453
CSC 115	ELEC 220	ELEC 260	ELEC 370	MATH 133	CENG 455
MECH 141	CSC 230	ELEC 260	ELEC 380	ELEC 220	ELEC 456
CSC 115	CENG 241	CSC 115	ELEC 403	MECH 141	ELEC 459
CSC 115	ELEC 250	ELEC 200	ELEC 404	ELEC 260	ELEC 460
CSC 115	STAT 254	PHYS 125	ELEC 407	CHEM 150	CENG 460

MECH 141	CENG 255	MATH 201	ELEC 410	MATH 133	ELEC 466
MECH 141	ELEC 260	STAT 254	MECH 410	MATH 201	ELEC 481
MECH 141	MECH 295	ELEC 220	ELEC 412	CSC 115	ELEC 482
ELEC 250	ELEC 300	MATH 201	ELEC 420	CENG 241	ELEC 483
ELEC 260	ELEC 310	CENG 241	CENG 421	CENG 241	ELEC 484
ELEC 250	ELEC 320	MECH 295	ELEC 426	MATH 101	ELEC 496
ELEC 260	ELEC 330	CHEM 150	ELEC 435	CSC 111	ELEC 499
ELEC 260	ELEC 340	CHEM 150	CENG 441		

Table 23 Selected Course Pairs for CourseX by Using P_i from Train2010-2013

CourseX	CourseY	CourseX	CourseY	CourseX	CourseY
MATH 100	MATH 200	MECH 141	ELEC 260	ELEC 220	ELEC 456
MATH 101	MATH 201	CHEM 150	ELEC 250	CSC 230	CENG 441
CSC 110	ELEC 200	CSC 160	ELEC 360	CENG 241	ELEC 310
CSC 111	CENG 241	ELEC 199	MATH 201	ELEC 250	CSC 349A
CSC 115	CENG 241	MATH 200	ELEC 310	STAT 254	ELEC 340
PHYS 122	MATH 200	ELEC 200	ELEC 434	CENG 255	ELEC 340
PHYS 125	MATH 200	MATH 201	ELEC 310	ELEC 260	ELEC 310
MATH 133	MATH 200	ELEC 216	ELEC 330	MECH 295	ELEC 320

Table 24 Selected Course Pairs for CourseY by Using P_i from Train2010-2013

CourseX	CourseY	CourseX	CourseY	CourseX	CourseY
PHYS 122	MATH 200	ELEC 250	CSC 349A	CENG 241	ELEC 452
MECH 141	ELEC 200	ELEC 260	ELEC 350	ELEC 200	ELEC 453
MATH 101	MATH 201	MECH 141	CENG 355	MATH 133	CENG 455
MECH 141	ELEC 216	ELEC 250	ELEC 360	ELEC 220	ELEC 456
MATH 133	ELEC 220	ELEC 260	ELEC 370	MECH 141	ELEC 459
MECH 141	CSC 230	ELEC 260	ELEC 380	ELEC 260	ELEC 460
CSC 115	CENG 241	MATH 201	ELEC 403	ELEC 220	CENG 460

CHEM 150	ELEC 250	ELEC 200	ELEC 404	ELEC 260	CENG 461
PHYS 122	STAT 254	PHYS 125	ELEC 407	MATH 133	ELEC 466
MECH 141	CENG 255	MATH 201	ELEC 410	MATH 201	ELEC 481
MECH 141	ELEC 260	MECH 295	ELEC 426	CENG 255	ELEC 482
MECH 141	MECH 295	MATH 101	ELEC 434	ELEC 199	ELEC 483
ELEC 250	ELEC 300	CENG 241	ELEC 435	CENG 241	ELEC 484
ELEC 260	ELEC 310	CHEM 150	CENG 441	MATH 101	ELEC 496
ELEC 250	ELEC 320	CSC 115	ENGR 446	MATH 201	ELEC 499
ELEC 260	ELEC 330	ELEC 250	ELEC 450		
ELEC 260	ELEC 340	MECH 295	CENG 450		

Table 25 Selected Course Pairs for CourseX by Using P_i from Train2010-2014

CourseX	CourseY	CourseX	CourseY	CourseX	CourseY
MATH 100	MATH 200	MECH 141	ELEC 260	ELEC 220	ELEC 456
MATH 101	MATH 201	CHEM 150	ELEC 250	CSC 230	CENG 441
CSC 110	ELEC 200	CSC 160	ELEC 466	CENG 241	ELEC 310
CSC 111	ELEC 330	ELEC 199	MATH 201	ELEC 250	CSC 349A
CSC 115	CENG 241	MATH 200	ELEC 310	STAT 254	ELEC 340
PHYS 122	MATH 200	ELEC 200	CENG 450	CENG 255	CENG 450
PHYS 125	MATH 200	MATH 201	CSC 349A	ELEC 260	ELEC 310
MATH 133	MATH 200	ELEC 216	ELEC 330	MECH 295	CENG 450

Table 26 Selected Course Pairs for CourseY by Using P_i from Train2010-2014

CourseX	CourseY	CourseX	CourseY	CourseX	CourseY
PHYS 122	MATH 200	ELEC 250	CSC 349A	ELEC 250	ELEC 450
MECH 141	ELEC 200	ELEC 260	ELEC 350	CENG 241	ELEC 452
MATH 101	MATH 201	MECH 141	CENG 355	ELEC 200	ELEC 453
MECH 141	ELEC 216	ELEC 260	ELEC 360	ELEC 250	CENG 455
MATH 133	ELEC 220	ELEC 260	ELEC 370	ELEC 220	ELEC 456

MECH 141	CSC 230	ELEC 260	ELEC 380	CHEM 150	ELEC 459
PHYS 122	CENG 241	MATH 201	ELEC 403	ELEC 260	ELEC 460
CHEM 150	ELEC 250	ELEC 250	ELEC 404	CSC 110	CENG 460
MATH 133	STAT 254	PHYS 122	ELEC 407	MATH 200	CENG 461
MECH 141	CENG 255	MATH 201	ELEC 410	CSC 160	ELEC 466
MECH 141	ELEC 260	ELEC 220	ELEC 412	MATH 201	ELEC 481
MECH 141	MECH 295	MECH 295	ELEC 426	STAT 254	ELEC 482
ELEC 260	ELEC 300	ELEC 199	ELEC 434	ELEC 199	ELEC 483
ELEC 260	ELEC 310	MECH 141	ELEC 435	CENG 241	ELEC 484
ELEC 250	ELEC 320	CHEM 150	CENG 441	MATH 101	ELEC 496
ELEC 260	ELEC 330	CSC 115	ENGR 446	MATH 201	ELEC 499
ELEC 260	ELEC 340	MECH 295	CENG 450		

Appendix 15 MAEs of Course Pairs Selected by Max(P_i)

The prediction MAEs of course pairs selected by the maximum of the combination of coefficient and enrolment from Train2010-2012, Train2010-2013 are shown in figures below.

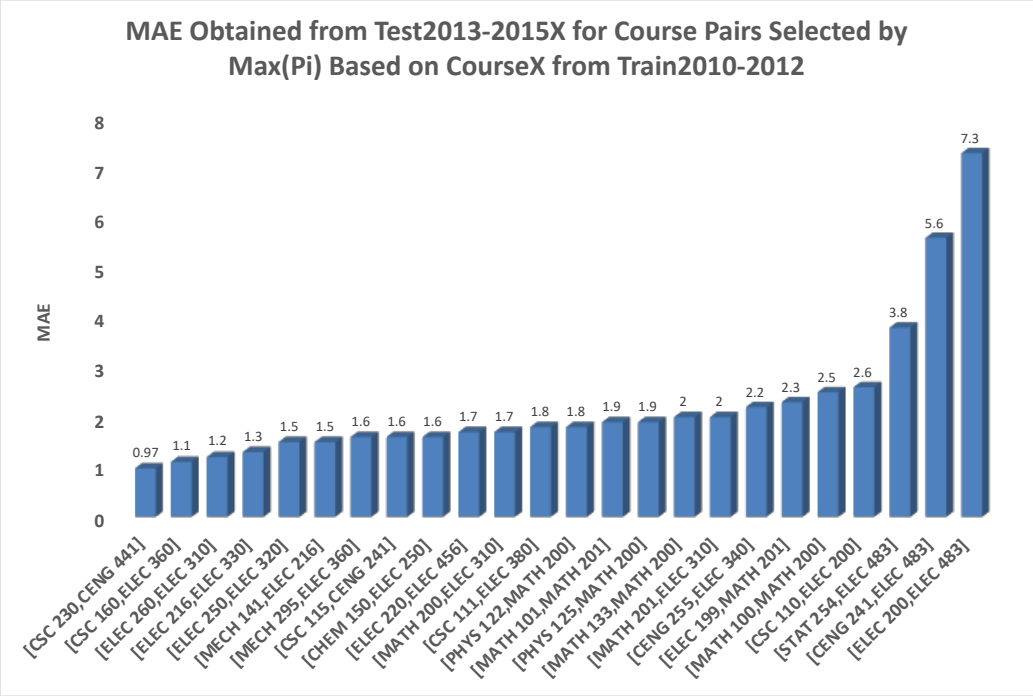

Figure 81 Prediction MAEs for Course Pairs Selected by Max(Pi) Based on CourseX from Train2010-2012 in Test2013-2015X

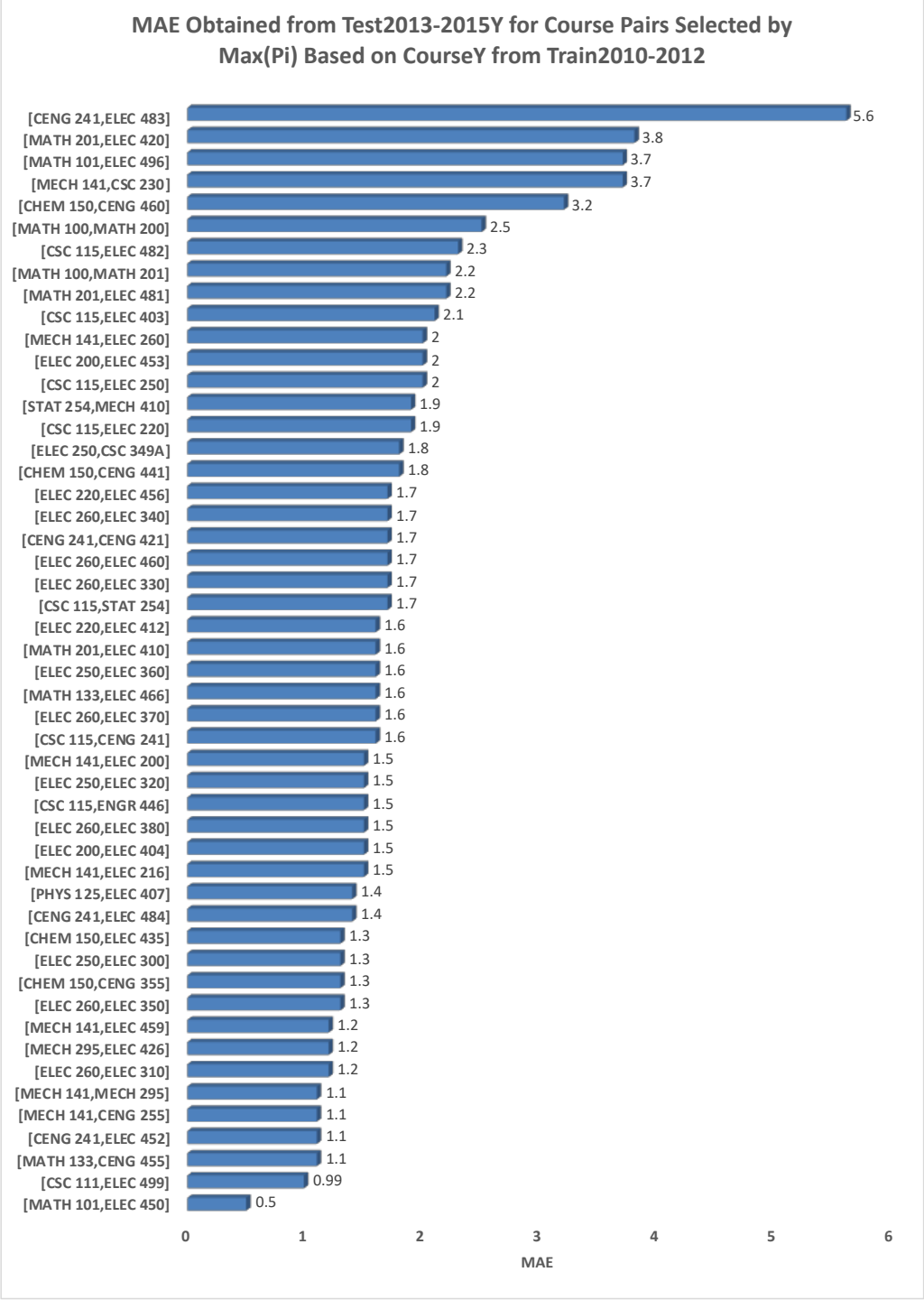

Figure 82 Prediction MAEs for Course Pairs Selected by Max(Pi) Based on CourseY from Train2010-2012 in Test2013-2015Y

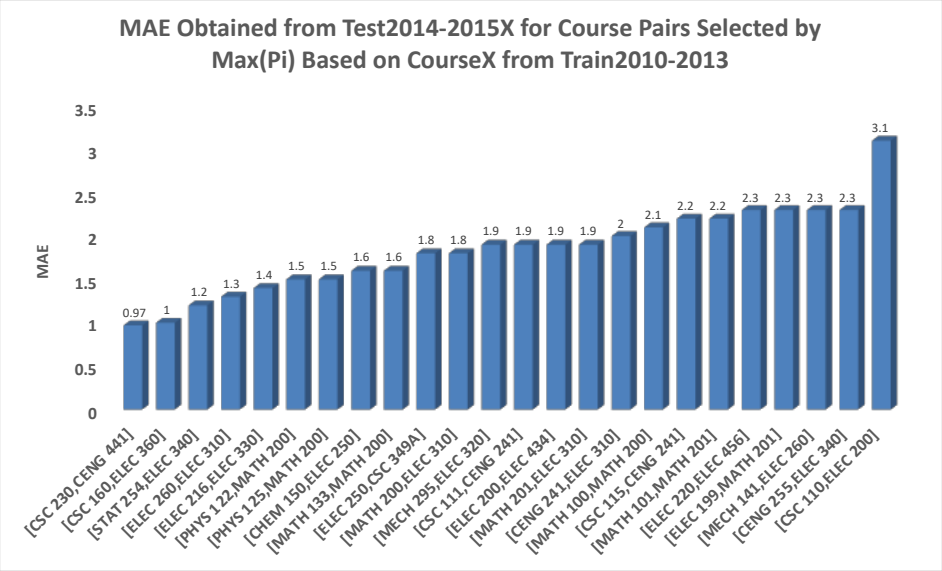

Figure 83 Prediction MAEs for Course Pairs Selected by Max(Pi) Based on CourseX from Train2010-2013 in Test2014-2015X

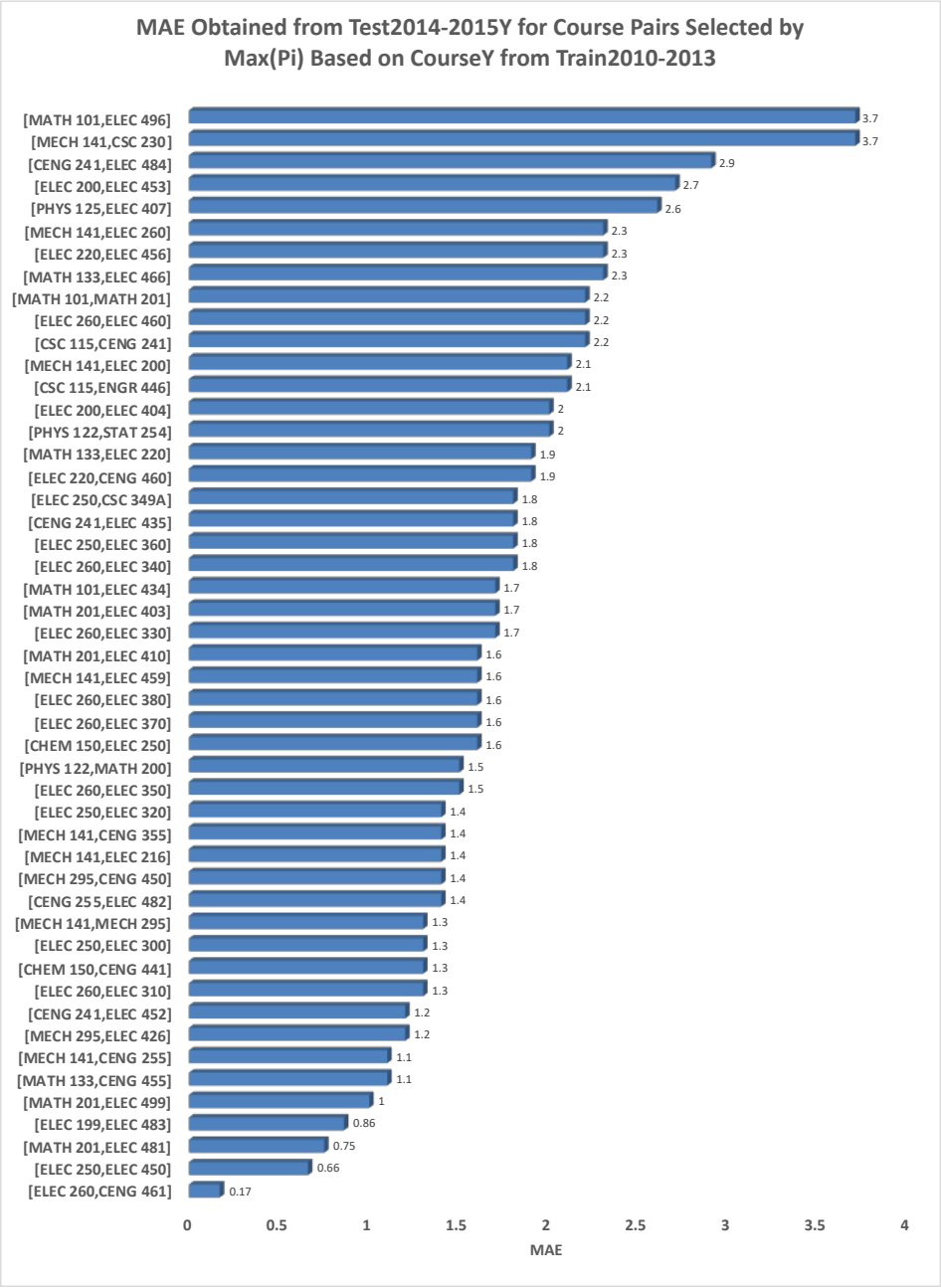

Figure 84 Prediction MAEs for Course Pairs Selected by Max(Pi) Based on CourseY from Train2010-2013 in Test2014-2015Y

Appendix 16 Precisions of Course Pairs Selected by Max(P_i)

The prediction precisions of course pairs selected by the maximum of the combination of coefficient and enrolment from Train2010-2012, Train2010-2013 are shown in figures below.

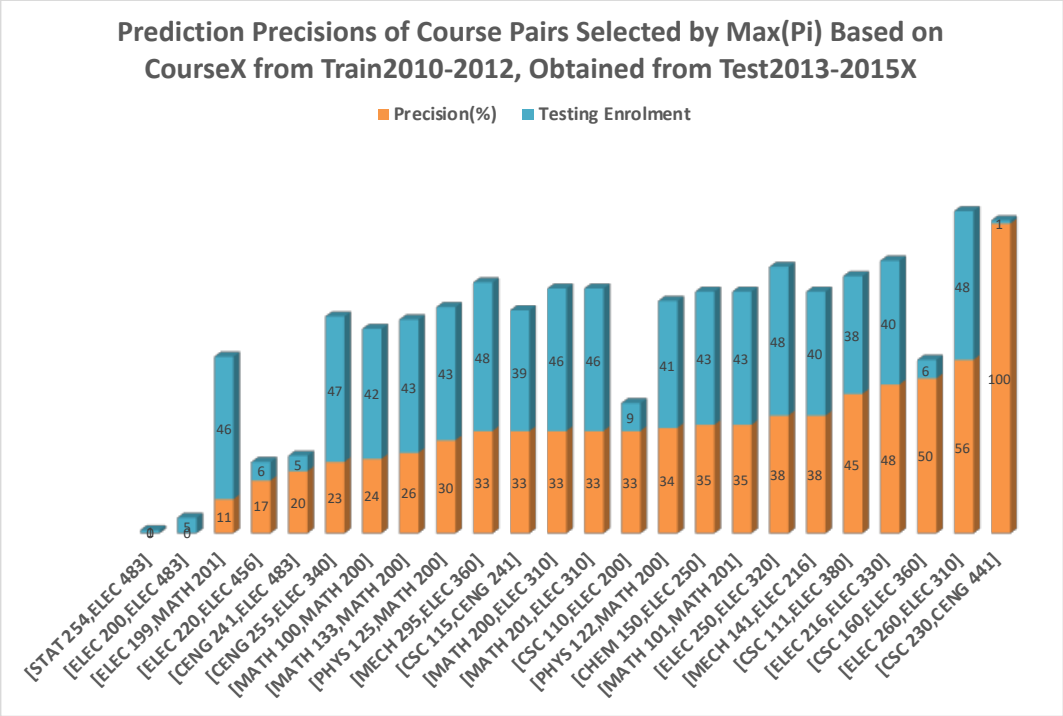

Figure 85 Prediction Precisions of Course Pairs in Test2013-2015X, Trained by Train2010-2012

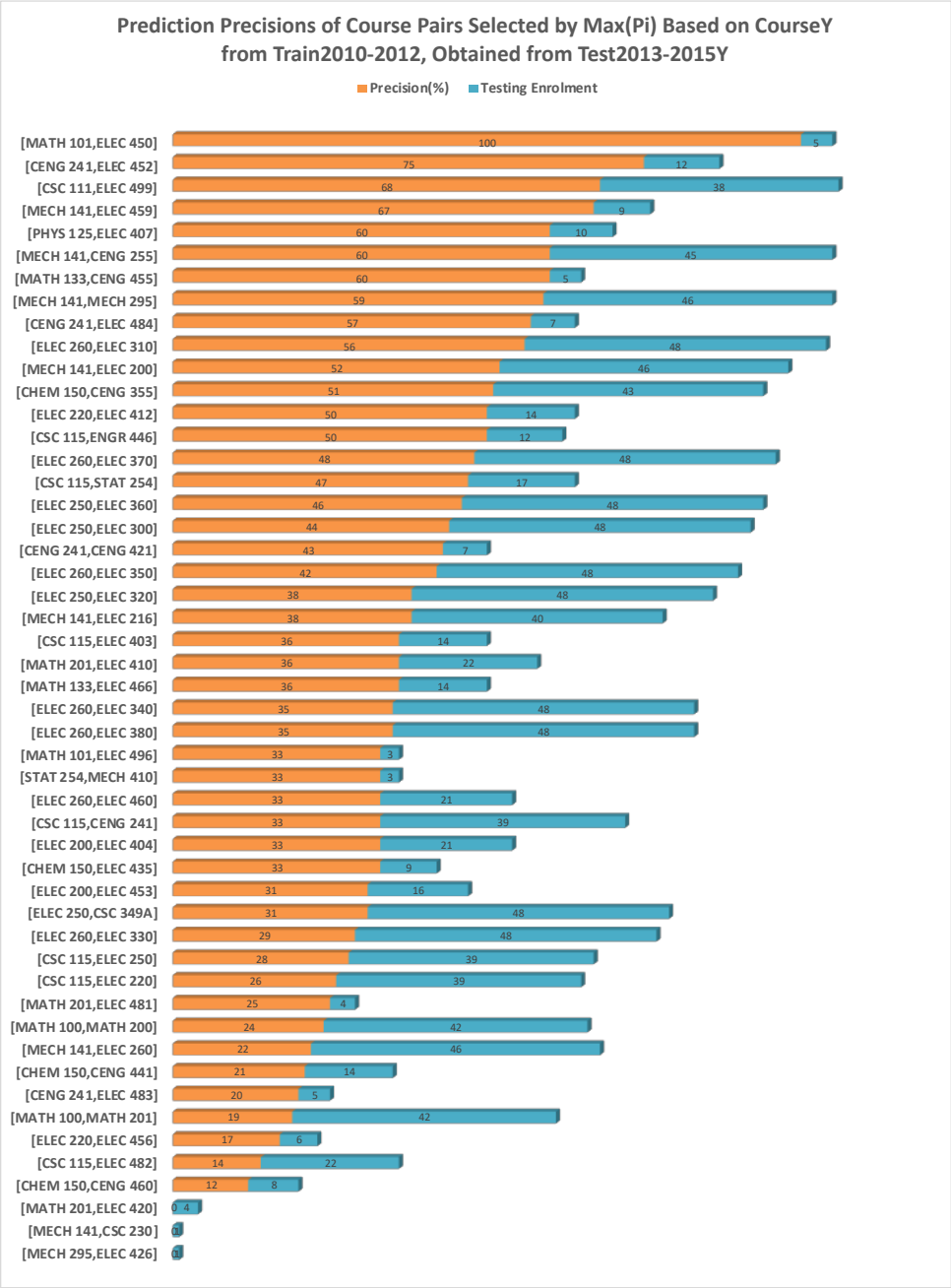

Figure 86 Prediction Precisions of Course Pairs in Test2013-2015Y, Trained by Train2010-2012

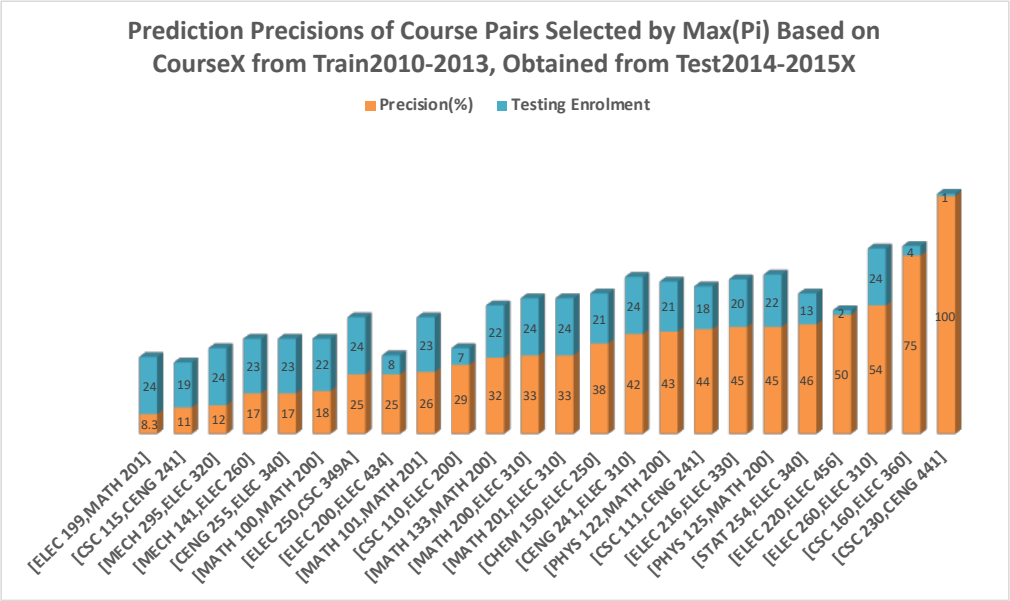

Figure 87 Prediction Precisions of Course Pairs in Test2014-2015X, Trained by Train2010-2013

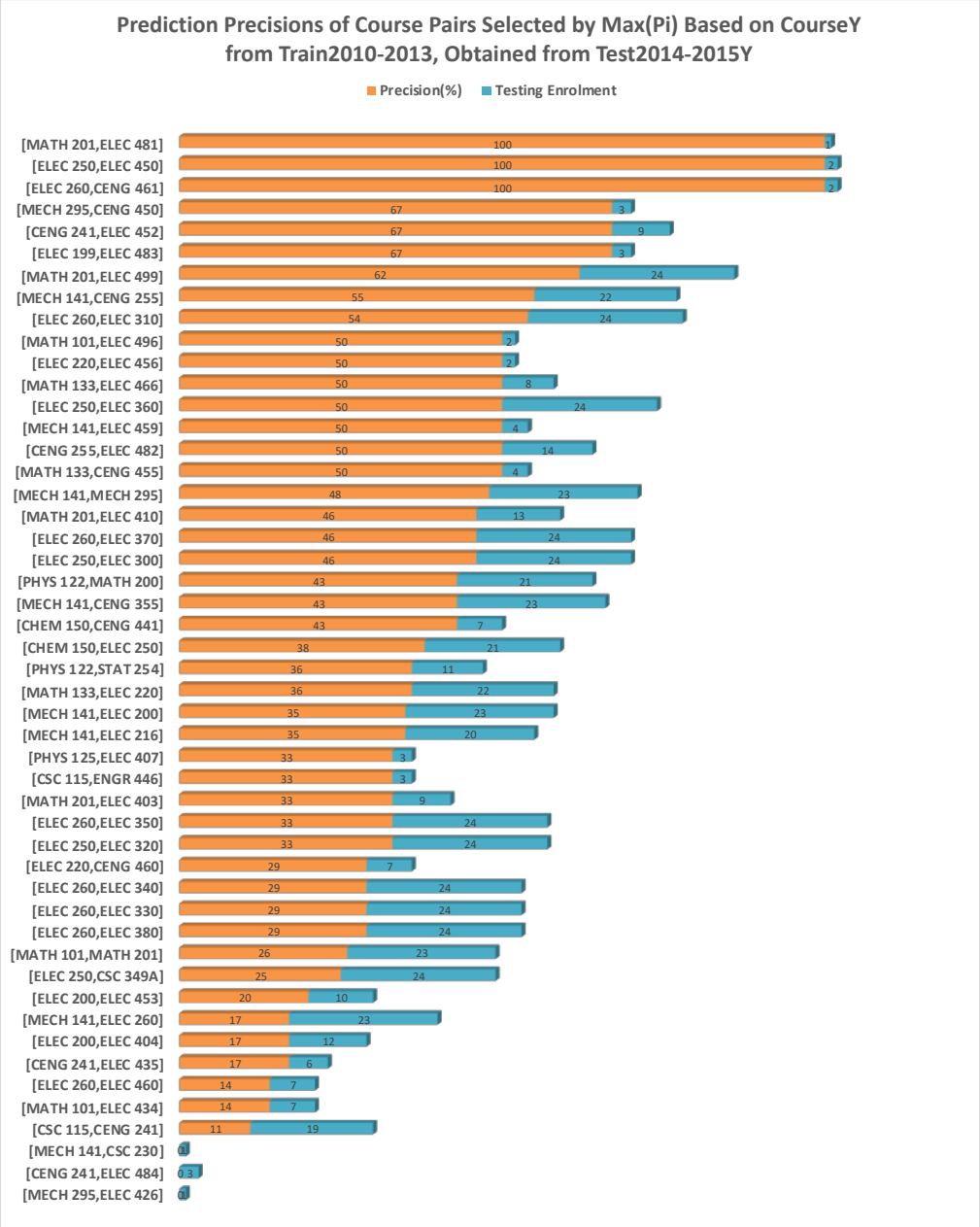

Figure 88 Prediction Precisions of Course Pairs in Test2014-2015Y, Trained by Train2010-2013