# Investigating Feature Importance in Educational Data, Towards Handling Data Missingness In Classification Tasks

by

Ndidi Mimi Obinwanne

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science
University of Alberta

# Abstract

The problem of missing data is unavoidable in many research fields, especially in education where data can be missing for justifiable reasons. Missing data causes bias in analysis, and traditional methods like complete case analysis and single imputation are suboptimal yet typically used to address the problem. These methods place emphasis on achieving complete datasets prior to attempting classification tasks. The consequences are a reduction in sample size, loss of statistical power, and loss of representation in the data. In this work, we investigate a simple approach to missing data and build upon the multiple imputation method. This simple approach avoids imputation and instead concatenates information about which features have missing values in an education dataset. This concatenation approach deprioritizes the estimation of values in order to provide an alternative to data completion. As a first attempt to demonstrate this approach is feasible, we conducted an investigation of how these methods for handling missing data affect two neural network architecture's ability to predict time to completion. To support this task, we first perform feature investigation using Structural Equation Modeling (SEM) to understand which features contain meaningful information. Results from this analysis showed that features containing data about student demography, high school performance details, English language skills, and university program details were important in understanding and explaining students' time to completion. We used SEM-identified features as input to a prediction task implemented with versions of the data that relied on current simple imputation techniques (zero imputation [ZNet], mean imputation [Mean], and iterative imputation [Iterative]) and one that used the non-imputation technique concatenation (Cat). We conducted training on two neural network architectures - SmallNets and MediumNets - and compared model performance across techniques. The results show that the non-imputation technique Cat,

achieved comparable or higher performance to that achieved by each of the three imputation techniques. Statistical tests in the SmallNets and MediumNets architecture showed differences existed between Cat and each of Mean and Iterative at different missingness levels. Cat outperformed Mean and Iterative with missingness levels at 10% and 80% in the SmallNets architecture. In the MediumNets architecture, it outperformed Mean and Iterative when missingness was at 40% and 80%. This indicates that Cat outperformed the imputation techniques Mean and Iterative at increasing missingness levels, and can perform better when used with a larger network. Our work provides a case study of the analysis and prediction of learner success even when data contains missingness, and it highlights that the simple concatenation approach might be sufficient for classification tasks with missing data.

# Preface

This thesis is an original work by Ndidi Mimi Obinwanne. The research project, of which this thesis is a part, received research ethics approval from the University of Alberta Research Ethics Board, Project Name “Program Improvement - Modelling Student Pathways”, No. Pro00083059, with an original approval date: 28 August 2018 and a current expiry date: 16 April 2024.

# Acknowledgments

I want to appreciate the astute supervision of my supervisors Carrie Demmans Epp and Martha White. Their guidance, support, advice, feedback and patience throughout the course of my study has been invaluable to my growth as a researcher. I thank members of the EdTekLA and RLAI labs for their encouragement and countless support in providing constructive feedback for my work. Special thanks to Lingwei Zhu, Andrew Patterson, Erfan Miahi, Daniela Teodorescu and Emma McDonald. My gratitude also goes to the University of Alberta, the Faculty of Science and the Computing Science department for funding my Master's program. Finally, I would like to thank my Mom for all her prayers and unwavering faith in me.

# Contents

# List of Tables

# List of Figures

# List of Abbreviations

CFA
: Confirmatory Factor Analysis

CFI
: Comparative Fit Index

ELP
: English Language Proficiency

ETS
: English Testing System

MAR
: Missing At Random

MCAR
: Missing Completely At Random

MI
: Multiple Imputation

MLE
: Maximum Likelihood Estimation

MNAR
: Missing Not At Random

MOOC
: Massive Open Online Courses

NaN
: Not a Number

RMSEA
: Root Mean Square Error Approximation

RQ
: Research Question

SEM
: Structural Equation Modelling

SRMR
: Standardized Root Mean Residual

# Chapter 1

# Introduction

In 2006, British mathematician Clive Humby declared that “data is the new oil”. This statement is better appreciated with the development of generative artificial intelligence tools like ChatGPT. This general-purpose conversation chatbot was powered by data and has shown potential benefits for many aspects of society, including education (Zhai, 2022). Another technological advancement that has powered learning through data is massive open online courses (MOOCs) (Romiszowski, 2013) with their potential to provide tailored content to students anywhere, at the student’s desired pace. Thus, we can agree with Clive Humby about the immense value data possesses. However, data is only as valuable as the insights it supports (Wickham, 2016). To achieve these insights, data must be analyzed.

Data analysis has become a cornerstone in every sector, providing insight into trends and predicting future outcomes. This reality is important because of the amount of data we generate and utilize daily. Global data generation within the next decade has been estimated to grow beyond 180 zettabytes (*Total Data Volume Worldwide 2010-2025*, n.d.). This growth in data generation has empowered many ambitious attempts at developing algorithms for performing data analysis and predictions. Among these ambitious attempts are tools with structural complexities that can learn patterns from data; also known as machine learning. Increased computational power is also a phenomenon that is pivotal to the advancement of machine learning (Hwang, 2018). Today, data analysis is considered integral and is conducted in several fields to gain insight and enhance robust decision making. To conduct rigorous data analysis, complete datasets are required but are rarely attainable as many real-world circumstances lead to data incompleteness also technically known as missingness in data (Enders, 2010).

Many analyses and research in education are extracted from data with missingness (Cox et al., 2014). In higher education and system design, the domains from which our dataset is generated, data can be missing for a number of reasons: human error, technical issues, or the transient

nature of learners in MOOCs. Many MOOCs are free, and they are typically designed to enable large numbers of geographically dispersed students to learn through web-based platforms. In this way, data about courses taken can become unavailable when a student attains proficiency in one program level and moves to a new level before completing said program or simply discontinues the course (Leach & Hadi, 2017). In both traditional classroom and non-traditional educational settings, student responses to test questions may be incomplete for a number of reasons: poor system design (van de Oudeweetering & Agirdag, 2018), misunderstanding the question, and difficulty navigating the system (Rhemtulla & Hancock, 2016) among others. Data missingness can also result from designing poor survey instruments that deprioritize information about minority or underrepresented groups (McKnight et al., 2007). Data missingness is therefore unavoidable.

Missing data can prevent precise analysis and cause a distortion in the generation of accurate results and insights. In complete case analysis, one consequence of missing data is a reduction in sample size which alters results and underestimates (or does not account for) causal factors. If there are missing inputs in variables such as student nationality, term admitted or courses taken, there will be underrepresentation when observations with missing inputs are deleted, which is known as complete case analysis (Enders, 2010). This increases the likelihood of drawing incorrect conclusions (Stavseth et al., 2019). Missing data also introduces possible biases which results in false estimates and errors (Barnes et al., 2008). Left unchecked, this will eventually render system support ineffectual.

Additionally, ethical considerations make it important to accommodate students who withhold identity information during self-disclosure procedures. This ensures that research analyses produce inclusive, accurate and quality results. A recent study (W. Li et al., 2022) which grouped student respondents by ethnicity and gender, found that students who self-identified as Black, responded the least to email surveys. Low engagements were due to a lack of institutional trust, volume of data collected versus “perceived benefits” and how data collected were used by instructors. Investigating optimal methods that learn from incomplete data as a result of survey participants' preference not to disclose certain information, becomes even more pertinent. The problem of missing data is an important one. It presents an opportunity to understand and

develop solution methods for achieving accuracy in statistical analysis even when there are missing or unavailable input values.

Developing methods to derive insights from data with missing inputs provides an opportunity to understand usability issues and learning challenges, regardless of data completeness. There are two traditional methods for handling the problem of missing data. These methods include complete case analysis (also known as deletion) and single imputation. With complete case analysis, there is a reduction in statistical power which causes underrepresentation and introduces bias (McKnight et al., 2007). Imputation, on the other hand, focuses on first completing the dataset before performing analysis or prediction tasks (Rubin, 1988). While data completion enables data analysis, single imputation focuses on estimating values that can be used as a proxy for true values (the literature commonly refers to this as a replacement of missing values). It does this in a round-robin fashion that makes assumptions about the missing data without any theoretical grounding to support those assumptions. This also leads to the introduction of bias (McKnight et al., 2007) and renders analysis defective. In our work, we attempt a novel approach that deemphasizes the need to estimate values to complete the dataset before conducting analyses.

Imputation techniques such as zero imputation (ZNet), mean imputation (Mean), and iterative imputation (Iterative) focus on data completion. Non-imputation techniques like concatenation (Cat) deprioritize data completion and instead, focus on using available data through the concatenation of incomplete data and indicator matrices. Concatenation provides us with the opportunity to achieve pattern recognition based on the missingness in the dataset. It does this by indicating missing values with ones, and available values with zeros. With this representation, during pattern recognition, a model can identify all missing values and focus on finding strong matches within values that are indicated as available. It essentially annotates the data with information about which values are missing and which are present so that the learned patterns can account for missing data. We apply pre-trained neural network models to the prediction task of classifying time to completion for undergraduate programs. The pre-trained models possess a representation of the data and capture valuable information about the input data such as patterns of missingness and meaningful features. From this representation, we assess performance of the

models on a specific binary classification task for our study. A comparison is made between the performance of imputation techniques and concatenation (Cat). Model performance is evaluated following training on neural networks. This is crucial because neural networks are considered "powerful modeling tools" (Féraud & Clérot, 2002) for classification problems. Notwithstanding, achieving optimal performance is challenging when utilizing incomplete data even with neural networks (Markey et al., 2006). There is a consensus that neural networks possess superior modeling techniques and increasingly better prediction accuracy, but are complicated because of their internal structures (Rosé et al., 2019).

Neural networks have demonstrated remarkable performance in various complex tasks and continue to evolve with improvements in accuracy and efficiency. While they can achieve great performance with complex classification tasks, their "blackbox" nature makes it challenging to understand how performance is achieved. Thus, classification tasks with missing data will be challenging, even for a neural network. For this reason, we take a different approach. Since understanding the internal structures of a neural network proves an uphill task, we shift our focus to understanding the structural relationships among features in the data itself. This presents us with an opportunity to test and measure theoretical models within the data. In this way, we gain a better understanding of the data. We conduct these analyses to understand whether there are underlying patterns within the data that can be recognized by a neural network so that we can use the identified features as input to the model.

To this end, we focus on investigating features in a dataset. We do this by understanding the structural relationships among features when the data is complete. Understanding structural patterns in a dataset requires a statistical method. We chose Structural Equation Modeling (SEM), a statistical analysis method that accommodates multiple variables in a single model. SEM models and analyses complex relationships among multiple features, to understand influence, identify unobserved (latent) constructs, and depict directional path diagrams amongst variables (Bowen & Guo, 2011). In this way, we understand relationships amongst variables within the data and discover features that provide useful inputs to a neural network model.

To summarize, data analysis is a monumental piece in the development of new technologies that advance progress in many sectors. Effective data analysis is not possible, however, when data is missing. Unfortunately, data missingness is unavoidable. Current handling methods are suboptimal because they make assumptions of the missing data. Without a novel method that ignores the need to complete data, the problem will continue to render data analysis defective and decision-making ineffectual. This can result in adverse policy formulation and practices that ostracize and, thereby, harm under-represented populations when there is a reduction in sample size. Defective analysis will impede advances in educational technologies such as MOOCs that provide education that is affordable and accessible (Li, 2019). With higher computational power and affordable storage, we argue that the objective of finding other methods that can generate insights from data, even with missing inputs is achievable. We posit that it is possible to learn from incomplete data, without first completing it by imputing values to compensate for its missingness. We are of the opinion that feature investigation will enhance prediction abilities and enable a different approach when handling missingness. The present thesis reports on a case study that provides a proof of concept for our ability to predict student outcomes in the presence of missingness.

The significance of this work is twofold: a) It provides a case study that uses structural equation modeling to analyze a real education dataset to identify important features and provide insight into student success as measured by time to completion. b) It highlights that a simple approach to missingness (Cat) is comparable to other established methods, providing an easy-to-use, but likely less biased, alternative to complete case analysis and imputation techniques.

This thesis has been structured to first provide a literature review on different contexts of missingness, their strengths and limitations (Chapter 2). Within this chapter, we also describe how missingness is represented in data, statistical and machine learning handling approaches, and provide a review of missingness in educational domains. In Chapter 3, we outline our methods for achieving the research goal of predicting student time to completion, and assessing the performance of imputation and non-imputation techniques. We begin by describing the data and the data cleaning steps we undertook. This was important because our study makes a case for conducting data mining when attempting prediction tasks, using non-data completion techniques

on missing data. In Chapter 4, we describe the feature investigation method we implemented (structural equation modeling [SEM]), the various SEM research questions we asked to identify meaningful features, our SEM implementation procedures, and a narration of how to interpret results from our tests. We also detail and interpret our results. Chapter 5 describes the task of making predictions using a dataset composed of features we identified through SEM. We introduce artificial missingness and obtain results from training the data on two neural network architectures - SmallNets and MediumNets. We also discuss our results from the prediction task and outline the ethical considerations and limitations of our work. Our conclusions and suggestions for future work are presented in Chapter 6.

# Chapter 2

# Literature Review

In this chapter, we review prior research on the mechanisms of missingness, different handling approaches - statistical and machine learning methods - missingness in educational data, and summarize with an outline of our research goal.

## 2.1 Mechanisms of Missingness

An understanding of the different contexts of missingness is important before investigating optimal handling approaches. Data can be missing within various contexts; MAR (Missing At Random), MCAR (Missing Completely At Random), and MNAR (Missing Not At Random). When data is MAR, the likelihood of missingness depends on observed data. When data is MCAR, there is an equal likelihood of missingness amongst all observed data. When data is MNAR, the likelihood of missingness depends on unobserved data (Papageorgiou et al., 2018).

The context of missingness helps in developing optimal solution methods. Various tests have been proposed to distinguish amongst data that are MAR, MCAR, and MNAR but the results of these tests may lead to bias because they depend on assumptions about the missing values and their relationships with observed data. Tests such as covariance tests and t-tests (or mean tests) have been proposed where mean and covariance comparisons are made to eliminate MCAR or MAR, which are easier to detect than MNAR. Each of these tests showed weaknesses that led to the conclusion that “mean comparisons do not provide a conclusive test of MCAR because MAR and MNAR mechanisms can provide missing data subgroups with equal means” (Enders, 2010).

## 2.2 Representation of Missingness in Data

It is important to acknowledge, from a theoretical standpoint, the distinction between zeros that are native to a complete dataset and zeros that are artificially added. In our work, the data contained both native and artificially added zeros. There is a subtle and important difference

between zeros that are native to the dataset and imputation as it is traditionally known: making replacements that are deemed to be correct guesses in the spots where a value is missing within an incomplete dataset. When we have missing values in a dataset, it typically presents as NaN (Not a Number). Table 2.1, shows a sample complete dataset which we edit in the succeeding Table 2.2 to represent missingess as encountered in the real world. Table 2.3 depicts the dataset with the representation of NaNs and Table 2.4 shows the data when NaNs are represented by zeros.

**Table 2.1:** Sample Complete Dataset

| Gender | Age | PremiumPupil | Confidence | QuestionId | ParentId | SubjectId | CorrectAnswer | AnswerValue | IsCorrect |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 17 | 1.0 | 100.0 | 12147 | 3 | 0 | 1 | 1 | 1 |
| 1 | 16 | 1.0 | 50.0 | 12147 | 3 | 0 | 1 | 1 | 1 |
| 2 | 17 | 0.0 | 25.0 | 12147 | 3 | 3 | 1 | 1 | 1 |
| 1 | 16 | 0.0 | 50.0 | 12147 | 3 | 2 | 1 | 1 | 1 |
| 1 | 16 | 1.0 | 100.0 | 12147 | 3 | 0 | 1 | 3 | 0 |

**Table 2.2:** Edited Sample Incomplete Dataset

| Gender | Age | PremiumPupil | Confidence | QuestionId | ParentId | SubjectId | CorrectAnswer | AnswerValue | IsCorrect |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 17 | 1.0 | 100.0 | 12147 | 3 | 0 | 1 | 1 | 1 |
| 1 | 16 | ? | ? | 12147 | 3 | 0 | 1 | 1 | 1 |
| 2 | 17 | ? | 25.0 | 12147 | 3 | 3 | 1 | 1 | 1 |
| 1 | 16 | ? | ? | 12147 | 3 | 2 | 1 | 1 | 1 |
| 1 | 16 | ? | ? | 12147 | 3 | 0 | 1 | 3 | 0 |

Note: ? indicates missing data

**Table 2.3:** Sample Incomplete Dataset (with NaNs)

| Gender | Age | PremiumPupil | Confidence | QuestionId | ParentId | SubjectId | CorrectAnswer | AnswerValue | IsCorrect |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 17 | 1.0 | 100.0 | 12147 | 3 | 0 | 1 | 1 | 1 |
| 1 | 16 | NaN | NaN | 12147 | 3 | 0 | 1 | 1 | 1 |
| 2 | 17 | NaN | 25.0 | 12147 | 3 | 3 | 1 | 1 | 1 |
| 1 | 16 | NaN | NaN | 12147 | 3 | 2 | 1 | 1 | 1 |
| 1 | 16 | NaN | NaN | 12147 | 3 | 0 | 1 | 3 | 0 |

Machine learning methods are unable to handle the presentation of NaNs in a dataset. Thus, when dealing with missingness through imputation, NaNs are typically replaced with zeros (0s).

**Table 2.4:** Sample Incomplete Dataset (with NaNs replaced with Zeros)

| Gender | Age | PremiumPupil | Confidence | QuestionId | ParentId | SubjectId | CorrectAnswer | AnswerValue | IsCorrect |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 17 | 1.0 | 100.0 | 12147 | 3 | 0 | 1 | 1 | 1 |
| 1 | 16 | 0.0 | 0.0 | 12147 | 3 | 0 | 1 | 1 | 1 |
| 2 | 17 | 0.0 | 25.0 | 12147 | 3 | 3 | 1 | 1 | 1 |
| 1 | 16 | 0.0 | 0.0 | 12147 | 3 | 2 | 1 | 1 | 1 |
| 1 | 16 | 0.0 | 0.0 | 12147 | 3 | 0 | 1 | 3 | 0 |

[1]Replacing NaNs with zeros presents a dataset to which machine learning methods can be applied. In Table 2.4, the "SubjectId" and "Confidence" columns have natural zeros that are valid values. Zeros can be native to a dataset when the value of a vector is 0. Survey participants' responses collected through a Likert–type questionnaire can be zero (0) if their response falls within that scale reference. For example, the response to an educational survey question such as "*On a scale of 0 to 5, with 5 being very confident and 0 being just guessing, how confident are you of your answer to question 1?",* can be 0. Thus, in this case, zero (0) is not a missing value but a valid input.

The data in Tables 2.2, 2.3, and 2.4 are used to illustrate these different data states and concerns. These data were taken from the NeurIPS 2020 education challenge dataset (*NeurIPS Education Challenge*, n.d.) - a competition where researchers around the world were invited to use machine learning methods to attempt various tasks. It was shared by Eedi - a leading London-based online educational platform that aims to personalize education for learners between the ages of 7 and 18. We give a brief description of the features of this dataset;

- DateOfBirth: year, month, and day as provided by the student.
- Gender: encoded numerical value that identifies the gender of each student (value is within [0,1,2,3] where 0 is unspecified, 1 is female, 2 is male and 3 is other.)
- PremiumPupil: a binary representation of whether the student has access to free school meals because of financial challenges or not - 0 indicates a Nonpremium pupil and 1 indicates a premium pupil.
- Confidence: the self-reported level of certainty a student provided for their answer to a question. Values are within [0 and 100], where 0 is a random guess and 100 is total certainty.
- QuestionId: a unique Id for each question.
- IsCorrect: a binary indicator that shows whether the student's answer is correct or incorrect (1 is correct, 0 is incorrect).

[1] Tables 2.1, 2.2, 2.3 and 2.4 contain educational data from the 2020 NeurIps Education Challenge *https://eedi.com/projects/neurips-education-challenge*

- CorrectAnswer: a numerical indicator for which multiple choice response item is correct. The responses are encoded as [A = 1, B = 2, C = 3, and D = 4].
- AnswerValue: answers provided by students to the multiple-choice questions which are encoded as [A = 1, B = 2, C = 3, and D = 4].

## 2.3 Handling Missing Data

In any study, it is important to understand the context of missingness because it informs the appropriate method that should be applied and the guidelines that direct implementation (Woods et al., n.d.). When analyzing data, we begin with understanding data characteristics through statistics, but accurate statistics are impossible when some input values are missing. Methodologists discourage the use of complete case analysis and single imputation methods for handling missingness (Wilkinson, 1999) because they lead to suboptimal analysis. Since statistical techniques are unable to analyze data with missing inputs, newer data handling techniques have emerged focusing on principled implementation approaches. Amongst these methods are multiple imputation (Rubin, 1988). Multiple imputation paved the way for other principled approaches like the maximum likelihood estimation which some studies (Dong & Peng, 2013) argue is a superior technique for handling data missingness. Two theoretical frameworks for missingness in data that provide solid foundations to explore solutions are Maximum Likelihood Estimation (MLE) and Multiple Imputation (MI) (Enders, 2010).

In applied research, the method of multiple imputation is not implemented as widely as expected due to a misunderstanding of the method itself and the contexts within which to apply it (Schafer & Graham, 2002). In a recent study (van Ginkel et al., 2020), these misunderstandings - phrased as misconceptions - were addressed with rebuttals that aimed at providing a better appreciation of the method to guide implementation. These studies (van Ginkel et al., 2020) have focused on clarifying that multiple imputation can be implemented under any context: MCAR, MAR and MNAR and that even when statistical tests show that data is not MAR, multiple imputation can be used. Notwithstanding, there still exists some hesitation in applied research to implementing multiple imputation because of the complexity of the method and the belief that it is only viable after complete case analysis has been attempted.

While MLE and MI are principled and widely accepted in the field, some studies have attempted to advance them with a combination of statistical methods and machine learning algorithms (Nelwamondo et al., 2007)  (Yadav & Roychoudhury, 2018). My review of their results showed improved performance in prediction when certain conditions were met and underperformance otherwise (Choudhury & Pal, 2019). Results also showed that imputation was done to first fill the missingness in most datasets used. This suggests that learning and prediction are achievable when some input vectors are missing (Le Morvan et al., 2020).

With the introduction of more rigorous methods for handling missingness, it seems rather counter-innovative that some foundational literature (Rubin, 1976) describe cases where it is acceptable to ignore missing data. Others (White & Carlin, 2010) suggest that there are scenarios where complete case analysis is an appropriate choice, even in epidemiological research. We argue that the current realities of our time - increased computational efficiency, exponential growth in data generation and affordable data storage - diminish grounds for ignoring missingness in data analysis. To support our position, we acknowledge that statistical regulatory guidelines (Wilkinson, 1999) discourage the use of traditional methods such as complete case analysis and single imputation. Consequently, the study, development, and implementation of principled approaches to handling missingness, provides a rationale for quality research to thrive (Dong & Peng, 2013). It is important to note that complete case analysis is still a valid method but only in studies where its application neither reduces sample size nor causes bias. These conditions are neither realistic nor achievable (Schafer & Graham, 2002).

In promoting principled approaches, one study (García-Laencina et al., 2010) categorized imputation methods for treating missingness into two groups - statistical imputation techniques and machine-learning-based imputation techniques. It is important to note that with this categorization, the size of the dataset and the volume of missing values account for the performance of imputation methods and their accuracy (Yadav & Roychoudhury, 2018).

In Table 2.5, we summarize the strengths and limitations of different contexts of missingness. We also provide an overview of handling approaches below.

- Complete case analysis: is mostly implemented when data is MCAR. An advantage to this handling approach is the simplicity it provides for analysis when incomplete observations are discarded. Conversely, it can result in loss of statistical power and lead to bias when the data mechanism is not MCAR (Rubin, 1976).
- Single imputation: describes replacement with single values that are estimated to be representative of the missing data and is typically used when data is MCAR. While it is useful when considering the simplicity of implementation, a disadvantage is the introduction of high bias and low variance in the dataset which makes it invalid when data is MAR (Rubin, 1988).
- Multiple imputation: describes multiple replacements of values that are estimated to be representative of the missing data from repeatedly sampling the dataset to achieve variation and account for the uncertainty of missingness. A benefit of this method is valid results through the production of less biased estimates when data is MAR, and the prevention of loss of statistical power. While this method provides a more rigorous process of ensuring bias is reduced, it is computationally expensive (Rubin, 1988).

**Table 2.5:** Contexts of Missingness, Applications, Strengths and Limitations

| **Methods** | **Strengths/Limitations** | **Type of Missingness** | | |
|---|---|---|---|---|
| | | **MCAR** | **MAR** | **MNAR** |
| Complete case analysis | Simplicity | ✓ | | |
| | Sample size reduction | ✓ | | |
| | Underrepresentation in data | ✓ | | |

| | | | | |
|---|---|---|---|---|
| | Compare across analyses | ✓ | | |
| | Can conduct analyses with incomplete data | ✓ | | |
| Single imputation | Potentially biased results | ✓ | | |
| | Ignores structural relationships between variables | ✓ | | |
| Multiple imputation | Accounts for variability due to sampling | ✓ | ✓ | ✓ |
| | Computationally expensive | ✓ | ✓ | ✓ |
| | Ability to handle errors | ✓ | ✓ | ✓ |

For context, consider a dataset showing late penalty fees for books returned to a library where we note that persons living farther from the library accrue higher fees than those who live close to the library. When distance is observed but fees are missing, we can estimate fees through observed data - distance - therefore, data in this context is missing at random (MAR). Additionally, if we determine that the rate of missingness is constant at 5% for every return, then data is missing completely at random (MCAR). When we are unable to estimate missing values based on observed data, we say that data is missing not at random (MNAR).

### 2.3.1 Imputation

Imputation is a missing data strategy focused on completing datasets for analyses. Two imputation methods as described by the literature are single and multiple imputation (Rubin, 1976). Single imputation follows a replacement strategy that replaces missing data with single values that are estimated to be representative of the missing data (Maydeu-Olivares, 2009). This

strategy does not account for uncertainty of the missing values. Other forms of single imputation that possess similar consequences are mean imputation, regression-based, and hot-deck imputation (Cox et al., 2014).

Multiple imputation is the state-of-the-art (Schafer & Graham, 2002) and is implemented by creating multiple versions of the data with estimated values. This approach allows it to account for uncertainty about the true values. Analyses are then conducted on the multiply imputed datasets, and the results present a statistically valid state that addresses inferences about the missing values (Patrician, 2002).

### 2.3.2 Statistical Approaches to Handling Missing Data

Multivariate Imputation by Chained Equations (MICE), is a statistical technique for multiple imputation that was introduced in 2011 (Buuren & Groothuis-Oudshoorn, 2011) and accounts for only linear relationships among parameters. The MICE technique has been successfully applied to datasets with 70% missingness where data is Missing At Random (MAR). It follows that MAR datasets are good candidates for multiple imputation methods (Buhi et al., 2008). A multivariate statistical method we employ in our work is Structural Equation Modeling (SEM). This method investigates observed (measured) and unobserved (latent) constructs within the data. It empowers our approach of ignoring the completion of data by providing useful information about important and unimportant features within the dataset. Using SEM enables the understanding of directional relationships amongst variables. SEM provides single analysis procedures to measure the direction and strength of observed and unobserved variables. It also employs confirmatory factor analysis and produces path diagrams that depict exact geometric measures of associations amongst variables. While SEM does not possess superior ability to depict causality amongst variables, it provides a strong basis for understanding associations and their strengths (Bowen & Guo, 2011). Implementing SEM requires the description of a hypothesis to support a previously defined theory. This necessitates the identification and reduction of features to support said defined theory. The principled approach to identification is Maximum Likelihood Estimation (MLE) (Bowen & Guo, 2011).

Research in the field of missing data has achieved many gains; from the development of principled approaches to the improvement of statistical software and packages (Buuren & Groothuis-Oudshoorn, 2011) that implement these principled approaches. Some studies have described and recommended best practices (Woods et al., n.d.) when dealing with missingness. The state-of-the-art approaches - Multiple Imputation (MI) and Maximum Likelihood Estimation (MLE) - provide solid foundations to build on. Even though these are solid bases for further work, they continue to focus on approximating values to fill in missing data inputs. In our work, we focused on methods that ignore completion of data and instead, find features within the incomplete data that can achieve learning tasks such as prediction. To achieve this objective, we emphasize the need to understand structural relationships among variables in the data.

### 2.3.3 Machine Learning and Other Approaches to Handling Missingness

A 2013 study (Nelwamondo et al., 2013) employed a combination of techniques comprising dynamic programming, genetic algorithms, and neural networks. The problem formulation was expressed as a set of possible states and corresponding actions and a reward for the accuracy of estimating missingness. In this way, the problem was structured as an optimality problem, a situation to which dynamic programming is typically suited (Nelwamondo et al., 2013). The category of neural networks employed in this work, known as auto-encoders, were preferred for their impressive ability for auto-association between variables in the input space (Thompson et al., 2003)**.** The base model (Abdella & Marwala, 2005) was developed by combining auto-encoders and genetic algorithms only. Results showed that the base model and dynamic programming technique assumed some level of correlation between variables. Understanding the degree to which each of these methods suppose a correlation effect amongst variables, would enable optimization of the policy being implemented. This work presents a different perspective to approaching missingness, through the explanation of variable correlations.

One study (Śmieja et al., 2018) presents a theoretically-grounded mechanism for training neural networks using incomplete data. The strategy is modeled on the uncertainty of the missing data using probabilistic function densities thereby eliminating the need for value imputations. The

probabilistic representation of the neural network is processed by considering the expected value, i.e, the average activation over imputations drawn from missing data density. In analyzing this model, especially because the model is trained on incomplete data, we note the possibility of loss of information. The study's theoretical analysis reports no loss in information and demonstrates this with a general probability density as opposed to density functions.

In 2018, a different handling approach, also using neural networks, implemented a novel architecture called Generative Adversarial Imputation Nets (GAIN) (Yoon et al., 2018). It implements imputation techniques through a "hint mechanism". The model provides hints to the neural network architecture to enable imputation. Through this, the architecture derives a generative model that assesses hint quality against results from post imputation. This model also handles challenges with the imputation method. Results from this work shows GAIN outperforming state-of-the-art imputation techniques that include MICE, MissForest, Auto-encoder, Matrix, and Expectation Maximization (EM).

In a recent study (Le Morvan et al., 2020), another principled architecture named NeuMiss networks was proposed. In the implementation of NeuMiss networks, a concatenation of incomplete data with an indicator dataset conditioned on missingness, is employed. Utilizing the concatenation technique creates non-linearity within the model and provides an opportunity for the model to scale to different contexts of missingness. NeuMiss networks require "medium-sized samples" (Le Morvan et al., 2020) of 1,000 to 100,000 observations to achieve the progress it provides.

## 2.4 Missingness in Educational Data Mining and Learning Analytics

In higher education research, missing data is often encountered. A 2004 review (James L. & Craig K., 2004) of studies published between 1999 and 2003 with missing data showed that in 2003, out of 545 studies examined, 229 had missing data. Many times the missing data were ignored and at other times, traditional methods like complete case analysis were applied. Even with the invention of more principled approaches like MLE and MI, research in this field has

substantially adopted traditional methods of handling the problem. This may be the case because many researchers are conflicted on the correct implementation of MLE and MI (van Ginkel et al., 2020). Another reason may be the prevalent expectation that the different contexts of missingness - MAR, MCAR, and MNAR - are mutually exclusive when in reality, all three can be present in a dataset at the same time (James L. & Craig K., 2004).

Some studies (Maydeu-Olivares, 2009) (Schafer & Graham, 2002) outline conditions for which traditional methods can be used to handle the problem but these conditions are rarely feasible in reality. It is therefore timely that recent reviews of missing data (van Ginkel et al., 2020), especially in education (Cox et al., 2014) (James L. & Craig K., 2004), acknowledge the need to prepare and circulate less technical communication procedures for employing MLE and MI to researchers. It is the hope that this will encourage more researchers in the field to adopt these principled approaches rather than traditional methods.

## 2.5 Our Research Goal

From our review of the literature, identifying the contexts of missingness is important in empowering research to find solutions to the problem. It is also evident that missingness is often encountered in educational data and is mostly handled with traditional methods. The literature posits the reasons as a misunderstanding of how to apply principled approaches, and the assumption that different contexts of missingness are mutually exclusive. Building on this understanding, we focus on the educational domain, simulate missingness, and employ simple imputation and non-imputation techniques to assess the performance of different methods for handling missingness within the Missing Completely At Random (MCAR) context. We aim to predict time to completion for undergraduate programs and we compare performance between imputation and non-imputation techniques to determine the ability of both methods to make predictions from incomplete data.

# Chapter 3

# Methods Overview and Data Description

In this chapter, with the aid of Figure 3.1, we illustrate the flow of our methods as they relate to our study objective. We also outline the steps we took to preprocess and describe the data. To prepare for SEM, we define variable categories for feature investigation. Methods specific to each modeling task are reported in their respective chapters. Chapter 4 contains specific details for SEM and Chapter 5 reports the details that are specific to prediction.

The approach to implementing techniques for our methods consists of applying simple neural network architectures for a binary-classification task. Since the goal is prediction with data that contain missingness, the concept of non-linearity (Le Morvan et al., 2020) becomes an important factor to consider. Acknowledging non-linearity is important in missing data. This is because direct relationships between dependent and independent variables become distorted when there are missing values in a dataset. It is therefore intuitive to adjust for non-linearity to provide statistical compensation for missingness. As a result, we explore imputation techniques that focus first on data completion. This category of techniques do not address the concept of non-linearity. We compare performance of these imputation techniques to a non-imputation technique that allows for the treatment of non-linearity within an incomplete dataset.

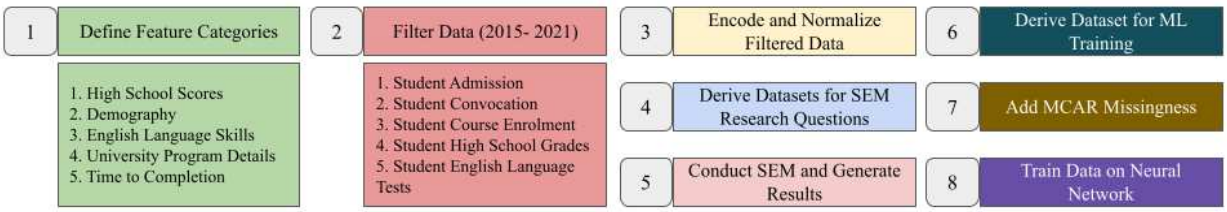
The current standard methodological approach for handling missing data is known as imputation; a technique where vectors containing missing data are first filled with estimated values in order to complete the data before attempting regression or classification tasks (Rubin, 1988). With imputation techniques, an assumption about the context of missingness is made. Traditional methods of handling missingness, therefore, are suboptimal but provide insights that suggest data completion can be ignored altogether when attempting to learn from an incomplete dataset. Some studies (Scheffer, 2002) have investigated thresholds, and conditions under which missingness levels can be ignored. Their results show that 5% and 10% missingness levels can be ignored with minimal adverse effects. As a result, more recent studies have attempted various techniques that aim at bypassing data completion. One study attempted a hint mechanism (Yoon et al.,

2018), while another attempted concatenation (Le Morvan et al., 2020). Results from these recent studies show that it is possible to find alternatives to data completion.

Since our data had been carefully curated, it had little missingness. To create an incomplete dataset, we randomly removed values in the train sets in incremental degrees and compared performance across imputation and non-imputation methods. Adding artificial missingness in this way created a context of missingness known as MCAR - Missing Completely At Random (Buhi et al., 2008). We chose the MCAR context of missingness as it is encountered in the real-world because MCAR is the simplest context of missingness to simulate (Rubin, 1976). The research on missing data typically attempts to recreate missingness through simulation. The context of missingness that is easily and most often simulated, is MCAR. Missingness within the context of MAR was not attempted because missingness in this context is conditional on something other than the dependent variable in focus (Scheffer, 2002). This contrasts with our objective of predicting a dependent variable conditioned on missingness within the data. Missingness within the MNAR context was also not attempted because this mechanism depends on the real value of the missing data. As a result, it is the most difficult condition to simulate and reproduce (Scheffer, 2002).

In our study with the University of Alberta Faculty of Science data, we preprocessed the dataset and employed Structural Equation Modeling (SEM) for the purpose of determining whether there are underlying patterns within the available data, when some values are missing. These underlying patterns may enable models to perform well without needing to first complete the dataset. Structural Equation Modeling was also important to determine informative features within an educational dataset that may be useful for predicting student outcomes, especially when student data is incomplete. We then applied imputation and non-imputation techniques and trained the models using two different neural network architectures. The use of two sizes of neural network was to observe whether performance improves with a larger neural network as is often observed (Wu et al., 2009).

**Figure 3.1:** Steps in the Cleaning and Pre-processing Procedures for the FoS Data




We investigated imputation and non-imputation techniques to identify the ability of both methods to handle missingness using the University of Alberta Faculty of Science (FoS) data. Educational institution datasets, such as FoS, may contain missing inputs for a number of reasons. Data could be missing because of attrition (a situation where students withdraw from a course or program of study causing a disruption in the time required for students to complete their programs), data not being captured properly, or students exercising their right to withhold information.

To determine machine learning methods that can predict the binary classes of time to completion for undergraduate programs, we implemented methods that attempted to get information about underlying patterns present within the available data. While we employed imputation techniques which first completed the data for the prediction task, we also explored a non-imputation technique that concatenates missing data with an indicator matrix of ones and zeros. To prepare the dataset for learning, we used a multivariate statistical analysis technique called Structural Equation Modeling (SEM) to understand structural relationships amongst variables within models. This procedure involved multiple regression analysis and examining the relationships between observed and unobserved (or latent) variables.

With structural equation modeling we investigated the data to understand how informative each feature was to the target variable and to other independent variables within the model for which it was defined. It also provided an opportunity to understand the degree to which each feature influenced variation in the target variable and in other variables within the model. We expressed this direct and indirect influence in models called mediation models.

Following structural equation modeling, we obtained a final dataset. The data contained within this dataset were identified through SEM. Data belonged to clearly defined categories:

Demography, High school details, English language skills, Program details, and Time to Completion. To determine the performance of imputation and non-imputation techniques for handling missingness, we trained the data for a binary-classification task. This task seeks to predict student time to completion for undergraduate programs.

## 3.1 Data Handling and Ethics

The Faculty of Science dataset was de-identified and contained in password protected excel files. We neither shared the data on any public platforms, nor gave access to unauthorized persons as stipulated by policies of the University of Alberta Research Ethics Board (REB) and those of the tri-council, which comprise the Canadian Institutes of Health Research (CIHR), the Natural Sciences and Engineering Research Council (NSERC), and the Social Sciences and Humanities Research Council (SSHRC). We only used authorized computing resources for computational processes and analysis. Authorized computing resources used include remote servers hosted by the Digital Research Alliance of Canada (formerly Compute Canada) of which the University of Alberta is a member.

## 3.2 Data

The FoS dataset is from the University of Alberta data warehouse and contains deidentified student data from 2011 through 2021. For the purpose of our study, we filtered and used data between the Fall 2015 to Fall 2021 terms. The data is largely categorical and was extracted into four main files which are described in the *Input section* below. We sampled features in the FoS datasets from the following categories: Demography, Program Details, English Language Skills, High school details, and Time to Completion for all students enrolled from 2015.

Following data cleaning, data comprised 104,231 total observations for this time period (2015 to 2021). There were 18,908 unique students who had an average age of *(M = 19.97, SD = 2.39).*

### 3.2.1 Input

*Registrations/Convocations*

This contains information about the terms in which students were registered, including data about their demographic and educational background and their current program. A student who,

for instance, attended for eight terms will have records for all eight terms and a column with the number of years it took for that student to complete their program.

*UofA Courses/Enrollment*

This contains information about the courses taken by students at the University of Alberta, such as the name and number of courses, grades, and credits earned.

*High School and/or Post Secondary Education (PSE) Transfer Courses*

This contains information on the courses students provided, including subject, level, grade, and grading scheme for the purpose of gaining admission or transfer credit toward their program.

*English Language and Other Competency Tests*

Information contained here describes different competency tests that students have taken and that were used in the admissions process.

## 3.3 Data Definition

Machine learning models require data in specific formats to learn and make predictions. It was essential that we preprocessed and transformed the largely categorical datasets into numerical formats that are suitable for machine learning classification tasks. Before preprocessing, we selected features and defined categories that will specify various models for the multivariate statistical modeling we perform. We defined the following categories: Student Identity, Demography, English Language Skills, Program Details, and Time to Completion.

For each category, we identified features that contained two properties: a) fit the category definition and b) contained possible values with low cardinality (less than or equal to 10) or values that can be aggregated to achieve low cardinality. A low cardinality column ensures that the data dimension can be managed appropriately during training by machine learning algorithms (Negi et al., 2020). With low cardinality, we create combination instances that are manageable for the algorithm to efficiently learn from. This prevents overfitting and allows for generalization to new or unseen data. We also reduce the amount of storage space required for computation and, in this way, reduce computational costs.

To assess and explain how various features influence the success of students in our dataset, we conducted Structural Equation Modeling (SEM) and tested the degree of variability for each category in the model we specified. The defined categories are as follows:

- Student Identity: a de-identified “Student ID” code that was used for identification and mapping purposes only.
- Demography: features that describe the characteristics of students, e.g., “Age”.
- Program Details: features that describe students’ program of study, e.g., “Course ID”.
- English Language Skills: features that describe students’ English language fluency, e.g., “IELTS Scores”.
- High school details: features that describe students’ high school performance, e.g., “High School Grades”.
- Time to Completion: features that describe the length of a student’s program, e.g., “Completion Term”.

## 3.4 Data Cleaning

In choosing features that were meaningful to the prediction goal of our study, we kept features within the contained cardinality of 10 to prevent the curse of dimensionality (Verleysen & François, 2005) and achieve efficient computation. This applied to all data as appropriate.

**Step 1:** Filter data in the High school details dataset.

The *high school details* dataset contained records of student performance in courses taken during high school. These scores were submitted as part of requirements by the University of Alberta for admission determination purposes. We sampled columns to derive 8 features: “Term year”, “Student_Id”, “High School Level”, “Academic Level”, “Grading Scheme”, “Course Grade”, and “Credits” and “Units Taken”. We replaced “Course Grade” column values with the equivalent 4.0 grade point values of the corresponding letter grades. The following are the

equivalent grade point values for the University of Alberta: A+ = 4.0, A = 4.0, A- = 3.7, B+ = 3.3, B = 3.0, B- = 2.7, C+ = 2.3, C = 2.0, C- = 1.7, D+ = 1.3, D = 1, and F = 0.

Following filtration and feature selection, total observations became 183,585 with 18,908 unique student identifiers. Each student took an average of 4.23 course units *(SD = 1.32, Min = 3, Max = 100)* where a unit represents approximately three hours of course work per week. Table 3.1 shows a sample of this filtered data.

**Table 3.1:** Sample Filtered High School Performance Dataset

| Term_Year | Student_Id | High_School_Level | Academic_Level | Grading_Scheme | Course_Grade | Credits | Units_Taken |
|---|---|---|---|---|---|---|---|
| 2015 | x1 | 90 | UNGRD | 7A | A+ | 4.0 | 3.0 |
| 2015 | x2 | 90 | UNGRD | 7A | A+ | 4.0 | 3.0 |
| 2015 | x3 | 90 | UNGRD | 7A | A+ | 4.0 | 3.0 |
| 2015 | x4 | 90 | UNGRD | 7A | A- | 3.7 | 3.0 |
| 2015 | x5 | 90 | UNGRD | 7A | A- | 3.7 | 3.0 |

The following were the grading schemes contained in the *high school details* dataset: '7A', 3, 'AP', 'IB', 'UBC', 1, 'CHA', 'CGP', 7, 'UG4', '7C', 'A','7D', 4. The corresponding grades for some of the grading schemes, such as 3, ‘AP’, ‘UBC’, 1, ‘CHA’, ‘CGP’, ‘UG4’ and 4, were in numerical forms either in percentages or in other standard point grade format (75, 56, 7) but others were not. Thus, we found conversion tables that would provide numerical equivalents to convert the letter grades to numerical formats. We also filtered out records with grades of “IP” (In Progress), “TR” (Transfer Credit) and “CR” (Credit) for high school details because there were no numerical equivalents for these grades. The conversion tables used are in the Appendix A section.

For Grading Schemes = ‘7A’ or ‘7’, values contained in the dataset showed both percentage grades and letter grades. We referenced the data dictionary and looked up the “External School” column to find the academic institution using this scheme and mapped the grades according to the institution’s categorization. We found that NorQuest College uses this scheme. The table of

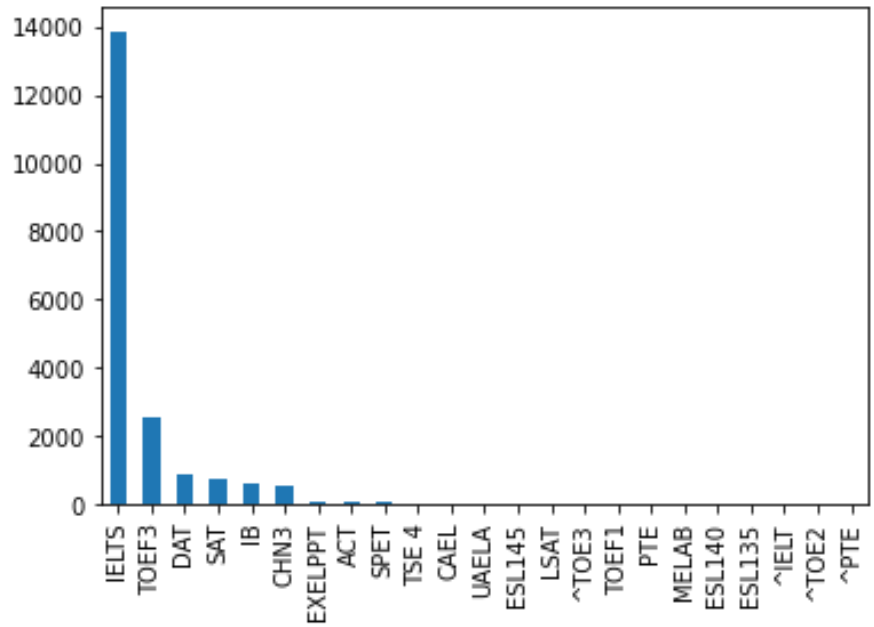
conversion is provided in the Appendix A *"grading scheme conversion"* section. We converted both percentage and letter grades to grade point values on a 4.0 scale based on the NorQuest College conversion guide.

**Step 2:** Filter data in the English Language Skills dataset.

In the *English language skills* dataset, there were two kinds of test results. The first set were for English language skills tests such as the International English Language Testing System (IELTS), Test of English as a Foreign Language (TOEFL), the Canadian English Language Test (CAEL), Pearson Test of English (PTE) and the University of Alberta English Language Assessment (UAELA). The use of specific English language skills requirements such as English as a Second Language 140 and 145 (ESL140, ESL145) were discontinued. The second set were professional and other standardized tests such as the Dentistry Admissions Test (DAT), the ACT College Entrance Exams (ACT), Graduate Record Exam (GRE), the Chinese University/college entrance exams (CHN3), and The Law School Admissions Test (LSAT). Figure 3 shows a bar plot for all tests and the corresponding number of students who sat for them.

We included test results that were for ELP only. The conversion table used to make standardizations across different test scores "IELTS", "TOEF3", "CAEL", "PTE", and "IB" can be found in the Appendix Section under *"Grading Scheme Conversion"*. The "IB" test identifier refers to the international Baccalaureate English diploma taken by students prior to admission to the University of Alberta. Conversion values can be found in the Appendix A *"grade conversion"* section. Figure 3.2 shows a bar plot of all tests.

**Figure 3.2:** A Bar Plot of All Tests




We filtered out all records that do not qualify as ELP requirements at the University of Alberta and conducted feature selection to derive the following 5 features: “Student_Id”, “TEST_ID”, “TEST_COMPONENT”, “SCORE”, and “LOADED_YEAR”. These variables align with earlier defined categories for high school details. We condensed similar records under the “TEST_COMPONENT” column so as to eliminate repetition. We achieved this by retaining “TOTAL” and dropping the variables “LISTENING”, “SPEAKING”, “WRITING” and “READING”. This ensured that we had an overall proficiency score across the different bands (Listening, Reading, Speaking and Writing) and tests that we could standardize. We also used the pandas Replace() function to replace “IBEXC” with “IBEX”, “TOTSC” with “TOTAL”, “^TOE3” with “TOEF3” as these were record entry mistakes and not different test components. Table 3.2 contains a sample of this data for one student who took the TOEFL test in 2015. Table 3.3 shows descriptive statistics for all selected test components.

**Table 3.2:** Sample Filtered English Language Skills Dataset

| Student_Id | Test_Id | Test_Component | Score | Year Assessed |
|---|---|---|---|---|
| x1 | TOEF3 | TOTAL | 120 | 2015 |
| x1 | TOEF3 | T LIS | 30 | 2015 |
| x1 | TOEF3 | T RD | 30 | 2015 |
| x1 | TOEF3 | T SPK | 30 | 2015 |
| x1 | TOEF3 | T WRT | 30 | 2015 |

**Table 3.3:** Statistics of Test Scores, for all English Language Skills Tests

| | IELTS | TOEF3 | IB | UAELA | CAEL | PTE | MELAB |
|---|---|---|---|---|---|---|---|
| **count** | 2041 | 339 | 338 | 5.00 | 4.00 | 3.00 | 3.00 |
| **mean** | 6.44 | 95.14 | 33.39 | 73.80 | 72.50 | 79.33 | 88.00 |
| **std** | 0.79 | 11.91 | 4.82 | 13.33 | 12.58 | 8.96 | 2.00 |
| **min** | 0.00 | 58.00 | 4.0 | 53.00 | 60.00 | 69.00 | 86.00 |
| **25%** | 6.00 | 88.00 | 31.00 | 73.00 | 67.50 | 76.50 | 87.00 |
| **50%** | 6.50 | 97.00 | 33.00 | 76.00 | 70.00 | 84.00 | 88.00 |
| **75%** | 7.00 | 105.00 | 37.00 | 77.00 | 75.00 | 84.50 | 89.00 |
| **max** | 8.50 | 120.00 | 45.00 | 90.00 | 90.00 | 85.00 | 90.00 |
| **Maximum possible value** | 9.00 | 120.00 | 45.00 | 90.00 | 90.00 | 90.00 | 90.00 |

**Step 3:** Get courses students enrolled in

Using data from the *Enrolment table* we created a filter with the column “Term Code” to ensure we were dealing with student course enrolment data for the correct time frame (from 2015). The

Term Code column contained encoded numerical values that identified each term. The filter used is - [(Enrolment data['Term Code'] > 1530)] where '> 1530' refers to the "Term Code" for semesters starting from Fall 2015. There were columns in this dataset with similar meanings to those in the *Registration table.* Thus, we dropped redundant features and sampled according to predefined categories. We sampled 5 out of 44 columns: 'Student_Id', 'Course Component', 'Course Hours', 'Grade', and 'Credits Earned'.

After filtration and feature selection, the enrolment data comprised 930,913 observations and 5 columns, with 4,248 unique student identities. Table 3.4 shows a sample of the dataset.

**Table 3.4:** Sample Filtered Course Enrolment Dataset

| Student_Id | Course_Component | Course_Id | Course_Hours | Credits_Earned | Grade |
|---|---|---|---|---|---|
| x1 | Lecture | 4032 | 3 | 3 | A- |
| x1 | Lecture | 603 | 3 | 3 | W |
| x1 | Lecture | 2134 | 4.5 | 3 | C+ |
| x1 | Laboratory | 2134 | NaN | 0 | NaN |
| x1 | Lecture | 231 | 4 | 3 | C+ |

We adhered to the definitions we created for our categories and ensured we selected features that were useful for structural equation modeling. The 5 sampled features from the enrolment table provided information about the course components students enrolled in, contact hours for these courses, credits taken for each of the courses, and grades attained.

Descriptive statistics for this filtered dataset comprise a total of 18,908 unique student identities. The feature for "course component" contains nominal data used to classify different types of courses. Table 3.5 shows these different course components and their corresponding total. The "Course Hours" column is a ratio feature which describes the total amount of time a course schedule is allotted. For example, most lectures are allotted 3 hours per schedule while some seminars are allotted 1.5 hours per schedule. We derived statistics for course hours *(M = 3.51,*

*SD = 4.19, Min = 0, Max = 4.5).* The feature for "total credits earned" is also a ratio feature. We derived statistics for this feature, per student *(M = 2.01, SD = 1.53, Min = 0, Max =* 8*).*

**Table 3.5:** Course Components and their Count

| Course_Component | Total |
|---|---|
| Lecture | 91,222 |
| Laboratory | 39,133 |
| Seminar | 9,154 |
| Lab-lecture | 390 |
| Lecture-lab | 103 |
| Credit by special assignment | 2 |

In Table 3.6, we provide an overview of "Grades", their definitions and numerical equivalence according to the University of Alberta.

**Table 3.6:** An Overview of the University of Alberta's Undergraduate Grading System

| Grade | Definition | Numerical Equivalence |
|---|---|---|
| A+ | Excellent | 4.0 |
| A | | 4.0 |
| A- | | 3.7 |
| B+ | Good | 3.3 |
| B | | 3.0 |
| B- | | 2.7 |
| C+ | Satisfactory | 2.3 |
| C | | 2.0 |
| C- | | 1.7 |

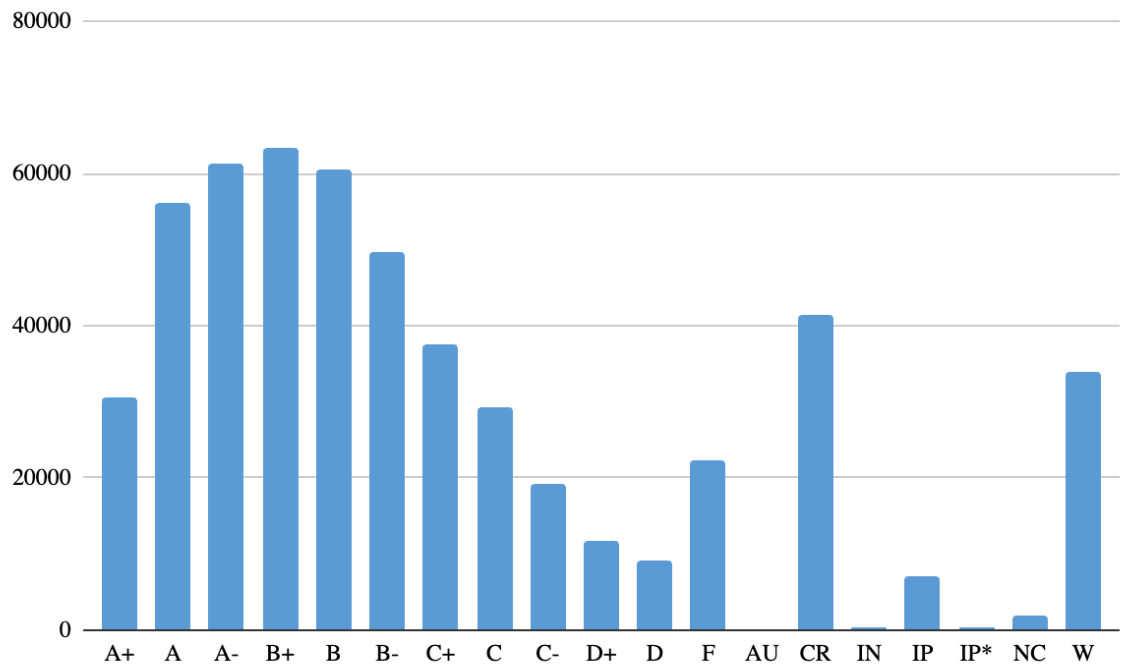
| | | |
|---|---|---|
| D+ | Poor | 1.3 |
| D | Minimal Pass | 1.0 |
| F | Failure | 0.0 |
| **Non-Numerical Grades** | | |
| AE | aegrotat standing | NA |
| AU | registered as an auditor | NA |
| AW | registered as an auditor and withdrew | NA |
| CR | completed requirements, no grade point value assigned | NA |
| EX | exempt | NA |
| IN | incomplete | NA |
| IP | course in progress (assigned to Part A of a Two-Term or One-Term A/B Course with the final grade assigned to Part B) | NA |
| IP* | withdrew from or failed course in progress (assigned to Part A of a Two-Term or One-Term A/B Course where the final grade assigned to Part B is a withdrawal or failure) | NA |
| NC | failure, no grade point value assigned | NA |
| W | withdrew with permission | NA |

We dropped observations that had no numerical equivalence such as “Grade” values; “IP”, “IP*”, “AU”, “W”, “AE”, “AW”, “CR”, “NC”, “IN”, “EX”. Courses with “IP” and “IP*” grades are for “In Progress” students who are still in between courses and are yet to graduate. We dropped

these observations to maintain consistency and ensure that student records were complete. It is possible for a student in high school taking an AP course to decide to enroll in another university, making their University of Alberta records, incomplete. For example, in Table 8, there are 5 rows for various enrolled courses for Student_Id "x1". There is 1 row that shows a withdrawn course, 1 row with Grade recorded as NaN and Credits_Taken recorded as "0.0" to indicate that this course component did not accrue credits. We dropped these two observations because there were no equivalent numerical values to define "Withdrawn" and Course components with "No Credits". It is important to note here that following withdrawal from a course, students typically take other courses to make-up the credits required to complete their programs. Additionally, this provides an opportunity for future work that analyzes the ratio of withdrawn courses alone or in combination with missing data, to understand its effect on program completion.

The grades column was converted to numeric values based on the University of Alberta's grade point system for undergraduates (*Grading System Explained*, n.d.). Figure 3.3 shows a bar plot of all grades.

**Figure 3.3:** A Bar Plot of the Grades Column



**Step 4:** Filter for students who were admitted in or after 2015.

From the *Registration table*, we focused only on students who were admitted to the university in or after 2015. The total number of observations from the registration table was 104,231 with 65 features. For filtration, we used the filters - [(Registration data reporting year >= 2015) & (New to the university flag == 'Yes')].

For dimensionality reduction, we conducted sampling to find features that fit categories we previously defined for SEM. We selected 8 columns using the data dictionary provided and the category definitions we created earlier: 'Student ID', ‘Admission Year’, 'Academic Load', 'Age’, ‘Gender’, ‘Legal Status’, ‘Credits Earned’, and ‘Completion Year’.

In our feature selection for these columns, we looked at the data dictionary, the corresponding values in each of the columns and eliminated columns that have similar meanings. Following filtration and feature selection, the registration table comprised 104,231 rows and 8 columns. A sample is shown in Table 3.7.

**Table 3.7:** Sample Filtered Registration Dataset

| Student_Id | Academic Load | Gender | Age | Legal Status | Credits Earned | Admission Year | Completion Year |
|---|---|---|---|---|---|---|---|
| x1 | Part-Time | Female | 21 | Canadian Citizen | 3 | 2019 | 2020 |
| x2 | Full-Time | Female | 22 | Canadian Citizen | 15 | 2020 | 2019 |
| x3 | Full-Time | Female | 21 | Canadian Citizen | 12 | 2019 | 2021 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| x4 | Full-Time | Female | 21 | Canadian Citizen | 12 | 2019 | 2019 |
| x5 | Full-Time | Female | 22 | Canadian Citizen | 15 | 2020 | 2018 |

This subset is comprised of 18,908 unique students and 104,231 observations. Table 3.8 shows the number of students admitted in each year. Students were 19.97 years of age on average *(SD = 2.39)* and they earned  an average of 11.35 credits each *(SD = 4.2, Min = 0 , Max =* 22.5*).*

**Table 3.8:** Total Count of Students and their Year of Admission

| Year | Total Number of Students Admitted |
|---|---|
| 2015 | 6,080 |
| 2016 | 6,217 |
| 2017 | 6,419 |
| 2018 | 6,350 |
| 2019 | 6,500 |
| 2020 | 7,373 |
| 2021 | 6,845 |

**Step 5:** Get program completion data for students admitted in 2015 or later.

We looked at the *Convocation table* to find student data of those who had convocated and were among the students identified in the filtered *Registration table*. The total number of observations from the convocation sheet was 8,617 with 48 features. We filtered records to find students who were admitted into the university from 2015 and had completed their programs by 2021. We did this by using the filter - [(Registration data reporting year >= 2015 && Convocation Year == 2021)]. At the time of data export, 4,651 students had graduated. We analyzed the data for students who had graduated and sampled features according to the earlier defined categories. We then dropped repeated columns that were also present in the registration data to achieve a total of 4 features: 'Student ID', ‘Completion Year’, ‘Admission Year’ and ‘Student Degree Number’. Table 3.9 shows a sample.

**Table 3.9:** Sample Filtered Convocation Dataset

| Student_Id | Completion_Year | Admission_Year | Student_Degree_Number |
|---|---|---|---|
| x1 | 2018 | 2015 | 1 |
| x2 | 2016 | 2015 | 2 |
| x3 | 2021 | 2017 | 1 |
| x4 | 2018 | 2015 | 3 |
| x5 | 2017 | 2016 | 4 |

Descriptive statistics for this filtered convocation dataset comprise 4,263 unique students who make up the total observation of 4,651 for the "Student_Id" column which is categorical. This suggests that some students enrolled in and completed more than one degree program within the period under consideration (2015 to 2021). The feature "Completion Year" is ordinal and constitutes the years 2016 to 2021. The "Student_Degree_Identifier" feature denotes an additional degree taken by students in conjunction with their undergraduate program degrees. This feature is of a nominal nature with values ranging from "1 to 4". Students with degree identifiers > 1 obtained multiple degrees at the time of their convocation. Table 3.10 shows a breakdown of each student degree identifier category and corresponding total number of students in them. Table 3.11 shows the years of completion and the total number of students who completed their programs in those years.

**Table 3.10:** Student Degree Identifiers and Total Number of Students in Each Category

| Student_Degree_Number | Total Number of Students |
|---|---|
| 1 | 4,264 |
| 2 | 343 |
| 3 | 39 |
| 4 | 5 |

**Table 3.11:** Completion Years and Total Number of Students in Each Year

| Completion_Year | Total Number of Students |
|---|---|
| 2021 | 1257 |
| 2020 | 1220 |
| 2019 | 992 |
| 2018 | 686 |
| 2017 | 347 |
| 2016 | 149 |

After merging the registration and convocation datasets, there were  38,407 total observations across 4,248 unique student identities. This indicates that out of 18,908 students admitted in 2015, as at 2021, 4,248 had graduated. The remaining 14,660 then consisted of a combination of students who withdrew from their programs and those whose programs extended beyond 2021.

The above cleaning steps prepared the data for analyses. We conducted two types of analyses: Structural Equation Modeling (SEM) and prediction. Each of these analyses required its own additional preprocessing. Further preprocessing and specific analysis methods will be covered in Chapter 4 (SEM) and Chapter 5 (Prediction with Missingness).

# Chapter 4

# Structural Equation Modeling (SEM)

## 4.1 Methods

In this chapter, we discuss feature investigation and its implementation. We conducted feature investigation using the multivariate statistical analysis technique known as Structural Equation Modeling (SEM). From earlier defined categories, we formulated research questions and assessed how informative individual category variables were to the target variable. The intent behind this investigation was to support the selection of features for the predictive models.

### 4.1.1 Investigating Feature Importance

In implementing SEM, we conducted feature investigation to understand the data, especially because we will later use the identified features from this investigation as part of a prediction task. This feature investigation provided evidence of individual feature importance and how useful they could be to pattern recognition. We created hypotheses and determined if there were influences amongst variables through testing. Results from these tests also provided information about the  degree to which variables influenced each other. We specified the below research questions:

- To what extent do High school and Program details explain variation in Time to Completion?
- To what extent do High school and Demography explain variation in Time to Completion?
- To what extent do High school and English language skills explain variation in Time to Completion?
- To what extent do Demography and Program details explain variation in Time to Completion?

- To what extent do English language skills and Program details explain variation in Time to Completion?
- To what extent do Demography and English language skills explain variation in Time to Completion?

The below steps were taken to obtain appropriate features to include in the final dataset that was used for machine learning. We maintained consistency and filtered rows based on the categories we defined earlier. We also normalized values in the final dataset.

### 4.1.2 Additional Pre-processing

It was important to further prepare the cleaned data to make it ready for feature investigation using SEM. An additional pre-processing step we took was to normalize values across the "Score" column obtained from the *English Language Skills* table and the "Course Grade" column obtained from the *high school details* table. We began by normalizing values within each Test_Id range. Normalization was done by first converting letter grades to their numerical equivalents on a 100 scale. Following this conversion, we divided the corresponding results by 100 to obtain values between 0 and 1. The goal was to find a balance across the different grading schemes and find equivalency among the different ELP tests and high school details. Test_Id and conversion values can be found in the Appendix A *"grade conversion"* section.

For English language skills tests, our conversion methods were consistent with the English Testing System (ETS) standards. We used the ETS conversion table that can be found in the Appendix A section (*Compare TOEFL iBT Scores*, n.d.) to normalize the data within each test identifier range (the grading schemes). We did this to find a balance across the different grading schemes and to put the various ELP tests on the same scale.

We then encoded categorical variables in columns such as "Course Component", "Gender", "Legal Status", and "Completion Year" which were obtained from the *Enrollment* table. Encoding was done using the LabelEncoder() method from the sklearn.preprocessing library (Pedregosa et al., 2011).

Following normalization and encoding, we created models using category definitions and filtered the cleaned data to obtain subsets that are targeted at answering specific SEM research questions. The models we created focused on generating results from the research inquiries we outline below. We merged different filtered datasets together to derive subsets containing 3 category definitions to enable us to conduct Structural Equation Modeling and answer our research questions.

For presentation purposes, we have maintained consistency in the variables used. It is important to note that during implementation in some models, some variables were excluded and others included. These are noted when providing the SEM model definition.

#### 4.1.2.1 How Do High School and Program Details Explain Time to Completion?

Our first SEM Research Question (RQ) aimed to determine the extent to which High school performance and university Program details explained variation in Time to Completion. The categories and corresponding variables required to answer this question are shown below. Table 4.1 shows a sample of the filtered data while Table 4.2 shows a sample of the encoded and normalized data.

- High school details - Student level, Grading Scheme, Course Grade and Total Credits Earned.
- Program Details - Credits Taken, Course Id, Grade, Course Component, and Course Hours.
- Time to Completion - Admission Year and Completion Year.

**Table 4.1:** Sample Filtered Dataset to Analyze SEM RQ1

| Student_Id | Student_Level | Grading_Scheme | Course_Grade | Total_Credits_Earned | Credits_Taken | Course_Id | Grade | Course_Component | Course_Hours | Admission_Year | Completion_Year |
|---|---|---|---|---|---|---|---|---|---|---|---|
| x1 | UNGRD | 7A | 88 | 3.0 | 3.0 | 6817 | A+ | Lecture | 3.0 | 2015 | 2018 |
| x2 | UNGRD | 7A | 88 | 3.0 | 3.0 | 11487 | A+ | Lecture | 3.0 | 2015 | 2018 |
| x3 | UNGRD | 7A | 89 | 3.0 | 3.0 | 6798 | A+ | Lecture | 3.0 | 2015 | 2018 |
| x4 | UNGRD | 7A | 89 | 3.0 | 3.0 | 6801 | A- | Lecture | 3.0 | 2015 | 2018 |
| x5 | UNGRD | 7A | 88 | 3.0 | 3.0 | 9595 | A- | Lecture | 3.0 | 2015 | 2018 |

**Table 4.2:** Sample Encoded and Normalized Dataset to Analyze SEM RQ1

| Student_Id | Student_Level | Grading_Scheme | Course_Grade | Total_Credits_Earned | Credits_Taken | Course_Id | Grade | Course_Component | Course_Hours | Admission_Year | Completion_Year |
|---|---|---|---|---|---|---|---|---|---|---|---|
| x1 | 0 | 3 | 0.88 | 3.0 | 3.0 | 35 | 4.0 | 1 | 3.0 | 1 | 3 |
| x2 | 0 | 3 | 0.88 | 3.0 | 3.0 | 49 | 4.0 | 1 | 3.0 | 1 | 3 |
| x3 | 0 | 3 | 0.89 | 3.0 | 3.0 | 38 | 4.0 | 1 | 3.0 | 1 | 3 |
| x4 | 0 | 3 | 0.89 | 3.0 | 3.0 | 39 | 3.7 | 1 | 3.0 | 1 | 3 |
| x5 | 0 | 3 | 0.88 | 3.0 | 3.0 | 50 | 3.7 | 1 | 3.0 | 1 | 3 |

### 4.1.2.2 How Do High School Details and Student Demography Explain Time to Completion?

Our second question was to understand the extent to which High school details and student Demography explained variation in Time to Completion. Below are the categories and corresponding variables we considered. Table 4.3 shows a sample of the filtered data while Table 4.4 shows a sample of the encoded and normalized data.

- High school details - Student level, Grading Scheme, Course Grade, and Total Credits Earned.
- Demography - Age, Legal Status, and Gender.
- Time to Completion - Admission Year and Completion Year.

**Table 4.3:** Sample Filtered Dataset to Analyze SEM RQ2

| Student_Id | Student_Level | Grading_Scheme | Course_Grade | Total_Credits_Earned | Age | Legal_Status | Gender | Admission_Year | Completion_Year |
|---|---|---|---|---|---|---|---|---|---|
| x1 | UNGRD | 7A | 100 | 5.0 | 22 | Canadian Citizen | Male | 2015 | 2020 |
| x2 | UNGRD | 7A | 71 | 5.0 | 22 | Canadian Citizen | Male | 2016 | 2021 |
| x3 | UNGRD | 7A | 71 | 5.0 | 23 | Canadian Citizen | Female | 2015 | 2019 |
| x4 | UNGRD | 7A | 90 | 5.0 | 21 | Canadian | Male | 2015 | 2020 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Citizen | | | |
| x5 | UNGRD | 7A | 77 | 5.0 | 26 | International Student | Female | 2017 | 2020 |

**Table 4.4:** Sample Encoded and Normalized Dataset to Analyze SEM RQ2

| Student_Id | Student_Level | Grading_Scheme | Course_Grade | Total_Credits_Earned | Age | Legal_Status | Gender | AdmissionYear | Completion_Year |
|---|---|---|---|---|---|---|---|---|---|
| x1 | 0 | 3 | 1 | 5.0 | 22 | 0 | 0 | 1 | 5 |
| x2 | 0 | 3 | 0.71 | 5.0 | 22 | 0 | 0 | 2 | 6 |
| x3 | 0 | 3 | 0.71 | 5.0 | 23 | 0 | 1 | 1 | 4 |
| x4 | 0 | 3 | 0.9 | 5.0 | 21 | 0 | 0 | 1 | 5 |
| x5 | 0 | 3 | 0.77 | 5.0 | 26 | 1 | 1 | 3 | 5 |

### 4.1.2.3 How Do High School Details and English Language Skills Explain Time to Completion?

With the third SEM research question, we aimed to understand the extent to which High school details and English Language Skills explained variation in Time to Completion. We considered the below categories and corresponding variables. Table 4.5 shows a sample of the filtered data while Table 4.6 shows a sample of the encoded and normalized data.

- High school details - Student level, Grading Scheme, Course Grade, and Total Credits Earned.

- English Language -Test Id and Score.
- Time to Completion - Admission Year and Completion Year.

**Table 4.5:** Sample Filtered Dataset to Analyze SEM RQ3

| Student_Id | Student_Level | Grading_Scheme | Course_Grade | Total_Credits_Earned | Test_Id | Score | Admission_Year | Completion_Year |
|---|---|---|---|---|---|---|---|---|
| x1 | UNGRD | 7A | 100 | 5.0 | IB | 4 | 2015 | 2020 |
| x2 | UNGRD | 7A | 71 | 5.0 | IB | 7 | 2016 | 2021 |
| x3 | UNGRD | 7A | 71 | 5.0 | IB | 6 | 2015 | 2019 |
| x4 | UNGRD | 7A | 90 | 5.0 | IB | 7 | 2015 | 2020 |
| x5 | UNGRD | 7A | 77 | 5.0 | IB | 4 | 2017 | 2020 |

**Table 4.6:** Sample Encoded and Normalized Dataset to Analyze SEM RQ3

| Student_Id | Student_Level | Grading_Scheme | Course_Grade | Total_Credits_Earned | Test_Id | Score | Admission_Year | Completion_Year |
|---|---|---|---|---|---|---|---|---|
| x1 | 0 | 3 | 1 | 5.0 | 2 | 0.65 | 1 | 5 |
| x2 | 0 | 3 | 0.71 | 5.0 | 2 | 0.9 | 2 | 6 |
| x3 | 0 | 3 | 0.71 | 5.0 | 2 | 0.85 | 1 | 4 |
| x4 | 0 | 3 | 0.9 | 5.0 | 2 | 0.9 | 1 | 5 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| x5 | 0 | 3 | 0.77 | 5.0 | 2 | 0.65 | 3 | 5 |

### 4.1.2.4 How Do Demography and Program Details Explain Time to Completion?

In the fourth SEM research question, the goal was to understand the extent to which Demography and Program details explained variation in Time to Completion. Table 4.7 shows a sample of the filtered data while Table 4.8 shows a sample of the encoded and normalized data. Below are the categories and corresponding variables we considered.

- Demography - Age, Legal Status, and Gender.
- Program Details - Credits Taken, Course Id, Grade, Course Component, and Course Hours.
- Time to Completion - Admission Year and Completion Year.

**Table 4.7:** Sample Filtered Dataset to Analyze SEM RQ4

| Student_Id | Age | Legal_Status | Gender | Credits_Taken | Course_Id | Grade | Course_Component | Course_Hours | Admission_Year | Completion_Year |
|---|---|---|---|---|---|---|---|---|---|---|
| x1 | 27 | Canadian Citizen | Male | 12 | 2341 | B- | Lecture | 3.0 | 2015 | 2018 |
| x2 | 27 | Canadian Citizen | Male | 12 | 232 | B- | Lecture | 3.0 | 2015 | 2018 |
| x3 | 28 | Canadian | Male | 12 | 4771 | B- | Lecture | 3.0 | 2015 | 2018 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Citizen | | | | | | | | |
| x4 | 27 | Canadian Citizen | Male | 12 | 321 | B- | Lecture | 3.0 | 2015 | 2018 |
| x5 | 27 | International Student | Male | 12 | 9281 | B- | Lecture | 3.0 | 2015 | 2018 |

**Table 4.8:** Sample Encoded and Normalized Dataset to Analyze SEM RQ4

| Student_Id | Age | Legal_Status | Gender | Credits_Taken | Course_Id | Grade | Course_Component | Course_Hours | Admission_Year | Completion_Year |
|---|---|---|---|---|---|---|---|---|---|---|
| x1 | 27 | 0 | 0 | 12 | 21 | 2.7 | 2 | 3.0 | 1 | 3 |
| x2 | 27 | 0 | 0 | 12 | 13 | 2.7 | 2 | 3.0 | 1 | 3 |
| x3 | 28 | 0 | 0 | 12 | 28 | 2.7 | 2 | 3.0 | 1 | 3 |
| x4 | 27 | 0 | 0 | 12 | 18 | 2.7 | 2 | 3.0 | 1 | 3 |
| x5 | 27 | 1 | 0 | 12 | 52 | 2.7 | 2 | 3.0 | 1 | 3 |

#### 4.1.2.5 How Do Program Details and English Language Skills Explain Time to Completion?

The fifth SEM research question sought to understand the extent to which Program Details and English Language Skills explained variation in Time to Completion. Below are the categories and corresponding variables we considered. Table 4.9 shows a sample of the filtered data while Table 4.10 shows a sample of the encoded and normalized data.

- Program Details - Credits Taken, Course Id, Grade, Course Component, and Course Hours.
- English Language - Test Id and Score.
- Time to Completion - Admission Year and Completion Year.

**Table 4.9:** Sample Filtered Dataset to Analyze SEM RQ5

| Student_Id | Credits_Taken | Grade | Course_Id | Course_Components | Course_Hours | Test_Id | Score | Admission_Year | Completion_Year |
|---|---|---|---|---|---|---|---|---|---|
| x1 | 9.0 | NaN | 93978 | Seminar | 3.0 | IELTS | 8.5 | 2015 | 2020 |
| x1 | 9.0 | C+ | 8777 | Lecture | 3.0 | IELTS | 8.5 | 2015 | 2020 |
| x1 | 0.0 | NaN | 1896 | Laboratory | NaN | IELTS | 8.5 | 2015 | 2020 |
| x1 | 0.0 | NaN | 1896 | Seminar | NaN | IELTS | 8.5 | 2015 | 2020 |
| x1 | 15.0 | B+ | 3010 | Lecture | 5.0 | IELTS | 8.5 | 2015 | 2020 |

**Table 4.10**: Sample Encoded and Normalized Dataset to Analyze SEM RQ5

| Student_Id | Credits_Taken | Grade | Course_Id | Course_Components | Course_Hours | Test_Id | Score | Admission_Year | Completion_Year |
|---|---|---|---|---|---|---|---|---|---|
| x1 | 9.0 | 0.0 | 12 | 1 | 3.0 | 3 | 0.85 | 1 | 5 |
| x1 | 9.0 | 2.3 | 3 | 2 | 3.0 | 3 | 0.85 | 1 | 5 |
| x1 | 0.0 | 0.0 | 21 | 0 | 0.0 | 3 | 0.85 | 1 | 5 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| x1 | 0.0 | 0.0 | 21 | 1 | 0.0 | 3 | 0.85 | 1 | 5 |
| x1 | 15.0 | 3.3 | 31 | 2 | 5.0 | 3 | 0.85 | 1 | 5 |

### 4.1.2.6 How Do Demography and English Language Skills Explain Time to Completion?

The sixth SEM research question tried to understand the extent to which Demography and English Language Skills explained variation in Time to Completion. Table 4.11 shows a sample of the filtered data while Table 4.12 shows a sample of the encoded and normalized data. We considered the below categories and corresponding variables.

- Demography - Age, Legal Status, and Gender.
- English Language - Test Id and Score.
- Time to Completion - Admission Year and Completion Year.

**Table 4.11:** Sample Filtered Dataset to Analyze SEM RQ6

| Student_ID | Age | Legal_Status | Gender | Test_ID | Score | Admission_Year | Completion_Year |
|---|---|---|---|---|---|---|---|
| x1 | 27 | International Student | Male | IELTS | 9 | 2015 | 2018 |
| x2 | 27 | International Student | Male | IELTS | 7.5 | 2015 | 2018 |
| x3 | 28 | International Student | Female | IELTS | 8.6 | 2015 | 2018 |
| x4 | 27 | International | Female | IB | 6 | 2015 | 2018 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | Student | | | | | |
| x5 | 27 | International Student | Male | IB | 7 | 2015 | 2018 |

**Table 4.12**: Sample Encoded and Normalized Dataset to Analyze SEM RQ6

| Student_Id | Age | Legal_Status | Gender | Test_Id | Score | Admission_Year | Completion_Year |
|---|---|---|---|---|---|---|---|
| x1 | 27 | 1 | 0 | 0 | 0.9 | 1 | 3 |
| x2 | 27 | 1 | 0 | 3 | 0.75 | 1 | 3 |
| x3 | 28 | 1 | 1 | 3 | 0.86 | 1 | 3 |
| x4 | 27 | 1 | 1 | 2 | 0.6 | 1 | 3 |
| x5 | 27 | 1 | 0 | 2 | 0.7 | 1 | 3 |

## 4.1.3 Implementing SEM

We conducted Structural Equation Modeling to enable the identification of structural relationships amongst variables in the SEM research questions we defined. An important part of our work was to establish grounds for conducting rigorous data mining to understand features in a dataset with missingness. It is our position that identifying features that enable pattern recognition and function mapping will empower research to approach the handling of missing data differently from the way it is currently being addressed in the field.

SEM models typically consist of two major components: structural models and confirmatory factor analysis (CFA). CFA is used to assess the fitness of measures of a construct and pre-defined theories about the structure of the construct. With CFA, a theory is described and latent constructs are used to test whether the understanding of this theory is consistent with the constructs identified. Models that align with these theories are introduced and tested to determine if the data fits the models. CFA focuses on identifying the number of factors within a model, the relationship between these factors and a set of observed variables, and the relationships among errors from the observed variables. In our work, we placed less emphasis on the CFA models because it was important to first identify the importance and strengths of the relationships amongst features. This objective provided a good foundation to select features that can enable prediction tasks when dealing with missing data (Salmanpour et al., 2021).

We began by importing the necessary packages: lavaan, SemPlot, MPsychoR, and corrplot in the statistical software R to enable model identification, estimation, and visualization. We also imported the derived datasets for all SEM research questions consisting of the categories Demography, Program Details, high school details, English Language Skills, and Time to Completion. We first explored variables that described pre-university characteristics and performance of students before examining variables that described their post-admission data.

The lavaan package enables the definition of the structural relationships existing within a model. When defining models, the single tilde (~) is used in the regression model definition to regress the target variable on the predictor variables. The double tilde (~~) is used to capture covariances and variances. Parameter labels (b1 and b2) are used to capture regression coefficients of the predictor variables. The parameter label (b3) is used to capture variance of one predictor variable with the other. Lavaan also provides the ability to capture the indirect effect of both predictor variables on the target variable. In the model definition, it can be represented with the new parameter "ind", which is the product of both regression coefficients b1 and b3.

During implementation, Maximum Likelihood (ML) estimation was used because data analysed with SEM are often not normally distributed. This was the case with our data which was also not normally distributed. With SEM, we are presented with a data distribution containing multiple variables. This scenario provides a basis to obtain parameter estimates and assess the models by maximising a likelihood function (Yuan & Bentler, 2007). In the context of SEM, degrees of freedom (DF) are calculated differently from typical statistical tests. DF are calculated by subtracting the number of parameters required for estimation from the number of variances and covariances in the model. DF relies on the number of variables available within the model and the number of relationships that exist between variables. There has to be a DF of at least 0 for the SEM to be identifiable (Bowen & Guo, 2011).

SEM analyzes the structural relationships between observed and unobserved (also known as latent) variables. To do this, it combines multiple regression and factor analyses to produce path analysis diagrams. Factor analysis path diagrams illustrate the presence of latent variables and test the strengths of the relationships between variables. Confirmatory Factor Analysis (CFA) is

then used to identify the presence of latent variables. Latent constructs are theoretical concepts that emanate from the field of psychology. They describe human traits that cannot be directly measured such as intelligence, self-esteem, and satisfaction. Traits such as intelligence, anxiety and creativity are examples of concepts of interest in the field of education. These traits are a combination of many social influences (El-Den et al., 2020) and as a result, are represented by latent variables which cannot be directly observed but can be estimated from observed variables (Warren, 1991). The latent variable “academic endeavor” represents the course load and program of students; it is a construct of interest in our study that can be estimated from the observed variables “Course_Id”, “Credits Taken”, and “Course Hours” in the “Program details” category. These two observed variables have measures that tell us how many credits were taken by students for each of their courses and the number of hours for each of the course components - lectures, laboratory, seminars, etc. Through these two observed variables we can infer that a student with high measures of course hours and credits taken, shows high academic endeavor.

### 4.1.4 Model Performance and Interpretation

Results from our SEM implementation provided useful path diagrams and identified latent variables. The regression model path diagrams showed directed measures of influence because our hypothesis began with variables that preceded the other. For example, students’ High school performance details preceded their University Program details. Thus, the ordering of our variables had an influence on the direction as it related to the target variable, Time to completion..

Path diagrams are one output of SEM. These diagrams consist of red and grey lines. Red lines indicate negative correlations while grey lines indicate positive correlations. Square (or rectangle) boxes are used to denote observed variables while circles are used to denote the presence of latent constructs. A single-headed directional arrow is used to denote a regression coefficient which regresses the dependent variable on the independent variable. Values along this path are the direct effect of the independent variable on the dependent variable. Path diagrams in SEM often contain residual variance which is characterized as the disturbance or the part of the dependent variable that is not explained by the predictor variable. Variances are denoted with one

single-headed arrow with a value beside the square box. A double-headed arrow denotes bi-directional correlation between variables. The path diagrams are also characterised by dashed lines that depict the significance and strength of their factor loadings. A continuous line indicates factor loadings with a significant relationship (at the .05 level) while a dashed line indicates a nonsignificant relationship. Even though a continuous line indicates significance, the strength of this loading can either be strong or weak. Factor loadings with values > .06 are estimated to be strong while those with values < .7 are estimated to be weak (Loehlin, 2003). Results from SEM's standardized path analysis can result in negative coefficients greater than 1. We interpret this correlation as we would when presented with any regression model where values are between -1 and 1 (Jöreskog, 1994).

To determine how acceptable the models are, some rules have been defined to assess model global fit (Lei & Wu, 2007). The rules state that the chi square must be non-significant ($\chi^2$, ns), comparative fit index (CFI) must be close to 1 (CFI ≥.95), root mean squared error approximation (RMSEA) must be less than 0.08, and standardized root mean squared residual (SRMR) must be less than 0.08. It is advised that more than one index be used to validate one's model. Researchers have also been advised to take these rules with a grain of salt because they are the result of evaluations made in very limited contexts (Steiger, 2007).

To assess model performance, we used multiple fit indices and focused on the structural models which identified how various factors relate to one another. We placed importance on whether the models identified latent constructs and whether they identified direct, indirect, or no relationships among variables.

Since there are many models in our study, the specification of each SEM is provided in the context of the results for the research question it aims to answer.

## 4.2 Results

For all *SEM* research questions, global fit indices were used to assess model acceptability: chi-square was non significant *($\chi^2$, ns)*, CFI was close to 1 *(cfi ≥.95)*, RMSEA was less than 0.08

*(RMSEA ≤ .08),* and SRMR was less than 0.08 *(SRMR≤ .08)* for all models. Features, their aliases, corresponding encodings, and comprehensive results for each *SEM RQ* can be found in the Appendix B *SEM Comprehensive Results* section.

## 4.2.1 Do High School and University Program Details Explain Time to Completion?

### 4.2.1.1 Model Specification

We hypothesized the structural relationships amongst all variables and provided the grammar to conduct modeling. Grammar definition involves regressing the target variable "Time to Completion" on the other two predictor variables - High school details and Program Details. Grammar definition also involved describing covariances and variances for each of the variables, where High school details covaries with itself, and Program details covaries with itself. The description of this model is shown below; where b1 and b2 refer to the regression coefficients of both predictor variables.

```
# Structural relations; Time_Complete = Time to Completion
      Time_Complete ~ b1*High_School + b2*Program_Details #
# Covariance and Variances
       High_School ~~ Program_Details
       Program_Details ~~  Program_Details
       High_School ~~ High_School
```

In a typical regression model, we specify only one outcome variable which, in this case, is Time to Completion. With SEM, we are able to test whether there are mediation models that have direct and indirect effects on the outcome variable. Thus, we specified a second model to determine whether there were unobserved variables in High school details that explain variations in Time to Completion, through Program details. A description of this model is shown below; where b3 represents the multiplied regression coefficient of Program and High school details. The results of this multiplication is stored in the variable "ind" which translates to "indirect". In

this way, we investigate whether a mediation exists from students' High school performance, through their university Program details, that would explain their Time to completion.

```
# Structural model; Time_Complete = Time to Completion
    Time_Complete ~ b1*High_School + b2*Program_Details
    High_School ~ b3*Program_Details
# Covariance structure of exogenous variables
    Program_Details ~~ Program_Details
# New parameter
    ind := b1*b3
```

### 4.2.1.2 Model Estimation & Path Analysis

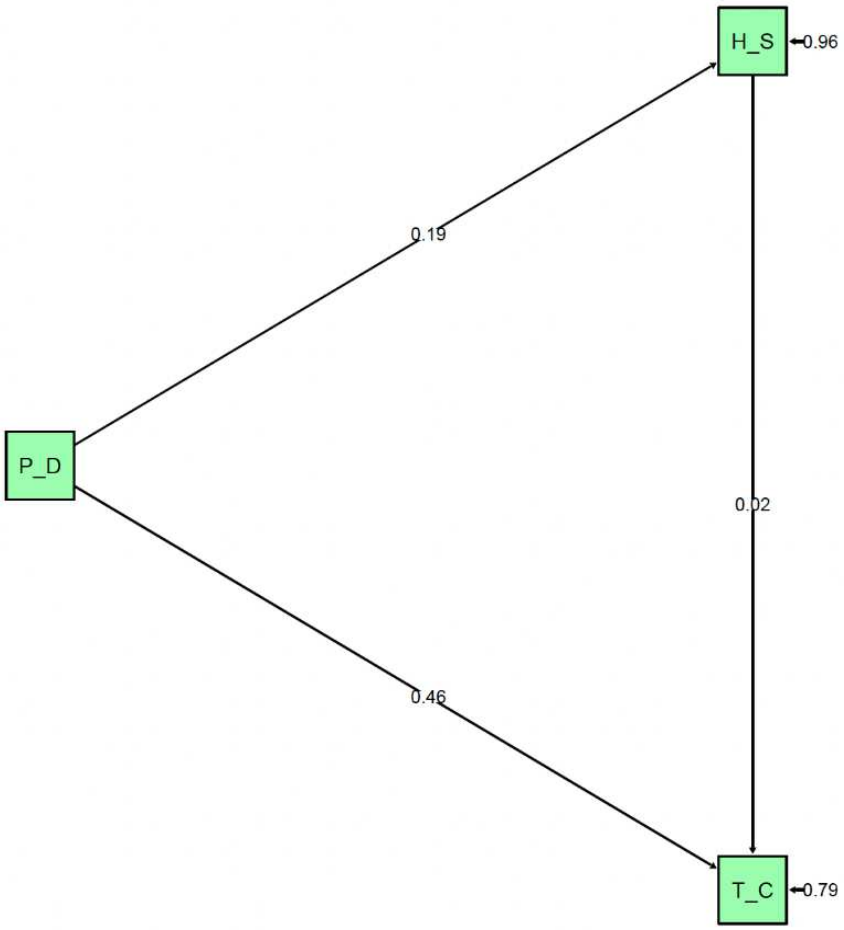
We fit the model specified to the data and obtained results, *($R^2$ = .21 $p < .001$)*. The r-square value was positive and showed that in this model, 21% of the variance in Time to completion can be explained by the two predictors High school performance and Program details. An r-square value < .03 is generally considered weak (Aiken, 2014) but in SEM, other indicators are used to describe the strength of an r-square value and its relevance to the hypothesis being tested. These indicators include the field of study, the number of exogenous constructs, how statistically significant the independent variables are, (F. Hair Jr et al., 2014) and the sample size of the data (Williamson, 2017).

Our field of study is education which falls under social and behavioral sciences. These fields inherently have greater levels of unexplainable variation because they try to explain human behavior, which is highly unpredictable (Whitley & Kite, 2012). As a result, a low r-square value is expected but does not constitute irrelevance of the model. Our model also involved two exogenous constructs from a sample size with 458 unique students. A higher sample size with more exogenous constructs will yield a higher r-square value (F. Hair Jr et al., 2014). Additionally, in this model, as shown in Figure 4, each of the independent variables - High school and Program details - explained the target variable - Time to completion. Thus, we interpret the r-square value of 21% as relevant for this hypothesis.

From the regression model specified, we obtained Figure 4.1 which shows a residual variance in High school $F(2,6) = .96$, $p < .001$ and in Time to completion $F(2,6) = .79$, $p < .001$ that is not explained by the model. There only exists grey lines in this regression path diagram which implies that there are only positive relationships among pairs of variables in this model. There is a positive correlation $r(3) = .19$, $p < .001$ between High school and Program details; a positive correlation $r(3) = .46$, $p < .001$ between Program details and Time to completion, and a positive correlation $r(3) = 0.02$, $p < .001$ between High school and Time to completion. This indicates that Time to completion can be explained by university Program details and High school performance details. The degree to which it is explained varies for both independent variables however. Program details explained Time to completion more than High school details in this model. The inequality of both influences is confirmed by a Wald test ($p < .001$).

The positive covariance ($F(2,6) = .19$, $p < .001$) obtained between High school performance and Program details suggests a positive relationship trend between both predictor variables. Both variables contain a combination of categorical variables such as “Course Id” and “Grading scheme” and quantitative variables such as “Grade”. We encoded the categorical variables but the number of courses taken by the students in High school are not equivalent to the number of courses they enrolled in for their university programs. Thus, we cannot describe a pattern with the courses. We can describe a pattern in the scores obtained. This positive covariance shows that students with high-scores in High school also obtained high grades in their university courses. This indicates a relevant positive association between measures of students’ High school performance and their university Program details.

**Figure 4.1**: SEM RQ1 Regression Path Diagram to Measure Influence of High School and Program Details




Note: P_D = Program Details, H_S = High School, T_C = Time to Completion. A comprehensive list of features and their aliases can be found in the Appendix B section.

The model was identified and showed enough pieces of information as parameters for estimation, which enabled us to perform Confirmatory Factor Analysis (CFA).

The mediation model specified generated a valid path diagram that showed the presence of latent constructs within the model. Figure 4.2 shows indirect influences between High school and Time to completion, and between Program details and Time to completion. The impact of High school details on Time to completion shows a positive correlation ($r(3) = .28$, $p < .001$). This suggests

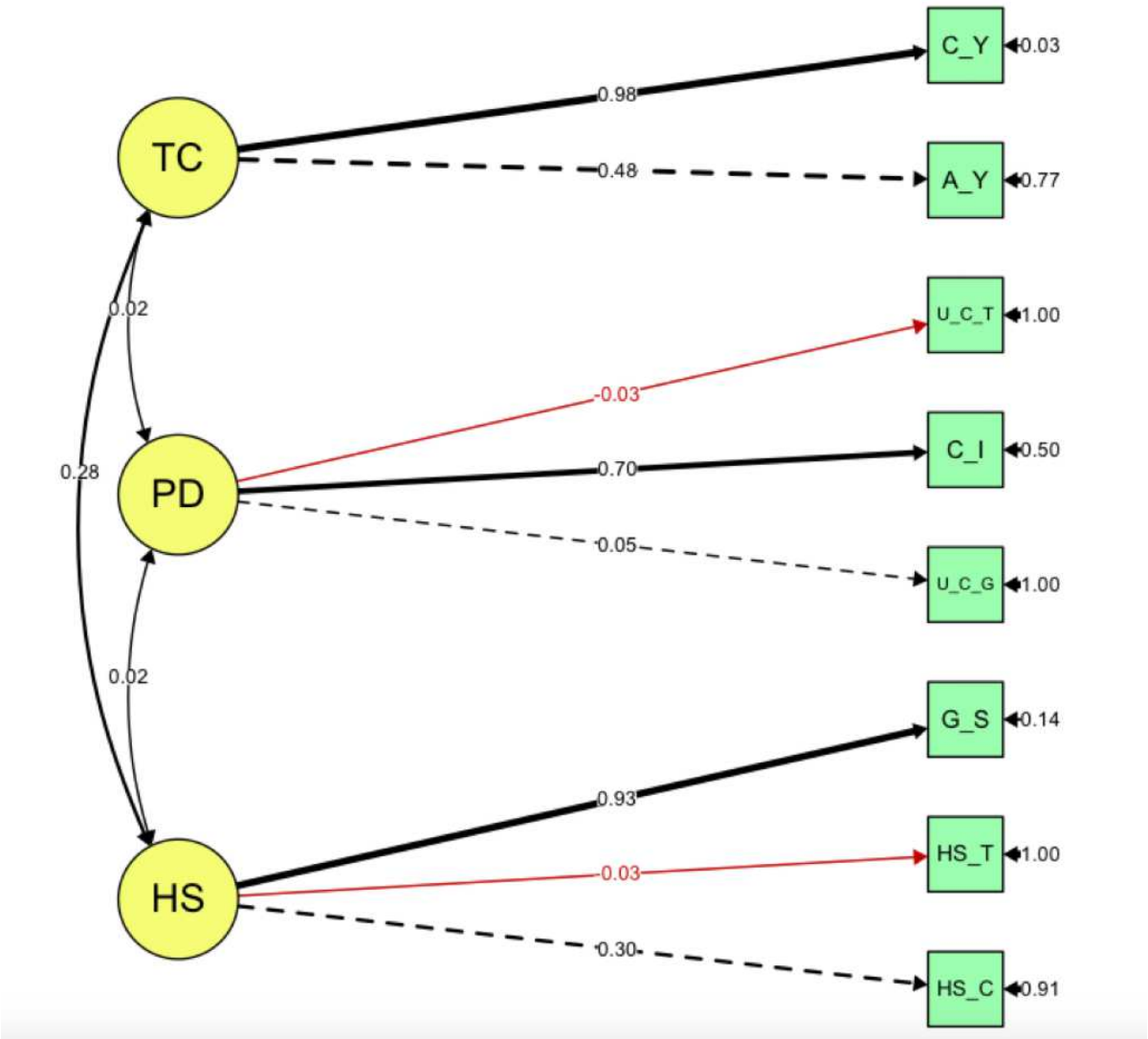
that a significant association exists between students' level of preparedness in High school, and students who completed their university programs within the time to completion durations we defined; less than or equal to 3 years and greater than 3 years. High school and Program details show a small positive correlation ($r(3) = .02, p < .001$) while Program details and Time to completion also show a small positive correlation ($r(3) = .02, p < .001$). This illustrates that a mediation exists between High school and Program details that explains Time to completion in a positive way. This mediated relationship is significant. We can therefore conclude that the positive correlation between these two pairs of variables provides evidence that Program details has an impact on Time to completion through High school.

The mediation model, as shown in Figure 4.2, illustrates the presence of latent constructs and the strengths of their effects. We defined the latent construct student preparedness for university, from the observed variables "High School Course Grades" (HS_C), "Grading Scheme" (G_S) and "High School Total Credits Earned" (HS_T) within the High school details category. We also defined academic endeavor from the observed variables "University Credits Taken" (U_C_T), "Course ID" (C_I) and "University Course Grade" (U_C_G) within the Program details category. From the Time to completion category, we defined the construct of speed from the observed variables "Admission Year" (A_Y) and "Completion Year" (C_Y). Student preparedness, academic endeavor and academic speed all showed covariations with one another. There was a significant positive relationship in student preparedness that was associated with the observed variable Grading Scheme (G_S). No relationship was observed with High School Course Grades (HS_C), while a weak negative relationship was observed in High School Total Credits Earned (HS_T). The academic endeavor construct showed a significant positive relationship with Course ID (C_I), no relationship was observed with University Course Grade (U_C_G) and a weak negative relationship was observed with University Credits Taken (U_C_T). For speed, we observed a strong positive relationship with Completion Year (C_Y) and no relationship with Admission Year (A_Y).

The implication of the covariations among constructs in this model signify that the latent variable within the High school details category explains Time to completion in combination with a latent variable within Program details. These latent variables do not explain the target variable to the

same degree however. High school performance provided higher explanatory power than Program details, in this model. The results of a Wald test ($p < .001$) confirmed this interpretation.

**Figure 4.2**: SEM RQ1 CFA and Factor Loading Path Diagram to Measure Influence of High School and Program Details




PD = Program Details, HS = High School, TC = Time to Completion. A comprehensive list of features and their aliases can be found in the Appendix B section.

## 4.2.2 Do High School Details and Student Demography Explain Time to Completion?

### 4.2.2.1 Model Specification

We specified two models. The first was a regression model used to regress the target variable Time to completion on Demography and High school details. We defined the grammar to obtain covariances and variances for each of the variables. This model provided covariances with variables in relation to itself and in relation to the target variable. The description of this model is shown below;

```
# Structural relations; Time_Complete = Time to Completion
        Time_Complete ~ b1*High_School + b2*Demography #
# Covariance and Variances
        High_School ~~ Demography
        High_School ~~ High_School
        Demography ~~ Demography
```

We specified a second model to determine whether High school can explain Time to completion through Demography. A description of this model is shown below;

```
# Structural model; Time_Complete = Time to Completion
      Time_Complete ~ b1*High_School + b2*Demography
      High_School ~ b3*Demography
# Covariance structure of exogenous variables
      Demography ~~ Demography
# New parameter
       ind := b1*b3
```

### 4.2.2.2 Model Estimation & Path Analysis

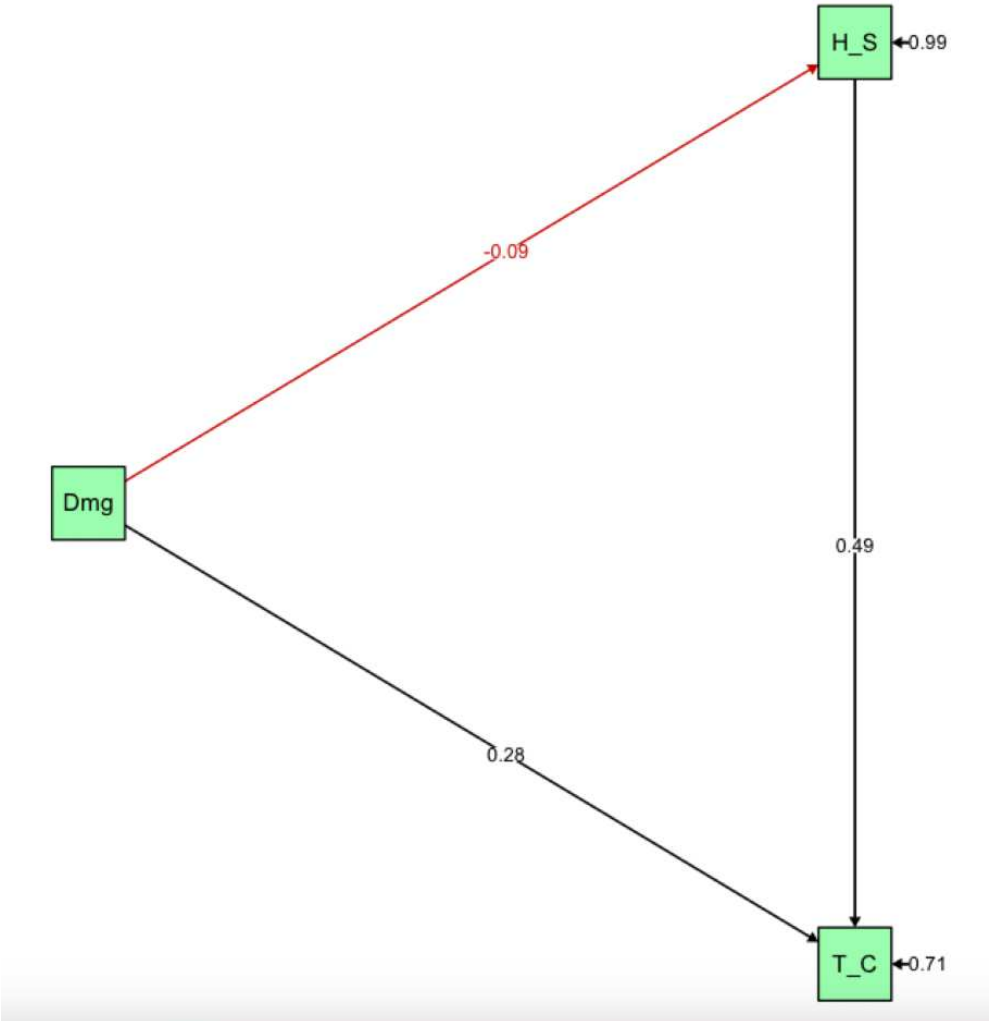
The model specified fit the data. It was identified and contained enough pieces of information to obtain parameters for estimation and perform CFA. We obtained regression results ($R^2 = .29$, $p <$

.001) that describe the estimated degree to which both predictor variables, Demography and High school details, explain the target variable Time to completion. This value is significant and positive.

No measurable covariance was found between High school and Demography ($F(2,6) = -0.09$, $p = .031$). As a result, we conclude that in this model, we cannot understand students' performance in high school through their identity as captured in Demography. This suggests that Demography does not influence students' High school performance as it relates to their grades, but influences their university time to completion.

Figure 4.3 shows residual variances in High school details, $F(2,6) = .99$, $p = <.001$, and in Time to completion, $F(2,6) = 0.71$, $p = <.001$, that is not explained by the model. Positive correlations ($r(3) = .49$, $p < .001$) exist between High school details and Time to completion, and ($r(3) = .28$, $p < .001$) between Demography and Time to completion. These positive correlations indicate significant associations that exist between individual predictor variables and the target variable. To describe these associations in terms of direction, we encoded the variable "Grading Scheme" to reflect whether students' high schools are domestic or international. In this way, we identified that a higher encoding for "Grading Scheme" reflected international students and low encodings reflected domestic students. Thus we can interpret the positive corelation between High school and Time to completion to mean that more students from international high schools completed their programs within stipulated timelines than domestic students. The influence of High school performance details on Time to completion is not equal to that of Demography on Time to completion as confirmed from a Wald test ($p < .001$). High school performance showed more influence on Time to completion than Demography.

**Figure 4.3:** SEM RQ2 Regression Path Diagram to Measure Influence of High School Details and Demography




Dmg = Demography, H_S = High School, T_C = Time to Completion. A comprehensive list of features and their aliases can be found in the Appendix B section.

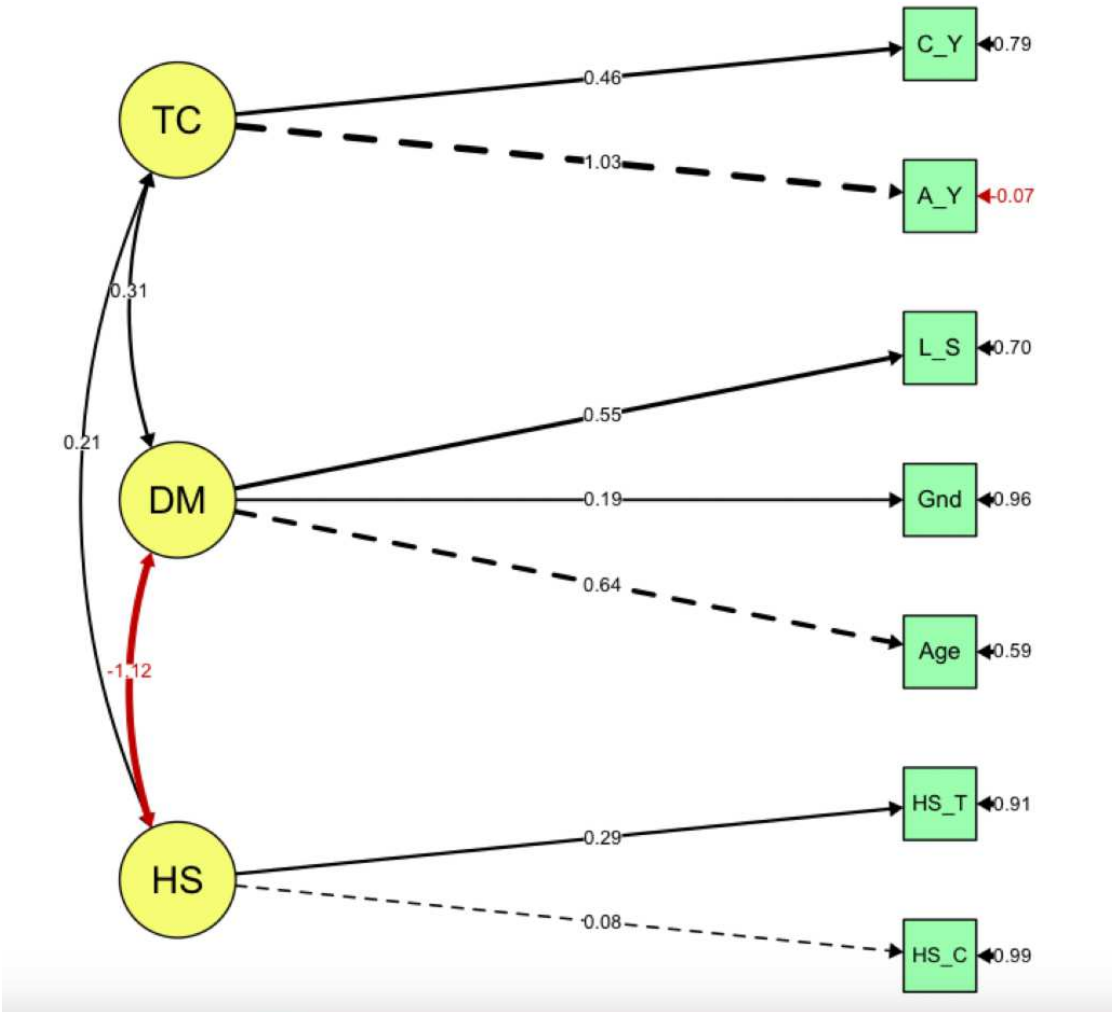
Figure 4.4 shows indirect influences between High school and Time to completion, and between Demography and Time to completion. A strong positive correlation ($r(3) = .21$, $p < .001$) is observed between High school and Time to completion which supports results obtained from *SEM* Research Question 1. We characterize 21% as strong because of the field of our study (education), our sample size (458) and the fact that we only have two exogenous constructs in this model. This correlation indicates a significant association between student preparedness and

the speed with which they complete their programs which we defined as thresholds between the first 3 years of their programs and the last 3 years of a potential 6 years. A positive correlation ($r(3) = .31$, $p < .001$) is also observed between Demography and Time to completion. This indicates a positive association between student identity and their Time to completion. A closer inspection showed that more international students completed their programs within the stipulated timelines than domestic students. This suggests that there may be economic influences or policies governing an international student's program at the University of Alberta that influence their time to completion. A negative correlation ($r(3) = -1.12$, $p = .032$) is observed between Demography and High school details. It is possible and acceptable to obtain negative coefficients greater than 1 in standardized path analysis (Jöreskog, 1994). A negative path loading of this nature is interpreted just like a negative correlation in a regression model would be; the predicted increase in the target variable for a one unit increase on the predictor, holding all other variables constant. This negative correlation is insignificant and therefore shows that there is no mediated influence on Time to completion that is explained by High school details, through Demography.

Figure 4.4 also shows that the mediation model specified generated valid path diagrams with the presence of latent constructs and the strengths of their effects. Latent constructs within this model all covary with one another. Existing latent constructs present in this model comprise student preparedness which we defined from the observed variables "High School Course Grades" (HS_C) and "High School Total Credits Earned" (HS_T). We also defined student identity from the observed variables "Age" (Age), "Gender" (Gnd) and "Legal Status" (L_S) within the Demography category. From the Time to completion category, we defined speed from the observed variable "Admission Year" (A_Y) and "Completion Year" (C_Y). Student preparedness, student identity and speed all show covariations with one another. There was a weak positive relationship between student preparedness and the observed variable High School Total Credits Earned (HS_T) while no relationship was observed with High School Course Grades (HS_C). The student identity construct showed weak positive relationships with Legal Status (L_S) and Gender (Gnd), while no relationship was observed with Age (Age). For speed, we observed a weak positive relationship with Completion Year (C_Y) and no relationship with Admission Year (A_Y).

Construct covariations in this model signify that the latent variable within the category High school details, explains Time to completion in combination with the latent variable within the Demography category. Their explanatory powers are not to the same degree. In this model, just like in the regression model, High school performance provided higher explanatory power than Demography. The results of a Wald test, ($p < .001$), confirmed this interpretation.

**Figure 4.4:** SEM RQ2 CFA and Factor Loading Path Diagram to Measure Influence of High School Details and Demography




DM = Demography, HS = High School, TC = Time to Completion. A comprehensive list of features and their aliases can be found in the Appendix B section.

## 4.2.3 Do High School Details and English Language Skills Explain Time to Completion?

### 4.2.3.1 Model Specification

The first model specified was a regression model that regressed the target variable Time to completion on measures of High school details and English Language Proficiency (ELP). The grammar to obtain covariances and variances for each of the variables were also defined. This model provided covariances of the predictor variables with themselves and in relation to the target variable. The description of this model is shown below:

```
# Structural relations; Time_Complete = Time to Completion
        Time_Complete ~ b1*High_School + b2*ELP #
# Covariance and Variances
        High_School ~~ ELP
        High_School ~~ High_School
        ELP ~~ ELP
```

The second model was specified to describe and test for mediations between predictor variables in relation to the target variable. From this second model, we obtained factor loadings that described the presence of latent constructs in the model. This second model also provided the basis to assess the extent to which High school details explained variations in Time to completion, through ELP. A description of this model is shown below;

```
# Structural model; Time_Complete = Time to Completion
      Time_Complete ~ b1*High_School + b2*ELP
      High_School ~ b3*ELP
# Covariance structure of exogenous variables
      ELP ~~ ELP
# New parameter
      ind := b1*b3
```

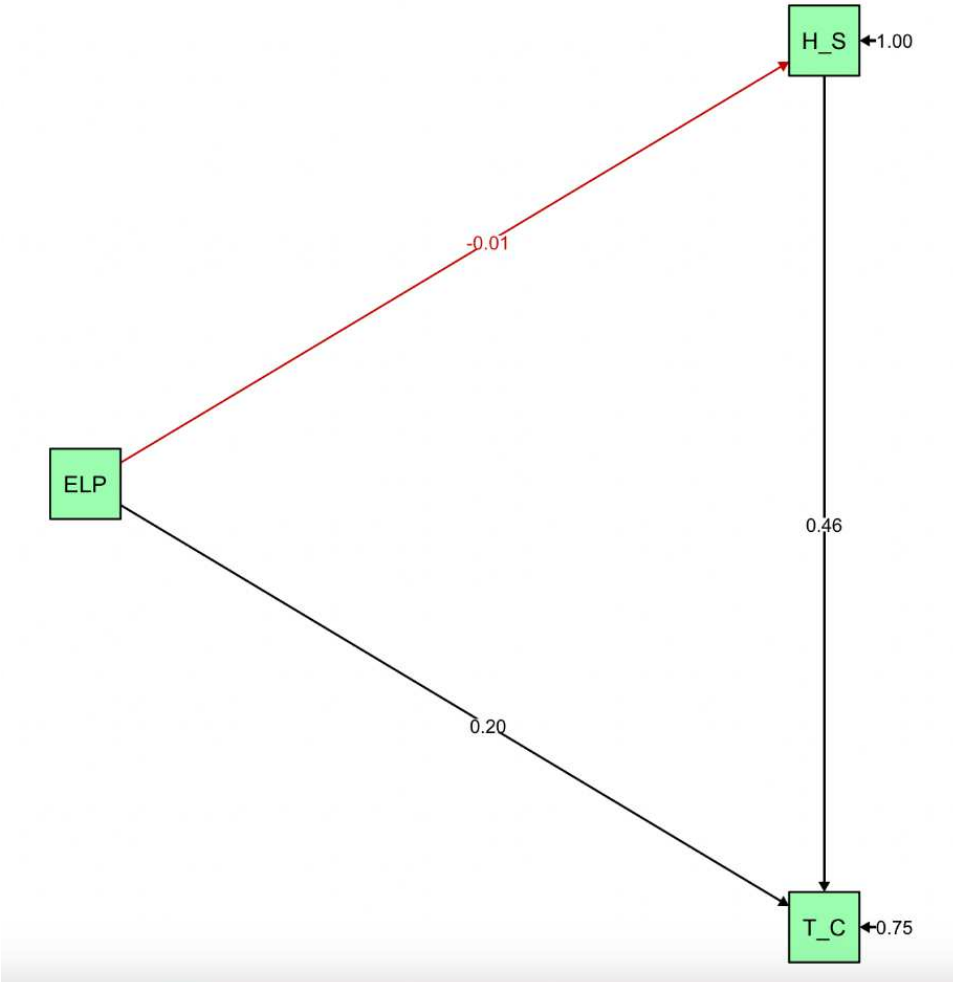
### 4.2.3.2 Model Estimation & Path Analysis

The regression model specified fit the data and showed enough pieces of information as parameters in the model to perform CFA. We obtained a strong positive coefficient of determination ($R^2 = .25$, $p < .001$) that describes the estimated degree to which High school details and ELP explain Time to completion. This indicates that data about students' high school performance and their proficiency in the English language, can help us understand their time to completion.

High school details has a non-significant covariance ($F(2,6) = -0.01$, $p = .052$) with ELP. Thus, we cannot understand students' English proficiency using details of their High school performance, or vice versa.

Figure 4.5 shows a residual variance in High school details ($F(2,6) = 1.00$, $p < .001$) and in Time to completion ($F(2,6) = .75$, $p < .001$) that is not explained by the model. There is a positive correlation ($r(3) = .46$, $p < .001$) between High school and Time to completion, and one ($r(3) = .20$, $p < .001$) between ELP and Time to completion. These results indicate significant influences from High school and ELP on Time to completion. While they can both explain the target variable, the extent to which they achieve this explanation is not the same for both variables. In this model, just as we saw in *SEM* RQ2, High school performance contained more explanatory power than ELP. This conclusion is confirmed by a Wald test ($p < .001$).

**Figure 4.5**: SEM RQ3 Regression Path Diagram to Measure Influence of High School Details and English Language Skills




ELP = English Language Proficiency, H_S = High School, T_C = Time to Completion. A comprehensive list of features and their aliases can be found in the Appendix B section.

The valid path diagram in Figure 4.6 shows mediations among latent constructs in all three categories. These mediations indicate indirect influences from one category through another to the target category. High school shows a significant positive correlation ($r(3) = .35$, $p < .001$) with Time to completion. This is consistent with results obtained from *SEM* RQ1 and *SEM* RQ2. This correlation indicates that high scoring students in High school take less time to complete their university programs. Indirect influence from ELP on Time to completion shows a minimal positive correlation ($r(3) = .04$, $p < .001$). The indirect influence between High school and ELP

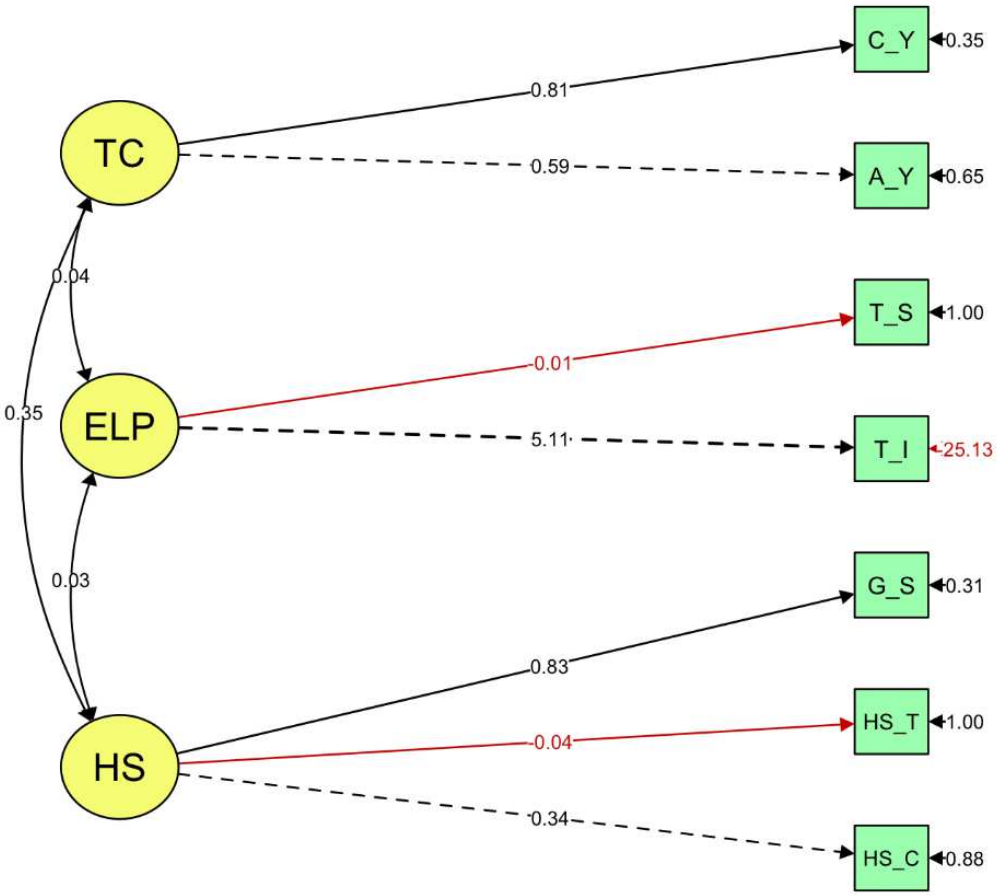
that contributes to explaining Time to completion also shows a minimal but significant positive correlation ($r(3) = .03$, $p < .001$). This indicates that we can understand student's time to completion from their level of preparedness in High school and their level of English proficiency. High scoring students from international high schools showed high English language proficiency and also showed speedy completion of their university programs. While there are direct and indirect influences from both predictor variables on the target variable, they do not explain Time to completion to the same degree. Students' High school details offer more explanatory power than ELP, as confirmed by a Wald test ($p < .001$).

Figure 4.6 illustrates the presence and strengths of latent constructs. Student preparedness is a latent construct within the High school category defined from "High School Course Grades" (HS_C), "Grading Scheme" (G_S) and "High School Total Credits Earned" (HS_T) variables. English proficiency is a construct we defined from the variables "Test Id" (T_I) and "Test Score" (T_S) within the ELP category. Speed is a construct we defined from "Admission Year" (A_Y) and "Completion Year" (C_Y) variables within the Time to completion category. All latent constructs showed covariations with one another. The strength and significance of the relationship of observed variables to each construct was shown in the factor loadings. There was a strong positive relationship in student preparedness that was observed with the variable Grading Scheme (G_S). No relationship was observed in High School Course Grades (HS_C), while a weak negative relationship was observed with High School Total Credits Earned (HS_T). This suggests that total course credits earned by students in high school did not help us understand how well prepared they were for university. Thus, students with fewer total credits could have been more prepared for university than students with higher total course credits. The English proficiency construct showed no relationship with Test Id (T_I) and a weak negative relationship with Test Score (T_S). For speed, we observed a strong positive relationship with Completion Year (C_Y) and no relationship with Admission Year (A_Y). This is consistent with results obtained from previous *SEM* research questions.

Both predictor variables do not explain the target variable to the same degree. Covariations among constructs in this model signify that the latent variable within High school details

explained Time to completion more than the latent variable within ELP. The results of a Wald test ($p < .001$) confirmed this interpretation.

**Figure 4.6**: SEM RQ3 CFA and Factor Loading Path Diagram to Measure Influence of High School Details and English Language Skills




ELP = English Language Proficiency, HS = High School, TC = Time to Completion. A comprehensive list of features and their aliases can be found in the Appendix B section.

## 4.2.4 Do Demography and Program Details Explain Time to Completion?

### 4.2.4.1 Model Specification

The first model specified was a regression model that aimed to determine structural relationships amongst variables Demography, Program Details and Time to Completion. The description of this model is shown below:

```
# Structural relations; Time_Complete = Time to Completion
        Time_Complete ~ b1*Program_Details + b2*Demography #
# Covariance and Variances
         Program_Details ~~ Demography
         Demography ~~ Demography
         Program_Details ~~ Program_Details
```

The second model specified was a mediation model that aimed to determine the presence and strengths of latent constructs within all three variable categories. Latent constructs may have direct and indirect effects on the outcome variable. This model will estimate the extent to which Demography explains variation in Time to Completion through Program details. A description of this model is shown below:

```
# Structural model; Time_Complete = Time to Completion
        Time_Complete ~ b1*Program_Details + b2*Demography
        Program_Details ~ b3*Demography
# Covariance structure of exogenous variables
         Demography ~~ Demography
# New parameter
         ind := b1*b3
```

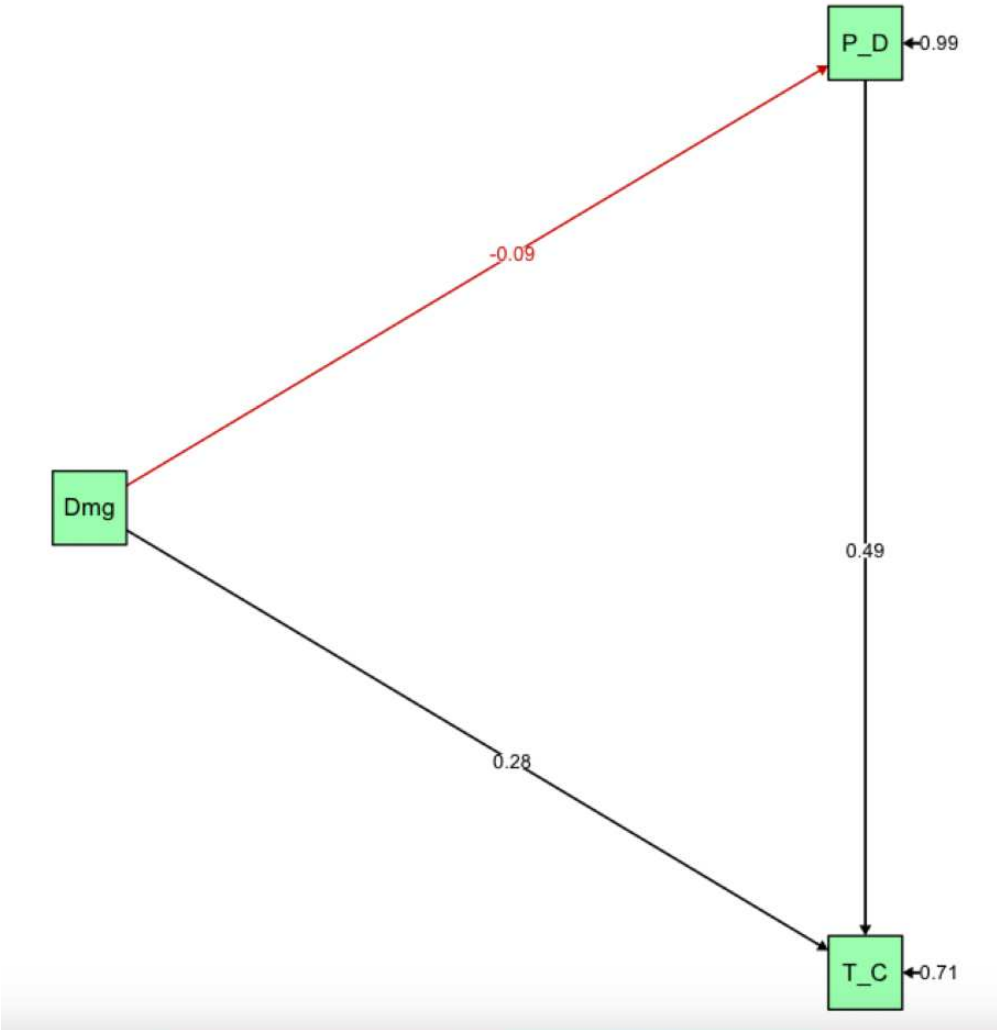
### 4.2.4.2 Model Estimation & Path Analysis

The data was sufficient enough to obtain parameter estimates for CFA and the model was exactly identified. From fitting the data to the model, we obtained a strong and positive regression result ($R^2 = .29,\ p < .001$). In practical terms, 29% of the variance in Time to completion can be explained by both predictor variables.

The covariance ($F(2,6) = -0.09$, $p = .025$) between Demography and Program details is not significant. Therefore we cannot understand student identity through their university Program details, or vice versa. We performed a Wald test to ascertain the equivalence of both predictor variables in explaining the target variable Time to completion. The result ($p < .001$) confirmed that Demography and Program Details do not explain Time to completion equally. Program details offered more explanatory power.

Figure 4.7 shows the path diagram for the regression model specified. There is a residual variance in Program details ($F(2,6) = .99$, $p < .001$) and in Time to completion ($F(2,6) = .71$, $p < .001$) that is not explained by the model. A significant and positive correlation ($r(3) = 0.49,\ p < .001$) exists between Program details and Time to completion. To put this in perspective, the target variable Time to completion can be understood by variables within the Program details category. This signifies that, in this model, Program details can explain Time to completion, and it supports the correlation result obtained in *SEM RQ1*. A significant and positive correlation also exists between Demography and Time to completion ($r(3) = .28$, $p < .001$) which indicates that Time to completion can be explained by Demography. Some of the Demographic features within this category cannot be said to increase or decrease. Features such as Gender and Legal status cannot increase or decrease; and Age cannot decrease even though it can increase. The identified relationships showed that more male international students completed their programs within 4 to 6 years than their domestic counterparts, while more female international students completed theirs within the first 3 years than domestic female students. Additionally, younger students, regardless of legal status, completed their programs within stipulated timelines more than older students.

**Figure 4.7**: SEM RQ4 Regression Path Diagram to Measure Influence of Demography and Program Details




Dmg = Demography, P_D = Program Details, T_C = Time to Completion. A comprehensive list of features and their aliases can be found in the Appendix B section.

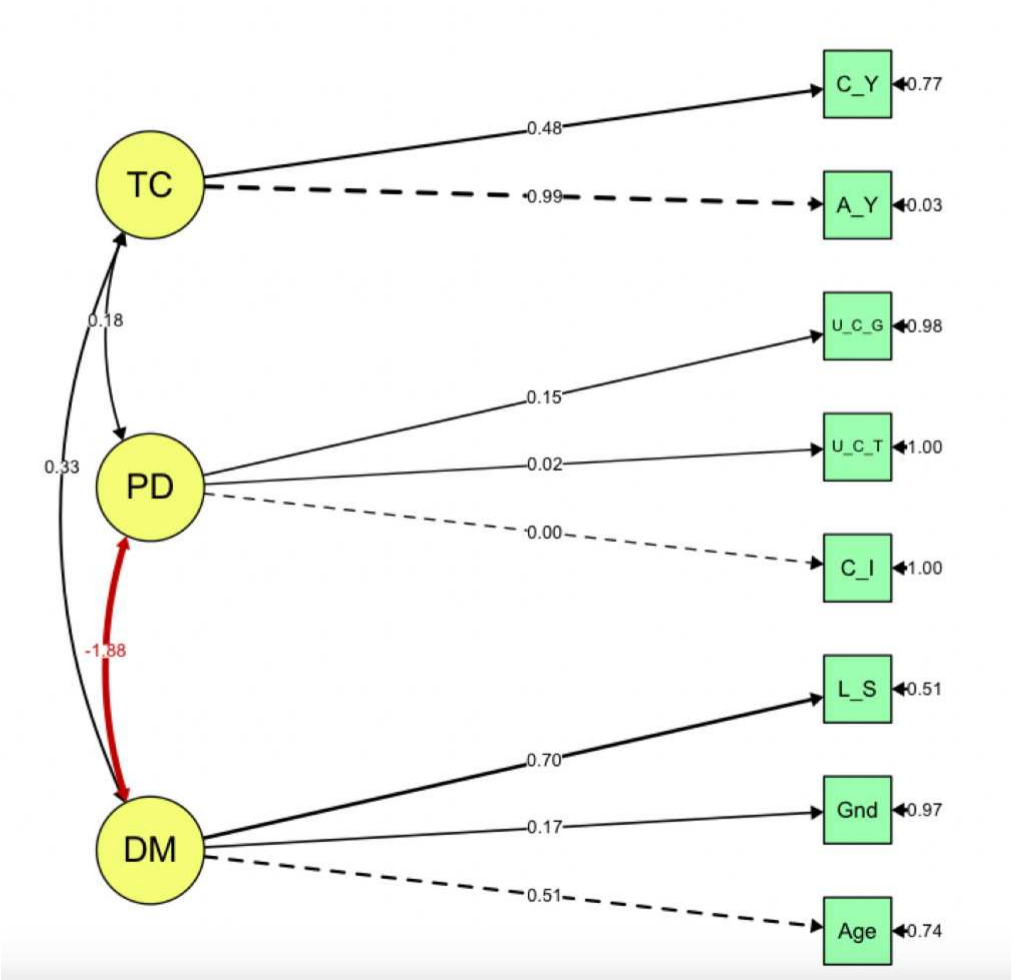
Figure 4.8 illustrates the mediation model specified with a valid path diagram that reflects the presence of latent constructs. There are direct influences between Demography and Time to completion and Program details and Time to completion. Indirect influences exist from Demography to Time to completion through Program details. The impact of Demography on Time to completion shows a strong positive correlation ($r(3) = .33$, $p < .001$). This implies that Demography can explain Time to completion. There is a weak positive correlation ($r(3) = .18$, $p < .001$) between Program details and Time to completion. This also suggests that Program details

can explain Time to completion but not to the same extent as Demography. A strong negative correlation ($r(3) = -1.88$, $p < .001$) exists between Demography and Program details. This illustrates that there is mediation between Demography and Program details in the explanation of Time to completion. The negative correlation seen between both predictor variables suggests that male international students had less course credits and had lower course grades than female domestic students. Younger students, regardless of legal status, had more course credits and scored higher in their course grades than older students. This inverse relationship trend influenced the Time to completion of both domestic and international students.

There are latent constructs within this model as shown in Figure 4.8. Student identity is defined from the observed variables "Age" (Age), "Gender" (Gnd) and "Legal Status" (L_S) within the Demography category. Academic endeavor is defined from the observed variables "University Credits Taken" (U_C_T), "Course ID" (C_I) and "University Course Grade" (U_C_G) within the Program details category. Speed is defined from two observed variables, "Admission Year" (A_Y) and "Completion Year" (C_Y), within the Time to completion category. Student identity, academic endeavor and speed all show covariations with one another. Student identity showed a strong positive relationship with Legal Status (L_S). A weak positive relationship was observed with Gender (Gnd) and a non-significant relationship was observed with Age (Age). This suggests that students' legal status is a more salient component of identity than their gender and that age was not an important component of their identity in this context. Academic endeavor showed weak positive relationships with University Course Grade (U_C_G) and University Credits Taken (U_C_T), and a non significant relationship with Course ID (C_I). This suggests that the grades obtained, and credits taken in their university programs explained students' academic endeavor more than their course codes. Speed showed a weak positive relationship with Completion Year (C_Y), and a non significant relationship with Admission Year (A_Y).

Covariations among constructs indicate that the latent variable within both predictor variable categories explain Time to completion to varying degrees. The results of a Wald test, ($p < .001$), confirmed this interpretation. Demography provided more explanation in this model.

**Figure 4.8:** SEM RQ4 CFA and Factor Loading Path Diagram to Measure Influence of Demography and Program Details




DM = Demography, PD = Program Details, TC = Time to Completion. A comprehensive list of features and their aliases can be found in the Appendix B section.

## 4.2.5 Do Program Details and English Language Skills Explain Time to Completion?

### 4.2.5.1 Model Specification

A regression model was specified and used to regress the target variable Time to completion on the two predictor variables - English language skills and Program details. The description of this model is shown below;

```
# Structural relations; Time_Complete = Time to Completion
        Time_Complete ~ b1*Program_Details + b2*ELP #
# Covariance and Variances
        Program_Details ~~ ELP
        Program_Details ~~ Program_Details
        ELP ~~ ELP
```

A mediation model was specified and used to test for mediations between predictor variables in relation to the target variable. A description of this model is shown below;

```
# Structural model; Time_Complete = Time to Completion
        Time_Complete ~ b1*Program_Details + b2*ELP
        Program_Details ~ b3*ELP
# Covariance structure of exogenous variables
        ELP ~~ ELP
# New parameter
        ind := b1*b3
```

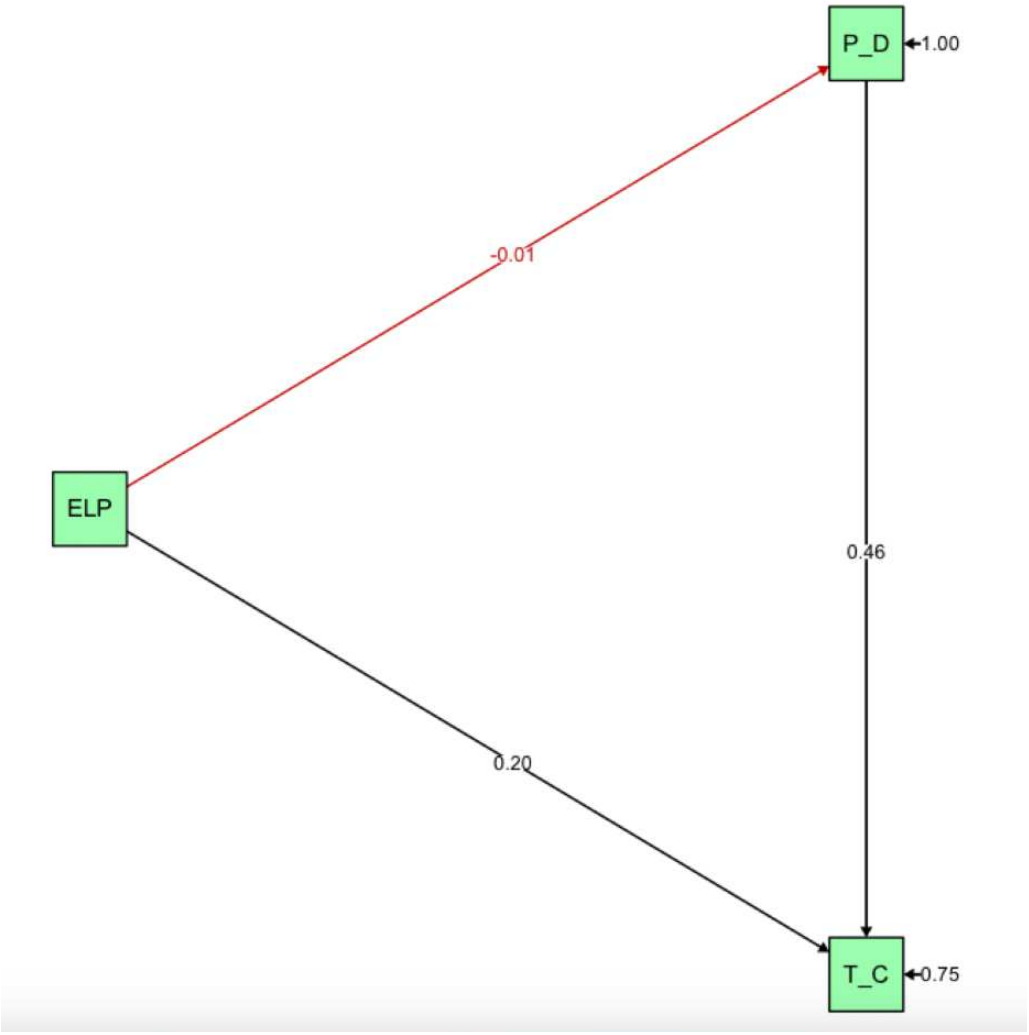
### 4.2.5.2 Model Estimation & Path Analysis

Regression coefficient ($R^2 = .25$, $p < .001$) obtained showed a strong positive estimated degree to which both English language skills and Program details can explain Time to completion. In this model, 25% variance in Time to completion was explained by both predictor variables implying that a substantial portion of students' Time to completing their university degrees, can be understood by course details of their university Program and English language skills. The result of a Wald test ($p < .001$) showed that the target variable Time to completion can be explained by both predictor variables but not to the same degree. In this model, Program details offered more explanatory power than ELP.

No covariance ($F(2,6) = -0.01$, $p = .041$) was found between English language skills and Program details. The regression model in Figure 4.9 shows residual variances ($F(2,6) = 1.00$, $p <$

.001) in Program details and ($F(2,6) = 0.75$, $p < .001$) in Time to completion that is not explained by the model. A strong positive correlation ($r(3) = .46$, $p < .001$) exists between Program details and Time to completion. This result suggests Time to completion can be explained by students' university Program details. There is a combination of quantitative and categorical variables within Program details. Quantitative variables include "University Course Grade"; categorical variables include "Course ID". This suggests, unsurprisingly, that students who attained high grades in their university courses, completed their programs within stipulated timelines. A positive correlation ($r(3) = .20$, $p < .001$) also exists between English language skills and Time to completion. Categorical variables within ELP include "Test_Id" while quantitative variables include "Test Score" This result suggests that students who scored highly in their English language tests, completed their university programs within the stipulated timelines.

**Figure 4.9***:* SEM RQ5 Regression Path Diagram to Measure Influence of English Language Skills and Program Details




ELP = English Language Proficiency, P_D = Program Details, T_C = Time to Completion. A comprehensive list of features and their aliases can be found in the Appendix B section.

Figure 4.10 illustrates the mediation model specified and shows the presence of latent constructs. There are direct influences between English language skills and Time to completion, and between Program details and Time to completion. The influence of English language skills on Time to completion shows a positive correlation ($r(3) = .41$, $p < .001$). This indicates that a significant association exists between a substantial number of students who earned high scores in their English language skills tests, and their Time to completion. There is also a direct influence

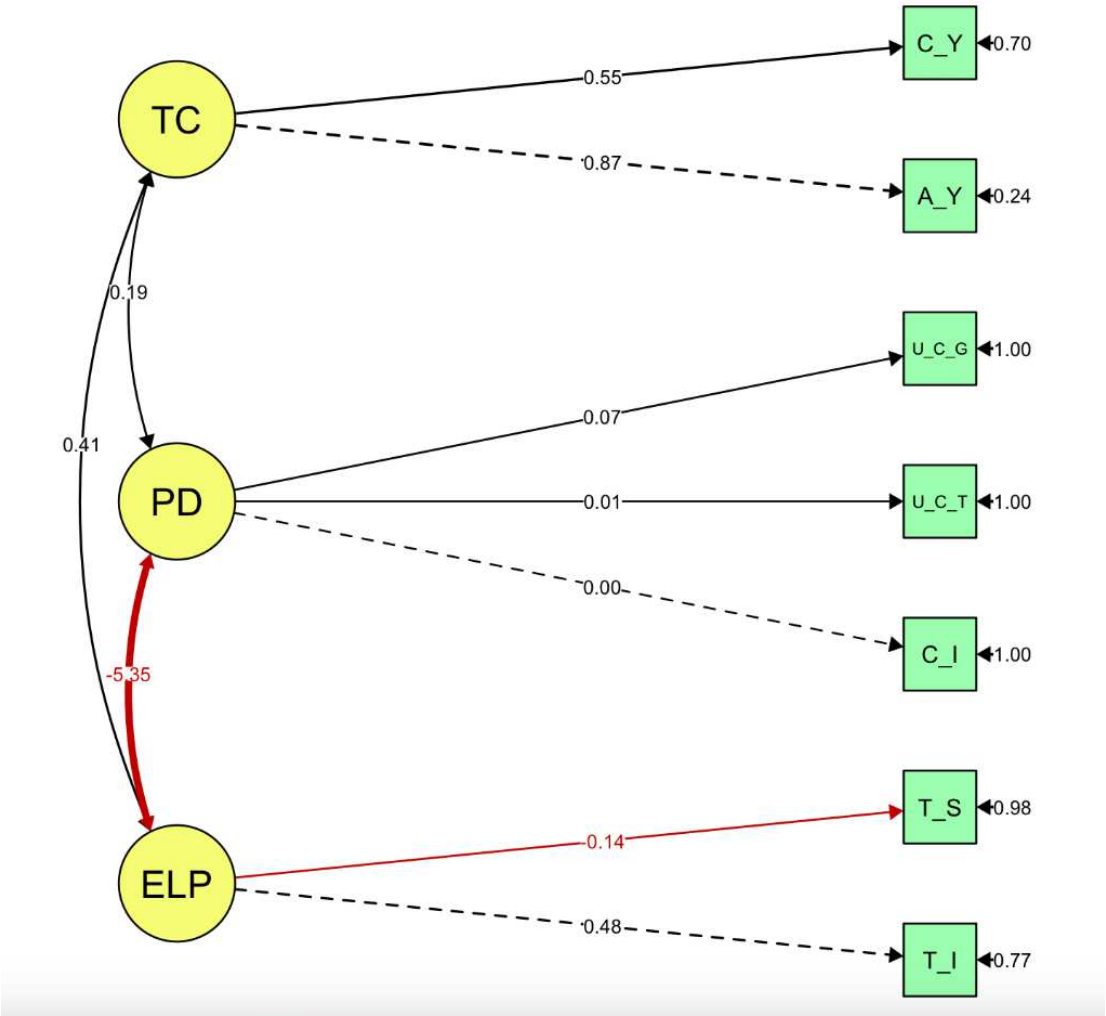
from Program details on Time to completion which also shows a significant positive correlation ($r(3) = .19$, $p < .001$). This implies a meaningful relationship exists between a substantial number of students with high grades in their program courses and their Time to completion. We interpret this relationship to mean that students with high grades and higher number of courses take longer to complete their programs than students with fewer courses. In contrast, English language skills and Program details showed no relationship: $r(3) = -5.35$, $p = .042$. This means that we cannot understand students' Time to completion from the combination of how proficient they are at the English language and their performance in university courses.

Figure 4.10 also depicts the presence of latent constructs and the strengths of their effects. Academic endeavor can be defined from the observed variables "University Credits Taken" (U_C_T), "Course ID" (C_I) and "University Course Grade" (U_C_G) within the Program details category. Academic speed can be defined from "Admission Year" (A_Y) and "Completion Year" (C_Y) within the Time to completion category. English proficiency is a construct defined from the variables "Test Id" (T_I) and "Test Score" (T_S) within the ELP category. All constructs within this model covary with one another. No relationship of English proficiency was observed with the measured variable Test Id (T_I). A weak negative relationship was observed with the measured variable Test Score (T_S). When we considered academic endeavor, weak positive relationships were observed with University Course Grade (U_C_G) and University Credits Taken (U_C_T). Additionally, no relationship was observed with Course ID (C_I) for the academic endeavor construct. Student's speed showed a weak positive relationship with Completion Year (C_Y) and no relationship was observed with Admission Year (A_Y). This suggests that students' completion year was a stronger indicator for how much time students took to complete their programs. Low values indicate less time taken for completion. While this is not surprising, it provides justification that the data is valid and consistent. These construct relationships illustrate the strengths and effects of latent variables in the same way the strengths and significance of observed variables can be measured.

The covariations among constructs in this model show few indirect effects between the predictor variables. Individually, both predictor variables influenced students' Time to completion. Jointly, in this model, students' ability to speak the English language did not help us understand their academic endeavor. That is, students' English proficiency had no effect on their academic

endeavors that could explain their time to completion. ELP showed higher influence on Time to completion, than Program details, in this model. The results of a Wald test, ($p > .001$), confirmed this interpretation.

**Figure 4.10***:* SEM RQ5 CFA and Factor Loading Path Diagram to Measure Influence of English Language Skills and Program Details




PD = Program Details, ELP = English Language Proficiency, TC = Time to Completion. A comprehensive list of features and their aliases can be found in the Appendix B section.

## 4.2.6 Do Demography and English Language Skills Explain Time to Completion?

### 4.2.6.1 Model Specification

The first model specified was a regression model used to regress the target variable Time to completion on English Language Skills and Demography. We also defined the grammar to obtain covariances and variances for each of the variables. The description of this model is shown below;

```
# Structural relations; Time_Complete = Time to Completion
    Time_Complete ~ b1*Demography + b2*ELP #
 # Covariance and Variances
    Demography ~~ ELP
    ELP ~~ ELP
    Demography ~~ Demography
```

The second model specified was a meditating model used to describe and test for mediations between Demography and English language skills in relation to Time to completion. We obtained measures called factor loadings that describe the presence of latent constructs. We assessed these latent constructs to determine the extent to which English language skills explain variations in Demography and Time to completion. A description of this model is shown below;

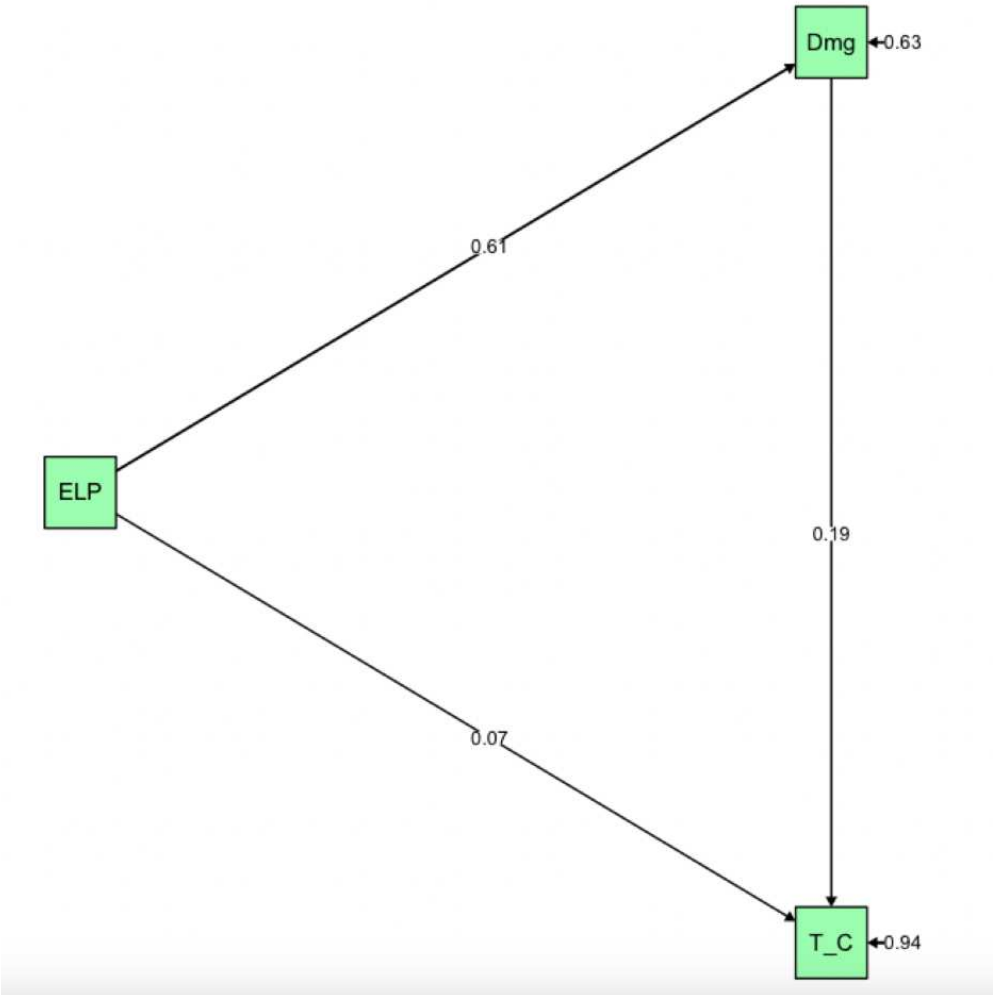
```
# Structural model; Time_Complete = Time to Completion
    Time_Complete ~ b1*Demography + b2*ELP
    Demography ~ b3*ELP
# Covariance structure of exogenous variables
    Demography ~~ Demography
# New parameter
    ind := b1*b3
```

### 4.2.6.2 Model Estimation & Path Analysis

The regression model specified fit the data and generated results ($R^2 = 0.10$, $p < .001$). The covariance ($F(2,6) = .61$, $p < .001$) obtained between English language skills and Demography was positive. Descriptors for Demography - Gender and Legal Status - all contain categorical values which we encoded. We then used those encodings to interpret the relationship. This suggests that the older the student, the more English proficiency they possess. It also suggests that male and international students possess high English language proficiencies.

Figure 4.11 depicts residual variances in Demography ($F(2,6) = .63$, $p < .001$) and in Time to completion ($F(2,6) = .94$, $p < .001$) that are not explained by the model. There are only positive correlations in this model; a positive correlation exists between English language skills and Time to completion ($r(3) = .07$, $p < .001$) , and between Demography and Time to completion ($r(3) = .19$, $p < .001$) . This supports the r-squared value obtained that shows significant and positive explanation for Time to completion from both predictor variables. Although both predictors have significant influence on Time to completion, a Wald test ($p < .001$) confirms that Demography has more influence than ELP.

**Figure 4.11**: SEM RQ6 Regression Path Diagram to Measure Influence of Demography and English Language Skills




ELP = English Language Proficiency, Dmg = Demography, T_C = Time to Completion. A comprehensive list of features and their aliases can be found in the Appendix B section.

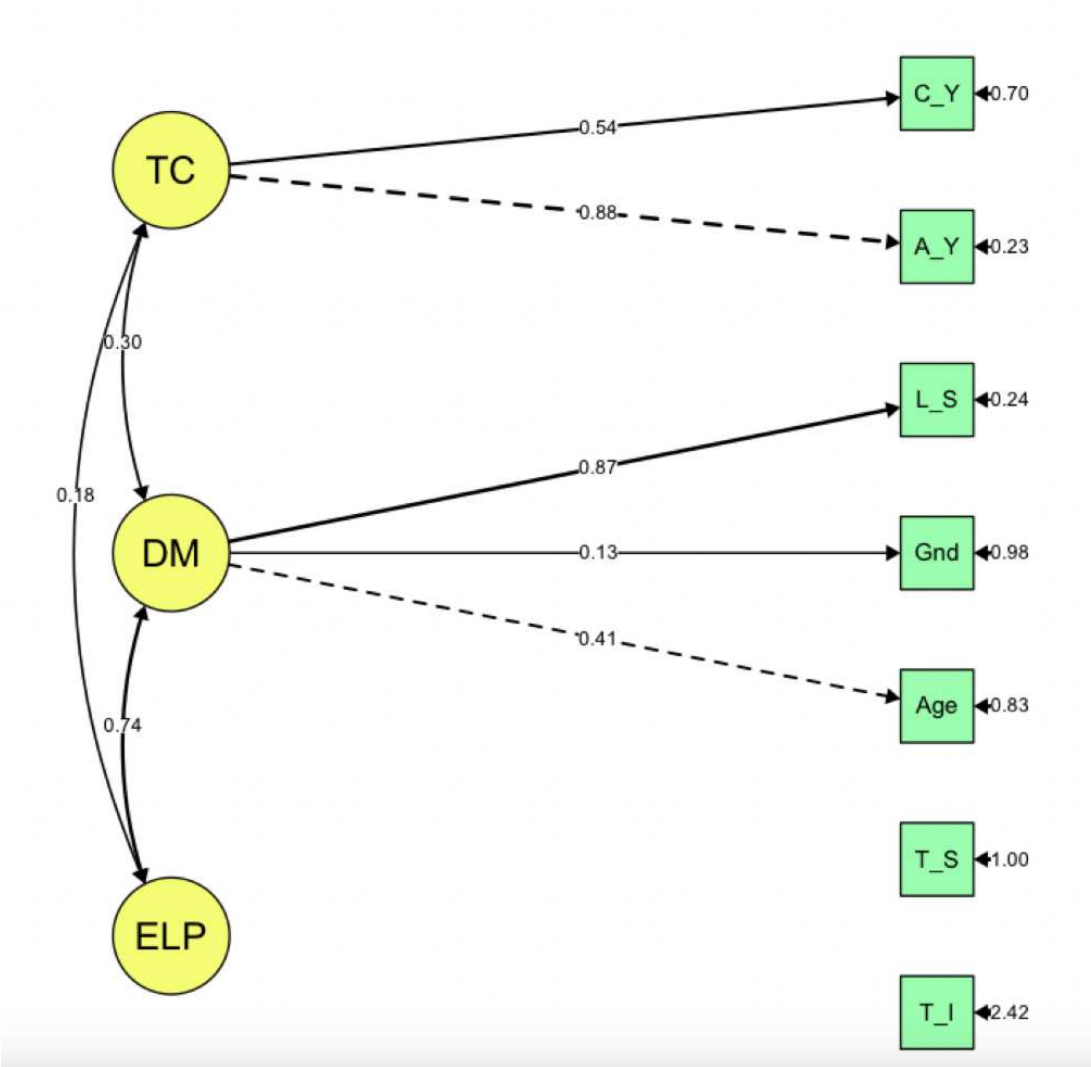
Figure 4.12 shows a path diagram for the mediation model specified. In fitting the data to this model, estimates were obtained that show factor loadings for latent constructs within only one predictor variable and the target variable. There are direct influences between English language skills and Time to completion, and between Demography and Time to completion. The impact of English language skills on Time to completion shows a positive correlation ($r(3) = .18$, $p < .001$). Demography also shows a significant positive correlation ($r(3) = .30$, $p < .001$) with Time to completion.

Within this model, latent constructs exist in all variable categories but only show factor loadings for Demography and Time to completion. Unsurprisingly, English language skills showed a strong positive correlation ($r(3) = .74$, $p < .001$) with Demography. Apart from the direct influence of English language skills on Time to completion, there is a mediated explanation that exists in combination with Demography for Time to completion. This influence indicates that English proficiency in combination with students' identity can be used to understand their Time to completion. The latent construct English proficiency showed no relationships with any observed variable within the English language skills category. This means that within this model, the effect and strength of this construct from observed variables cannot be determined, but its presence is identified (Loehlin, 2003). Latent construct influence also exists between Demography and Time to completion implying that students' identity can be used to directly explain their Time to completion.

From Figure 4.12, we can observe the factor loadings of latent constructs which we defined from observed variables. Student identity is a construct we defined from "Age" (Age), "Gender" (Gnd) and "Legal Status" (L_S) within the Demography category. We also defined English proficiency from the observed variables "Test Id" (T_I) and "Test Score" (T_S) within the English language skills category. From the Time to completion category, we defined the construct of speed from the observed variable "Admission Year" (A_Y) and "Completion Year" (C_Y). Student identity, English proficiency and speed all covary with one another. Student identity showed a weak positive relationship with the observed variable Gender (Gnd). A strong positive relationship was observed with Legal Status (L_S) and no relationship was observed with Age (Age). This result aligns with prior SEM investigations for student identity. It suggests that student identity is more defined by their legal status and gender, than by age. For speed, we observed a weak positive relationship with Completion Year (C_Y) and a non significant relationship with Admission Year (A_Y). No factor loadings were obtained for the English proficiency construct in this model.

Covariations among constructs in this model signify that the latent variable within the Demography category explains Time to completion in combination with the latent variable within English language skills. Demography offered more explanatory power than ELP (Wald test $p < .001$).

**Figure 4.12***:* SEM RQ6 CFA and Factor Loading Path Diagram to Measure Influence of Demography and English Language Skills




DM = Demography, ELP = English Language Proficiency, TC = Time to Completion. A comprehensive list of features and their aliases can be found in the Appendix B section.

## 4.3 Discussion of SEM results by RQ

### 4.3.1 High School Details and Program Details

In *SEM RQ1*, High school and Program details jointly showed significant but different influences on Time to completion. There was a strong positive relationship observed between student's preparedness for university as measured by their High school details, and the amount of academic endeavor they exhibited in their university Program details. From the results, it was clear that students who achieved past successes in High school, also demonstrated appreciable levels of academic endeavor in their university programs. Individually, and not surprisingly, program details showed more influence on Time to completion, directly interpreting students' academic endeavor in university as a strong indicator of their potential to complete their programs on time.

In a 2019 study (Galla et al., 2019) where the evaluation was done between high school grades and university admissions test scores, results found that high school grades contained higher predictive ability than admissions test scores. Our study evaluates high school performance and university program details. Thus, while we do not have admissions test scores as a variable in our work because Canada does not require a secondary assessment for university admissions, this study (Galla et al., 2019) supports results from this *SEM* investigation about the predictive ability of high school performance.

Even though direct regression analysis showed a weaker effect of high school details on time to completion, the mediation model provided evidence of a strongly mediated influence on time to completion from the combination of high school and program details. This leads us to suggest that High school performance details as captured in our dataset may be insufficient in fully understanding students' on time completion of their university programs. A 2007 study (Geiser & Santelices, 2007) argued that high school performance consistently provided the highest predictive ability for four year programs, across all disciplines. It also argued that high school performances are less disadvantageous for minority and underrepresented student populations than standardized tests. This reinforces our position that high school performance details as

captured in the FoS dataset, may be inadequate in analysing its predictive abilities toward understanding student time to completion.

### 4.3.2 High School Details and Demography

For our second research question, SEM results showed significant joint influence of Demography and High school details on Time to completion. Individually, both variables showed they can explain Time to completion. This supports results from other studies (Bradley & Renzulli, 2011) (Pong & Ju, 2000) where Demography has been shown to affect the time to completion of students from various cultural backgrounds. It also supports results from studies (Galla et al., 2019) (Vulperhorst et al., 2018) where High school details influenced students' Time to completion. Some studies (Mccoy, 2005) have also shown that student's performance in High school can be understood by demographic factors. In our study however, we observe no significant influence of Demography on High school which indicates that demographic descriptors of students - age, gender and legal status - as captured in the FoS dataset, do not aid our understanding of their performance in high school. We suggest that it is possible that other descriptors, or a combination, could provide more explanation as shown in other studies (Bruno & Dženana, 2014) where descriptors such as "place of birth", "mother's education level", and "mother's profession" were all contributors to students' academic success leading to on-time completion. In some other studies, (Baker & Hawn, 2022) demography details have been shown to add bias to algorithms used in analysing student performance. This study also showed that having relevant performance data may aid the performance of algorithmic models and reduce bias introduced by demographic descriptors.This indicates that a gap exists between having insufficient demography data and having too much.

The most obvious explanation for a relationship between students' backgrounds and their High school performance will be influenced by many factors especially educational and socio-economic standards in their home countries (Battle & Lewis, 2002). Better demographic descriptors that describe the socio-economic statuses of students provide evidence that demography can influence performance in High school (Bradley & Renzulli, 2011) and,

invariably, how prepared students are for university. Collecting the right descriptors and the right amount for the FoS dataset might show a trend in student backgrounds and their level of preparedness. This will aid better understanding of their time to completion. Additionally, since more students most likely completed High school in their home countries without the influence of migration policies on their education, it follows that their identities would not affect their grades but might affect their time to completion in a university situated outside of their home countries.

### 4.3.3 High School Details and English Language

Results from this model showed that there was a significant level of influence from High school details and English language skills on Time to completion. In this model, students' preparedness as demonstrated by their high school details, provided an indication of their potential to complete their programs on time. Students who achieved success in their High school courses, were also able to successfully complete their university programs in time. In one study, components of student preparedness were distinguished to determine which factor contained more predictive ability between high school grade point average (GPA) and the scores of three core subjects (Vulperhorst et al., 2018). Results showed that achieving academic success depended on the set of credentials students were admitted with. This supports recommendations by a 2007 study (Geiser & Santelices, 2007) to place priority on high school performance scores rather than standardized tests, when attempting to understand learning outcomes. With this approach, the study posits that there will be less adverse effects on minority and underrepresented students.

Results from the *SEM* implementation of this model also revealed that students who tested well in their English language proficiency assessments also completed their programs in time. It is important to distinguish between one's English language proficiency and their expression of cognitive abilities required for them to achieve academic success and complete their programs in time. English language proficiency is not indicative of intelligence (Genesee, 1976). The theory of multiple intelligences (Gardner, 1993) outlines different types of "intelligences" adopted by students in different scenarios of learning. Other factors that may provide a better explanation according to another work (Shute et al., 2015), include persistence, student's ability to access resources, and their ability to express knowledge acquired. These factors - persistence, ability to

access resources and express knowledge - were analysed using SEM to understand learning outcomes. Thus, the influence of ELP achievements on time to completion we observed in our results when combined with this literature suggests that there are other unobserved factors that might provide a better explanation for academic on time completion. The mediation model specified provided evidence of latent variables within High school details and English language skills that may provide a better explanation of their time to completion

No relationship between ELP and High school performance was observed. This indicates that we cannot explain students' high school achievements by their English language skills. From our results, how well students perform in their high school courses cannot be determined by their level of English language proficiency. Considering that cognitive abilities are acquired via various means, there are some challenges with identifying the relationship between ELP and academic performance. It is therefore not surprising that while High school and ELP individually influence students' on-time completion of their programs, they show no effects between each other. Some studies (Graham, 1987) contend that individual institutions, and perhaps subject departments, have a minimum threshold below which inadequacy as stipulated by their standards, will affect students' academic success and on-time completion.

### 4.3.4 Demography and Program Details

In this model, student Demography and Program details jointly and individually showed strong positive influences on Time to completion. Student preparedness showed direct influence on Time to completion. This supports results from *SEM RQ1* which also showed that program details could explain Time to completion. Demography also showed direct influence on Time to completion in alignment with results obtained from *SEM RQ2*. The influence of Demography on time to completion in this investigation showed that younger students, irrespective of legal status, completed their programs faster than older students. This is a reasonable trend because older students are more likely to have other life engagements and responsibilities such as marriage, full time work and kids. These factors are more likely to cause them to take longer periods of time completing their programs.

There was no relationship between Demography and Program details however. This demonstrates that students' performance in their university courses cannot be explained by details regarding where they are from, their gender or age, as seen in other studies (Vulperhorst et al., 2018). This means that in this model, students were not more or less likely to take 12 hour or 9 hour course credits in a week based on their country of origin, age or gender. It is important to note however, that international students at the UofA, face the possibility of losing their visa statuses if they withdraw from too many courses. This is not the case for domestic students.

To get a better sense of what other factors may explain the relationship between students' demography and how well they will perform in their university course, we note the latent constructs within both predictor variables in the mediation model. This mediation model is evidence that there are other factors that may explain the relationship between student demography and their university performance. The University of Alberta international student admissions policy stipulates a complete academic history which includes international curricula (*Admission Requirements | Undergraduate Admissions & Programs*, n.d.). We posit that the international curricula received from international students are not wholly comparable to the university's standards and suggest an understanding of the other components that are focused on in these other countries.

### 4.3.5 Program Details and English Language Skills

In this investigation, we observed a significant positive influence between Program details and Time to completion. This supports our results from *SEM RQ1* where academic endeavor also provided the strongest indicator for their Time to completion. From this we observed that students with high grades and high number of courses take longer to complete their programs than students with fewer courses. This is reasonable because students with more courses tend to be double major students or students enrolled in programs that require a longer duration to complete. Additionally, ELP showed a strong positive influence on Time to completion as observed in *SEM RQ3*.

There were latent constructs in the mediation relationship among student's ELP, their Program details and their time to completion that indicate the presence of other factors which could better explain this relationship. In order to complete a university program in the stipulated timeframe, students had to successfully conclude all courses and acquire the minimum grades required. To achieve the minimum grades required for completion, a certain level of cognitive ability is required. Some studies show an influence of cognitive abilities on second language acquisition (Genesee, 1976) but cognitive abilities cannot truly be measured (Sternberg, 1996) because they cannot be said to be one thing, hence Gardner's theory of many intelligences (Gardner, 1993). It is for reasons such as this that our study focuses on understanding structural relationships amongst variables that are observed and those that are unobserved. In this way, we are given an opportunity to ask the question differently and consider other factors. These factors may provide a better understanding for evidence of latent variables within students' ELP and Program details that covary with one another but do not show an observable influence in the regression model.

In this SEM investigation we also observed no relationship between ELP and Program details. It is clear, from this model, that a student's ability to write, read and speak the English language is no indicator of how much academic endeavor they would demonstrate in their university studies. This lack of relationship may be the result of requirements for international students to maintain a certain number of courses per term in order to keep their visa status. Being able to speak the English language does not tell us how many courses students will take in their programs. Additionally, Canada's visa policies for international students stipulates that they provide English Proficiency test scores in order to obtain study permits. While many international students seek to acquire higher education from Canada, they also consider migrating and remaining in Canada after their education. As a result, English language proficiency does not explain how much academic endeavor students will exhibit in their programs.

### 4.3.6 Demography and English Language Skills

With research question 6, results from the regression model showed both ELP and student demography could influence and explain students' time to completion. Individually however, Demography showed more influence than ELP. This is consistent with results from *SEM RQ2* where Demography showed a positive influence on Time to completion. It is also consistent with

results obtained from *SEM RQ3* where ELP showed a positive influence on Time to completion. A strong positive relationship exists between ELP and Demography. This supports studies (Sharkey & Layzer, 2000) that show that academic success of second language learners are impacted by their status as second language learners. This means that a student's demographic descriptors can explain their academic performance and this will in turn, impact their time to completion. Acknowledging that cognitive abilities can be developed through many things and cannot truly be measured (Sternberg, 1996), it was important to see that in the mediation model, Demography and ELP showed significant interrelationships that suggest the presence of unobserved variables. This means that factors within students' ability to read, write and speak the English language, in combination with their nationalities, gender and age, contain other information that may provide a better explanation for the time it takes them to complete their university programs.

We contrast the results in this model between ELP and Demography to results from *SEM RQ3* between ELP and High school details. *SEM RQ3* results showed no relationship between ELP and High school details. There was a lack of variability in the ELP scores as many students obtained scores greater than 70%. This indicates that High school performance cannot be understood by student's English proficiency. Results in this *SEM RQ6* model showed strong positive associations between ELP and Demography. This is a logical association because many international students are required to provide English proficiency test scores as a requirement for admission.. All three variable categories contain features that describe the formative timespan that prepared students for university. Observing the strong positive relationship between ELP and Demography in this model confirms the validity of this hypothesis.

## 4.4 SEM Summary

Results from the SEM models showed important relationships amongst many variables: students' past achievements in high school which is a proxy for their level of preparedness for university, their ability to read, write and speak the English language, their countries of origin, gender, age, and their academic endeavor when taking university courses. These relationships were described using models that aimed at understanding the degree to which two combinations of different

variables influenced and explained student time to completion. We observed direct significant influences from all predictor variables on the target variable, in all models. We also observed the presence of latent constructs that mediated influences through one another in explaining the target variable. This implies that data capturing for these sets of variables could be improved to include other descriptors that can better explain the target variable. We recommend the inclusion of data such as socioeconomic factors, during data capturing and preprocessing. We also recommend the inclusion of socioeconomic descriptors (Vulperhorst et al., 2018) that may provide more insight into students' academic success and inform researchers on learning support recommendations.

It is clear that all variables were important in explaining student's time to completion, albeit in different models. To provide a baseline for future work, and in line with rigorous data mining, it was important for us to begin our hypotheses with two predictor variables. In this way, we ensured that we investigated simple models that provided insights at a granular level, before progressing to complex ones. Results from these investigations provided justification to include all variables for training, and simulated artificial missingness to determine if neural networks can make predictions with the available data.

# Chapter 5

# Prediction With Missingness

To determine if there was potential for our proposed approach, we conducted a classification task using features identified from our SEM implementation. This task aimed to predict the amount of time students take to complete their degrees (Time to Completion).

Before predicting student Time to Completion, we implemented SEM to obtain information about meaningful features within hypothesized models and the extent to which latent variables in these features explained the target variable (see Chapter 4). We sampled variables based on the SEM results. Features in this dataset align with the category definitions we created: High School, English Language Skills, Demography, Program Details, and Time to Completion. Within these categories, we identified features that were important for understanding the target variable. In *SEM RQ2, RQ4* and *RQ6,* we consistently identified "Legal Status" and "Gender" from the Demography category. In *SEM RQ3* and *RQ5,* we identified "ELP Test Id" and "Test Score" respectively from the ELP category. In *SEM RQ1* and *RQ3,* we consistently identified "Grading Scheme", and identified "Total Credits Earned" in *SEM RQ2,* from the High School category. In *SEM RQ4* and *RQ5,* the features "Grades" and "Credits Taken" from the Program Details category, were identified. The identified features provided us with the training data required for our prediction task. Some of the variables sampled and their composition is shown in Table 5.1.

**Table 5.1:** Sample FoS Dataset for Prediction with Missingness

| Student_Id | Gender | Legal_Status | Grades | Credits_Taken | ELP_Test_Id | Test_Score | Grading_Scheme | Total_Credits_Earned | Time_Complete |
|---|---|---|---|---|---|---|---|---|---|
| x1 | 1 | 0 | 4 | 3.0 | 0 | 0.88 | 4 | 12.00 | 0 |
| x1 | 1 | 0 | 4 | 3.0 | 0 | 0.88 | 4 | 12.00 | 0 |
| x1 | 1 | 0 | 4 | 3.0 | 0 | 0.88 | 4 | 12.00 | 0 |
| x1 | 1 | 0 | 4 | 3.0 | 0 | 0.88 | 4 | 12.00 | 0 |
| x1 | 1 | 0 | 4 | 3.0 | 0 | 0.88 | 4 | 12.00 | 0 |

This classification task was turned into a binary one where 1 to 3 years was encoded as 0 and 4 to 6 years was encoded as 1, shown as Time_Complete in Table 5.1. There were 123,522 total observations in this dataset with 458 unique student identities. Observations for students whose time to completion fall within class 0 was 73,533 and 49,989 for those whose time to completion fall within class 1. Consequently, class 0 constituted 60% of the entire dataset while class 1 constituted 40%.

We then introduced missingness and trained two neural network architectures (SmallNets and MediumNets). Our prediction task is a complex one because it involves the use of data with missingness (Twala & Cartwright, 2010). Since neural networks have demonstrated remarkable performance with complex classification tasks (Wu et al., 2009), we are motivated to employ them regardless of their "blackbox" (Féraud & Clérot, 2002) nature. Traditionally, standard imputation techniques are applied to complete the data when carrying out prediction tasks with missing data. These techniques make an assumption about the context of missingness, which introduces bias and a margin of error that cannot be quantified, (Patrician, 2002) in the data. In our study, we explored feature investigation to understand whether a neural network can recognize any underlying patterns within the data. We also applied a non-imputation technique to observe whether performance is comparable to that of imputation techniques.

# 5.1 Methods

We used three imputation techniques (ZNet, Mean, and Iterative) and one non-imputation technique - Cat. To generate a train and test split, we used the *train_test_split* function from the Sklearn library to obtain an 80% train set and a 20% test set. The train/test split was stratified on the target variable, **Y**, to ensure a balance of classes in both train and test sets.

## 5.1.1 Introducing Artificial Missingness

Missingness was added artificially within the MCAR context to simulate missingness as it is encountered in the real-world (Rubin, 1976). Other contexts of missingness also occur in the real-world. Sometimes they occur simultaneously with MCAR (James L. & Craig K., 2004), and other times, they occur more often than MCAR (Rubin, 1976). Moreover, the MCAR pattern of

missingness is typically the starting point for researchers when simulating missingness (Rubin, 1988). In this way, we reproduce the problem and apply methods that can be deployed to real-world settings of data missingness.

Missingness was added by randomly removing data points from Demography, High school, Program details, and English language skills. We evaluated multiple amounts of missingness, removing data in double increments starting from 5%. Thus, we introduced missingness starting from 5% and increased to 10%, 20%, 40%, and 80%. We started with a 5% missingness baseline as implemented in some studies (Scheffer, 2002) because it closely mirrors missingness levels in the real-world (Scheffer, 2002).

## 5.1.2 Imputation Techniques

### 5.1.2.1 ZNet - Zero Imputation

As with every model to which machine learning methods are applied, we ensured we had input variable **X** and target variable **Y** to which we could map a function for learning. We first differentiated between independent and dependent variables. We achieved this by distinguishing the target variable as a separate vector with the label **Y**. The remaining data was labeled **X**. We then employed the zero imputation strategy by creating a *zero_impute* function to replace the missing values with zeros. With zeros in places where there were previously different values, the dataset became incomplete and simulated real-world scenarios where vectors could be missing for various reasons.

### 5.1.2.2 Mean - Mean Imputation

We explored a second model we aliased as Mean. With this model, we separated input variable **X** from output variable **Y**. We normalized **X** to achieve standardization and created a function *make_missing* that randomly removed original input values and generated an incomplete dataset. With random zeros in this newly generated incomplete dataset we computed the mean of each column of **X** and imputed the resulting means in positions with missing values. By replacing missing values with the averages of each column with missingness, we derived the Mean model.

With categorical data, imputation is done using the most frequently occuring value (also called Mode) (Memon et al., 2023).

#### 5.1.2.3 Iterative - Round-Robin Imputation

The iterative algorithm we employed is from Sklearn (Pedregosa et al., 2011). It is an imputation technique that uses a multivariate mechanism for imputation. Its strategy for imputation models features with missing values as a function of other values. Imputation happens sequentially or in a round-robin fashion until all missing values are imputed. The particular approach used by this algorithm is the Bayesian Ridge regression estimation model. In this model, a regression problem is created and a probability estimation is performed. Again, we note that the iterative model focuses first on data completion before executing a prediction task.

### 5.1.3 Non-Imputation Technique

#### 5.1.3.1 Cat - Concatenation

This method is called concatenation, aliased as Cat. It is inspired by other work that focused on ignoring completion of the dataset before learning (Yoon et al., 2018). Another factor that made concatenation a suitable option is the property of non-linearity it possesses. For predictions with missing values, it is intuitive to assume the condition of non-linearity among features and values and provide statistical compensation for missingness (Le Morvan et al., 2020). The concept of non-linearity is important because there is a disruption in the linear relationships between predictor and target variables when values in a dataset are missing. This disruption in linearity introduces complexity when a function mapping is attempted in order to achieve classification in a prediction task. This level of complexity makes neural networks a suitable pathway to develop solutions to the problem of missingness (Le Morvan et al., 2020).

In implementing Cat, we began by deriving an indicator matrix that is modeled on the incomplete dataset. The indicator matrix consists of ones and zeros, where ones represent missing data inputs and zeros indicate available data.This modeling was important because we wanted performance to be conditioned on the missingness in the data. We then created a function to concatenate the now incomplete dataset with the derived indicator matrix of ones and zeros.

The purpose of this concatenation is to present the neural network with a dataset that is a combination containing indicators of missingness and available data. With this combination, the expectation is that nodes in the network layers will find mappings within the available data and indicator matrix and condition learning based on patterns in the missingness.

### 5.1.4 Experiment Setting

Our experiments were set within two neural network architectural layouts - SmallNets and MediumNets. Each architecture contained specific layer sizes. SmallNets was of size 16x16 while MediumNets was of size 64x64.

We defined activation functions for the hidden layers and output layers. *Relu* for the output layer and *sigmoid* as an activation function for the hidden layers. Training was done for 200 epochs, with learning rates of 0.001, 0.0001, and 0.00001. Two optimization techniques: Stochastic Gradient Descent (SGD) and Adam were used to optimize performance during training. Accuracy results were calculated by averaging seeds from the last 5% of epochs where the lowest losses were obtained, for all learning rates.

We employed the early stopping strategy for regularization of performance to combat overfitting. The early stopping technique involved training in batches and generating a hold-out set that was used for validation. Depending on the performance of the validation set, the training continued while improvement was observed. Training terminated when no further improvement was observed. This regularization technique cushioned the effect of overfitting.

Results from hyperparameter tuning on SmallNets produced minimized losses for different missingness levels from different learning rates. Loss values indicate how close the prediction probabilities of an observation are to the actual value. The higher the loss, the farther away the prediction probability is from the actual value. We take an average to represent this probability. We averaged loss values for each learning rate for different missingness levels. The 5% missingness averaged loss values for learning rates - 0.001, 0.0001 and 0.00001 - to produce a loss of 0.70. Missingness at 10% averaged loss values for all learning rates to produce a loss of 0.69 while missingness at 20%, 40%, and 80% produced loss values of 0.60, 0.58 and 0.50,

respectively. In the MediumNets architecture, averaged loss values for all learning rates at 5% missingness produced a loss of 0.64. Missingness at 10% produced a loss value of 0.60, while missingness levels of 20%, 40% and 80% produced loss values of 0.60, 0.58 and 0.50, respectively. These values also account for the measure of dispersion for loss values which is within the the range of 0.2, and 10 for epochs.

### 5.1.5 Performance Metrics

To assess model performance, we measured accuracy by collecting cross entropy loss values. Cross entropy measures the performance of a classification model whose output is a probability value between 0 and 1. We also employed classification metrics that include precision, recall and F1-score to represent the performance of all models. Precision calculates the number of correctly predicted classes divided by the total number of correctly predicted classes and incorrectly predicted classes. The recall metric calculates the number of true positives divided by the summation of true positives and false negatives. For recall, the focus is to assess and minimize false negatives. The F1 score weights both precision and recall equally in order to represent with a single measure the number of times a model made an accurate prediction across the entire dataset. As this was a binary classification task, we also report accuracy which divides the number of correct predictions by the total number of predictions and multiplies this value by 100 to get a percentage. We used the Scikit-Learn module which includes the accuracy_score package.

### 5.1.6 Comparing Model Performance

Our study compared the performance of imputation and non-imputation techniques using the same data. To evaluate performance across models, we employed the McNemar test. This is a statistical test used to determine whether there is a significant difference between paired data, especially for situations where the same subjects or data-points are measured or categorized twice (Dietterich, 1998). We compared the calculated McNemar statistic and p-value from contingency tables made from the predictions of each pair of classifiers, in relation to the groundtruth. When the p-value was less than 0.05 and the McNemar statistic was greater than 3.4

(Dietterich, 1998), we rejected the null hypothesis and concluded that there was a significant difference in the predictions made between both models.

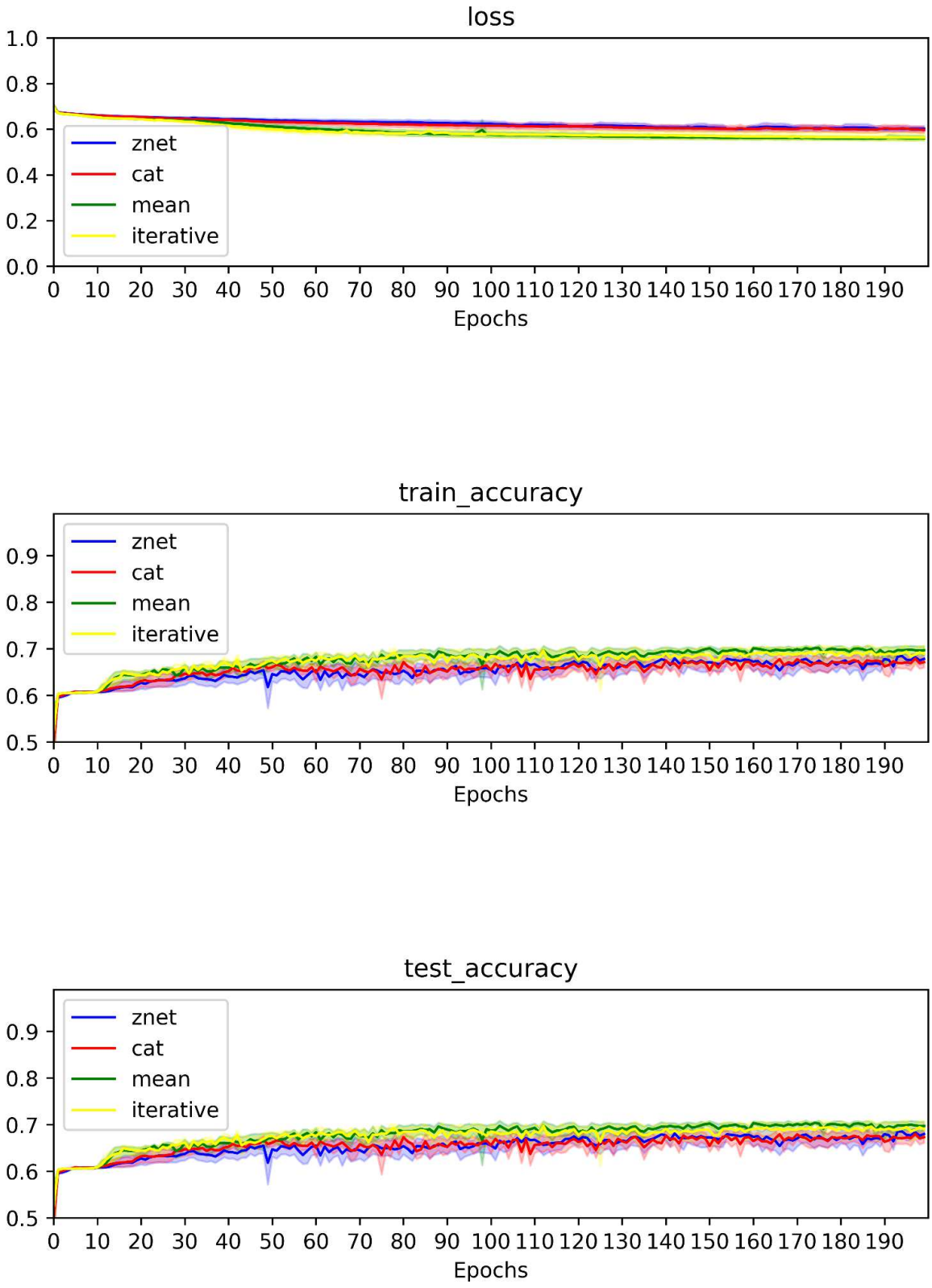
## 5.2 Results

### 5.2.1 SmallNets

Figure 5.1 shows results obtained from training the FoS data on the SmallNets architecture of 16x16 units. There are three plots which depict loss values, train and test accuracy for all models during a 200 epoch training. Loss values for all models terminated at the 200th epoch. Loss values continued to drop for all models until the last epoch. They started at a loss of 0.7 and dropped to 0.5. Train and test accuracies improved for all models, showing that the non-concatenation technique - Cat - can achieve meaningful prediction.

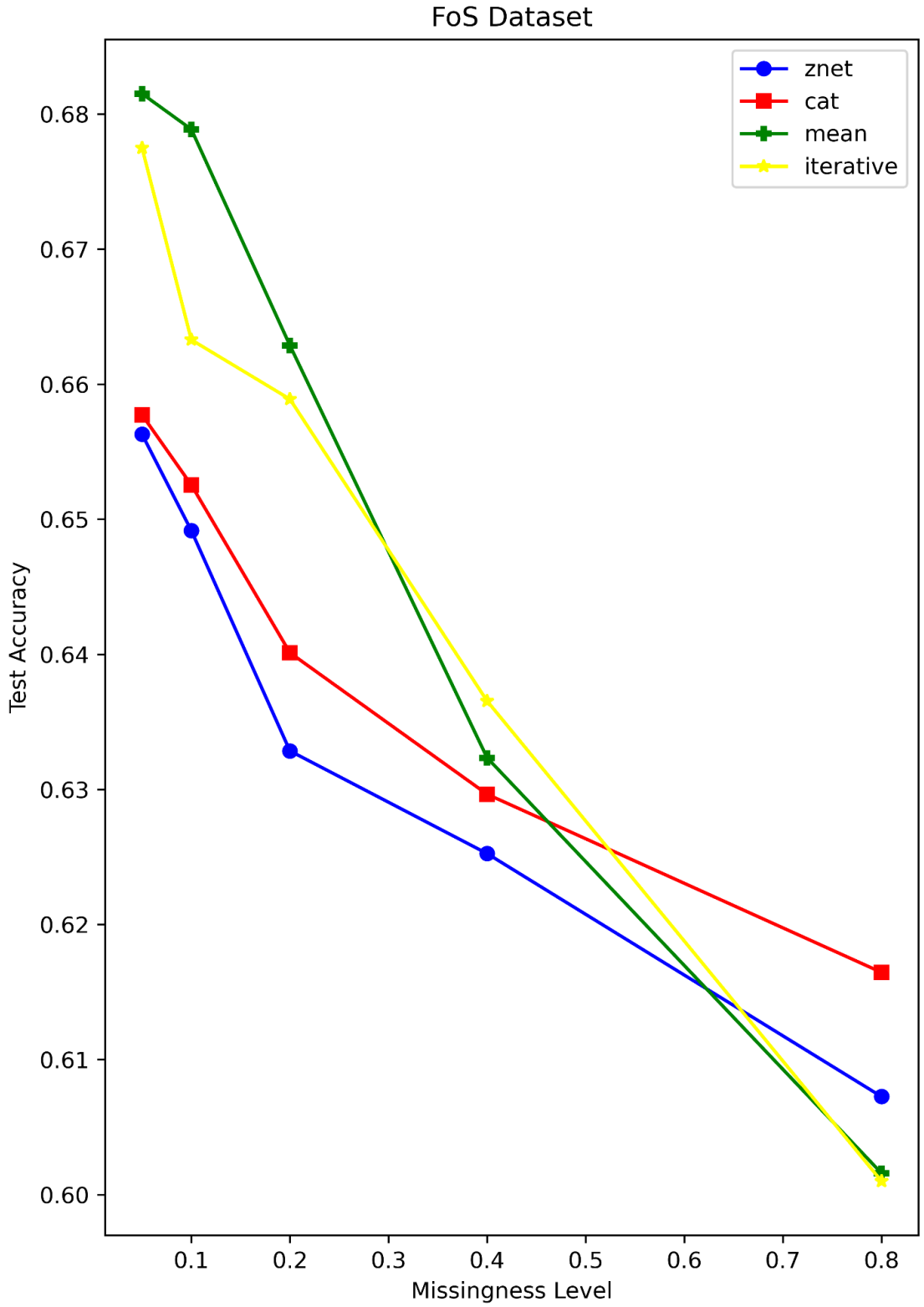
Figure 5.1 focuses on epochs and not missingness levels because we wanted to observe the computational efficiency of all models. While it is computationally efficient for a model to quit training at an early epoch rather than at a later one, if learning continues to happen, it is expected that the model should optimize for best performance. Cat achieved similar learning rates before the 200th epoch and thus yielded more value for the time it took to complete training. Imputation techniques - ZNet, Mean, and Iterative - also achieved similar accuracy levels with Cat but continued to train for longer even when performance did not improve.

**Figure 5.1**: SmallNets Loss, Train and Test Accuracies for Predicting Time to Completion



In Figure 5.2, test accuracy is plotted against missingness levels to show the pattern of learning across the various degrees of missingness. Figure 5.2 shows test accuracies attained by all models across different missingness levels. Test accuracy started at around 68% for Mean and Iterative, and at around 66% for ZNet and Cat. It dropped sharply for all models and continued to drop at this rate until missingness was at 40%. After 40% missingness, accuracy drop rate slowed down, and stopped at around 60% for Mean and Iterative when missingness was at 80%. For ZNet, accuracy dropped to around 61% and reached 62% for the non-imputation technique, Cat. This showed that Cat outperformed all imputation models, achieving the highest test accuracy even when missingness was at 80%.

**Figure 5.2**: SmallNets Test Accuracy of all Models versus Missingness level for Predicting Time to Completion



To determine prediction performance for each model, we calculated binary classification scores for precision, recall, and F1. Table 5.2 contains these scores. These scores were calculated across all missingness levels to provide a baseline. In a future work, we aim to investigate their performance for different missingness levels.

**Table 5.2:** SmallNets Classification Metrics - F1 scores, Precision and Recall - for both Classes, for all Models Across all Missingness Levels

| **Class: 0 "Completed <= 3years"** | **Model** | **Precision** | **Recall** | **F1-Score** |
|---|---|---|---|---|
| | Cat | 0.64 | 0.90 | 0.74 |
| | ZNet | 0.63 | 0.93 | 0.75 |
| | Mean | 0.62 | 0.97 | 0.76 |
| | Iterative | 0.62 | 0.98 | 0.76 |
| **Class: 1 "Completed > 3years"** | **Model** | **Precision** | **Recall** | **F1-Score** |
| | Cat | 0.67 | 0.30 | 0.41 |
| | ZNet | 0.65 | 0.18 | 0.27 |
| | Mean | 0.88 | 0.14 | 0.21 |
| | Iterative | 0.90 | 0.14 | 0.30 |

To determine whether there were statistical differences between model performances in the SmallNets architecture, we conducted a McNemar test for each missingness level. Table 5.3 contains the results. Statistical differences were observed in 10% and 80% missingness levels between the predictions of Cat and Mean, Cat and Iterative. This shows that the test accuracy achieved by Cat was significantly better than those achieved by Mean and Iterative when the data contained 10% and 80% missingness. Statistical differences also existed between the test accuracies of Cat and ZNet when the missingness level was at 80%. This also shows that Cat significantly outperformed ZNet. Between ZNet and Mean, and ZNet and Iterative, we also observed statistical differences when missingness was at 80%. This shows that ZNet outperformed both Mean and Iterative.

**Table 5.3:** McNemar Statistical Test Results between Models in SmallNets, for Each Missingness Level

| Models | % Missingness | McNemar's Chi-square (1.0) | p | Cramer's phi |
|---|---|---|---|---|
| Cat vs Mean | 5% | 1.800 | 0.179 | 0.244 |
| | **10%** | **7.000** | **0.008** | **0.500** |
| | 20% | 1.285 | 0.256 | 0.207 |
| | 40% | 1.285 | 0.256 | 0.207 |
| | **80%** | **10.000** | **0.001** | **0.000** |
| Cat vs ZNet | 5% | 0.142 | 0.705 | 0.071 |
| | 10% | 3.000 | 0.083 | 0.327 |
| | 20% | 0.333 | 0.563 | 0.105 |
| | 40% | 1.000 | 0.317 | 0.182 |
| | **80%** | **3.600** | **0.047** | **0.346** |
| Cat vs Iterative | 5% | 0.333 | 0.563 | 0.105 |
| | **10%** | **3.571** | **0.048** | **0.357** |
| | 20% | 0.142 | 0.705 | 0.069 |
| | 40% | 0.400 | 0.527 | 0.115 |
| | **80%** | **10.000** | **0.001** | **0.000** |
| Mean vs Iterative | 5% | 1.000 | 0.317 | 0.182 |
| | 10% | 0.666 | 0.414 | 0.149 |
| | 20% | 1.600 | 0.205 | 0.230 |
| | 40% | 0.333 | 0.563 | 0.105 |
| | 80% | 0.006 | 0.519 | 0.102 |
| ZNet vs Mean | 5% | 0.666 | 0.414 | 0.154 |
| | 10% | 1.285 | 0.256 | 0.207 |
| | 20% | 2.000 | 0.157 | 0.258 |
| | 40% | 0.666 | 0.414 | 0.149 |
| | **80%** | **4.000** | **0.045** | **0.000** |
| ZNet vs | 5% | 0.000 | 1.000 | 0.000 |

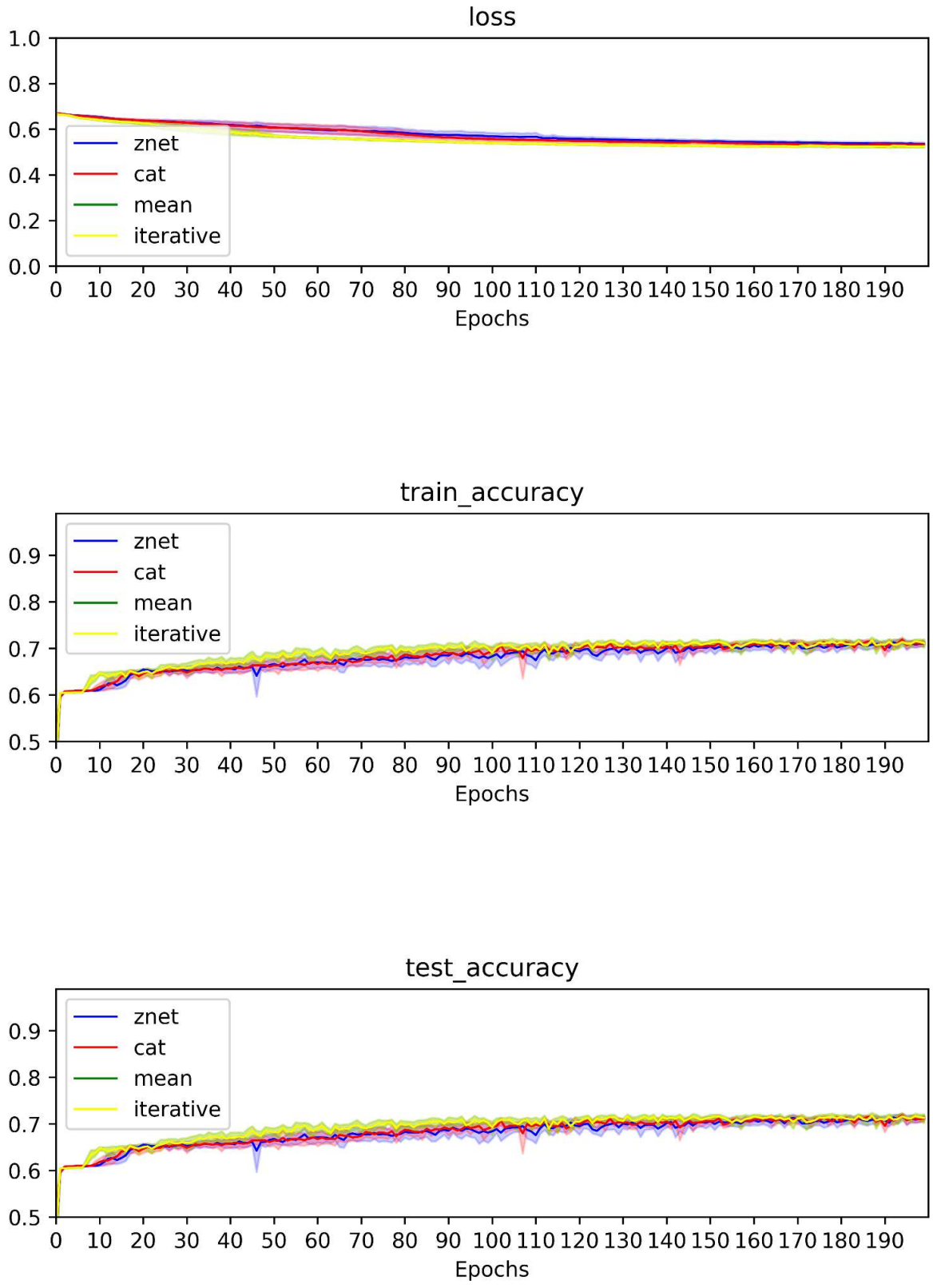
| | | | | |
|---|---|---|---|---|
| Iterative | 10% | 0.200 | 0.654 | 0.081 |
| | 20% | 0.000 | 1.000 | 0.000 |
| | 40% | 0.111 | 0.738 | 0.060 |
| | **80%** | **4.000** | **0.045** | **0.000** |

### 5.2.2 MediumNets

Figure 5.3 shows results obtained on the MediumNets architecture of 64x64 units. There are three plots that illustrate loss values, train and test accuracy for all models during a 200 epoch training. Loss values drop at similar rates for all techniques until the last epoch. Loss values started at 0.63 and dropped to 0.5. Training for all models terminated at the 200th epoch. Train and test improved for all models, showing that the non-concatenation technique - Cat - can achieve meaningful prediction.

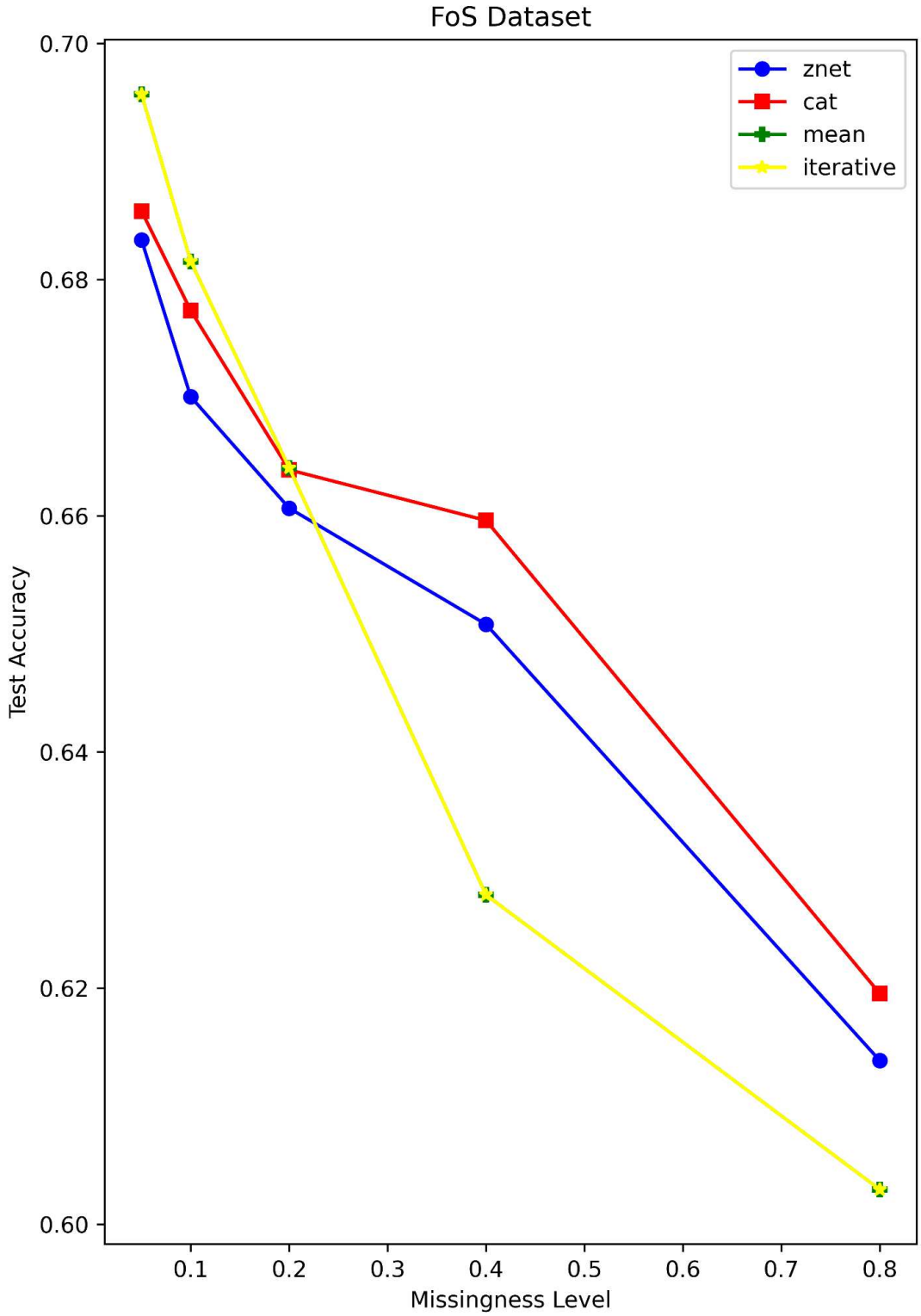
Figure 5.3 also addresses computational efficiency by focusing on epochs and not missingness levels. All models continued to train and achieve steady incremental accuracies till the last epoch. This suggests that with a longer training epoch (greater than 200) and larger nodes in a neural network, all models can achieve more accuracies and attain more computational efficiency.

**Figure 5.3**: MediumNets Loss, Train and Test Accuracies for Predicting Time to Completion



In Figure 5.4, test accuracy was plotted against missingness from 5% to 80%. This plot shows the pattern of learning from all algorithms in a neural network architecture of 64x64. We observed that test accuracies for all techniques started at around 69% for Mean and Iterative but started to drop immediately afterwards. Test accuracies for ZNet and Cat started at around 68% and also dropped sharply afterwards. As missingness increased, accuracy dropped even faster for all techniques. The non-imputation technique - Cat - outperformed all other techniques, achieving 62% accuracy even with 80% missingness. All imputation techniques achieved accuracy levels of less than 62% with 80% missingness. In this architecture, just as with SmallNets, we saw learning curves drop as missingness levels increased. The non imputation technique retained the highest accuracy level once the missingness rate was greater than or equal to 20%.

**Figure 5.4**: MediumNets Test Accuracy of all Models, versus Missingness level for Predicting Time to Completion



Classification metrics were calculated to determine the prediction performance of each model across all missingness levels. Table 5.4 contains the results of this calculation. These scores were calculated across all missingness levels to provide a baseline.

**Table 5.4:** MediumNets Classification Metrics - F1 scores, Precision and Recall - for both Classes, for all models Across all Missingness Levels

| **Class: 0 "Completed <= 3years"** | **Model** | **Precision** | **Recall** | **F1-Score** |
|---|---|---|---|---|
| | Cat | 0.64 | 0.95 | 0.76 |
| | ZNet | 0.64 | 0.94 | 0.76 |
| | Mean | 0.63 | 0.96 | 0.76 |
| | Iterative | 0.63 | 0.96 | .076 |
| **Class: 1 "Completed > 3years"** | **Model** | **Precision** | **Recall** | **F1-Score** |
| | Cat | 0.75 | 0.22 | 0.34 |
| | ZNet | 0.74 | 0.22 | 0.33 |
| | Mean | 0.86 | 0.18 | 0.27 |
| | Iterative | 0.86 | 0.18 | 0.27 |

To determine whether there were statistical differences between model performances in the MediumNets architecture, we conducted a McNemar test. Table 5.5 contains the results. Statistical differences were found between Cat and Mean, and Cat and Iterative when missingness levels were at 40% and 80%. This shows that Cat outperformed Mean and Iterative when the data had 40% and 80% missingness. Statistical differences were also found between ZNet and Mean, and ZNet and Iterative when missingness levels were at 40% and 80%. No statistical differences were found between Cat and ZNet, and between Mean and Iterative.

**Table 5.5:** McNemar Statistical Test Results between Models in MediumNets, for Each Missingness Levels

| Models | % Missingness | McNemar's Chi-square (1.0) | p | Cramer's phi |
|---|---|---|---|---|
| Cat vs Mean | 5% | 0.200 | 0.654 | 0.084 |
| | 10% | 0.000 | 1.000 | 0.000 |
| | 20% | 0.200 | 0.654 | 0.081 |
| | **40%** | **4.500** | **0.033** | **0.387** |
| | **80%** | **9.000** | **0.002** | **0.000** |
| Cat vs ZNet | 5% | 0.000 | 1.000 | 0.000 |
| | 10% | 0.000 | 1.000 | 0.000 |
| | 20% | 0.000 | 1.000 | 0.000 |
| | 40% | 1.000 | 0.317 | 0.182 |
| | 80% | 2.000 | 0.157 | 0.258 |
| Cat vs Iterative | 5% | 0.200 | 0.654 | 0.084 |
| | 10% | 0.000 | 1.000 | 0.000 |
| | 20% | 0.142 | 0.705 | 0.069 |
| | **40%** | **4.500** | **0.033** | **0.387** |
| | **80%** | **9.000** | **0.002** | **0.000** |
| Mean vs Iterative | 5% | 0.000 | 1.000 | 0.000 |
| | 10% | 0.000 | 1.000 | 0.000 |
| | 20% | 0.000 | 1.000 | 0.000 |
| | 40% | 0.000 | 1.000 | 0.000 |
| | 80% | 0.000 | 1.000 | 0.000 |
| ZNet vs Mean | 5% | 0.200 | 0.654 | 0.084 |
| | 10% | 0.000 | 1.000 | 0.000 |
| | 20% | 0.200 | 0.654 | 0.081 |
| | **40%** | **3.571** | **0.048** | **0.345** |
| | **80%** | **5.000** | **0.025** | **0.000** |
| ZNet vs | 5% | 0.200 | 0.654 | 0.081 |

| | | | | |
|---|---|---|---|---|
| Iterative | 10% | 0.000 | 1.000 | 0.000 |
| | 20% | 0.333 | 0.563 | 0.109 |
| | **40%** | **3.571** | **0.048** | **0.345** |
| | **80%** | **5.000** | **0.025** | **0.000** |

# Chapter 6

# Discussion

Results from SmallNets and MediumNets showed that the non-imputation technique - Cat - achieved the highest test accuracy when the highest level of missingness was present- 80%. The McNemar statistical tests conducted for SmallNets confirmed statistical differences when data missingness was at 10% and 80% between Cat and Mean, and Cat and Iterative. This informs us that even with 80% data missingness, a neural network can recognize patterns and learn from the available data. Cat did not outperform ZNet at 10% however. This suggests that the ZNet imputation technique can achieve comparable performance with Cat in a small network architecture. Within this architecture, ZNet also outperformed Mean and Iterative when data missingness was at 80%. In the MediumNets architecture, there were significant differences between Cat, Mean and Iterative when missingness was at 40% and 80%. ZNet also outperformed Mean and Iterative when data missingness was at 40% and 80%. This suggests that the amount of missingness and the technique employed in prediction tasks play a role in the performance of the models. These results indicate that a larger network with more layers and units can enhance performance in non imputation techniques. Similar performances were observed for Cat and ZNet which is not surprising because missingness at random (MCAR) for both techniques will behave similarly since zeros are both involved. This also provides some evidence that expert layers can be formed within layers of a neural network to learn missingness patterns and maximize information from available data. The motivation for this theory is informed by the fact that when handling missingness, the concept of non-linearity becomes a focal point (Yoon et al., 2018). To determine patterns and distinguish missing vectors from available ones, a neural network can learn a function mapping. Function mappings can be linear and non-linear. The latter is the case when missingness exists in the data. It is, therefore, our position that some level of expert networks can be built within the layers of the network to find information. An expanded view of this scenario could lead to layers within the network also learning different contexts of missingness. This could enable researchers to derive guidelines for implementing prediction techniques that are optimal, for data missingness. Since the internal structures of neural networks are still unknown (and are referred to as blackbox models), this

would be an ambitious objective but could prove useful in understanding non-linearity and missingness, through neural networks.

In our study, we attempt to find information from our results that can enable the understanding of missingness at various levels. Some literature (Scheffer, 2002) suggests that missingness levels at 5% and 10%, can be ignored under certain conditions, as they may not potend adverse effects when handling missingness. This suggests that complete case analysis can be applied when missingness levels are at 5% and 10% with few consequences. From this, we understand that the literature has attempted to find an acceptable threshold for missingness on which to safely ignore the handling of missingness by deleting observations with missing values. It is important to note that the conditions stated for ignoring missingness are rarely feasible in reality. In both SmallNets and MediumNets, test accuracies dropped at a faster rate for missingness levels up to 40%, than missing levels after 40%. This might indicate that function mappings within the neural network attain stable prediction capacity after 40%. A 2019 study (Choudhury & Pal, 2019) implemented imputation using missing data and recorded significant underperformance of their proposed model when missingness was less than 50%. In another study, (Śmieja et al., 2018) performance increased when missingness exceeded 25%. It is therefore our position that developing a neural network architecture that is able to adapt to varying missingness levels, will provide the foundation for making reasonable predictions with missing data. Training with larger networks and more learning rates for longer epochs may provide substance to this theory

The performance of non-imputation technique - Cat - provides grounds for research with missing data to deprioritize data completion when implementing prediction tasks. Imputations such as Mean and zero (ZNet) implement imputation by replacing missing values with the mean and zeros, respectively. With Iterative imputations, a sequential replacement is done using values that are a function of available data. When data is categorical, one-hot encoding is first implemented before replacement is done. In this scenario however, with one-hot representation, when data is missing, that representation is treated as a NaN. In this way, encoded features are handled as independent features for an Iterative imputation. Regardless of the imputation technique used, data completion poses the danger of introducing more bias to an already biased dataset. Our study aspires to motivate research in the field to deemphasize data completion.

An overview of the model performances shows that non-imputation technique Cat is comparable to all imputation techniques. This notwithstanding, the imputation technique ZNet succeeded Cat, performing better than Iterative and Mean imputation on both neural network architectures. ZNet, which is a zero-imputation technique, showed improvement with MediumNets. Its loss values closely followed that of Cat, ending at a training epoch just before 200. Compared to performance in SmallNets, the MediumNets architecture provided more layers and units for pattern recognition. It is a positive sign therefore, that with an imputation technique like ZNet, pattern recognition improved in the MediumNets to achieve more test accuracy when predicting student time to completion. It is worthy to note that typical imputation techniques fill in missing values with estimates but with ZNet's zero-imputation, we filled missing vectors with zeros. Improved performance observed by ZNet confirms the MCAR context of missingness created, because the expectation was that mean imputation (Mean) would perform better than ZNet. Cat was also less likely to perform better than ZNet under MCAR, since there is no structural pattern to learn. A future research endeavor will be to try larger neural network architectures, more tuning and other non-imputation techniques.

## 6.1 Ethics

To ensure that we conducted ethical research and adhered to principles that protect the dignity, rights and welfare of research participants, we carried out a number of measures. It was advantageous that our data was archival data and thus, we did not need to recruit participants into our study. This archival data was anonymised by de-identifying terms within the data. In this way, we were able to protect student identity. We obtained proper consents to interact with the data which was password protected. We also signed appropriate Non-Disclosure Agreements (NDAs) and did not share the data with unauthorised persons digitally, or otherwise.

## 6.2 Limitations

### 6.2.1 SEM

To accommodate future work based on ours, it is important that we outline the limitations that exist in this study. We made assumptions when specifying our SEM models. These assumptions include model accuracy which refers to a presumption that the research questions we hypothesized were valid and without error. Another assumption is that our research questions may have been too narrow and should have involved more than two predictor variables in each hypothesis. This notwithstanding, our study provides basis to consider the formulation of hypotheses that aim to investigate missing data features for the purpose of making predictions.

To implement SEM, data has to be large enough to obtain reliable results. The FoS data contained 123,522 samples from 458 unique students. Notwithstanding, research within the educational domain that can provide enough substantiation for policy formulation requires sample sizes that are larger (Williamson, 2017). Our study and its sample size provides a starting point to begin to make useful feature selections that can aid prediction tasks toward the provision of educational support. Even though we employed statistical software, the implementation of SEM involves complex mathematical models that can be difficult to construct and interpret. Often times, model fits are not obtained to measure the overall fit of the model to the data (Yuan & Bentler, 2007) but provide useful path diagrams to understand direct and indirect structural relationships. In this way, our study illustrates how to implement and interpret SEM, and identify useful parameters, when investigating features for prediction tasks.

### 6.2.2 Prediction with Missingness

The task of predicting time to completion using the FoS dataset was straightforward because the dataset consisted of 123,522 total observations and a 60/40 split between both classes being predicted. While a common limitation of supervised machine learning is lack of data (Jiang et al., 2020), the addition of missingness raises the level of difficulty of this task and potentially introduces an amount of bias that we did not estimate. Regardless, our study is relevant because

of the prior feature investigation we performed to understand structural relationships. In this way, even with missingness, the prediction abilities of the models are empowered because meaningful features are being utilized.

The network architectures we employed to train various models also contained some constraints; the small network was of size 16x16 and medium network was of size 64x64. Results obtained from the small network showed no statistical difference between pairs of models but the medium network contained models with statistically different predictions. Thus, our study provides basis to explore network architectures of larger sizes to obtain better performance in prediction tasks

# Chapter 7

# Conclusion

In many sectors, data missingness is inevitable, making it a problem with solution methods that can compel positive differences especially within education. Applying standard methods that focus on imputation causes bias and loss of representation in the data. Many principled approaches to these imputation techniques exist but in many cases, introduce more uncertainty into the data. With missing input vectors in a dataset, statistical analysis is impaired and performing prediction tasks with machine learning methods becomes challenging.

In this study, we investigated the importance of features and their ability to influence performance when machine learning methods using neural networks were applied to data with missingness. The features in our dataset comprised variables within five categories: Demography, Program Details, High school details, English Language Skills, and Time to Completion. All categories contained multiple variables that were investigated to determine how important each one was to the target variable Time to completion. We explored the possibility of achieving optimal performance when classifying students time to completion, using a dataset from the University of Alberta's Faculty of Science.

Results from the SEM feature investigation showed that all category variables were important towards predicting time to completion. Some features were more important than others, within certain models. As a result, we were able to identify these features from our implementation of SEM, and utilize them for the prediction task. Every SEM investigation showed significant presence of latent constructs. Further investigation with more complex models, comprising more than two predictor variables, will benefit prediction tasks with missingness and provide a better understanding of the hierarchy of importance. Within a hierarchy, an ablation study may then show which combination of variables can provide the most optimal performance.

Our structural equation modeling aimed to determine which features were important within the simple models we hypothesized. From this, we obtained a dataset with these identified features

and introduced missingness such that if values within these important features were missing, we could still achieve meaningful classification. Artificial missingness was added under MCAR to present incompleteness.
This incomplete data was then trained using two neural network architectures: SmallNets with 16x16 units and MediumNets with 64x64 units. Neural network training results showed that the non-imputation technique Cat outperformed all imputation techniques ZNet, Iterative, and Mean. Of all imputation techniques, ZNet showed the highest performance. With larger neural network architectures, investigating other non-imputation strategies could prove beneficial in handling missing data, without first completing the data.

With educational data, such as the University of Alberta's Faculty of Science data, missingness can occur for valid reasons. While investigating features may not diagnose causal factors, it can provide insights into learning patterns, because it enables the identification of features with structural relationships that can optimize prediction tasks. In this way, even when data incompleteness results from students exercising their rights not to respond to survey questions, learning support can still be inclusive by catering to student diversity, and reduce bias.

It is our position that identifying features that enable pattern recognition and function mapping will empower research to approach the handling of missing data differently from the way it is currently being addressed in the field. Additionally, observing the performance of different neural network architectures provides an opportunity to build expert layers within the network that may signal to various nodes how to learn a function to maximize available data for prediction.

# Bibliography

Abdella, M., & Marwala, T. (2005). The use of genetic algorithms and neural networks to approximate missing data in database. *IEEE 3rd International Conference on Computational Cybernetics, 2005. ICCC 2005.*, 207–212. https://doi.org/10.1109/ICCCYB.2005.1511574

*Admission Requirements | Undergraduate Admissions & Programs*. (n.d.). Retrieved December 13, 2023, from https://www.ualberta.ca/admissions/international/admission/admission-requirements/index.html?

Aiken, P. C., Patricia Cohen, Stephen G. West, Leona S. (2014). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (2nd ed.). Psychology Press. https://doi.org/10.4324/9781410606266

Baker, R. S., & Hawn, A. (2022). Algorithmic Bias in Education. *International Journal of Artificial Intelligence in Education*, *32*(4), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Barnes, S. A., Mallinckrodt, C. H., Lindborg, S. R., & Carter, M. K. (2008). The impact of missing data and how it is handled on the rate of false-positive results in drug development. *Pharmaceutical Statistics*, *7*(3), 215–225. https://doi.org/10.1002/pst.310

Battle, J., & Lewis, M. (2002). The Increasing Significance of Class: The Relative Effects of Race and Socioeconomic Status on Academic Achievement. *Journal of Poverty*, *6*(2), 21–35. https://doi.org/10.1300/J134v06n02_02

Bowen, N. K., & Guo, S. (2011). *Structural Equation Modeling*. Oxford University Press.

Bradley, C. L., & Renzulli, L. A. (2011). The Complexity of Non-Completion: Being Pushed or Pulled to Drop Out of High School. *Social Forces*, *90*(2), 521–545.

https://doi.org/10.1093/sf/sor003

Bruno, T., & Dženana, Đ. (2014). *Determining the impact of demographic features in predicting student success in Croatia*. https://ieeexplore.ieee.org/abstract/document/6859754/

Buhi, E. R., Goodson, P., & Neilands, T. B. (2008). Out of Sight, Not Out of Mind: Strategies for Handling Missing Data. *American Journal of Health Behavior*, *32*(1), 83–92. https://doi.org/10.5993/AJHB.32.1.8

Buuren, S. van, & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*, 1–67. https://doi.org/10.18637/jss.v045.i03

Choudhury, S. J., & Pal, N. R. (2019). Imputation of missing data with neural networks for classification. *Knowledge-Based Systems*, *182*, 104838. https://doi.org/10.1016/j.knosys.2019.07.009

*Compare TOEFL iBT Scores*. (n.d.). Retrieved October 12, 2023, from https://www.ets.org/toefl/score-users/ibt/compare-scores.html

Cox, B. E., McIntosh, K., Reason, R. D., & Terenzini, P. T. (2014). Working with Missing Data in Higher Education Research: A Primer and Real-World Example. *The Review of Higher Education*, *37*(3), 377–402. https://doi.org/10.1353/rhe.2014.0026

Dietterich, T. G. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, *10*(7), 1895–1923. https://doi.org/10.1162/089976698300017197

Dong, Y., & Peng, C.-Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, *2*, 222. https://doi.org/10.1186/2193-1801-2-222

El-Den, S., Schneider, C., Mirzaei, A., & Carter, S. (2020). How to measure a latent construct:

Psychometric principles for the development and validation of measurement instruments. *International Journal of Pharmacy Practice*, *28*(4), 326–336. https://doi.org/10.1111/ijpp.12600

Enders, C. K. (2010). *Applied Missing Data Analysis*. Guilford Press.

F. Hair Jr, J., Sarstedt, M., Hopkins, L., & G. Kuppelwieser, V. (2014). Partial least squares structural equation modeling (PLS-SEM): An emerging tool in business research. *European Business Review*, *26*(2), 106–121. https://doi.org/10.1108/EBR-10-2013-0128

Féraud, R., & Clérot, F. (2002). A methodology to explain neural network classification. *Neural Networks*, *15*(2), 237–246. https://doi.org/10.1016/S0893-6080(01)00127-7

Galla, B. M., Shulman, E. P., Plummer, B. D., Gardner, M., Hutt, S. J., Goyer, J. P., D'Mello, S. K., Finn, A. S., & Duckworth, A. L. (2019). Why High School Grades Are Better Predictors of On-Time College Graduation Than Are Admissions Test Scores: The Roles of Self-Regulation and Cognitive Ability. *American Educational Research Journal*, *56*(6), 2077–2115. https://doi.org/10.3102/0002831219843292

García-Laencina, P. J., Sancho-Gómez, J.-L., & Figueiras-Vidal, A. R. (2010). Pattern classification with missing data: A review. *Neural Computing and Applications*, *19*(2), 263–282. https://doi.org/10.1007/s00521-009-0295-6

Gardner, H. (1993). *Multiple intelligences: The theory in practice* (pp. xvi, 304). Basic Books/Hachette Book Group.

Geiser, S., & Santelices, M. V. (2007). Validity of High-School Grades in Predicting Student Success beyond the Freshman Year: High-School Record vs. Standardized Tests as Indicators of Four-Year College Outcomes. Research & Occasional Paper Series: CSHE.6.07. In *Center for Studies in Higher Education*. Center for Studies in Higher Education. https://eric.ed.gov/?id=ED502858

Genesee, F. (1976). The Role of InTELligence in Second Language Learning1. *Language Learning*, *26*(2), 267–280. https://doi.org/10.1111/j.1467-1770.1976.tb00277.x

*Grading System Explained*. (n.d.). Retrieved September 2, 2023, from https://www.ualberta.ca/registrar/examinations/assessment-and-grading/grading-system-explained

Graham, J. G. (1987). English Language Proficiency and the Prediction of Academic Success. *TESOL Quarterly*, *21*(3), 505–521. https://doi.org/10.2307/3586500

Hwang, T. (2018). Computational Power and the Social Impact of Artificial Intelligence. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3147971

James L., P., & Craig K., E. (2004). *Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement*. https://journals.sagepub.com/doi/abs/10.3102/00346543074004525?casa_token=sTrIRpWO_UcAAAAA:brFvVnZEYErUpUmtZ2hHpEB-0pj9Lf49b0QmhN_LaAl4Y69hT4kW7q7uMQzsaapBYtnaFvHwOzJDKOk

Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020). Supervised Machine Learning: A Brief Primer. *Behavior Therapy*, *51*(5), 675–687. https://doi.org/10.1016/j.beth.2020.05.002

Jöreskog, K. G. (1994). Structural equation modeling with ordinal variables. In *Multivariate analysis and its applications* (Vol. 24, pp. 297–311). Institute of Mathematical Statistics. https://doi.org/10.1214/lnms/1215463803

Le Morvan, M., Josse, J., Moreau, T., Scornet, E., & Varoquaux, G. (2020). NeuMiss networks: Differentiable programming for supervised learning with missing values. *Advances in Neural Information Processing Systems*, *33*, 5980–5990. https://proceedings.neurips.cc/paper/2020/hash/42ae1544956fbe6e09242e6cd752444c-Abstract.html

Leach, M., & Hadi, S. M. (2017). Supporting, categorising and visualising diverse learner behaviour on MOOCs with modular design and micro-learning. *Journal of Computing in Higher Education*, *29*(1), 147–159. https://doi.org/10.1007/s12528-016-9129-6

Lei, P.-W., & Wu, Q. (2007). Introduction to Structural Equation Modeling: Issues and Practical Considerations. *Educational Measurement: Issues and Practice*, *26*(3), 33–43. https://doi.org/10.1111/j.1745-3992.2007.00099.x

Li, W., Sun, K., Schaub, F., & Brooks, C. (2022). Disparities in Students' Propensity to Consent to Learning Analytics. *International Journal of Artificial Intelligence in Education*, *32*(3), 564–608. https://doi.org/10.1007/s40593-021-00254-2

Li, Y. (2019). *MOOCs in Higher Education: Opportunities and Challenges*. 48–55. https://doi.org/10.2991/ichssr-19.2019.10

Loehlin, J. C. (2003). *Latent Variable Models: An Introduction to Factor, Path, and Structural Equation Analysis* (4th ed.). Psychology Press. https://doi.org/10.4324/9781410609823

Markey, M. K., Tourassi, G. D., Margolis, M., & DeLong, D. M. (2006). Impact of missing data in evaluating artificial neural networks trained on complete data. *Computers in Biology and Medicine*, *36*(5), 516–525. https://doi.org/10.1016/j.compbiomed.2005.02.001

Maydeu-Olivares, A. (2009). *The SAGE Handbook of Quantitative Methods in Psychology*. 1–800. https://www.torrossa.com/en/resources/an/4913732

Mccoy, L. P. (2005). Effect of Demographic and Personal Variables on Achievement in Eighth-Grade Algebra. *The Journal of Educational Research*, *98*(3), 131–135. https://doi.org/10.3200/JOER.98.3.131-135

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing Data: A*

*Gentle Introduction*. Guilford Press.

Memon, S. MZ., Wamala, R., & Kabano, I. H. (2023). A comparison of imputation methods for categorical data. *Informatics in Medicine Unlocked*, *42*, 101382. https://doi.org/10.1016/j.imu.2023.101382

Negi, P., Marcus, R., Mao, H., Tatbul, N., Kraska, T., & Alizadeh, M. (2020). Cost-Guided Cardinality Estimation: Focus Where it Matters. *2020 IEEE 36th International Conference on Data Engineering Workshops (ICDEW)*, 154–157. https://doi.org/10.1109/ICDEW49219.2020.00034

Nelwamondo, F. V., Golding, D., & Marwala, T. (2013a). A dynamic programming approach to missing data estimation using neural networks. *Information Sciences*, *237*, 49–58. https://doi.org/10.1016/j.ins.2009.10.008

Nelwamondo, F. V., Golding, D., & Marwala, T. (2013b). A dynamic programming approach to missing data estimation using neural networks. *Information Sciences*, *237*, 49–58. https://doi.org/10.1016/j.ins.2009.10.008

Nelwamondo, F. V., Mohamed, S., & Marwala, T. (2007). Missing data: A comparison of neural network and expectation maximization techniques. *Current Science*, *93*(11), 1514–1521. https://www.jstor.org/stable/24099079

*NeurIPS Education Challenge*. (n.d.). Retrieved July 27, 2023, from https://eedi.com/projects/neurips-education-challenge

Papageorgiou, G., Grant, S. W., Takkenberg, J. J. M., & Mokhles, M. M. (2018). Statistical primer: How to deal with missing data in scientific research?†. *Interactive CardioVascular and Thoracic Surgery*, *27*(2), 153–158. https://doi.org/10.1093/icvts/ivy102

Patrician, P. A. (2002). Multiple imputation for missing data†‡. *Research in Nursing & Health*,

*25*(1), 76–84. https://doi.org/10.1002/nur.10015

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in {P}ython}. *Journal of Machine Learning Research*, *12*, 2825--2830. https://scikit-learn/stable/modules/generated/sklearn.impute.IterativeImputer.html

Pong, S.-L., & Ju, D.-B. (2000). The Effects of Change in Family Structure and Income on Dropping Out of Middle and High School. *Journal of Family Issues*, *21*(2), 147–169. https://doi.org/10.1177/019251300021002001

Rhemtulla, M., & Hancock, G. R. (2016). Planned Missing Data Designs in Educational Psychology Research. *Educational Psychologist*, *51*(3–4), 305–316. https://doi.org/10.1080/00461520.2016.1208094

Romiszowski, A. J. (2013). Topics for Debate: What's Really New about MOOCs? *Educational Technology*, *53*(4), 48–51.

Rosé, C. P., McLaughlin, E. A., Liu, R., & Koedinger, K. R. (2019). Explanatory learner models: Why machine learning (alone) is not the answer. *British Journal of Educational Technology*, *50*(6), 2943–2958. https://doi.org/10.1111/bjet.12858

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1988). An Overview of Multiple Imputation. *In Proceedings of the Survey Research Section, American Statistical Association*, 79–84.

Salmanpour, M. R., Shamsaei, M., & Rahmim, A. (2021). Feature selection and machine learning methods for optimal identification and prediction of subtypes in Parkinson's disease.

*Computer Methods and Programs in Biomedicine*, *206*, 106131. https://doi.org/10.1016/j.cmpb.2021.106131

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147

Scheffer, J. (2002). *Dealing with missing data*. https://mro.massey.ac.nz/handle/10179/4355

Sharkey, J., & Layzer, C. (2000). Whose Definition of Success? Identifying Factors That Affect English Language Learners' Access to Academic Success and Resources. *TESOL Quarterly*, *34*(2), 352–368. https://doi.org/10.2307/3587961

Śmieja, M., Struski, Ł., Tabor, J., Zieliński, B., & Spurek, P. (2018). Processing of missing data by neural networks. *Advances in Neural Information Processing Systems*, *31*. https://proceedings.neurips.cc/paper/2018/hash/411ae1bf081d1674ca6091f8c59a266f-Abstract.html

Stavseth, M. R., Clausen, T., & Røislien, J. (2019). How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data. *SAGE Open Medicine*, *7*, 2050312118822912. https://doi.org/10.1177/2050312118822912

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences*, *42*(5), 893–898. https://doi.org/10.1016/j.paid.2006.09.017

Sternberg, R. J. (1996). Myths, Countermyths, and Truths About Intelligence. *Educational Researcher*, *25*(2), 11–16. https://doi.org/10.3102/0013189X025002011

Thompson, B. B., Marks, R. J., & El-Sharkawi, M. A. (2003). On the contractive nature of autoencoders: Application to missing sensor restoration. *Proceedings of the International Joint Conference on Neural Networks, 2003.*, *4*, 3011–3016 vol.4.

https://doi.org/10.1109/IJCNN.2003.1224051

*Total data volume worldwide 2010-2025*. (n.d.). Statista. Retrieved June 8, 2022, from https://www.statista.com/statistics/871513/worldwide-data-created/

Twala, B., & Cartwright, M. (2010). Ensemble missing data techniques for software effort prediction. *Intelligent Data Analysis*, *14*(3), 299–331. https://doi.org/10.3233/IDA-2010-0423

van de Oudeweetering, K., & Agirdag, O. (2018). Demographic data of MOOC learners: Can alternative survey deliveries improve current understandings? *Computers & Education*, *122*, 169–178. https://doi.org/10.1016/j.compedu.2018.03.017

van Ginkel, J. R., Linting, M., Rippe, R. C. A., & van der Voort, A. (2020). Rebutting Existing Misconceptions About Multiple Imputation as a Method for Handling Missing Data. *Journal of Personality Assessment*, *102*(3), 297–308. https://doi.org/10.1080/00223891.2018.1530680

Verleysen, M., & François, D. (2005). The Curse of Dimensionality in Data Mining and Time Series Prediction. In J. Cabestany, A. Prieto, & F. Sandoval (Eds.), *Computational Intelligence and Bioinspired Systems* (pp. 758–770). Springer. https://doi.org/10.1007/11494669_93

Vulperhorst, J., Lutz, C., de Kleijn, R., & van Tartwijk, J. (2018). Disentangling the predictive validity of high school grades for academic success in university. *Assessment & Evaluation in Higher Education*, *43*(3), 399–414. https://doi.org/10.1080/02602938.2017.1353586

Warren, B. (1991). Concepts, Constructs, Cognitive Psychology, and Personal Construct Theory. *The Journal of Psychology*, *125*(5), 525–536. https://doi.org/10.1080/00223980.1991.10543316

White, I. R., & Carlin, J. B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine*, *29*(28), 2920–2931. https://doi.org/10.1002/sim.3944

Whitley, B. E., & Kite, J., Mary E. (2012). *Principles of Research in Behavioral Science: Third Edition* (3rd ed.). Routledge. https://doi.org/10.4324/9780203085219

Wickham, H. (2016). Data Analysis. In H. Wickham (Ed.), *Ggplot2: Elegant Graphics for Data Analysis* (pp. 189–201). Springer International Publishing. https://doi.org/10.1007/978-3-319-24277-4_9

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604. https://doi.org/10.1037/0003-066X.54.8.594

Williamson, B. (2017). *Big Data in Education: The digital future of learning, policy and practice*. 1–256. https://www.torrossa.com/en/resources/an/5017810

Woods, A. D., Gerasimova, D., Van Dusen, B., Nissen, J., Bainter, S., Uzdavines, A., Davis-Kean, P. E., Halvorson, M., King, K. M., Logan, J. A. R., Xu, M., Vasilev, M. R., Clay, J. M., Moreau, D., Joyal-Desmarais, K., Cruz, R. A., Brown, D. M. Y., Schmidt, K., & Elsherif, M. M. (n.d.). Best practices for addressing missing data through multiple imputation. *Infant and Child Development*, *n/a*(n/a), e2407. https://doi.org/10.1002/icd.2407

Wu, C. L., Chau, K. W., & Li, Y. S. (2009). Methods to improve neural network performance in daily flows prediction. *Journal of Hydrology*, *372*(1), 80–93. https://doi.org/10.1016/j.jhydrol.2009.03.038

Yadav, M. L., & Roychoudhury, B. (2018). Handling missing values: A study of popular imputation packages in R. *Knowledge-Based Systems*, *160*, 104–118. https://doi.org/10.1016/j.knosys.2018.06.012

Yoon, J., Jordon, J., & Schaar, M. (2018). Gain: Missing data imputation using generative adversarial nets. *International Conference on Machine Learning*, 5689–5698.

Yuan, K.-H., & Bentler, P. M. (2007). 17—Robust Procedures in Structural Equation

Modeling**This research was supported by NSF grant DMS04-37167, the James McKeen Cattell Fund, and grants DA01070 and DA00017 from the National Institute on Drug Abuse. In S.-Y. Lee (Ed.), *Handbook of Latent Variable and Related Models* (pp. 367–397). North-Holland. https://doi.org/10.1016/B978-044452044-9/50020-3

Zhai, X. (2022). *ChatGPT User Experience: Implications for Education* (SSRN Scholarly Paper 4312418). https://doi.org/10.2139/ssrn.4312418

# Appendix A

## Grading Scheme Conversion

**Figure A.1**: NorQuest College Grading Scheme of '7A' and '7'

| Letter Grade (Post-secondary programs) | Grade Point Value (Post-secondary programs) | Percentage (Alberta Education courses/preparatory) | Descriptor |
|---|---|---|---|
| A+ | 4.0 | 95-100 | Excellent |
| A | 4.0 | 90-94 | |
| A- | 3.7 | 85-89 | |
| B+ | 3.3 | 80-84 | Very Good |
| B | 3.0 | 75-79 | |
| B- | 2.7 | 70-74 | |
| C+ | 2.3 | 67-69 | Satisfactory/ Acceptable |
| C | 2.0 | 64-66 | |
| C- | 1.7 | 60-63 | |
| D+ | 1.3 | 55-59 | |
| D | 1.0 | 50-54 | Pass |
| F | 0.0 | 0-49 | Fail |

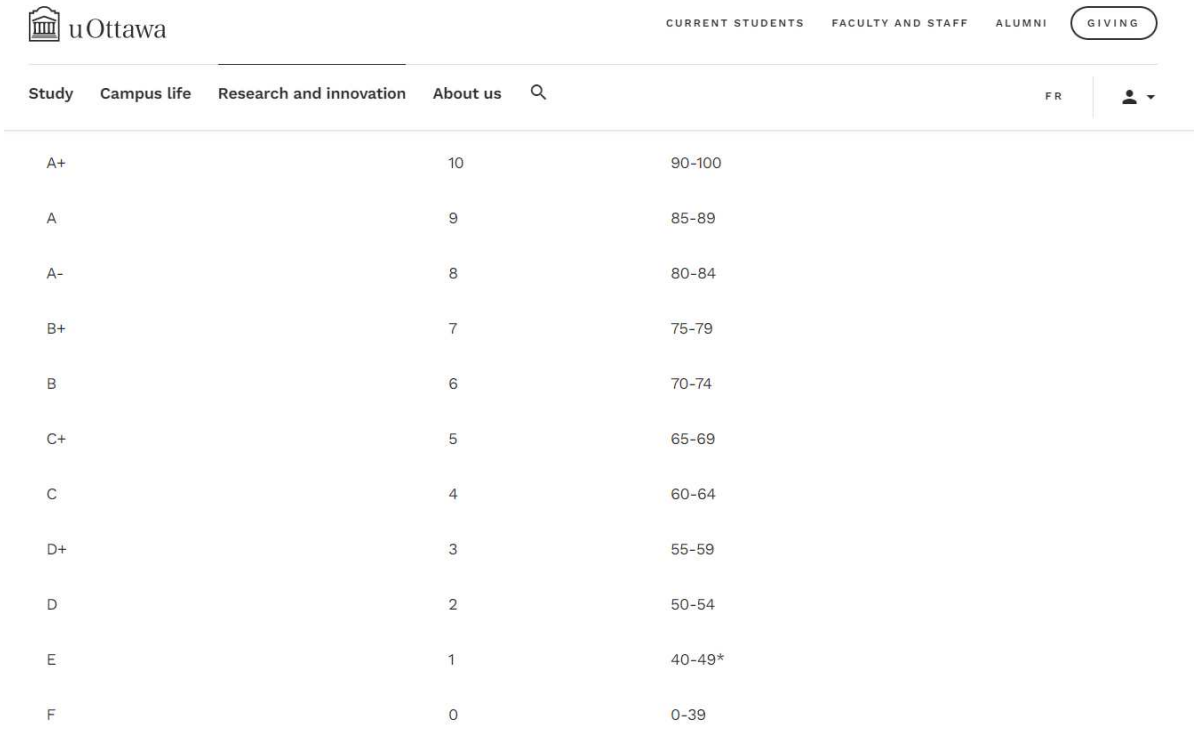
**Figure A.2**: Canadian Mennonite University Grading Scheme of '7C'

## Grade Points (UG)

Grade points are assigned to each letter grade as follows:

| Letter Grade | Grade Points | |
|---|---|---|
| A+ | 4.5 | Exceptional |
| A | 4.0 | Excellent |
| B+ | 3.5 | Very Good |
| B | 3.0 | Good |
| C+ | 2.5 | Satisfactory |
| C | 2.0 | Adequate |
| D | 1.0 | Marginal |
| F | 0 | Failure |
| P | NA | Pass |

**Figure A.3**: University of Ottawa Grading Scheme of '7D'




| | | |
|---|---|---|
| A+ | 10 | 90-100 |
| A | 9 | 85-89 |
| A- | 8 | 80-84 |
| B+ | 7 | 75-79 |
| B | 6 | 70-74 |
| C+ | 5 | 65-69 |
| C | 4 | 60-64 |
| D+ | 3 | 55-59 |
| D | 2 | 50-54 |
| E | 1 | 40-49* |
| F | 0 | 0-39 |

**Figure A.4**: IB Grading Scheme Conversion Table

**IB Grades and GPA Calculator**

| IB Grade | Description | GPA Equivalent | Letter Grade |
|---|---|---|---|
| 7 | Excellent | 4+ | A+ |
| 6 | Very Good | 4 | A |
| 5 | Good | 3 | B |
| 4 | Satisfactory | 2 | C |
| 3 | Mediocre | 1 | D |
| 2 | Poor | 0 | E |
| 1 | Very Poor | 0 | F |

**Table A.1**: UAlberta AP Grading Scheme Conversion Table

| AP Result | Percent Equivalent |
|---|---|
| 5 | 96% |
| 4 | 86% |
| 3 | 76% |
| 2 | 65% |
| 1 | Not accepted for admission |

## English Language Test Conversion Table

**Table A.2**: Conversion standards across English Language Tests

| Scores | IELTS | TOEF3 | IB | CAEL | PTE |
|---|---|---|---|---|---|
| 9 | 9 | 118-120 | - | >90 | N/A |
| 8.5 | 8.5 | 115-117 | - | >85 | >89 |
| 8 | 8 | 110-114 | - | >80 | >84 |
| 7.5 | 7.5 | 102-109 | >98 | >75 | >76 |
| 7 | 7 | 94-101 | - | >70 | >66 |
| 6.5 | 6.5 | 79-93 | - | >60 | >56 |
| 6 | 6 | 60-78 | >90 | >50 | >46 |
| 5.5 | 5.5 | 46-59 | - | >40 | >36 |
| 5 | 5 | 35-45 | >82 | >35 | >29 |
| 4.5 | 4.5 | 32-34 | >73 | >30 | >23 |
| 4 | 0-4 | 0-31 | - | 0-29 | <23 |

| | | | | | |
|---|---|---|---|---|---|
| 3 | - | - | >55 | - | - |
| 2 | - | - | - | - | - |
| 0-1 | - | - | - | - | - |

# Appendix B

## SEM Comprehensive Results

### Features and Aliases

**Table B.1**: Features, Variables and Aliases

| Category | Alias | Feature | Alias |
|---|---|---|---|
| **Demography** | Dmg/DM | Legal Status | L_S |
| | | Age | Age |
| | | Gender | Gnd |
| **High School Details** | H_S/HS | Grading Scheme | G_S |
| | | High School Course Grades | HS_C |
| | | High School Total Credits Earned | HS_T |
| **Program Details** | P_D/PD | Course ID | C_I |
| | | University Credits Taken | U_C_T |
| | | University Course Grade | U_C_G |
| **English Language Skills** | ELP | Test ID | T_I |
| | | Test Score | T_S |
| **Time to Completion** | T_C/Time_Complete | Admission Year | A_Y |
| | | Completion Year | C_Y |

### Features and Encodings

**Table B.2**: Variables, Types and Encodings

| **Feature** | Legal Status | | Age | Gender | |
|---|---|---|---|---|---|
| **Type** | International Student | Domestic Student | Not encoded | Female | Male |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Encoding** | 1 | 0 | NA | | | | 1 | 0 |
| **Feature** | Grading Scheme | | High School Course Grades | | | | High School Total Credits Earned | |
| **Type** | International High Schools | Domestic High Schools | Not Encoded | | | | Not Encoded | |
| **Encoding** | 1 | 0 | NA | | | | NA | |
| **Feature** | Course ID | | University Credits Taken | | | | University Course Grade | |
| **Type** | Not encoded | | Not encoded | | | | Not encoded | |
| **Encoding** | NA | | NA | | | | NA | |
| **Feature** | Test ID | | | | | | Test Score | |
| **Type** | IELTS | TOEF | IB | UEALA | CEAL | PTE | MELAB | Not encoded |
| **Encoding** | 0 | 1 | 2 | 3 | 4 | 5 | 6 | NA |
| **Feature** | Admission Year | | | | | Completion Year | | |

| **Type** | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Encoding** | 0 | 1 | 2 | 3 | 4 | 5 | 0 | 1 | 2 | 3 | 4 | 5 |

# SEM 1

**Table B.3**: Summarized Results from fitting SEM RQ1, Regression Model

| | |
|---|---|
| Estimator | ML |
| Number of model parameters | 6 |
| Number of observations | 123522 |
| Model Test User Model: | |
| Test statistic | 0.000 |

| | | | |
|---|---|---|---|
| Degrees of freedom | | | 0 |
| Model Test Baseline Model: | | | |
| Degrees of freedom | | | 3 |
| P-value | | | 0.000 |
| User Model versus Baseline Model: | | | |
| Comparative Fit Index (CFI) | | | 1.000 |
| Tucker-Lewis Index (TLI) | | | 1.000 |
| Root Mean Square Error of Approximation: | | | |
| RMSEA | | | 0.000 |
| Standardized Root Mean Square Residual: | | | |
| SRMR | | | 0.000 |
| Regressions: | | | |
| | Estimate | | Std.all |
| Time_Complete ~ | | | |
| High_Schl (b1) | 0.013 | | 0.016 |
| Progrm_Dtl (b2) | 0.000 | | 0.458 |
| Covariances: | | | |
| | Estimate | | Std.all |
| High_School ~~ | | | |
| Progrm_Details | 2610.597 | | 0.192 |
| Covariance matrix: | | | |
| | Tm_Cmp | Hgh_Sc | Prgr_D |
| Time_Complete | 0.851 | | |
| High_School | 0.115 | 1.423 | |

| | | | |
|---|---|---|---|
| Program_Details | 48371000 | 2.610597e+03 | 1.292854e+08 |
| Variances: | | | |
| | Estimate | | Std.all |
| High_School | 1.432 | | 1.000 |
| Progrm_Details | 129285352.108 | | 1.000 |
| .Time_Complete | 0.670 | | 0.787 |
| R-Square: | | | |
| | Estimate | | |
| Time_Complete | 0.213 | | |

**Table B.4:** Summarized Results from fitting SEM RQ1, Mediation Model

| | |
|---|---|
| Estimator | ML |
| Number of model parameters | 19 |
| Number of observations | 123522 |
| Model Test User Model: | |
| Degrees of freedom | 17 |
| P-value (Chi_square) | 0.000 |
| Model Test Baseline Model: | |
| Degrees of freedom | 28 |
| P-value | 0.000 |
| User Model versus Baseline Model: | |

| | | |
|---|---|---|
| Comparative Fit Index (CFI) | | 0.924 |
| Tucker-Lewis Index (TLI) | | 0.875 |
| Root Mean Square Error of Approximation: | | |
| RMSEA | | 0.044 |
| Standardized Root Mean Square Residual: | | |
| SRMR | | 0.028 |
| Latent Variables | | |
| | Estimate | Std.all |
| HS =~ | | |
| HS_Crse_Grade | 1.000 | 0.304 |
| HS_Ttl_Crdt_Er | -3.094 | -0.032 |
| Grading_Scheme | 41.117 | 0.927 |
| PD =~ | | |
| Uni_Crse_Grade | 1.000 | 0.047 |
| Course_ID | 3691.107 | 0.705 |
| Uni_Credts_Tkn | -0.198 | -0.029 |
| TC =~ | | |
| Admit_Year | 1.000 | 0.484 |
| Completion_Yer | 1.012 | 0.983 |
| Covariances: | | |
| | Estimate | Std.all |
| HS ~~ | | |
| PD | 0.000 | 0.018 |
| TC | 0.006 | 0.277 |

| | | |
|---|---|---|
| PD ~~ | | |
| TC | 0.001 | 0.021 |
| Variances: | | |
| | Estimate | Std.all |
| HS | 0.001 | 1.000 |
| PD | 0.001 | 1.000 |
| TC | 0.464 | 1.000 |
| .HS_Crse_Grade | 0.011 | 0.907 |
| .HS_Ttl_Crdt_Er | 10.746 | 0.999 |
| .Grading_Scheme | 0.315 | 0.141 |
| .Uni_Crse_Grade | 0.638 | 0.998 |
| .Course_ID | 19463.308 | 0.503 |
| .Uni_Credts_Tkn | 0.066 | 0.999 |
| .Admit_Year | 1.512 | 0.765 |
| .Completion_Yer | 0.016 | 0.033 |
| R-Square: | | |
| | Estimate | |
| HS_Crse_Grade | 0.093 | |
| HS_Ttl_Crdt_Er | 0.001 | |
| Grading_Scheme | 0.859 | |
| Uni_Crse_Grade | 0.002 | |
| Course_ID | 0.497 | |
| Uni_Credts_Tkn | 0.001 | |
| Admit_Year | 0.235 | |

| | |
|---|---|
| Completion_Yer | 0.967 |

# SEM 2

**Table B.5**: Summarized Results from fitting SEM RQ2, Regression Model

| | | |
|---|---|---|
| Estimator | | ML |
| Number of model parameters | | 6 |
| Number of observations | | 123522 |
| Model Test User Model: | | |
| Test statistic | | 0.000 |
| Degrees of freedom | | 0 |
| Model Test Baseline Model: | | |
| Degrees of freedom | | 3 |
| P-value | | 0.000 |
| User Model versus Baseline Model: | | |
| Comparative Fit Index (CFI) | | 1.000 |
| Tucker-Lewis Index (TLI) | | 1.000 |
| Root Mean Square Error of Approximation: | | |
| RMSEA | | 0.000 |
| Standardized Root Mean Square Residual: | | |
| SRMR | | 0.000 |
| Regressions: | | |
| | Estimate | Std.all |
| Time_Complete ~ | | |

| | | | |
|---|---|---|---|
| High_Schl (b1) | 0.000 | | 0.488 |
| Demography (b2) | 0.372 | | 0.285 |
| Covariances: | | | |
| | Estimate | | Std.all |
| High_School ~~ | | | |
| Demography | -1005.855 | | -0.0942 |
| Covariance matrix: | | | |
| | Tm_Cmp | Hgh_Sc | Demgrph |
| Time_Complete | 0.851 | | |
| High_School | 6449.236 | 229826380.118 | |
| Demography | 0.156 | -1005.855 | 0.500 |
| Variances: | | | |
| | Estimate | | Std.all |
| High_School | 229826380.118 | | 1.000 |
| Demography | 0.500 | | 1.000 |
| .Time_Complete | 0.602 | | 0.707 |
| R-Square: | | | |
| | Estimate | | |
| Time_Complete | 0.293 | | |

**Table B.6**: Summarized Results from fitting SEM RQ2, Mediation Model

| | |
|---|---|
| Estimator | ML |

| | | |
|---|---|---|
| Number of model parameters | | 17 |
| Number of observations | | 123522 |
| Model Test User Model: | | |
| Degrees of freedom | | 11 |
| P-value (Chi_square) | | 0.000 |
| Model Test Baseline Model: | | |
| Degrees of freedom | | 21 |
| P-value | | 0.000 |
| User Model versus Baseline Model: | | |
| Comparative Fit Index (CFI) | | 0.877 |
| Tucker-Lewis Index (TLI) | | 0.765 |
| Root Mean Square Error of Approximation: | | |
| RMSEA | | 0.083 |
| Standardized Root Mean Square Residual: | | |
| SRMR | | 0.055 |
| Latent Variables | | |
| | Estimate | Std.all |
| HS =~ | | |
| HS_Crse_Grade | 1.000 | 0.085 |
| HS_Ttl_Crdt_Er | 102.267 | 0.292 |
| Grading_Scheme | 41.117 | 0.927 |
| DM =~ | | |
| Age | 1.000 | 0.640 |
| Gender | 0.099 | 0.188 |

| | | |
|---|---|---|
| Legal_Status | 0.549 | 0.549 |
| TC =~ | | |
| Admit_Year | 1.000 | 1.034 |
| Completion_Yer | 0.222 | 0.461 |
| Covariances: | | |
| | Estimate | Std.all |
| HS ~~ | | |
| DM | -0.010 | -1.118 |
| TC | 0.003 | 0.211 |
| DM ~~ | | |
| TC | 0.423 | 0.309 |
| Variances: | | |
| | Estimate | Std.all |
| HS | 0.000 | 1.000 |
| DM | 0.885 | 1.000 |
| TC | 2.112 | 1.000 |
| .HS_Crse_Grade | 0.012 | 0.993 |
| .HS_Ttl_Crdt_Er | 9.842 | 0.915 |
| .Age | 1.274 | 0.590 |
| .Gender | 0.238 | 0.965 |
| .Legal_Status | 0.618 | 0.699 |
| .Admit_Year | -0.136 | -0.069 |
| .Completion_Yer | 0.387 | 0.788 |
| R-Square: | | |

| | Estimate |
|---|---|
| HS_Crse_Grade | 0.007 |
| HS_Ttl_Crdt_Er | 0.085 |
| Age | 0.410 |
| Gender | 0.035 |
| Legal_Status | 0.301 |
| Admit_Year | NA |
| Completion_Yer | 0.212 |

# SEM 3

**Table B.7**: Summarized Results from fitting SEM RQ3, Regression Model

| Estimator | ML |
|---|---|
| Number of model parameters | 6 |
| Number of observations | 123522 |
| Model Test User Model: | |
| Test statistic | 0.000 |
| Degrees of freedom | 0 |
| Model Test Baseline Model: | |
| Degrees of freedom | 3 |
| P-value | 0.000 |
| User Model versus Baseline Model: | |
| Comparative Fit Index (CFI) | 1.000 |

| | | | |
|---|---|---|---|
| Tucker-Lewis Index (TLI) | | | 1.000 |
| Root Mean Square Error of Approximation: | | | |
| RMSEA | | | 0.000 |
| Standardized Root Mean Square Residual: | | | |
| SRMR | | | 0.000 |
| Regressions: | | | |
| | Estimate | | Std.all |
| Time_Complete ~ | | | |
| High_Schl (b1) | 0.000 | | 0.463 |
| ELP (b2) | 0.563 | | 0.196 |
| Covariances: | | | |
| | Estimate | | Std.all |
| High_School ~~ | | | |
| ELP | -32.493 | | -0.009 |
| Covariance matrix: | | | |
| | Tm_Cmp | Hgh_Sc | ELP |
| Time_Complete | 0.851 | | |
| High_School | 4836.927 | 129277336.007 | |
| ELP | 0.057 | -32.493 | 0.103 |
| Variances: | | | |
| | Estimate | | Std.all |
| High_School | 129277336.007 | | 1.000 |
| ELP | 0.103 | | 1.000 |
| .Time_Complete | 0.638 | | 0.749 |

| R-Square: | |
|---|---|
| | Estimate |
| Time_Complete | 0.251 |

**Table B.8**: Summarized Results from fitting SEM RQ3, Mediation Model

| Estimator | ML |
|---|---|
| Number of model parameters | 18 |
| Number of observations | 123522 |
| Model Test User Model: | |
| Degrees of freedom | 11 |
| P-value (Chi_square) | 0.000 |
| Model Test Baseline Model: | |
| Degrees of freedom | 21 |
| P-value | 0.000 |
| User Model versus Baseline Model: | |
| Comparative Fit Index (CFI) | 0.810 |
| Tucker-Lewis Index (TLI) | 0.638 |
| Root Mean Square Error of Approximation: | |
| RMSEA | 0.097 |
| Standardized Root Mean Square Residual: | |
| SRMR | 0.055 |

| Latent Variables | | |
|---|---|---|
| | Estimate | Std.all |
| HS =~ | | |
| HS_Crse_Grade | 1.000 | 0.340 |
| HS_Ttl_Crdt_Er | -3.551 | -0.041 |
| Grading_Scheme | 32.916 | 0.828 |
| ELP_ =~ | | |
| Test_Id | 1.000 | 5.111 |
| Test_Score | -0.000 | -0.014 |
| TC =~ | | |
| Admit_Year | 1.000 | 0.591 |
| Completion_Yer | 0.679 | 0.805 |
| Covariances: | | |
| | Estimate | Std.all |
| HS ~~ | | |
| ELP_ | 0.004 | 0.034 |
| TC | 0.011 | 0.349 |
| ELP_ ~~ | | |
| TC | 0.099 | 0.037 |
| Variances: | | |
| | Estimate | Std.all |
| HS | 0.001 | 1.000 |
| ELP_ | 10.607 | 1.000 |
| TC | 0.691 | 1.000 |

| | | |
|---|---|---|
| .HS_Crse_Grade | 0.011 | 0.885 |
| .HS_Ttl_Crdt_Er | 10.739 | 0.998 |
| .Grading_Scheme | 0.702 | 0.315 |
| .Test_Id | -10.201 | -25.127 |
| .Test_Score | 0.014 | 1.000 |
| .Admit_Year | 1.284 | 0.650 |
| .Completion_Yer | 0.173 | 0.351 |
| R-Square: | | |
| | Estimate | |
| HS_Crse_Grade | 0.115 | |
| HS_Ttl_Crdt_Er | 0.002 | |
| Grading_Scheme | 0.685 | |
| Test_Id | NA | |
| Test_Score | 0.000 | |
| Admit_Year | 0.350 | |
| Completion_Yer | 0.649 | |

# SEM 4

**Table B.9**: Summarized Results from fitting SEM RQ4, Regression Model

| | |
|---|---|
| Estimator | ML |
| Number of model parameters | 6 |
| Number of observations | 123522 |
| Model Test User Model: | |

| | | | |
|---|---|---|---|
| Test statistic | | | 0.000 |
| Degrees of freedom | | | 0 |
| Model Test Baseline Model: | | | |
| Degrees of freedom | | | 3 |
| P-value | | | 0.000 |
| User Model versus Baseline Model: | | | |
| Comparative Fit Index (CFI) | | | 1.000 |
| Tucker-Lewis Index (TLI) | | | 1.000 |
| Root Mean Square Error of Approximation: | | | |
| RMSEA | | | 0.000 |
| Standardized Root Mean Square Residual: | | | |
| SRMR | | | 0.000 |
| Regressions: | | | |
| | Estimate | | Std.all |
| Time_Complete ~ | | | |
| Prgrm_Dtl (b1) | 0.000 | | 0.488 |
| Demogrphy (b2) | 0.372 | | 0.285 |
| Covariances: | | | |
| | Estimate | | Std.all |
| Program_Details ~~ | | | |
| Demography | -754.620 | | -0.094 |
| Covariance matrix: | | | |
| | Tm_Cmp | Prgr_D | Dmgrph |
| Time_Complete | 0.851 | | |

| | | | |
|---|---|---|---|
| Program_Details | 4837.100 | 129285349.564 | |
| Demography | 0.156 | -754.620 | 0.500 |
| Variances: | | | |
| | Estimate | | Std.all |
| Program_Detals | 129285349.564 | | 1.000 |
| Demography | 0.500 | | 1.000 |
| .Time_Complete | 0.602 | | 0.707 |
| R-Square: | | | |
| | Estimate | | |
| Time_Complete | 0.293 | | |

**Table B.10**: Summarized Results from fitting SEM RQ4, Mediation Model

| | |
|---|---|
| Estimator | ML |
| Number of model parameters | 19 |
| Number of observations | 123522 |
| Model Test User Model: | |
| Degrees of freedom | 17 |
| P-value (Chi_square) | 0.000 |
| Model Test Baseline Model: | |
| Degrees of freedom | 28 |
| P-value | 0.000 |
| User Model versus Baseline Model: | |
| Comparative Fit Index (CFI) | 0.373 |

| | | |
|---|---|---|
| Tucker-Lewis Index (TLI) | | -0.033 |
| Root Mean Square Error of Approximation: | | |
| RMSEA | | 0.146 |
| Standardized Root Mean Square Residual: | | |
| SRMR | | 0.093 |
| Latent Variables | | |
| | Estimate | Std.all |
| DM =~ | | |
| Age | 1.000 | 0.507 |
| Gender | 0.114 | 0.170 |
| Legal_Status | 0.885 | 0.701 |
| PD =~ | | |
| Course_ID | 1.000 | 0.002 |
| Uni_Credts_Tkn | 0.015 | 0.016 |
| Uni_Crse_Grade | 0.428 | 0.146 |
| TC =~ | | |
| Admit_Year | 1.000 | 0.987 |
| Completion_Yer | 0.244 | 0.483 |
| Covariances: | | |
| | Estimate | Std.all |
| DM ~~ | | |
| PD | -0.383 | -1.877 |
| TC | 0.343 | 0.332 |
| PD ~~ | | |

| | | |
|---|---|---|
| TC | 0.069 | 0.182 |
| Variances: | | |
| | Estimate | Std.all |
| DM | 0.554 | 1.000 |
| PD | 0.075 | 1.000 |
| TC | 1.925 | 1.000 |
| .Age | 1.605 | 0.743 |
| .Gender | 0.240 | 0.971 |
| .Legal_Status | 0.450 | 0.509 |
| .Course_ID | 19337.715 | 1.000 |
| .Uni_Credts_Tkn | 0.066 | 1.000 |
| .Uni_Crse_Grade | 0.626 | 0.979 |
| .Admit_Year | 0.050 | 0.025 |
| .Completion_Yer | 0.377 | 0.767 |
| R-Square: | | |
| | Estimate | |
| Age | 0.257 | |
| Gender | 0.029 | |
| Legal_Status | 0.491 | |
| Course_ID | 0.000 | |
| Uni_Credts_Tkn | 0.000 | |
| Uni_Crse_Grade | 0.021 | |
| Admit_Year | 0.975 | |
| Completion_Yer | 0.233 | |

# SEM 5

**Table B.11**: Summarized Results from fitting SEM RQ5, Regression Model

| | | |
|---|---|---|
| Estimator | | ML |
| Number of model parameters | | 6 |
| Number of observations | | 123522 |
| Model Test User Model: | | |
| Test statistic | | 0.000 |
| Degrees of freedom | | 0 |
| Model Test Baseline Model: | | |
| Degrees of freedom | | 3 |
| P-value | | 0.000 |
| User Model versus Baseline Model: | | |
| Comparative Fit Index (CFI) | | 1.000 |
| Tucker-Lewis Index (TLI) | | 1.000 |
| Root Mean Square Error of Approximation: | | |
| RMSEA | | 0.000 |
| Standardized Root Mean Square Residual: | | |
| SRMR | | 0.000 |
| Regressions: | | |
| | Estimate | Std.all |
| Time_Complete ~ | | |
| Prgrm_Dtl (b1) | 0.000 | 0.463 |

| | | | |
|---|---|---|---|
| ELP (b2) | 0.563 | | 0.195 |
| Covariances: | | | |
| | Estimate | | Std.all |
| Program_Details ~~ | | | |
| ELP | -32.310 | | -0.009 |
| Covariance matrix: | | | |
| | Tm_Cmp | Prgr_D | ELP |
| Time_Complete | 0.851 | | |
| Program_Details | 4837.100 | 129285353.271 | |
| ELP | 0.057 | -32.310 | 0.103 |
| Variances: | | | |
| | Estimate | | Std.all |
| Program_Detals | 129285353.271 | | 1.000 |
| ELP | 0.103 | | 1.000 |
| .Time_Complete | 0.638 | | 0.749 |
| R-Square: | | | |
| | Estimate | | |
| Time_Complete | 0.251 | | |

**Table B.12**: Summarized Results from fitting SEM RQ5, Mediation Model

| | |
|---|---|
| Estimator | ML |
| Number of model parameters | 17 |

| | | |
|---|---|---|
| Number of observations | | 123522 |
| Model Test User Model: | | |
| Degrees of freedom | | 11 |
| P-value (Chi_square) | | 0.000 |
| Model Test Baseline Model: | | |
| Degrees of freedom | | 21 |
| P-value | | 0.000 |
| User Model versus Baseline Model: | | |
| Comparative Fit Index (CFI) | | 0.060 |
| Tucker-Lewis Index (TLI) | | -0.794 |
| Root Mean Square Error of Approximation: | | |
| RMSEA | | 0.175 |
| Standardized Root Mean Square Residual: | | |
| SRMR | | 0.099 |
| Latent Variables | | |
| | Estimate | Std.all |
| ELP_ =~ | | |
| ELP_Test_Id | 1.000 | 0.477 |
| ELP_Score | -0.053 | -0.136 |
| PD =~ | | |
| Course_ID | 1.000 | 0.001 |
| Uni_Credts_Tkn | 0.024 | 0.013 |
| Uni_Crse_Grade | 0.384 | 0.068 |
| TC =~ | | |

| | | |
|---|---|---|
| Admit_Year | 1.000 | 0.873 |
| Completion_Yer | 0.312 | 0.546 |
| Covariances: | | |
| | Estimate | Std.all |
| ELP_ ~~ | | |
| PD | -0.230 | -5.353 |
| TC | 0.152 | 0.408 |
| PD ~~ | | |
| TC | 0.032 | 0.186 |
| Variances: | | |
| | Estimate | Std.all |
| ELP_ | 0.092 | 1.000 |
| PD | 0.020 | 1.000 |
| TC | 1.504 | 1.000 |
| .ELP_Test_Id | 0.313 | 0.772 |
| .ELP_Score | 0.014 | 0.981 |
| .Course_ID | 19337.731 | 1.000 |
| .Uni_Credts_Tkn | 0.066 | 1.000 |
| .Uni_Crse_Grade | 0.637 | 0.995 |
| .Admit_Year | 0.471 | 0.239 |
| .Completion_Yer | 0.345 | 0.702 |
| R-Square: | | |
| | Estimate | |
| ELP_Test_Id | 0.228 | |

| | |
|---|---|
| ELP_Score | 0.019 |
| Course_ID | 0.000 |
| Uni_Credts_Tkn | 0.000 |
| Uni_Crse_Grade | 0.005 |
| Admit_Year | 0.761 |
| Completion_Yer | 0.298 |

# SEM 6

**Table B.13**: Summarized Results from fitting SEM RQ6, Regression Model

| | |
|---|---|
| Estimator | ML |
| Number of model parameters | 6 |
| Number of observations | 123522 |
| Model Test User Model: | |
| Test statistic | 0.000 |
| Degrees of freedom | 0 |
| Model Test Baseline Model: | |
| Degrees of freedom | 3 |
| P-value | 0.000 |
| User Model versus Baseline Model: | |
| Comparative Fit Index (CFI) | 1.000 |
| Tucker-Lewis Index (TLI) | 1.000 |
| Root Mean Square Error of Approximation: | |
| RMSEA | 0.000 |

| | | | |
|---|---|---|---|
| Standardized Root Mean Square Residual: | | | |
| SRMR | | | 0.000 |
| Regressions: | | | |
| | Estimate | | Std.all |
| Time_Complete ~ | | | |
| Demogrphy (b1) | 0.254 | | 0.195 |
| ELP (b2) | 0.210 | | 0.073 |
| Covariances: | | | |
| | Estimate | | Std.all |
| Demography ~~ | | | |
| ELP | 0.138 | | 0.608 |
| Covariance matrix: | | | |
| | Tm_Cmp | Dmgrph | ELP |
| Time_Complete | 0.851 | | |
| Demography | 0.156 | 0.500 | |
| ELP | 0.057 | 0.138 | 0.103 |
| Variances: | | | |
| | Estimate | | Std.all |
| Demography | 0.500 | | 1.000 |
| ELP | 0.103 | | 1.000 |
| .Time_Complete | 0.800 | | 0.939 |
| R-Square: | | | |
| | Estimate | | |
| Time_Complete | 0.101 | | |

**Table B.14**: Summarized Results from fitting SEM RQ6, Mediation Model

| | | |
|---|---|---|
| Estimator | | ML |
| Number of model parameters | | 17 |
| Number of observations | | 123522 |
| Model Test User Model: | | |
| Degrees of freedom | | 11 |
| P-value (Chi_square) | | 0.000 |
| Model Test Baseline Model: | | |
| Degrees of freedom | | 21 |
| P-value | | 0.000 |
| User Model versus Baseline Model: | | |
| Comparative Fit Index (CFI) | | 0.951 |
| Tucker-Lewis Index (TLI) | | 0.906 |
| Root Mean Square Error of Approximation: | | |
| RMSEA | | 0.081 |
| Standardized Root Mean Square Residual: | | |
| SRMR | | 0.050 |
| Latent Variables | | |
| | Estimate | Std.all |
| ELP_ =~ | | |
| ELP_Test_Id | 1.000 | NA |
| ELP_Score | 0.009 | NA |

| | | |
|---|---|---|
| DM =~ | | |
| Age | 1.000 | 0.410 |
| Gender | 0.107 | 0.130 |
| Legal_Status | 1.364 | 0.873 |
| TC =~ | | |
| Admit_Year | 1.000 | 0.876 |
| Completion_Yer | 0.309 | 0.544 |
| Covariances: | | |
| | Estimate | Std.all |
| ELP_ ~~ | | |
| DM | 0.338 | 0.738 |
| TC | 0.171 | 0.182 |
| DM ~~ | | |
| TC | 0.222 | 0.300 |
| Variances: | | |
| | Estimate | Std.all |
| ELP_ | -0.578 | NA |
| DM | 0.362 | 1.000 |
| TC | 1.517 | 1.000 |
| .ELP_Test_Id | 0.984 | 2.425 |
| .ELP_Score | 0.014 | 1.003 |
| .Age | 1.797 | 0.832 |
| .Gender | 0.243 | 0.983 |
| .Legal_Status | 0.210 | 0.238 |

| | | |
|---|---|---|
| .Admit_Year | 0.459 | 0.232 |
| .Completion_Yer | 0.346 | 0.704 |
| R-Square: | | |
| | Estimate | |
| ELP_Test_Id | -1.425 | |
| ELP_Score | -0.003 | |
| Age | 0.168 | |
| Gender | 0.017 | |
| Legal_Status | 0.762 | |
| Admit_Year | 0.768 | |
| Completion_Yer | 0.296 | |