Methodological Advances for Deepening the Interrogation of Data from Education Evaluations

by
Alexander Charles O'Riordan

Thesis presented in fulfilment of the requirements for the degree of
Master of Commerce in the Department of
Economics at Stellenbosch University

Supervisor: Professor Servaas van der Berg

March 2021

Stellenbosch University https://scholar.sun.ac.za

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: March 2021

Name: Alexander Charles O'Riordan

Copyright © 2021 Stellenbosch University
All rights reserved

Stellenbosch University https://scholar.sun.ac.za

Summary

The ability to read is arguably the most important skill in most academic settings. All present and future learning is dependent fundamentally on the ability to appropriately interpret text and internalise its meaning. Students that do not possess the level of literacy skill necessary to decode text will be incapable of keeping pace with an education curriculum that is based primarily on the sequential acquisition of certain skills. It is a serious concern that a large number of students within the South African education system do not possess adequate literacy ability; this illiteracy is omnipresent in poorer schools. Therefore, the link between poverty, inadequate education quality, and low reading ability becomes a binding constraint on the ability of poor students to perform academically. This link is a means through which limited labour market prospects - and in many cases, poverty - become entrenched.

While the improvement of education system is the central focus for policymakers, the inability to adequately model and explain the determinants of reading ability is a fundamental constraint on research. And as well-designed policy is dependent on accurate research, it is imperative that research in this area be performed as well as is possible. Contemporary work in the field of the economics of education centres on the use of well-established methods that have remained relatively unchanged. This thesis takes the stance that commonly used methods may be inappropriate in certain contexts.

The broad aim of this thesis is twofold. First, to demonstrate that value exists in the use of novel techniques that differ from what is currently common. Second, in doing so, derive findings that can be useful in guiding both current and future policy and research. The approaches taken in this thesis borrow extensively from methods used in the educational psychological and statistical learning literature. Two broad empirical themes are emphasised - latent constructs (factor modelling) and non-linear modelling. The thesis includes three empirical chapters (2, 3, and 4), each of which considers a unique aspect of the application of non-linear and factor modelling to large education-based datasets.

Chapter 2 investigates the potentially harmful effects of negative item responses using a latent construct framework. The intention of Chapter 2 is to use procedures of factor modelling to investigate whether the nature and design of survey questions can influence the information contained in the derived measured variable. It finds evidence that negatively worded survey questions capture information that is inconsistent with that derived from equivalent positively worded survey questions. Chapter 3 analyses how the choice of factor estimation technique affects the magnitude of measurement error transmitted to the factor from the underlying features. That is, Chapter 3 compares different methods of dimension reduction with regard to their relative ability to extract meaningful variation from N features while minimising the extraction of noisy variation. It finds significant performance differences between alternative methods of dimension reduction in the presence of erroneously measured variation, with evidence that Exploratory Factor Analysis is a superior approach. Chapter 4 fits a non-linear model using measured features and estimated factors to determine their relationships with one another and with reading performance. While Chapters 2 and 3 are methodological studies, Chapter 4 is an empirical analysis that models and interprets functional relationships that determine reading ability among South African grade 4 students. It finds several policy-relevant outcomes that are discussed in the chapter.

Stellenbosch University https://scholar.sun.ac.za

Opsomming

Die vermoë om te kan lees is waarskynlik die belangrikste vaardigheid om te hê in die meeste akademiese omgewings. Alle huidige en toekomstige geleerdheid is fundamenteel afhanklik van die vermoë om teks op die gepaste manier te analiseer en die betekenis daarvan te internaliseer. Leerders wat nie die vlakke van geletterdheid het om teks te ontleed nie sal nie in staat wees om by te bly met ‘n opvoedkundige kurrikulum wat hoofsaaklik gefokus is op die sekwensiële verkryging van sekere vaardighede nie. Dit wek ernstige kommer dat ‘n groot aantal studente in die Suid Afrikaanse onderwys stelsel nie genoegsame geletterdheid het nie. Hierdie ongeletterdheid is alomteenwoordig in armer skole. Daarom word die verbintenis tussen armoede, onvoldoende opvoedkundige kwaliteit en lae leesvermoë ‘n bindende beperking op die vermoë van armer studente om akademiese sukses te behaal. Hierdie verbintenis is hoe beperkte arbeidsmark vooruitsigte en in baie gevalle, armoede, verskans word.

Terwyl die verbetering van die onderwysstelsel die sentrale fokus is van beleidmakers, is die onvermoë om die faktore wat leesvermoë bepaal te modelleer en verduidelik op ‘n voldoende manier ‘n fundamentele beperking op navorsing. En omdat beleide wat goed ontwerp is afhanklik is van akkurate navorsing is dit noodsaaklik dat navorsing in hierdie gebied so goed as moontlik gedoen word. Huidige werk in die veld van ekonomie van opvoeding is gesentreer rondom die gebruik van goed gevestigde metodes wat relatief onveranderd gebly het. Hierdie tesis neem die standpunt dat algemeen gebruikte metodes onvanpas kan wees onder sekere omstandighede. Die breë doelwit van hierdie tesis is tweedelig. Eerstens, om te demonstreer dat daar waarde is in die gebruik van nuwe tegnieke wat verskil van die huidige algemene metodes. Tweedens, deur dit te demonstreer, bevindinge te kry wat bruikbaar kan wees om beide huidige en toekomstige beleide te beïnvloed.

Die benaderinge wat in hierdie tesis gebruik word maak baie staat op metodes wat in opvoedkundige sielkunde gebruik word sowel as die statistiese onderrig literatuur. Twee breë empiriese temas word beklemtoon – latente konstrukte (faktor modellering) en nie-liniêre modellering. 1 Die tesis sluit drie empiriese hofstukke in (2,3 en 4) waar elkeen ‘n unieke aspek van die toepassing van nie-liniêre en faktor modellering tot groot opvoedkundige gebaseerde datastelle oorweeg.

Hoofstuk 2 ondersoek die moontlike skadelike effek van negatiewe item reaksies wat ‘n latente konstrukte raamwerk gebruik. Die doel van Hoofstuk 2 is om prosedures van faktor modellering te gebruik om te ondersoek of die natuur en ontwerp van vrae in ‘n vraelys die inligting omvat in die afgeleide ondersoekte veranderlike kan beïnvloed. Dit vind bewyse dat negatief bewoorde vrae in ‘n vraelys inligting opneem wat strydend is met inligting wat verkry is van ‘n ekwivalente vraag wat positief bewoord is. Hoofstuk 3 analiseer hoe die keuse van ‘n faktor estimasie tegniek die grootte van die metings fout oorgedra na die faktor vanaf die onderliggende eienskappe affekteer. Dit wil se, Hoofstuk 3 vergelyk verskillende metodes van dimensie vermindering met betrekking tot hul relatiewe vermoë om betekenisvolle variasie vanaf N eienskappe te kry terwyl die ekstraksie van raserige variasie geminimaliseer word. Dit vind beduidende verskille in die optrede tussen alternatiewe metodes van dimensie vermindering in die aanwesigheid van verkeerdelik gemete variasie met bewyse dat Eksploratiewe Faktor Analise ‘n beter benadering is. Hoofstuk 4 pas ‘n nie-liniêre model toe met die gebruik van gemete eienskappe en geskatte faktore om te bepaal wat hul verhoudings is met mekaar en met lees prestasie. Terwyl Hoofstuk 2 en 3 metodologiese studies is, is Hoofstuk 4 ‘n empiriese analise wat die funksionele verhoudinge wat leesvermoë onder Suid Afrikaanse Graad 4 leerders beïnvloed modelleer en interpreteer. Dit vind verskeie beleid relevante uitkomstes wat bespreek word in die hoofstuk.

Stellenbosch University https://scholar.sun.ac.za

Thesis Content

Chapter 1: Introduction and Thesis Overview

Chapter 2: Negative Item Response Bias in Education-Based Surveys - a Latent Construct Estimation Approach

Chapter 3: Measurement Error in the Presence of Noisy Data - an Investigation into Procedures of Noise Reducing Factor Estimation

Chapter 4: Investigating the Determinants of Reading Ability Using Gradient Boosted Regression

Stellenbosch University https://scholar.sun.ac.za

Chapter 5: Conclusions and Implications

Stellenbosch University https://scholar.sun.ac.za

Figures and Tables

Note: Figures and tables are numbered according to their relative section in each chapter. Therefor, figure numbers may repeat but are added to this directory as: *Chapter, chapter section, figure number within that section.*

Chapter 2

Chapter 3

Stellenbosch University https://scholar.sun.ac.za

Chapter 4

Stellenbosch University https://scholar.sun.ac.za

Stellenbosch University https://scholar.sun.ac.za

Chapter 1: Thesis Introduction and Overview

Alexander C. O'Riordan[a]

[a]*Department of Economics, Stellenbosch University*

Abstract

Chapter 1 introduces the thesis, the individual chapters, and certain important concepts that are used extensively throughout the thesis. Furthermore, it outlines the purpose of the thesis, which can be broadly defined as an application of novel statistical techniques to large education-based datasets to highlight several important aspects of the data. In so doing, this thesis introduces new knowledge to the literature and demonstrates the value that exists in the application of novel techniques. It borrows extensively from the educational psychology and statistical learning literature respectively and employs throughout a theoretical framework based on the concept of latent constructs. The centrality of latent constructs to the approach employed in this thesis is outlined in this chapter. Moreover, Chapter 1 includes a broad outline of the theoretical and empirical methodology, limitations of the research, and the individual chapter abstracts.

Keywords: Latent Variable Approach, Non-Linear Modelling, Reading Performance, PIRLS

JEL classification L250, L100

1. Introduction

The ability to read is arguably the most important skill in most academic settings. All present and future learning is dependent fundamentally on the ability to appropriately interpret text and internalise its meaning. Put simply, a student must first learn to read before reading to learn. Students that do not possess the level of literacy skill necessary to decode text will be incapable of keeping pace with an education curriculum that is based primarily on the sequential acquisition of certain skills. From the above, the centrality of reading in the education and learning context becomes apparent. Therefore, it is a serious concern that a large number of students within the South African education system do not possess adequate literacy ability; this illiteracy is omnipresent in poorer schools. The reason behind this is twofold: first, these students receive an extremely low-quality education; and second, reading is often not a common practice in poorer households. As a result, reading ability is cultivated neither at school nor at home. Therefore, the link between poverty, inadequate education quality, and low reading ability becomes a binding constraint on the ability of poor students to perform academically. This link is a means through which limited labour market prospects - and in many cases, poverty -

*Corresponding author: Alexander C. O'Riordan
Email address: 18484212@sun.ac.za (Alexander C. O'Riordan)

Stellenbosch University https://scholar.sun.ac.za

become entrenched. Moreover, due to the racial and spatial dimensions of South African poverty, this link also enforces and perpetuates current patterns of inequality.

The link between poor education quality and undesirable labour market outcomes highlights the centrality of education in the context of South African poverty. The extent to which the dichotomous education system perpetuates inequality in South Africa warrants action. Importantly, this action must be guided by rigorous research and well-targeted policy. This is the broad rationale for this thesis which is guided by the need to make contributions to our ability to overcome extant binding constraints to the educational progression and ultimate labour market success of poor South Africans. At the centre of the binding constraints on the education of the poor is limited literacy. Reading must form the basis of any policy designed to improve educational outcomes, specifically early-stage reading skills. It is vital to ensure that students are capable of reading to prevent poor learning outcomes in later grades and the perpetuation of this binding constraint on the education of the poor.

While the education system is the central focus for policymakers, the inability to adequately model and explain the determinants of reading ability is a fundamental constraint on research. And as well designed policy is dependent on accurate research, it is imperative that research in this area be performed as well as is possible. Contemporary research in the field of the economics of education centres on the use of linearly additive production functions. This approach generally assumes that some linear combination of observable variables constitutes the primary component of the data generating process that underlies the analysed educational outcome (literacy ability in this case).

This method is insufficient and inappropriate for several reasons. First, it assumes a constant relationship between all values of the input variables and the outcome variable - it is linear. While linear modelling is a convenient means of deriving a quantitative measure of relationships, it can lead to erroneous conclusions. Moreover, such linear projections are less useful when human behaviour forms a major component of the analysed data generating process. Second, linear modelling limits the theoretical framework of research to an additive production function approach. While this approach is well-grounded in economic theory, it is not the appropriate theory to use in this context. The additive production function approach typically does not make considerations for non-parametric interactions and non-linear relationships which are useful when analysing human behaviour and preferences. Third, the strict insistence on ceteris paribus interpretation is not entirely appropriate in this setting. When human behaviour, preferences, and opinions are major components of what is being studied, it is not useful to know the effect of a change in one input while all others are held constant. In practice, it is almost impossible to change one input without impacting any others. Having ceteris paribus interpretations and forming policy-relevant conclusions from such findings can lead to poorly designed policy. Such policy will be vulnerable to unintended consequences and may not be as effective as the research upon which it is based may indicate. The failings of OLS in this context necessitate an alternative approach, one that is more suited to uncover functional and interactive relationships that can factor in complementarities with other variables. Importantly, the approaches taken in this thesis do not

Stellenbosch University https://scholar.sun.ac.za

ignore ceteris paribus interpretation. Rather, they add to it in the form of functional relationships that consider correlations with other features and interactive effects in the prediction of the outcome variable.

The broad aim of this thesis is twofold. First, it aims to demonstrate that value exists in the use of novel techniques that differ from what is currently common. Second, in doing so, it derives findings that can be useful in guiding both current and future policy and research. Put simply, new methods are applied to uncover possible flaws in methodologies that are more common. The approaches taken in this thesis borrow extensively from methods used in the educational psychological and statistical learning literature. Two broad empirical themes are emphasised - latent constructs (factor modelling) and non-linear modelling.[1] The thesis includes three empirical chapters (2, 3, and 4), each of which considers a unique aspect of the application of non-linear and factor modelling to large education-based datasets. Broadly, each chapter has the following purpose. The first chapter looks at the identification of latent constructs. The second chapter looks at the estimation of latent constructs. The third chapter uses estimated factors (a modelled latent construct) in a non-linear modelling framework.[2]

The latent construct approach is appealing for several reasons, both empirically and theoretically. In large datasets, it is common to find correlations among measured variables (features). This correlation can result from direct or indirect causation or joint dependence on other features. While correlation among two features is relatively easy to understand and does not cause empirical complications, a strong correlation among $N > 2$ features can present a challenge. A grouping of N highly correlated features can contain complex information and interrelationships. In many cases, such clusters of highly correlated features can complicate estimation procedures with multicollinearity and problems of high dimensionality. The latent construct approach provides a useful framework within which to ease the challenges presented by high correlations among features. Moreover, the use of latent constructs enables the application of specifications and procedures of estimation that draw from well established qualitative models. In addition to aiding inference, it enables a better fit of non-linear models through an improved theoretical understanding of the analysed relationships.

The non-linear modelling approach is operationalized mainly by the use of gradient boosted regression, an ensemble-based method that relies on many individual regression trees to estimate model parameters. Moreover, hierarchical cluster analysis and Loess regression are also used. The latent construct approach is operationalized using several methods broadly related to factor modelling and dimension reduction. Methods used include exploratory factor analysis, confirmatory factor analysis, principal component analysis, kernel principal component analysis, and neural network autoencoders. The fundamental premise of the latent construct approach is that measured variables are an observable measure of some unobservable process. For example, a researcher can observe the number of cars

[1] Non-linear modelling in this context includes machine learning algorithms and non-parametric procedures.
[2] These concepts are explained fully in the Methodology section of this chapter.

Stellenbosch University https://scholar.sun.ac.za

owned by a family. This observable measure is revealing the unobservable propensity of that family to own cars. In this case, the propensity to own cars is the latent construct and the observed number of cars is the measured variable. Another way of understanding latent constructs is that they are unobservable processes that determine observable outcomes. From this perspective, the propensity of a family to own cars is what determines how many cars they have. The main idea of factor modelling is to use measured variables to quantify the unobservable processes from which they are derived - a latent construct. An estimated latent construct is called a factor.

The thesis contains five chapters in total: this introductory chapter, a concluding chapter, and three empirical chapters. Each of the three empirical chapters includes a unique piece of research, one that is distinct from the others but ultimately related by the objective of applying non-linear and factor modelling approaches to education-based research. The broad outline of the thesis is as follows. This chapter introduces the thesis, outlines the important concepts, and provides a brief description of each chapter. Chapter 2 investigates the potentially harmful effects of negative item responses using a latent construct framework. The intention of Chapter 2 is to use procedures of factor modelling to investigate whether the nature and design of the survey question can influence the information contained in the derived measured variable. Chapter 3 analyses how the choice of factor estimation technique affects the magnitude of measurement error transmitted to the factor from the underlying features. That is, Chapter 3 compares different methods of dimension reduction with regard to their relative ability to extract meaningful variation from N features while minimising the extraction of noisy variation. Chapter 4 fits a non-linear model using measured features and estimated factors to determine their relationships with one another and with reading performance. While Chapters 2 and 3 are methodological studies, Chapter 4 is an empirical analysis that models and interprets functional relationships that determine reading ability among South African grade 4 students. Chapter 5 concludes the thesis: here the main findings from the thesis are collated and main research and policy contributions are summarized.

The rest of this chapter is structured as follows. First, the overarching methodology is discussed and important definitions are provided. Second, the limitations of the study are discussed. Finally, the abstracts for the three empirical chapters are provided.

2. Broad Research Methodology and Approach

This section outlines several important concepts that are used throughout the thesis. While the Introduction section does briefly define most of the concepts used, it is useful to provide a more substantial description here before using them in later sections of the thesis.

A latent construct is an unobservable process or phenomenon, for example, intelligence, determination, or grit. These concepts cannot be directly measured, but there do exist observable variables that reveal

Stellenbosch University https://scholar.sun.ac.za

information about their existence and magnitude. For example, the achieved result of a maths test can provide an observable measure of innate and unobservable mathematical ability. An important distinction to note is that between a latent construct and a factor. A latent construct is unobservable. A factor is not, it is an estimated measure of an unobservable latent construct. In practice, a factor is estimated using several highly correlated variables that broadly relate to a single latent construct. For example, the results of several exams can be used to estimate a factor that relates to the latent construct of academic ability. Another important distinction to note is that a latent construct is *identified* - a researcher will select certain informationally homogenous measured variables that reflect a specified latent construct. This is the process of identification. Thereafter, a factor is *estimated* using these measured variables and one of many existing methods of factor estimation (Principal Component Analysis, for example).[3]

The approach taken in this thesis emphasises non-parametric methods. One outcome of this is that model estimates do not contain exact quantitative measures of relationships, such as the coefficient estimates of OLS. Importantly, the approach taken here is not designed to do so as the purpose of this work is to uncover and theorise new relationships, investigate the potential value that new methods may provide, and uncover possible inadequacies that exist in common contemporary methods and research. This thesis attempts to add to the literature by presenting research based fundamentally on psychological theory operationalised by methods of statistical learning. Ultimately, this complement of multifarious approaches will be collected and condensed into useful economic and statistical interpretation that will ideally contribute positively to the literature.

3. Limitations

The main limitation to the impact of this study derives from the nature of the data used. When fitting models using cross-sectional data, it is often impossible to confidently assign causality to identified relationships. The findings of this thesis, specifically those of Chapter 4, are to be treated as descriptive. However, by relying on previous literature and extant theory, it is possible to speculate on possible causality. Therefore, while this study is performed with full knowledge of this data-driven limitation, sufficient confidence is placed in these and related previous findings to provide several policy-relevant insights and proposals, as well as contributions to future research.

[3] Another important point to note is that the terms *variable* and *feature* are used interchangeably throughout the thesis. A *variable* is a measured item derived from, in this case, the response to a question on a survey - as is a *feature*. The term feature is more commonly used in the statistical learning literature.

Stellenbosch University https://scholar.sun.ac.za

4. Chapter Abstracts

4.1. Chapter 2

In applied survey-based research, it is common to encounter responses based on both positively and negatively worded questions. In practice, responses are typically recoded to ensure that the numerical values attached to the responses of positively and negatively worded questions are aligned. This is done under the assumption that the responses to negatively worded questions are perfectly reversed reflections of responses to identical or similar positively worded questions - that the variation is inversed. This chapter tests this assumption within a framework of factor modelling. It finds significant differences in the degree to which the different question orientations capture information about single latent constructs that a specific group of questions is designed to capture. And thus, a failure of the assumption.

4.2. Chapter 3

In applied survey-based research, it is common that individual responses are captured with error. This can result from limited memory, misinterpretation or misunderstanding of the questions asked. Moreover, in research that uses large survey-based datasets, procedures of dimension reduction are often employed as an initial step to improve the performance of subsequent estimation procedures. These two phenomena - measurement error and dimension reduction - can combine in such a way that is detrimental to empirical work. This chapter investigates whether or not certain techniques of dimension reduction outperform others with regard to mitigating the effects of measurement error in the features used in the procedure. This is done by leveraging the theory of attenuation bias and the findings of Chapter 2. The results indicate that there are differences in the degree to which certain methods of factor estimation mitigate the effects of measurement error in the underlying features. The analysis yields several findings, perhaps most useful of which is that exploratory factor analysis outperforms the much more commonly used procedure of principal component analysis in terms of the aforementioned metric.

4.3. Chapter 4

This chapter investigates the determinants of reading performance among South African grade 4 students. The approach taken is novel in terms of both the statistical methodology and the theoretical framework. The empirical analysis makes extensive use of gradient boosted regression, a statistical learning technique that enables the analysis of complex nonlinear and interactive relationships. The analysis yields several interesting and policy-relevant findings. First, psychological processes are found to be important predictors of reading performance, specifically negative social interaction and

Stellenbosch University https://scholar.sun.ac.za

self-efficacy. Moreover, the importance of these measured factors is not consistent over the wealth distribution. Second, the importance of the household as a centre for learning is highlighted. Findings indicate that children with parents that are more willing and able to help with their learning tend to perform better than those without helpful or involved parents. Third, the composition of students within the classroom in terms of relative reading ability is a strong predictor of performance. The remedial requirements of some students in the class, likely due to learning backlogs, can negatively impact on the learning process of other students in the class. These findings all lend themselves to possible policy interventions.

Stellenbosch University https://scholar.sun.ac.za

Chapter 2: Negative Item Response Bias in Education-Based Surveys - a Factor Modelling Approach

Abstract

In applied survey-based research, it is common to encounter responses based on both positively and negatively worded questions. In practice, responses are typically recoded to ensure that the numerical values attached to the responses of positively and negatively worded questions are aligned. This is done under the assumption that the responses to negatively worded questions are perfectly reversed reflections of responses to identical or similar positively worded questions - that the variation is inversed. This chapter tests this assumption within a framework of factor modelling. It finds significant differences in the degree to which the different question orientations capture information about single latent constructs that a specific group of questions is designed to capture. And thus, a failure of the assumption.

Keywords: Latent Construct Estimation, Negatively Item Response, Confirmatory Factor Analysis, Hierarchical Cluster Analysis

JEL classification A21, C81, C83, I21, O12

1. Introduction

In survey-based research, data derives directly from individual responses to the items included in a specifically and intentionally designed questionnaire. While it is generally well understood that the sample of respondents needs to adhere to several requirements, such as random selection and broad representation of the population, the same is not true to the same extent for the specified format and orientation of questionnaire response items. This chapter seeks to highlight the importance of the design and orientation (specifically, positive or negative wording) of survey items, and how these questionnaire item characteristics can affect the ultimate outcomes of empirical research. Whether used alone as an explanatory variable in regression, or as an input into a factor modelling procedure, it is vital that the information contained within each derived feature is reliable and accurately reflects the intended component of the specific data generating process as per its design.

It is common for questionnaire construction specialists to include negatively worded items in surveys. Moreover, it is also common to encounter several items, possibly of different design and orientation, that are intended to capture the same underlying process - many questions broadly for one piece of information. This is done, in part, to disrupt response sets and thus maintain active response engagement by the respondent (Marsh, 1986). The premise is that including items that are slightly dissimilar to those preceding it will induce the respondent to pause and think, and thus provide more accurate and holistic information. However, this is done under the assumption that responses to items of either orientation are perfectly aligned with one another - the observable variation equally

Stellenbosch University https://scholar.sun.ac.za

well reflects underlying processes.[1] The broad purpose of this chapter is to test this assumption and uncover a potential source of measurement error - differing question orientations.

The factor modelling approach is ideal for this purpose as its implicit intention is to model unobservable phenomena that determine observable outcomes - the unobservable processes that underlie observable responses. The basic premise of factor estimation is to use a group of informationally homogenous observable features to estimate a specified number of factors, typically a single factor, as is the case in this chapter. Importantly, these procedures provide information on how well each individual feature, within the specified group of informationally homogenous features, fits the single specified factor. Therefore, we are able to quantify the relative strength with which each feature reveals information about the single specified factor. Moreover, this approach differs from contemporaneously common tests for sampling adequacy in that it emphasises group structure in relation to an underlying factor, rather than only the interrelationships within and among the group's individual features.[2]

The analysis employs Confirmatory Factor Analysis (CFA) to test the factor structure of two example latent constructs - self-efficacy and the enjoyment of mathematics. Two samples of response item groups are drawn from the South African 2016 Trends in International Mathematics and Science Study (TIMSS) grade 8 dataset, with each corresponding to one of the two studied latent constructs. CFA is a useful method in that it allows the researcher to impose a factor structure upon an identified latent construct and then test the adequacy of the hypothesised structure (Brown, 2015). That is, identification of the latent construct is discretionary. Methods such as Explanatory Factor Analysis and Principal Component Analysis, while very similar, are used primarily to uncover factor structure rather than test the adequacy of imposed structure. They are therefore less useful in this context. In addition to CFA, a thorough hierarchical clustering exercise is applied to the group of features associated with each latent construct. This is valuable in addition to CFA in that cluster analysis allows the data to speak for itself without the imposition of hypothesised structure or *a priori* design or consideration for underlying latent constructs. It is merely a statistical grouping exercise, one that can demonstrate which features share similar characteristics and which do not. Moreover, traditional measures of sampling adequacy such as Cronbach's Alpha are used. The idea supporting this addition is to investigate whether or not tests that are typically applied in practice will also reveal possible informational differences based on question orientation.

The chapter is structured as follows. First, the theoretical and practical effects of the inclusion of negatively worded items are discussed. Second, the data used is described. Third, the employed methodology is explained. Fourth, the empirical results are provided and discussed. Finally, conclu-

[1] given that the responses to negatively worded items have appropriately been adjusted and recoded to ensure that the scale ordering is comparable.

[2] Sampling adequacy testing is the broad term given to procedures used to test the degree of homogeneity among a group of features (what we are essentially doing in this chapter). It is a common step in procedures of factor estimation and modelling (Cerney & Kaiser, 1977). In this chapter, we compare the results of the Confirmatory Factor Analysis to more typical methods of sampling adequacy in order to uncover any possible differences in the results. As traditional tests of sampling adequacy are generally relied upon in empirical work, any differences in the results would be a concerning finding.

Stellenbosch University https://scholar.sun.ac.za

sions are provided.

2. Theoretical Framework: The use of Negatively Worded Response Items

For the use of features based on negatively worded items (referred to as negative features hereafter) to be justified, their inclusion must first, introduce minimal additional noise in the form of measurement error into the data, and second, ideally provide beneficial effects in the form of response-set disruption and more accurate variation. That is, they must at a minimum have no negative effect and ideally have some positive effect. Using a factor estimation approach, it is possible to empirically investigate whether or not an individual feature is accurately measuring the underlying construct that it is designed to measure. Moreover, we can also use this approach to compare the extent to which individuals in a group of features reflect a single underlying construct. There are two main phenomena by which negative features can have harmful noise inducing effects, specifically when using responses provided by young adolescent students or those with underdeveloped literacy skills (Chiavaroli, 2017).[3] Each is discussed in turn.[4]

First, for positively and negatively worded items to be measuring the same underlying opinions and processes as positively worded items, respondents need to be capable of employing double negative logic. The mental processing required to understand and apply the logic of the English double negative is complex (Hunt, 1978). Such logic requires a relatively high level of verbal reasoning, which may not yet have developed for young students. For example, if a given student does enjoy mathematics, the question "I do *not* enjoy mathematics" requires a response of "I do *not* agree" while the question "I enjoy mathematics" requires a response of "I agree". If the respondent is incapable of applying the appropriate logic, their given responses may capture this inability in the form of measurement error that is essentially impossible to identify by analysing the response in isolation. Therefore, the responses to negatively worded items may be noisy reversed scores of the responses given to similar positively worded items. This noise results from the confusion and uncertainty experienced when answering a negatively worded item and the inappropriate or erroneous application of double negative logic.

The potential for noise, and its distribution within the sample, created by this limited verbal reasoning is exacerbated in the South African context, one in which the distribution of reading ability is extremely unequal (Spaull, 2013). Students in poorer schools generally receive a low-quality education and will have more difficulty applying English double negative logic to appropriately answer a negatively worded item than would their wealthier peers. While not all students in the sample are tested in English, it is assumed that the double negative logic required in other testing languages is similarly distinct from that used when interpreting positively worded text. Therefore, if it is true that negatively worded items do induce noisy responses, their inclusion in questionnaires could induce a systematic

[3] In the South African context, in many cases there exists a disconnect between student age and literacy skills. Many are several years behind the level of development typically associated with their age (Spaull & Hoadley, 2017)
[4] See Wong & Rindfleisch (2003); Weijters & Baumgartner (2013); and Hartly (2014).

Stellenbosch University https://scholar.sun.ac.za

source of bias into the data in the South African case.

The second phenomenon is relatively simple, It might be the case that respondents are simply ignorant to the nuance of negatively worded items and not notice their distinction (MacDonald, 2013). In this case, response scores will be perfectly reversed versions (once recoded) of the response scores for positively worded items. If this is true, the proposed positive effect of negatively worded items will be eliminated. However, and more importantly, if these responses are *exactly* opposite to those intended by the orientation of the item, as seems plausible, the recoding (reversing of the scores) of the relevant features should completely remove this inaccurate measurement. Conversely, if the aforementioned first phenomenon does occur, it is not necessarily the case that simply recoding the data will solve the problem. Moreover, the second phenomenon will not cause systematic bias as will the first. Therefore, of these two phenomena, the first is more concerning and is the central point of interest of this chapter.

3. Data

3.1. Description of Data

The South African 2016 Trends in International Mathematics and Science Study (TIMSS) grade 8 dataset is used for the empirical analysis (N = 10370). Specific features are selected to estimate two factors - one, a measure of the enjoyment of mathematics, the other, a measure of positive self-efficacy toward mathematics. TIMSS is preferred for this chapter due to its extensive and available student questionnaire information. Moreover, the grade 8 dataset is preferred to the grade 4 dataset as the slightly older students should be better suited for the purpose of this study. The two specific studied factors are chosen because each has nine features that can be used in their identification. Among the nine features used to estimate the enjoyment factor, two are negative. Of the nine features used to estimate the efficacy factor, five are negative.

The definition of a negatively worded item used here is relatively simple. These are items that, for a given response to have the same intended meaning, must be answered with an opposite scoring than would be used when answering a similar positively worded item. For example, if a student thoroughly enjoys mathematics, she would answer the following two example questions - 1) I do enjoy mathematics, 2) I do *not* enjoy mathematics - as follows. Question 1 would be answered with a response of "I agree", while question 2 would be answered with a response of "I do *not* agree". Importantly, on the Likert type scale used in the TIMSS questionnaire, these responses would be on opposite ends of the response scale. Question 2 is the negatively worded item and question 1 is the positively worded item.

Here it is again valuable to note an important distinction. The main premise of this chapter is not to investigate whether the inclusion of negative features will negatively affect the estimation of factors. It obviously will, the estimation procedures used are not able to distinguish between the responses to items of different orientation and design. They see only the responses to items in numerical terms and are based broadly on the correlation structure that exists among a specified group of features. Therefore, negative features must be recoded to reverse the numerical order of the values attached

Stellenbosch University https://scholar.sun.ac.za

to the qualitative responses. Importantly, this chapter does not investigate whether ignorance of this necessary recoding of features will affect factor estimation. Rather, it investigates the effect of appropriately recoded negative features.

3.1.1. Latent Constructs - Enjoyment of Mathematics and Self-Efficacy

Each of the two factors is estimated using nine features, each of which is based on a response item that pertains to the self-reported degree of either the enjoyment of mathematics (first factor), or a measure of mathematical self-efficacy (second factor). The items are answered along a Likert-type scale that ranges from 1 to 4 as follows; 1 - Agree a lot, 2 - Agree a little, 3 - Disagree a little, 4 - Disagree a lot. Table 3.1 lists the nine items from which the features used to estimate the enjoyment factor are derived. It also provides the code of the feature used in the analysis that follows. A code with a "p" indicates a positive feature while an "n" indicates a negative feature. Table 3.2 lists the nine items from which the features used to estimate the efficacy factor are derived.

Code	Item
p1	I enjoy learning mathematics
n1	I wish I did not have to study mathematics
n2	Mathematics is boring
p2	I learn many interesting things in mathematics
p3	I like mathematics
p4	I like any schoolwork that involves numbers
p5	I like to solve mathematics problems
p6	I look forward to mathematics class
p7	Mathematics is one of my favorite subjects

Table 3.1: Enjoyment Factor - Features and Codes

Code	Item
p1	I usually do well in mathematics
n1	Mathematics is more difficult for me than for many of my classmates
n2	Mathematics is not one of my strengths
p2	I learn things quickly in mathematics
n3	Mathematics makes me nervous
p3	I am good at working out difficult mathematics problems
p4	My teacher tells me I am good at mathematics
n4	Mathematics is harder for me than any other subject
n5	Mathematics makes me confused

Table 3.2: Efficacy Factor - Features and Codes

3.2. Correlation Analysis

Figure 3.1 displays correlation heatmaps for the nine features included in the two factor estimation procedures. The top two plots are for the efficacy factor, the lower two, the enjoyment factor. The

Stellenbosch University https://scholar.sun.ac.za

figures to the left show the heatmaps for features that have not been recoded while the figures to the right include negative features that have been recoded. Positive features are noted with a "p", negative features are noted with an "n". A darker blue shade indicates a stronger positive correlation while a darker red shade indicates a stronger negative correlation.

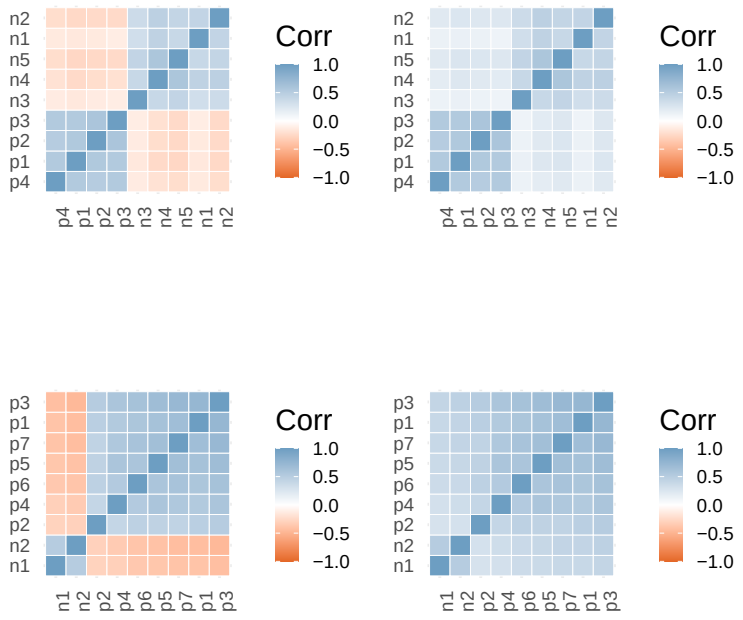


Figure 3.1: Correlation Heatmap - Efficacy Construct (Top) and Enjoyment Construct (Bottom)

From the top two plots of figure 3.1, it is evident that the negative features are negatively correlated with the positive features before being recoded. This is a simple representation of the aforementioned assertion that responses to similar items of differing orientation will be on opposite ends of the response scale. Viewed in isolation, this does reveal that most students in this sample are aware, to a certain extent, of the negative orientation of these items, thus indicating that the second phenomenon of ignorance to negatively worded items is not pervasive. A more interesting insight is revealed by the figure to the top right, which plots the correlations including the recoded negative features. It is evident that, even once recoded, the negative features remain clustered together and only weakly correlated with the positive features while being relatively strongly correlated with only one another.[5] Therefore, both the magnitude of correlation and the groupings made by the clustering algorithm indicate that these features are characteristically different to the others. That is, responses to items of differing orientation appear to be imperfectly aligned.

[5]The ordering of the features on the heatmaps in figures 3.1 is determined by a clustering algorithm. Therefore, features that demonstrate similar characteristics are positioned near to one another on the heatmap.

Stellenbosch University https://scholar.sun.ac.za

From the bottom two plots of figure 3.1, it is evident that the negative features are clustered together and negatively correlated with the positive features. Again, the more interesting finding is that once recoded, the negative features remain clustered together. Moreover, they are correlated strongly only with one another. If all nine features, associated with either factor, are truly capturing the same latent construct, there should be no noticeable difference in the degree of correlation or the grouping by the cluster analysis.

This initial look at the data reveals that the recoding of negative features may not be sufficient. The features with different orientations do appear to be consistently different from one another. In the following section, a more extensive clustering exercise is applied to the data to better understand the group structures that exist. Importantly, clustering can provide information about group structure and which features are similar to one another along certain dimensions. It does not at all provide information about factor structure and the relative ability of a group of features to reflect a specific latent construct or estimate a specific factor.

3.3. Hierarchical Cluster Analysis

Clustering is a statistical partitioning technique that groups a set of features based on their characteristics (Maimon & Rokach, 2010). In this case, correlation clustering is used, the correlation coefficient is the relevant measure of similarity between variables. Broadly, hierarchical clustering differs from other methods such as k-means clustering in that the number of final clusters is not specified *a priori*. Hierarchical clustering can be subdivided into two broad categories, agglomerative nesting (AGNES) and divisive analysis (DIANA). AGNES is a bottom-up approach in which each feature initially represents a distinct cluster. These individual clusters are iteratively merged into larger clusters until only a single large cluster exists and the full hierarchical structure is obtained. DIANA is a top-down approach that works in the exact opposite manner. The procedure starts with a single cluster which encompasses all of the features. This single cluster is iteratively sub-divided until each feature is its own distinct cluster (Maimon & Rokach, 2010).

The technique used here is a complete linkage agglomerative nesting hierarchical cluster algorithm based on Euclidean distances within the correlation matrix. Complete linkage is based on maximum inter-cluster dissimilarity, that is, the similarity of two clusters is the similarity of their two most dissimilar members. In simpler terms, in an agglomerative approach in which the algorithm starts with N clusters, the first two clusters to merge will be the two that are most similar, the two that are the closest by euclidean distance. There are now *N-1* clusters in total with one cluster containing two features. Let's assume that the next iteration results in the two features closest to the cluster formed in the first iteration to merge into one. There are now two clusters that each contain two features and *N-4* clusters that contain only one feature. The concept of complete linkage is that the distance between these two clusters each with two features is a measure of the distance between the two features that are the furthest apart from one another within the two respective clusters. Therefore, the next iteration of cluster merging is determined by the relative closeness in Euclidean distance of the two most dissimilar features within each cluster. This iterative process will continue until one all-encompassing cluster exists.

Stellenbosch University https://scholar.sun.ac.za

3.3.1. Enjoyment Latent Construct

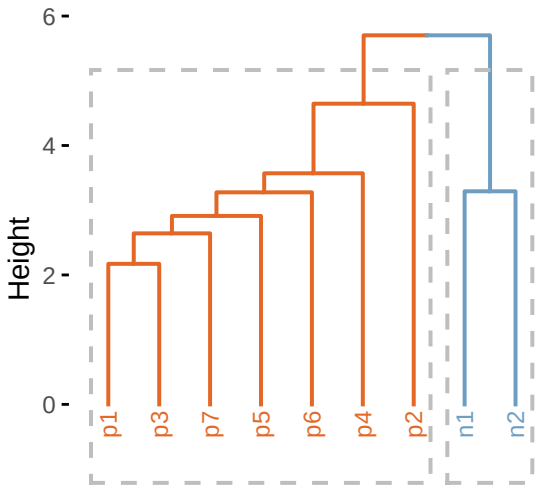


Figure 3.2: Enjoyment Latent Construct - Dendrogram

Figure 3.3 displays the dendrogram for the results of the cluster analysis of the nine features used to estimate the enjoyment factor. It is evident that the two negative features are grouped separately from the seven positive features. Boxes (partitions) are drawn around the dendrogram at a cutoff specified at two clusters. The cutoff need not be two, however, this serves to demonstrate the main partition in the data which separates the features associated with positively and negatively worded response items. This box partition reveals how the data would have been clustered if two final clusters were explicitly specified.

It is worth referring back to the procedure of agglomerative nesting, which is bottom-up. In creating partitions in the data, the clustering algorithm kept the features associated with positively and negatively worded items separate from one another, joining them only at the final iteration in which the final two clusters are forced to merge into one. Therefore, it is along the positive-negative orientation that the features are most dissimilar. Another meaningful insight can be gained by comparing the heights at which clusters are iteratively joined. The height measure indicates the relative distance between the clusters on the two branches joined at that particular node. It is evident that the two negative features join one another only after p1, p3, p7, p5, and p6 have merged into one cluster. This indicates that the two negative features are more dissimilar to one another than are the first five most similar positive features. **Therefore, not only are negative features dissimilar to the positive features, they are also relatively dissimilar to one another.**

This significant separation created by the clustering algorithm is demonstrated more extremely by the cluster plot in figure 3.4. From Figure 3.4 it is evident first that there exist two main clusters, one

Stellenbosch University https://scholar.sun.ac.za

comprised entirely of positive features, the other, entirely of negative features.[6] More importantly, these two clusters are far apart, and in clustering terms, highly dissimilar. While cluster 1 appears upon initial inspection to be large and spread out, with the exception of p2, its points are relatively close to one another along the x-axis, the dimension upon which the majority of the information exists. It is only feature p2 that shows significant dissimilarity to the other positive features. Referring back to Table 3.1, it is evident that this particular positive item is slightly dissimilar to the others. Both clusters are compact and relatively far from one another. **Features derived from items with the same orientation are similar to one another while being dissimilar to features derived from items with the opposite orientation.** Therefore, these features could be considered to have different characteristics.

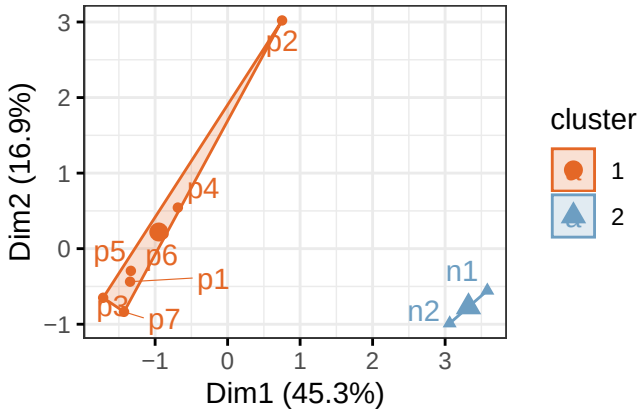


Figure 3.3: Enjoyment Latent Construct - Cluster Plot

3.3.2. Efficacy Latent Construct

The efficacy construct differs from the enjoyment construct in that it incorporates a larger number of negative features. Therefore, a cluster analysis applied to the features used for its identification can provide additional results that are not possible with features of the enjoyment construct. Figure 3.5 displays the results of the clustering algorithm in the form of four dendrograms. While the dendrograms themselves are identical, each of the four differs in the specified number of box partitions. There are specifications of 2, 3, 4, and 5 box partitions. Again, these box partitions reveal what the individual components of n clusters would be if n final clusters were to be specified.

[6] Note here that two clusters are explicitly specified. Therefore, it is not the number of clusters that is of interest, rather, it is the composition of each cluster that is important.

Stellenbosch University https://scholar.sun.ac.za

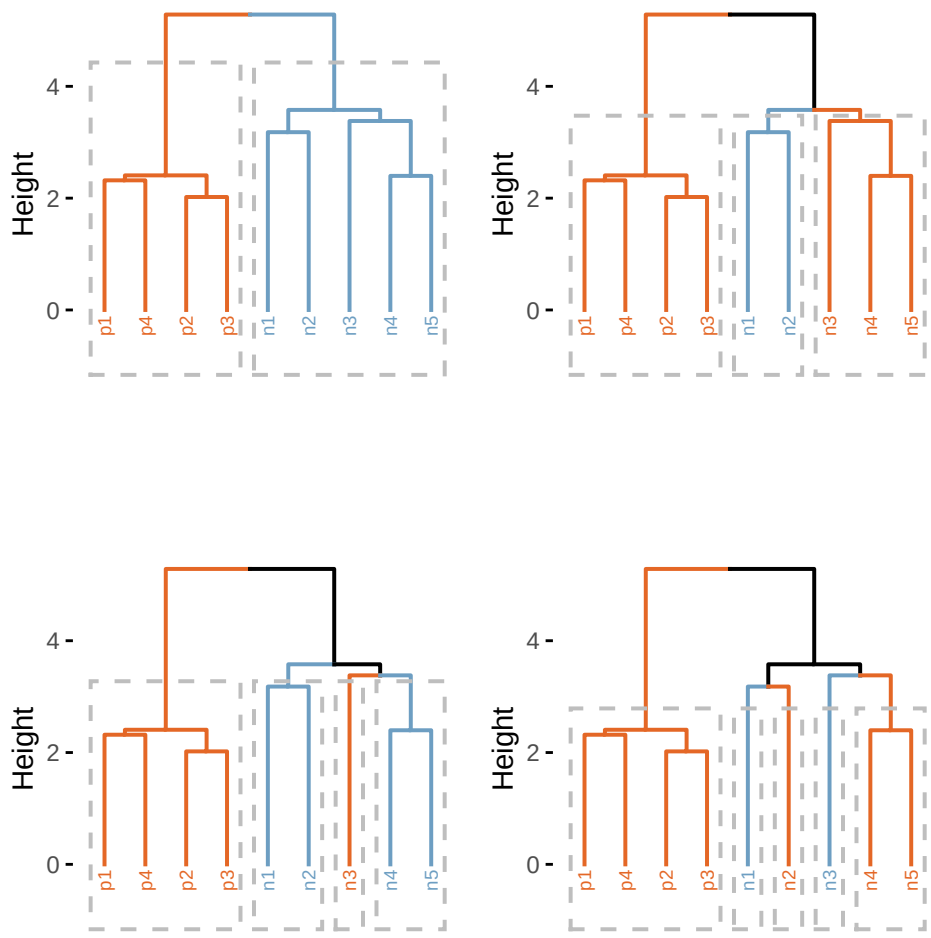


Figure 3.4: Efficacy Latent Construct - Dendrogram

From the dendrograms in figure 3.5, it is evident that the positive features remain in a single cluster, regardless of the box partition specification. This indicates that strong similarity exists between these features. Conversely, the negative features are split into separate clusters at each successively higher box partition specification. Therefore, within the initial single cluster that contains only negative features, the individual features are relatively dissimilar. This finding is corroborated by the height at which the four positive features merge into a single cluster. It is evident that the four positive features merge into a single cluster before even the first cluster containing two negative features is merged. What this indicates is that the two most dissimilar positive features are more similar to one another than are the two most similar negative features. This finding is in itself interesting. It indicates that not only are negative features dissimilar to positive features, they are also relatively dissimilar to one another. **Therefore, the information provided by the group of negative features appears to be internally inconsistent.** The results shown in Figure 3.6, which have the same interpretation as those in Figure 3.4, further corroborate the interpretation that negative features are dissimilar to

Stellenbosch University https://scholar.sun.ac.za

positive features, and are also relatively dissimilar to one another.

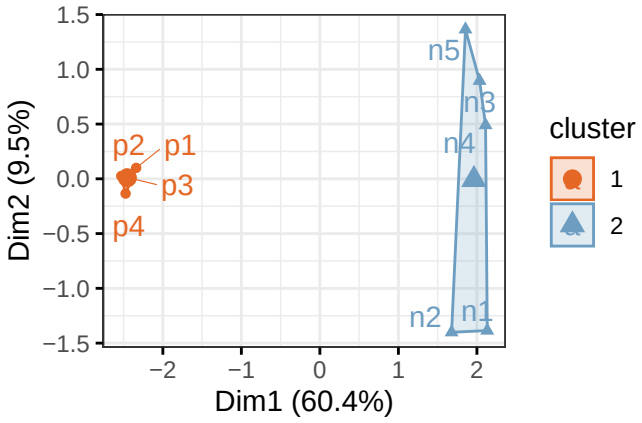


Figure 3.5: Efficacy Latent Construct - Cluster Plot

4. Methodology

4.1. Confirmatory Factor Analysis (CFA)

CFA is a statistical technique used to test and verify a proposed factor structure among a group of features (Suhr, 2006). It achieves this by estimating the common sources of covariance among N features (limited to $N = 1$ in this chapter). The implicit assumption underlying the use of CFA is that the observable features are determined by a number of unobservable processes. Simply, observable features are outcomes of unobservable latent constructs. For example, a student will answer a question of mathematical enjoyment positively if she does enjoy mathematics. In this way, the positive response that we observe is an outcome of her innate unobservable enjoyment of mathematics.

In contrast to Principal Component Analysis, which explains maximum total variance, CFA explains maximum shared variance (covariance) among a set of features (Babyak & Green, 2010). Moreover, CFA makes a distinction between variance that is common to all N features (shared variance), and that that is unique to each feature (idiosyncratic variance). Therefore, CFA is a correlation-focused approach in which factors represent the common variance of N features, and the variance not explained is defined as feature-specific (idiosyncratic) variation.

CFA allows one to test for the existence of an *a priori* specified relationship among a set of features and their proposed common underlying latent construct. It does this by assigning factor loadings to each employed observable feature. These loadings are measures of the degree to which individual features are determined by estimated factors - how much does each unobservable construct influence each observable feature. Therefore, if feature y loads highly onto factor x, we can infer that the observable response to the item from which feature y is derived is largely determined by the latent

Stellenbosch University https://scholar.sun.ac.za

construct underlying, and represented by, the estimated factor x. Moreover, if features y, w and b all load highly onto factor x, we can infer that these three features share a common underlying latent construct - they are determined by the same underlying process or phenomenon. As indicated, in this chapter, we investigate whether or not the observable features derived from positively and negatively worded items are equally determined by the same underlying unobservable latent construct.

The empirical procedure underlying CFA is based primarily on the Common Factor Model (Thurstone, 1947). The Common Factor Model is premised on the notion that there exist two types of latent constructs that influence observed item responses, and their derived features - shared and idiosyncratic (MacCallum, 2009). The CFA procedure models features as a linear function of shared and idiosyncratic influence (variance) by a specified number of factors. The following outlines the fundamental equation of CFA as originally proposed by Joreskog (1967).

$$\mathbf{y} = \mathbf{\Lambda x} + \mathbf{z} \tag{1}$$

The $\mathbf{y}$ vector contains the N observable features. $\mathbf{x}$ is a vector of m common factor scores, the single unobserved latent construct in this case ($m = 1$). $\mathbf{z}$ is a vector of N unique scores - idiosyncratic variance. $\mathbf{\Lambda}$ is an $N \times m$ matrix containing the factor loadings for each feature. From the above, it is evident that CFA decomposes each feature y into variance that is shared, and variance that is unique. Therefore, the results of CFA reveal the degree to which each of the N observed features is influenced by the unobserved common factor. The above fundamental equation depends on three critical assumptions.

$$E(x) = E(z) = 0 \tag{2}$$

$$E(xx') = \mathbf{\Phi} \tag{3}$$

$$E(zz') = \mathbf{\Psi} \tag{4}$$

The dispertion matrix of y is defined as

$$E(yy') = \mathbf{\Sigma\Lambda\Phi\Lambda'} + \mathbf{\Psi} \tag{5}$$

The equations listed above succinctly describe the fundamental premise of CFA. The dispersion matrix of y demonstrates the diagonalization procedure used to obtain the factor loadings, $\mathbf{\Lambda}$. The estimation of CFA is performed by maximum likelihood, with the maximization of the following likelihood function

$$F_{ml} = ln|\mathbf{\Sigma}| + tr(\mathbf{S\Sigma}^{-1}) - ln|\mathbf{S}| - m \tag{6}$$

Where matrix $\mathbf{S}$ contains estimates of variances and covariances of the components of y.

5. Empirical Results

In this chapter, CFA is used to test hypotheses about factor structure using different model specifications based on feature orientation. Therefore, it is useful to think in terms of restricted and unrestricted

Stellenbosch University https://scholar.sun.ac.za

models. In this case, an unrestricted model is one that contains all of the features associated with each factor, including the recoded features based on negatively worded items. The restricted models are those that contain a limited number of features of a single orientation. In total, five CFA models are estimated. Two models are based on the enjoyment factor and three models are based on the efficacy factor. The two models fit to the enjoyment factor are: one unrestricted model that uses all nine features, and one restricted model that uses only the seven positive features. Of the three models fit to the efficacy factor, one is unrestricted and two are restricted. Of the two restricted models, one uses only positive features while the other uses only negative features. Again, the unrestricted model uses all of the features. The performance of the individual features is investigated using measures of fit and the factor loadings. Appendix B contains information about the employed measures of fit.[7]

5.1. Estimated Factor - Enjoyment of Mathematics

Table 5.1 displays the measures of fit for the one restricted and one unrestricted CFA model fit to the features of the enjoyment factor. It is evident that the restricted model outperforms the unrestricted model across every measure of fit. This finding indicates that the model which includes only positive features outperforms the model that includes features of both orientations, with regard to estimating the factor of mathematical enjoyment. In line with the findings of earlier sections of this chapter, this result indicates that the inclusion of negative features can be detrimental to factor estimation and that their inclusion in construct identification should be done with caution. That is, the two distinct feature orientations do not equally well reflect the underlying construct of mathematical enjoyment.

Fit_Measure	Unrestricted	Resctricted
Comparative Fit Index	0.962	0.988
Tucker-Lewis Index	0.949	0.983
Loglikelihood user model	-106906.5	-78618.96
Akaike Information Criterion	213849.01	157265.92
Bayesian Information Criterion	213922.25	157322.89
Root-Mean square error	0.080	0.057
Standardized Root Mean Square Residual	0.040	0.017
model chi-square	1838.670	489.535

Table 5.1: Enjoyment CFA - Measures of Fit

Table 5.2 displays the factor loadings of the features onto the enjoyment factor. Results for both the unrestricted and restricted models are shown. As is common, the loading of the first feature is scaled to 1. All loading scores are statistically significant though this information is not shown. From

[7] In addition to the CFA results, Appendix A includes the results for several traditional measures of sampling adequacy - Chronbach's Alpha (both raw and standardized), Guttman's Lambda, and the Kaiser-Meyer-Olkin measure (Guttman, 1945; Cronbach, 1951; Kaiser & Rice, 1974). From the results of these tests, presented in tables 8.1 and 8.2, it is evident that they reveal no distinction between positive and negative features - that they are informationally identical. This is true for both constructs. Moreover, the overall measures, those that are used to test the sampling adequacy of a group of features as a whole, indicate that each test measure is well within the acceptable range. This is highly concerning as, in practice, a researcher would generally apply these tests and base their subsequent actions on their results alone. However, these results do ensure that the findings of this chapter take on increased significance.

Stellenbosch University https://scholar.sun.ac.za

Table 5.2 it is evident that the two negative features have a similar and relatively weak loading onto the enjoyment factor. This indicates that these two features are determined to a lesser degree by the underlying construct of mathematical enjoyment than are the positive features. Interestingly, the feature coded as p2, which is based on the "I learn many interesting things in mathematics" item has the lowest loading. Recall this particular feature was also the most dissimilar positive feature according to the cluster analysis. This is likely due to the slightly different nature and design of the item which does not relate as directly as do the others to the actual enjoyment of mathematics. The findings shown in Table 5.2 indicate that the negative features do not load onto the enjoyment factor as well as the other features do, with the exception of feature p2. Therefore, it appears that there exists a weaker relationship between the negative features and the underlying latent construct of mathematical enjoyment when compared to the positive features. Again, the two distinct feature orientations do not equally well reflect the underlying construct of mathematical enjoyment.

Factor_Loadings	Unrestricted	Resctricted
p1	1.000	1.000
n1	0.839	-
n2	0.830	-
p2	0.743	0.742
p3	1.218	1.219
p4	1.004	1.009
p5	1.150	1.153
p6	1.084	1.083
p7	1.335	1.335

Table 5.2: Enjoyment CFA - Factor Loadings

5.2. Estimated Factor - Self-Efficacy

Table 5.3 shows the measures of fit for the one unrestricted and two restricted CFA models fit to the efficacy factor. The interpretation of the results is the same for those shown in table 5.1. The two restricted models are, in this case, one that contains only positive features and one that contains only negative features. In this case, the unrestricted model performs significantly worse than the two restricted models. An interesting characteristic of this model to consider is the relatively equal share of positive and negative features. The poor performance of the unrestricted model seems to corroborate the story that positive and negative features do not equally reflect the same underlying latent factor, noting again that the model is specified with a single factor. **Therefore, the poor performance is possibly the result of forcing the model to estimate a single factor when the features used in the model are determined by two distinct underlying constructs captured by items of different orientation.** This narrative is supported by the superior measures of fit for the two restricted models.

Stellenbosch University https://scholar.sun.ac.za

Fit.Measure	Unrestricted	Resctricted_Pos	Resctricted_Neg
Comparative Fit Index	0.655	0.995	0.978
Tucker-Lewis Index	0.540	0.984	0.956
Loglikelihood user model	-125701.29	-49829.69	-71689.85
Akaike Information Criterion	251438.58	99675.39	143399.7
Bayesian Information Criterion	251511.81	99707.95	143440.39
Root-Mean square error	0.191	0.062	0.076
Standardized Root Mean Square Residual	0.140	0.013	0.025
model chi-square	10220.532	80.542	301.142

Table 5.3: Efficacy CFA - Measures of Fit

Comparing only the two restricted models allows for further interpretations. While both restricted models outperform the unrestricted model on every measure, the restricted model with only positive features strongly outperforms the restricted model with only negative features. Here it is important to remember that we cannot explicitly specify the factor that is being estimated, we are only able to use theory to identify the latent construct and specify the number of factors. The CFA then estimated the single most important factor. However, if we assume that both sets of differently orientated items are designed to measure self-efficacy, which does seem very plausible, it appears that negative features perform significantly worse than do positive features with regard to reflecting the underlying construct of self-efficacy. This finding again corroborates previously posed interpretations that recoded negative features are a noisy reflection of positive features, and their grouping provides internally inconsistent information. This is possible evidence of confusion in response that is caused by negatively worded items.

Table 5.4 shows the factor loadings for one unrestricted and two restricted CFA models fit to the efficacy factor data. The interpretation of the results is the same for those shown in table 5.2. As before, the orientation of the features is revealed by the codes containing either “p” or “n”. From the unrestricted model, it is evident that all five negative features load onto the efficacy factor less strongly than do the four positive factors. In some cases, such as n1 and n3, this loading is significantly weaker. The restricted models, which each include only factors of a single orientation, have much higher factor loadings on average.

This finding indicates that, when features are separated into their distinct orientations, the CFA models perform better. When viewed in combination with earlier findings in this chapter, the results in Table 5.4 appear to indicate that the combination of positive and negative features is harmful to the estimation of a single factor. That is, positive and negative features do not reflect the same underlying latent construct equally well. A stronger interpretation is that items of different orientations are not capturing the same underlying latent construct, the one may be capturing a noisy version of the other. It is a superior strategy to use features based on only one orientation of response items. A bolder interpretation would be that only positive features should be used. Importantly, One cannot generalise these findings indiscriminately. However, these findings do demonstrate that researchers should approach empirical work and factor modelling with an *a priori* understanding that similar issues may exist in the data that they plan to use.

Stellenbosch University https://scholar.sun.ac.za

Factor_Loadings	Unrestricted	Resctricted_Pos	Resctricted_Neg
p1	1.000	1.000	-
n1	0.588	-	1.000
n2	0.857	-	1.144
p2	1.123	1.159	-
n3	0.594	-	0.994
p3	1.157	1.211	-
p4	1.157	1.192	-
n4	0.871	-	1.416
n5	0.899	-	1.320

Table 5.4: Efficacy CFA - Factor Loadings

6. Conclusion

The results contained in this chapter indicate that homogenous positive and negative features differ with regards to their ability to capture information of the same underlying latent constructs. That is, the two distinct orientations of questionnaire response items do not illicit identically aligned responses. Rather, the responses to negatively worded items appear to be noisy reversed reflections of those to positively worded items. This finding is of particular significance in the field of latent construct estimation, where the accurate estimation of factors and capturing of underlying latent constructs is essential for appropriate inference. Moreover, this chapter finds that traditional measures of sampling adequacy fail to reveal this informational distinction based on response item orientation.

These findings reveal that the common practice of including negatively worded response items in survey questionnaires should be done with caution. Moreover, the findings reveal that the equally common practice of recoding features based on negatively worded items and continuing with subsequent estimation procedures may result in biased outcomes. This bias will result from the measurement error contained in features that, due to the nature and orientation of the item upon which they are based, do not adequately capture the information that they are intended to capture as per their design. Therefore, researchers should carefully consider the effects that the use of such features could have before proceeding with estimation. These results appear robust, at least in the South African context. The central findings are consistent across basic correlation analysis, hierarchical cluster analysis, and confirmatory factor analysis. However, one shortcoming of the research of this chapter is the limited focus on only the mathematical performance of South African grade nine learners, and the use of only two example latent constructs.

There exists vast scope for future research that can leverage the methods used, and the results found, in this chapter. For example, further study can include a wider array of example latent constructs, or the complexity of response item wording can be adjusted. Another productive area to research would be cross-country comparisons. South Africa has a very unique education system, therefore, one cannot easily generalise results found in the South African context. Furthermore, research can better identify the source of the dissimilarity between responses to items of different orientations. There is much more that can be done with this research direction. Importantly, the findings from such research will have practical implications that can easily be incorporated into future work.

Stellenbosch University https://scholar.sun.ac.za

7. Appendix A: Measures of Sampling Adequacy

Code	Raw_Alpha	Std_Alpha	Lambda	KMO
p1	0.88	0.88	0.88	0.93
n1	0.90	0.90	0.89	0.89
n2	0.89	0.90	0.89	0.90
p2	0.89	0.89	0.89	0.96
p3	0.87	0.87	0.87	0.91
p4	0.88	0.89	0.89	0.95
p5	0.88	0.88	0.88	0.94
p6	0.88	0.88	0.88	0.95
p7	0.87	0.88	0.88	0.93
Overall	0.89	0.9	0.9	0.93

Table 7.1: Enjoyment Construct - Measures of Sampling Adequacy

Code	Raw_Alpha	Std_Alpha	Lambda	KMO
p1	0.78	0.78	0.80	0.86
n1	0.79	0.80	0.81	0.85
n2	0.78	0.78	0.80	0.88
p2	0.78	0.78	0.80	0.84
n3	0.80	0.80	0.82	0.87
p3	0.79	0.78	0.80	0.83
p4	0.79	0.79	0.80	0.85
n4	0.78	0.78	0.80	0.82
n5	0.77	0.78	0.80	0.83
Overall	0.8	0.81	0.83	0.85

Table 7.2: Efficacy Construct - Measures of Sampling Adequacy

The value in analysing these traditional and more commonly used tests of sampling adequacy lies in their comparison with the novel methods used in this chapter. In practice, researchers will generally perform one or more of the above tests to investigate whether or not each feature within a group of features derived from similar response items contains homogenous information (Cerney & Kaiser, 1977). It is a common step in factor modelling. A practical example is asset index creation. It is common for researchers to input several features based on response items regarding asset ownership into a PCA, and use its first component as a feature in subsequent analysis. Before running this PCA, it is common to perform a KMO test for sampling adequacy to investigate whether or not the proposed grouping of features is appropriately homogenous for use in PCA (Kaiser, 1974; Cerney & Kaiser, 1977). If the KMO statistic is within an acceptable range, the researcher can be assured that their proposed grouping of features can be used in PCA, and that the use of the first component in the subsequent analysis is justified and appropriate.[8] The test results in Tables 8.1 and 8.2, and results throughout the chapter as a whole, indicate that this approach may be flawed.

[8]Note here that the definition of this acceptable range is inconsistent. Some sources indicate that a value greater than 0.5 is sufficient (Cerney & Kaiser, 1977). However, it is more commonly accepted that values between 0.75 - 0.9 are adequate.

Stellenbosch University https://scholar.sun.ac.za

From the results in Tables 8.1 and 8.2, it is evident that all nine features used to identify either of the two latent constructs are sufficiently homogenous according to all included tests for sampling adequacy. The results in these tables indicate that using these features in procedures of factor estimation, or interchangeably in regression, is appropriate and statistically justified. Simply, the traditional tests of sampling adequacy find no internal inconsistencies among either group of nine features. Importantly, inference regarding the use of either group of features as a whole is done using the overall measure. Each overall measure, for each group and test statistic, is above 0.8. Interestingly, the overall values for the group of features associated with the efficacy factor are lower than those of the enjoyment factor, noting that this grouping contains more negative features. However, viewed in isolation, the overall test statistics from the efficacy feature grouping do not raise concern. Therefore, these overall test statistic values are misleading when we consider the findings in the other sections of this chapter.

The misleading nature of these overall test statistic values forms part of the rationale for the work done in this chapter. Relying on these measures of sampling adequacy alone may be inappropriate in certain circumstances. The typical process of estimating a selected test of sampling adequacy and then continuing with regression or factor estimation is not sufficient, a more thorough examination of data should be performed. Furthermore, the test statistic values for the individual features reveal no distinction between the measures of sampling adequacy for features derived from items of different orientation. From these tests alone, there is no notable evidence that features may be informationally different from one another based on their orientation.

Stellenbosch University https://scholar.sun.ac.za

8. Appendix B: Measures of Fit

8.1. Model Test Statistic

This is the most basic test statistic of the CFA model, several other fit measures are based on some function of this measure. It is a test statistic derived as a function of the sample size and the fit function (F_{ml}). Therefore, the test captures the dispersion matrix, the variance-covariance matrix, and a trade-off with the sample size. Therefore, it is a measure of the overall fit and the discrepancy between the sample and fitted covariance matrices. It is a test for perfect fit, with a smaller value indicating a better fit. A value of zero would indicate a perfect fit. A perfect fit would indicate that the variance of all included features is perfectly explained by a single factor of underlying covariance. The model test statistics is calculated as

$$T = (n-1)F_{ml} \tag{7}$$

In asymptotically large samples, and given a sufficiently large m, T follows a χ^2 distribution with degrees of freedom equal to the number of unique variance and covariance in the variance-covariance matrix of the observed variables. The degrees of freedom are calculated as follows, where q is the number of freely estimated parameters and m is the number of features.

$$df = \frac{m(m+1)}{2} - q \tag{8}$$

8.2. Comparative Fit Index

The comparative fit index, as the name suggests, can only be used to compare the fit of two competing models, it is not an absolute measure. The Comparative Fit Index compares the fit of the estimated model with that of the null model. The null model, in this case, is the model with the worst possible fit. In the null model, features have zero covariance. The null model would be the model with the maximum possible Model Test Statistic described above. Therefore, it is expected that the estimated model will have a better fit than the null model, the Comparative Fit Index is a measure of how much better the estimated model is than the null model. A higher value is prefered. The Comparative Fit Index is calculated as

$$CFI = \frac{(T_{null} - df_{null}) - (T_{estimated} - df_{estimated})}{T_{null} - df_{null}} \tag{9}$$

8.3. Tucker-Lewis Index

The Tucker-Lewis Index is a comparative fit index that improves upon the omitted Bentlre-Bonett index in that it penalises additional parameters. Therefore, it is a fit index that favours a more parsimonious model. The Tucker-Lewis index, as well as the abovementioned Comparative Fit Index, provides information on how the estimated model fit improves upon the fit of the null model. A fit

Stellenbosch University https://scholar.sun.ac.za

measure of 0.95, for example, indicates that the estimated model improves upon the fit of the null model by 95%. The measure depends on the average correlations among the set of features. If the average correlation is low, the fit measure will be small. Therefore, a large value is prefered. The Tucker-Lewis Index is calculated as

$$TLI = \frac{(\frac{T}{df})_{null} - (\frac{T}{df})_{estimated}}{\frac{T}{df}_{null}} \tag{10}$$

8.4. Root-Mean Square Error

The Root-Mean Square Error is an absolute fit measure that assesses the lack of fit of a model. If two individual models are run, one a restricted version of the other, this measure can then be used to compare the relative fit of the two estimated models. As this measure indicates the lack of fit, a smaller value is prefered. A smaller value is prefered. The Root-Mean Square Error is calculated as follows, where N is the number of observations.

$$RMSE = \sqrt{\frac{(\frac{T_{estimated}}{N-1})}{df_{estimated}} - \frac{1}{N-1}} \tag{11}$$

8.5. Standardized Root Mean Square Residual

The Standardized Root Mean Square Residual is an absolute measure of fit, it is defined as the square-root of the difference between the residuals of the sample covariance matrix and the hypothesized model. It can be interpreted as the average standardized residual covariance (Maydeu-Olivares. 2017). Therefore, it is a measure of the covariance that remains after the influence of the single factor has been partialed out. A smaller value is preferred. The Standardized Root Mean Square Residual is calculated as

$$SRMR = \sqrt{\frac{S}{\frac{m(m+1)}{2+m}}} \tag{12}$$

8.6. Log-likelihood of Estimated Model

The maximum log-likelihood value is derived from the maximum likelihood estimation procedure of CFA. This measure is useful only when compared to that of competing models, it is a comparative and not an absolute measure. It is the value at which the numerical optimisation procedure applied to the log-likelihood function reaches its final iteration. This is the value at which the likelihood function is optimised. A blunt interpretation of this measure is that it reveals how likely it is that the observed features are produced by the fitted model. Therefore, a larger value is prefered.

Stellenbosch University https://scholar.sun.ac.za

8.6.1. Akaike Information Criterion

The Akaike Information Criterion is a measure of fit that is widely used and has application beyond the CFA literature. It is a purely comparative measure that has no interpretative value when viewed in isolation. It is a useful measure of fit in that it punishes over-parameterization, it is a measure that considers the trade-off between fit and parsimony. In this case, we would expect the measure to be biased toward the restricted models, as they will provide a relatively similar fit but with considerably more degrees of freedom. A smaller value is prefered. The Akaike Information Criterion is calculated as follows, where q is the number of freely estimated parameters and L is the log-likelihood value.

$$AIC = 2q + 2ln(L) \tag{13}$$

8.7. Bayesian Information Criterion

The Bayesian Information Criterion is theoretically very similar to the Akaike Information Criterion. However, the Bayesian Information Criterion places a stronger penalty of model complexity, it is more inclined to favour a parsimonious model than is the Akaike Information Criterion. A smaller value is prefered. The Bayesian Information Criterion is calculated as

$$BIC = ln(N)q - 2ln(L) \tag{14}$$

8.8. Measures of Fit Summary Table

	Fit_Measure	Preferred_Value
1	Comparative Fit Index	Larger
2	Tucker-Lewis Index	Larger
3	Loglikelihood user model	Larger
4	Akaike Information Criterion	Smaller
5	Bayesian Information Criterion	Smaller
6	Root-Mean square error	Smaller
7	Standardized Root Mean Square Residual	Smaller
8	model chi-square	Smaller

Table 8.1: Measures of Fit Summary

Stellenbosch University https://scholar.sun.ac.za

Chapter 3: Measurement Error in the Presence of Noisy Data - an Investigation into Procedures of Noise Reducing Factor Estimation

Abstract

In applied survey-based research, it is common that individual responses are captured with error. This can result from limited memory, misinterpretation or misunderstanding of the questions asked. Moreover, in research that uses large survey-based datasets, procedures of dimension reduction are often employed as an initial step to improve the performance of subsequent estimation procedures. These two phenomena - measurement error and dimension reduction - can combine in such a way that is detrimental to empirical work. This chapter investigates whether or not certain techniques of dimension reduction outperform others with regard to mitigating the effects of measurement error in the features used in the procedure. This is done by leveraging the theory of attenuation bias and the findings of Chapter 2. The results indicate that there does exist a difference in the degree to which certain methods of factor estimation mitigate the negative effects of measurement error in the underlying features. The analysis yields several findings, perhaps most useful of which is that exploratory factor analysis outperforms the much more commonly used procedure of principal component analysis.

Keywords: Latent Construct Estimation, PCA, EFA, non-linear Kernel PCA, Neural Network Autoencoder

JEL classification C14, C18, C45, I21

1. Introduction

The use of theoretical latent constructs and their associated estimated factors in applied empirical research is broadly beneficial for two reasons. First, it allows for an improved theoretical framework to be applied to estimation procedures. Second, it reduces the dimensionality of data and thus aids in overcoming the issues that high dimensions present (eg: curse of dimensionality (Friedman, 1997)). The extent to which these benefits are realised depends critically on the precision with which the latent constructs are identified and estimated. While the previous chapter focused mainly on the identification of latent constructs, the focus of this chapter is exclusively on the statistical estimation of factors. The estimation of factors introduces an additional step into the broader process, one that, if performed incorrectly, can result in erroneous variation in subsequent procedures of estimation.

When using a factor modelling approach, the fundamental premise is to condense N features into $L < N$ factors. The overarching concept is dimension reduction - projecting variation from a higher to a lower-dimensional space. In this chapter, only reduction from N dimensions to 1 dimension is considered. That is, only a single factor is extracted from each group of informationally homogenous features.[1] It is important to note that feature extraction and dimension reduction are not exactly

[1] In this case, informationally homogenous features are those that are derived from similar survey response items

Stellenbosch University https://scholar.sun.ac.za

identical concepts. However, in the manner that they are applied in this chapter, they are identical.[2]

The specific focus of this chapter is on the effect that measurement error in the N underlying features has on the single estimated factor, and the relative extent to which alternative methods of dimension reduction retain this error in their projections. The need for the research contained in this chapter derives in part from the findings of chapter 2 of this thesis - some features contain more measurement error than others. Variables that are based on survey responses and that reveal information about psychological processes are often measured with error (Marsh, 1986). Given this measurement error, it is necessary when using a latent construct approach to use techniques of factor estimation that reduce its negative effects as much as possible. This chapter compares the ability of several techniques to reduce the erroneous variation contained within their respective unidimensional projections.

There exist many methods that can be used to estimate factors. Each has characteristics that are desirable in certain circumstances, given the nature of the data and broader methodology. The empirical analysis in this chapter investigates the performance of several techniques with regard to the extraction of meaningful unidimensional variation while minimising the simultaneous extraction of noisy variation. Specifically, Principal Component Analysis (PCA), Exploratory Factor Analysis (EFA), Kernal Principal Component Analysis (KPCA), and Neural Network Autoencoders (AE) are compared. This selection of methods provides a diverse grouping of techniques, specifically along the linear and non-linear divide. Moreover, they differ substantially in the complexity of their implementation. The trade-off between complexity and performance is important and is considered in the analysis. For example, PCA is a purely non-parametric technique and is the simplest, while Autoencoders are complex neural networks that contain many parameters and are the most complex.

This analysis is complicated by the unobservable nature of latent constructs. It is fundamentally impossible to accurately measure the fit of a factor as it is an estimate of an unobservable latent construct. As a result, it is not possible to identify the relative share of meaningful and erroneously measured variation contained in each factor. Therefore, to draw inference we rely on extant econometric theory as well as the findings of chapter 2 of this thesis - that features derived from negatively worded items contain more measurement error than do those derived from positively worded items. It can be shown that measurement error results in attenuation bias, a systematic bias by which OLS point estimates are pushed toward zero (Wooldridge, 2002). The empirical strategy employed in this chapter is designed to leverage the effects of attenuation bias to make inferences about the presence and magnitude of the measurement error retained in unidimensional projections.

This strategy is operationalised by fitting OLS regressions using only factors estimated with one of the four abovementioned methods of factor estimation. The relative magnitudes of OLS point estimates are compared, and thus differences in measurement error are inferred. Steps are taken to ensure accurate comparison is possible. Moreover, a simulation-based analysis is performed in order to generate controlled results which are used to aid the inference of the real analysis.

[2]When used in the manner as is the case in this chapter, the terms “dimension reduction”, “feature extraction”, and “factor estimation” can be used interchangeably.

Stellenbosch University https://scholar.sun.ac.za

The chapter is structured as follows. In section 2 the theoretical framework upon which the empirical analysis is based is described. Section 3 broadly describes the techniques of dimension reduction that are used. Section 4 describes and summarizes the data used. Section 5 presents and discusses the empirical results and Section 6 concludes.

2. Theoretical Framework

This section first outlines the econometric theory underlying the effects of measurement error on OLS point estimates (attenuation bias). Thereafter, the application of, and reliance upon, the theory of attenuation bias to the particular context of this chapter is discussed.

2.1. Measurement Error and Attenuation Bias

This analysis focuses only on measurement error in the explanatory variables. In this case, the explanatory variables are factors that have been estimated using N features as inputs. As shown in Chapter 2, some of these features are likely to be measured with error, particularly those derived from negatively worded response items. Importantly, some of this measurement error may be retained in the unidimensional projection of the N features (the factor). The following sub-section outlines the broad econometric theory of measurement error in the explanatory variables, and the effect that it has on OLS point estimates.

To see the effects of measurement error in the explanatory variables, consider the following model with a single explanatory variable measured with error x^*.

$$y = \beta_0 + \beta_1 x^* + z \tag{1}$$

where x^* is an observed measure of x, one measured with error. It is assumed that z, the residual term, has a zero mean and is uncorrelated with both x^* and x. If it were possible to regress x on y, consistent estimates of β_1 could be obtained using OLS. However, if the above model is estimated, OLS does not necessarily produce consistent point estimates of β_1. The properties of OLS estimates when x^* is used in place of x depend critically on the nature of the measurement error e, which is defined as $e = x^* - x$. Note that, as z is uncorrelated with both x^* and x, it is also uncorrelated with e.

There are two assumptions regarding measurement error that are typically emphasised in the econometrics literature (Wooldridge, 2002). The first is that the measurement error is uncorrelated with the noisy observed measure x^*. That is, $cov(x^*, e) = 0$. If this is true, it must be the case that the measurement error is correlated with the unobserved measure x, that is, $cov(x, e) \neq 0$. To understand the properties of OLS under this assumption, the model is re-written as follows, noting that the measurement error will be captured in the residual term (Wooldridge, 2002)

$$y = \beta_0 + \beta_1 x^* + (z - \beta_1 e) \tag{2}$$

Stellenbosch University https://scholar.sun.ac.za

Referring back to the earlier assumption that both z and e are uncorrelated with $x*$, it is evident that OLS will provide consistent estimates of β_1 in the above case. To see this, we expand the above equation further.

$$y = \beta_0 + \beta_1 x^* + (z - \beta_1(x^* - x)) \tag{3}$$

$$y = \beta_0 + \beta_1 x^* + (z - \beta_1 x^* + \beta_1 x) \tag{4}$$

$$y = \beta_0 + \beta_1 x + z \tag{5}$$

The only outcome of measurement error under the first assumption is an increase in the error variance. To see this, note again that z is uncorrelated with e. Therefore, the error variance in equation 2 is

$$var(z - \beta_1 e) = \sigma_z^2 + \beta_1^2 \sigma_e^2 > \sigma_z^2 \tag{6}$$

The properties of OLS under this first assumption are such that point estimates of β_1 will be consistent.

The second assumption regarding measurement error is more general to the econometrics literature, and is typically the implied assumption in discussions of measurement error. This is the classical error in variables (CEV) assumption (Hyslop & Imbens, 2001), Under this assumption, measurement error is uncorrelated with the unobserved explanatory variable x. That is, $cov(x, e) = 0$. An alternative interpretion of this assumption is gained by writing x^* out as $x^* = x + e$ and noting that the assumption implies that the two components of the observed x^* are uncorrelated. If the CEV assumption holds, it is then the case that $cov(x^*, e) = \sigma_e^2 \neq 0$ must be true.[3] Therefore, under the CEV assumption, the covariance between x^* and e is equal to the variance of the measurement error e. Because x^* and z are uncorrelated , the covariance between the composite error $(z - \beta_1 e)$ and x^* is

$$cov(x^*, (z - \beta_1 e)) = -\beta_1 cov(x^*, e) = -\beta_1 \sigma_e^2 \tag{7}$$

Therefore, under the CEV assumption, OLS does not provide consistent point estimates of β_1 due to the non-zero correlation between the composite error and x^* (Wooldridge, 2002). Moreover, the magnitude and direction of the bias can be characterised as

$$plim(\hat{\beta}_1) = \beta_1 \frac{\sigma_x^2}{\sigma_x^2 + \sigma_e^2} \tag{8}$$

The bias that results under the CEV assumption is known as attenuation bias (Hyslop & Imbens, 2001). From the above, it is evident that the strictly positive non-zero denominator $(\sigma_x^2 + \sigma_e^2 > \sigma_x^2)$ will push the point estimate $\hat{\beta}_1$ toward zero.The following sub-section outlines the application of the theory to the particular empirical purpose of this chapter.

[3] $cov(x^*, e) = cov(x + e, e) = cov(e, e) = var(e) = \sigma_e^2 \neq 0$

Stellenbosch University https://scholar.sun.ac.za

2.2. Application of Attenuation Bias Theory

The purpose of this chapter is to leverage the theory of attenuation bias to analyse the relative magnitudes of measurement error retained within unidimensional projections estimated using different techniques. This empirical strategy is based on several important assumptions. First, among the groupings of features used to estimate each factor, all do contain some measurement error, while some contain more than others. To discern between features with more or less measurement error, the results of Chapter 2 are relied upon. That is, it is assumed that features derived from negatively worded response items (negative features) contain more measurement error than do those derived from positively worded items. Moreover, this assumption extends to the relative number of negative features within a grouping of N total features used to estimate a factor. A larger number of negative features among the N total features is assumed to ensure more total erroneous variation is present within the grouping of features.

An important point to note again is this, measured features are not being used directly for estimation in the OLS regressions. Rather, they are being used indirectly. That is, the explanatory variables in our regressions are not observable features. Rather, we are using factors as explanatory variables in the regressions. These factors have themselves been estimated using observable features, some of which contain significant measurement error. Therefore, there is an intermediate step in the process, one that could either exacerbate or improve the issue of measurement error present in the used features.

There are two channels through which measurement error can ultimately bias OLS point estimates in this context. First, the measurement error that is contained within the underlying features used to estimate the factors. If this error is retained in estimated unidimensional projections, it will indirectly affect the OLS point estimates. The second channel is from the unidimensional projections themselves, which will determine the magnitude of measurement error that is retained in the factors. Therefore, in the latent construct context, the measurement error present in ultimate procedures of estimation is a function of both the choice of features and the choice of method for factor estimation. This highlights the importance of appropriate identification of latent constructs and estimation of their associated factors.

The empirical analysis in this chapter depends fundamentally on an appropriate reliance on the theory of attenuation bias. To leverage this theory to perform meaningful inference, it must be the case that the Classical Error in Variables (CEV) assumption is met. That is, $cov(x, e) = 0$ and consequently, $cov(x^*, e) \neq 0$. It must be true that the measurement error is correlated with the observable features. As shown in chapter 2 of this thesis, this assumption is arguably valid in this case. The design and orientation of response items affect the degree to which the measured response contains error. Features that are based on negatively worded items are more likely to be measured with error than those based on positively worded items. Therefore, the orientation of response items is, in this case, a source through which there exists a correlation between observable features and measurement error - and ultimately a source through which reliance on the theory of attenuation bias is validated.

Stellenbosch University https://scholar.sun.ac.za

3. Methodology

The implicit aim of latent factor estimation is to uncover the fundamental unidimensional information that is contained within several related and highly correlated observable features. This implicit aim is made explicit by statistical methods of dimension reduction - reducing the dimensionality of a large number of correlated features by projecting the variation from a high to a lower-dimensional space. Five factors are used in this analysis: Rejection, Motivation, Enjoyment, Efficacy, and Belonging.[4] Of these five factors, three are identified and estimated using only positive features: Belonging, Rejection, and Motivation. Two are identified and estimated using both positive and negative features: Enjoyment and Efficacy. The theory underlying these factors and their identification is beyond the scope of this analysis. What does matter here is the relative use of positive and negative features in the estimation of each factor. The analysis focuses first on the comparison across all five factors estimated using all of the features with which they are identified. Thereafter the analysis limits its focus to only Efficacy and Enjoyment and compares these two facotrs estiamted using both positive and negative features (unrestricted) and only positive features (restricted).

There exist many methods of dimension reduction. Certain methods such as Principal Component Analysis and Confirmatory Factors Analysis are well established in applied statistics and are used across a wide variety of applications. The recent phenomena of big data and statistical learning, in addition to increases in computational capacity, have resulted in the emergence of many complex methods of dimension reduction. Methods such as Uniform Manifold Approximation and Projection (UMAP) are highly complex mathematical procedures that rely on intricate geometric and topological theory (McInnes, 2018). These techniques are capable of projecting unstructured multidimensional data onto lower-dimensional spaces. However, for this chapter, the data is relatively well structured and the use of such complex techniques is superfluous.

This analysis is restricted to four techniques of dimension reduction, chosen specifically to provide sufficiently varied results. These techniques each offer unique characteristics that could result in superior performance when estimating a unidimensional projection. The four methods used are Principal Component Analysis (PCA), Exploratory Factor Analysis (EFA), Kernel Principal Component Analysis (KPCA), and a Neural Network Autoencoder (AE). The main line of distinction is along the characteristic of linearity - PCA and EFA are linearly additive methods, KPCA and the Autoencoder are based on projection using non-linear mappings. The following sub-sections provide brief outlines of each method individually. Additionally, the employed OLS strategy is discussed.

3.1. Principal Component Analysis (PCA)

PCA is an established and widely used method of dimension reduction (MacCallum, 2009). Its popularity derives, in part, from the simplicity of its execution. PCA is a non-parametric orthogonal linear transformation based on an eigen-decomposition of the dispersion matrix. It returns N orthogonal

[4]See Appendix C for an outline of the features used to estimate each factor

Stellenbosch University https://scholar.sun.ac.za

outputs, or components. Each component contains a part of the total variation contained within the N input features, with the first component containing the most and the rest following in order of magnitude of variation explained. Therefore, PCA provides a relatively simple and robust mathematical basis for constructing a set of N new orthogonal components that explain, in decreasing order, the source of variation, with the majority of variation compressed into the first component.

3.2. Exploratory Factor Analysis (EFA)

The idea underlying EFA is that its estimated factors are related to unobservable real-world latent constructs, such as intelligence.[5] This differs from PCA, which estimates components that are simple geometric abstracts that do not necessarily map onto identified latent constructs. Moreover, unlike PCA, EFA analyses only the shared variance. Therefore, EFA is a correlation-focused approach in which factors represent the common variance among a group of features. Variance not explained is defined as feature-specific (idiosyncratic) variation. This is an important distinction, while PCA does not entirely ignore correlations and covariances, it concentrates mainly on explaining total variance rather than shared variance.[6]

The premise of EFA can be explained as follows, where n_i is an individual feature

$$n_i = \Sigma_{k=1}^{M} \beta_i . f_k + e_i \tag{9}$$

β_i is the factor loading of feature n_i onto each of the M factors f_k. e_i is the idiosyncratic variance that is not correlated with the common factors f_k. From the above (assume $M = 1$), the central thesis of EFA is apparent. EFA assumes that the observable variance of each feature is a weighted linear combination of variance that is shared with other features and also that part that is not shared (idiosyncratic). The strategy used in this chapter is to restrict the factors to one (M=1), and use the N observable features to estimate the single factor, and repeat this for all latent constructs. A maximum likelihood procedure is used to estimate the EFA loadings.

3.3. Kernel Principal Component Analysis (KPCA)

KPCA is a non-linear extension of PCA that better exploits the complex spatial structure of a high-dimensional feature space (Schölkopf, Smola & Müller, 1998). Standard principal components are obtained from an eigen-decomposition of the dispersion matrix and reveal the direction of maximum variance. In Kernel PCA, original features are expanded with non-linear transformations, after which standard PCA is then applied to the transformations within the new feature space. The utility of KPCA becomes clear when one considers that, while it may not be possible to appropriately linearly separate N points from $d \leqslant N$ dimensions, it is almost always possible to linearly separate N points

[5] That is, EFA is designed to reveal specific underlying processes, rather than just mathematically segment variation based on direction.
[6] There is no consideration for idiosyncratic variation in PCA.

Stellenbosch University https://scholar.sun.ac.za

from $d \geqslant N$ dimensions. This implies that the first step of KPCA is to map x_i onto a higher dimensional space using some function Φ. Importantly, the Φ function creates linearly independent vectors in the new feature space. In more simple terms, KPCA first expands the original feature space to enable general PCA to better identify patterns in the data.

With general PCA, it is possible to compute principal components from the inner-product matrix by taking its eigen-decomposition (Friedman, Hastie, & Tibshirani, 2001). KPCA relies upon this same process. As the dimensions of the feature space created by Φ can be arbitrarily large, the KPCA procedure specifies a kernel function inner-product matrix as $\mathbf{K} = [K\langle x_i, x_i'\rangle] = (\Phi(x_i), \Phi(x_i)')$, and finds its eigenvectors by eigendecomposition. The kernel $\mathbf{K}$ specifies the space onto which the original features are projected, the dimensional space in which general procedures of PCA are then applied.

3.4. Neural Network Autoencoder

An Autoencoder is a Neural Network that is trained to learn efficient representations of the input data (Vincent, 2011). Importantly, the dimensional space onto which efficient representations are projected can be specified by the researcher. Autoencoders are fundamentally feedforward deep learning models and learn patterns contained within the data through successive iterations through the neural network. Unlike the other three methods, it is here useful to rely on a more technical outline of the method.

An Autoencoder maps an input vector $\mathbf{x}^d \in [0,1]'$ onto a hidden representation $\mathbf{y}'^d \in [0,1]'$ through a deterministic mapping function $\mathbf{y} = f_\theta(\mathbf{x}) = s(\mathbf{W}\mathbf{x} + \mathbf{b})$. With autoencoders, all inputs are encoded into the model as either 0 or 1. The exact details are beyond the scope of this methodological description but the intuition is this, either the signal sent by an input feature is strong enough and the node is activated (1), or it is not (0). Autoencoders use activation functions that determine whether or not the signal from a particular input feature is strong enough. In the terminology of neural networks, the neuron is activated by a sufficiently strong signal from a particular feature, and this information is passed through the mapping function to the final representation.

This mapping function is parameterized by $\theta = [\mathbf{W}, \mathbf{b}]$, where $\mathbf{W}$ is a d' x d matrix of weights and $\mathbf{b}$ is a bias vector. This representation $\mathbf{y}$ is then mapped onto a reconstructed vector $\mathbf{z}^d \in [0,1]'$ where $\mathbf{z} = g_{\theta'}(\mathbf{y}) = s(\mathbf{W}'\mathbf{y} + \mathbf{b}')$ with $\theta' = [\mathbf{W}', \mathbf{b}']$. Therefore, each x_i is mapped to a corresponding y_i which is then mapped onto the reconstruction z_i. The parameters are opimized by minimizing the avergage reconstruction error as follows

$$\theta^*, \theta'^* = \underset{\theta,\theta'}{\operatorname{argmin}} \frac{1}{N} \Sigma_i^N L(x_i, z_i) \tag{10}$$

where L is the loss function, specified here as the mean squared error $L(\mathbf{x}, \mathbf{z}) = ||\mathbf{x} - \mathbf{z}||^2$. From the above, the premise of autoencoders becomes apparent. They are function approximators that optmimize the parameters of the functions f_θ and $g_{\theta'}$ in order to minimise the information loss when mapping the data from input $\mathbf{x}$ to output $\mathbf{z}$. In this case, f_θ learns the patterns and $g_{\theta'}$ learns to represent these patterns. The autoencoders in this chapter use a mean squared error loss function

Stellenbosch University https://scholar.sun.ac.za

and a tanh activation function. They are specified with three hidden layers, allowing for non-linear patterns to be learned and mapped onto the output representation.

One shortcoming of both KPCA and Autoencoders is that they require hyper-parameter tuning. This would generally be done with cross-validation or a large grid-search. However, traditional methods of tuning are not easily applicable in this case as what is being fit is unobservable. Moreover, the process of tuning introduces substantial complexity into the procedure of estimating either KPCA or an Autoencoder. Therefore, parameters are left un-tuned here for two reasons. First, un-tuned parameters ensure a more accurate comparison with PCA and EFA in terms of the complexity of implementation. Second, tuning parameters in this case would not have the effect that tuned parameters would typically have given, that what is being fit is based on an unobservable latent construct.

3.5. OLS Estimation Strategy

In total, eight OLS regressions are fit using the estimated factors as covariates. Four of the eight include both positive and negative features (Table 5.1), while the other four use only positive features (Table 5.2). This allows for a comparison of OLS models fit using covariates with differing amounts of measurement error. The OLS regressions are estimated using the following function.

$$y_i = \hat{\beta}_i \hat{X}_{i,w,n} + e_i \tag{11}$$

where y_i is an observed measure of mathematical performance for student i. $\hat{X}_{i,w}$ is a vector containing the factor scores for student i, and $w \in (PCA, EFA, KPCA, Autoencoder)$, and $n \in (positive\&negative, positive)$. Therefore, each regression fits the outcome variable using all five of the factors obtained using one of the four studied methods of dimension reduction. The magnitude of the $\hat{\beta}_i$ estimate relative to those obtained from the other specifications in which W changes but n is held constant is the central focus for the analytical inference.

Several steps are taken to ensure comparability of the point estimates across regressions. The sample is limited to only the first three socioeconomic quintiles to ensure a homogenous group of students (Spaull, 2013)[7]. For all regressions, robust standard errors are estimated and post-stratification survey weights are used. Moreover, all estimated factors are standardized to have zero mean and unit variance. In addition to the abovementioned measures, a simulation-based analysis is performed to generate results from a controlled setting.[8]

3.6. Simulation Analysis

The simulated data is generated along a five-point scale to replicate observed data that would be obtained when using a Likert-scale survey response design. Similarly to the real analysis, the purpose

[7]The first three quintiles refer to the bottom 60% of students based on an asset index created at the school level. Therefore, these are students that attend the poorest 60% of schools.
[8]See appendix A for the simulation results.

Stellenbosch University https://scholar.sun.ac.za

of this analysis is to simulate measured features and use these features to estimate factors using four methods - PCA, EFA, KPCA, and Autoencoder. In addition to the features used as inputs into the procedures of factor estimation, a reading performance measure is also simulated. Thereafter, estimated factors are used in OLS regression as explanatory variables to infer the relative measurement error retained in each unidimensional projection (again relying on the theory of attenuation bias). The simulated reading performance measure is used as the outcome variable in these regressions. The simulation analysis is subdivided into four scenarios, each with a different amount of measurement error. A complete outline of the simulation analysis and its results is presented in Appendix A.

To facilitate the concept of a single underlying latent construct, each simulated feature is based upon a single normal distribution to which another normal distribution with a unique variance is added as

$$x_i = X + e \tag{12}$$

$$x_i = \mathcal{N}(500, 100) + \mathcal{N}(0, j) \tag{13}$$

Where $e = \mathcal{N}(0, j)$ is the simulated measurement error and underlying latent construct is simulated as $X = \mathcal{N}(500, 100)$. j is adjusted to analyse the performance of each method of factor estimation as the simulated measurement error increases. Simply, j is the magnitude of simulated measurement error and it is adjusted to alter the magnitude of measurement error in each simulated feature. Importantly, each simulated feature has a non-zero measure of j. This design is intended to ensure that each feature reflects part of the underlying latent construct while not perfectly reflecting it, and to ensure some variation between features. For the features that are considered to be noisy measures, substantially larger values of j are used compared to those that are not noisy features ($e \neq 0, \forall i$). Another important component of e is the zero mean, which is a typical assumption in studies of measurement error. Furthermore, as the change in j is exogenous, and is entirely uncorrelated with x but is feature-specific, the classical error in variables assumption is valid.

4. Data

The 2016 grade 9 South African Trends in International Mathematics and Science Study (TIMSS) is used in this chapter. The outcome variable of interest is the first plausible value of overall mathematical performance. Only selected features are included, and observations with missing values or those above socioeconomic quintile 3 are eliminated. After this procedure, 6092 valid TIMSS observations are retained. From this data, 5 latent constructs are identified and 20 factors are estimated - four for each construct corresponding to the four methods of dimension reduction.[9] Moreover, there are 8 additional factors estimated from the constructs of Enjoyment and Efficacy using only positive features. This is further explained in section 5.2. In this Data section, the focus is solely on the already estimated factors. A description of the features used to estimate each factor is provided in Appendix C.

[9] the latent constructs used are: Belonging, Rejection, Enjoyment, Efficacy, and Motivation.

Stellenbosch University https://scholar.sun.ac.za

Figure 4.1 displays the kernel density distributions of the outcome variable and each estimated factor. Moreover, their joint distribution is displayed in the central contour plot of each figure. From the four figures, it is evident that the distribution of mathematical performance is normal, with a slight degree of right skewness. This is not true for several of the estimated factors. The distributions of the factors estimated using KPCA are significantly non-normal. Rather, they appear mostly to be bimodal. Except for the KPCA factors, most other factors appear to be relatively normally distributed. There also appears to be a pattern of long tails in either direction. This indicates that, for most factor estimates, outliers are biased in one direction.

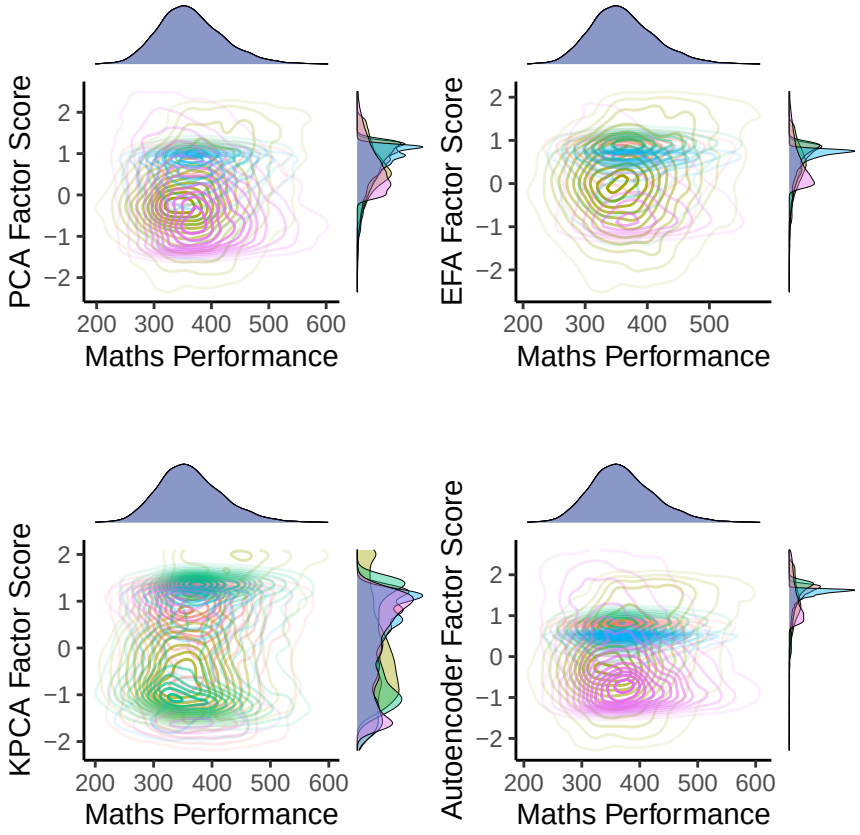


Figure 4.1: TIMSS Factor Distribution Plots - PCA (Top Left), EFA (Top Right), KPCA (Bottom Left), Autoencoder (Bottom Right)

With regard to the OLS estimation, the finding that not all factor estimates are normal is concerning. This information must be used when drawing inference from the OLS analysis. However, one positive finding from the above figures is the relatively tight range of the factors estimated by each respective method of dimension reduction. This will negate any concerns for the influence of different scales on the OLS point estimates.[10]

[10]See Appendix B for a correlation analysis of the factors.

Stellenbosch University https://scholar.sun.ac.za

5. Results[11]

The results section is divided into two subsections. The first (5.1) contains the OLS regression results obtained using factors that are estimated using both negative and positive features. This sub-section uses all five of the estimated factors: Belonging, Rejection, Enjoyment, Efficacy, and Motivation. The purpose of this sub-section is to analyse the differences in coefficient magnitudes between different factors, some of which are estimated using only positive features (Belonging, Rejection, Motivation) and some which are estimated using both positive and negative features (Enjoyment and Efficacy).

The second sub-section (5.2) uses only the two factors that are identified using both positive and negative features: Enjoyment and Efficacy. In this sub-section, the main comparison is on the factors of Efficacy and Enjoyment estimated using: first, both positive and negative features; second, only positive features. That is, the results discussed in this section are the OLS coefficients obtained using the Enjoyment and Efficacy factors estimated using both positive and negative features, and only positive features. The purpose of this sub-section differs from that of 5.1 in that the same factors are compared (enjoyment and Efficacy). Again the point of comparison is along the lines of factors estimated using both positive and negative features, and factors estimated using only positive features. The results of interest in this sub-section are the coefficients for the Enjoyment and Efficacy factors estimated using only positive features. The coefficients for these two same factors when estimated using both positive and negative features are also included in this sub-section for comparison.

5.1. Including Both Positive and Negative Features

Model	M1	M2	M3	M4
Factors	PCA	EFA	Autoencoder	KPCA
Belonging	-1.852	-5.348***	-0.564	0.935
	(1.148)	(1.148)	(1.203)	(1.090)
Rejection	-11.025***	-13.884***	-9.464***	-7.603***
	(0.981)	(0.959)	(0.976)	(1.025)
Enjoyment	4.385***	4.498***	0.750	10.231***
	(1.564)	(1.635)	(1.517)	(1.334)
Efficacy	13.905***	8.520***	16.148***	12.829***
	(1.311)	(1.447)	(1.245)	(1.228)
Motivation	8.376***	11.483***	10.841***	3.266***
	(1.135)	(0.999)	(1.022)	(1.180)
Constant	346.338***	347.468***	346.850***	345.744***
	(1.014)	(1.007)	(1.015)	(1.012)
Observations	5,298	5,298	5,298	5,298
R-squared	0.140	0.145	0.142	0.132

Table 5.1: TIMSS OLS Regression Results

[11] See Appendix A for the results and discussion of the simuation analysis.

Stellenbosch University https://scholar.sun.ac.za

Table 5.1 displays the results of four OLS regressions models, each fit using five estimated factors. Each model contains factors estimated using one of the four methods of dimension reduction. Inference is limited to only those for which more than two of the point estimates are statistically significant. Therefore, the Belonging factor is ignored. Before interpreting the results, it is necessary to note that the Motivation and Rejection latent constructs are identified using only positive features. The Enjoyment and Efficacy constructs are identified with both positive and negative features, with the Efficacy construct containing more negative features (5) than the Enjoyment construct (2). Moreover, it is also necessary to note again that point estimates are being compared in absolute values. Attenuation bias refers to point estimates being pushed toward zero in absolute terms.

From the point estimates of the four constructs of interest, it is evident that there does exist a pattern along the line of those factors that include negative features and those that do not. From models M1 and M2, it is evident that for the two constructs that include negative features – Enjoyment and Efficacy, the most attenuated point estimates belong to the PCA and EFA factor estimates respectively. The two largest point estimates for these constructs belong to the KPCA and Autoencoder factor estimates respectively. From models M3 and M4, it is evident that the largest point estimates for these two factor estimates belong to the Autoencoder and KPCA estimates. The results for the two remaining constructs – Rejection and Motivation – display the exact opposite result. That is, their most attenuated point estimates belong to the factor estimates of the Autoencoder and the KPCA respectively while the largest point estimates both belong to the factor estimates of EFA.

From the results in table 5.1, it appears that the linear methods of dimension reduction - PCA and EFA - retain more of the measurement error contained within the underlying features than do the non-linear methods of dimension reduction - Autoencoder and KPCA. Specifically when there is a known substantial amount of measurement error contained in the inputs to the procedure of dimension reduction, negative features in this case. Conversely, it appears that the non-linear methods of dimension reduction perform better with regard to discarding erroneously measured variation. This deduction is derived by noting that the two largest coefficients for the constructs of Enjoyment and Efficacy belong to the factor estimates of KPCA and the Autoencoder respectively while the most attenuated belong to the estimates of the PCA and EFA respectively.

Moreover, it appears that linear methods of dimension reduction outperform their non-linear counterparts when the factor being estimated contains only positive features, and thus less measurement error. This is inferred by noting that the most attenuated point estimates for the two factors that contain only positive features - Rejection and Motivation - belong to the KPCA model (M4) while the least attenuated belongs to the EFA model (M2). Therefore, the findings displayed in Table 5.1 appear to demonstrate that non-linear methods of factor estimation perform well when underlying features contain measurement error. Conversely, linear methods of factor estimation perform well when the underlying features contain relatively little measurement error.

Stellenbosch University https://scholar.sun.ac.za

5.2. Including only Positive Features

The results in Table 5.2 include the coefficients for Efficacy and Engagement that are presented in Table 5.1 (the top two lines in Table 5.2). In addition to this, Table 5.2 contains the results of OLS regressiosn fit using the Enjoyment and Efficacy factors that are estimated using only positive features. Therefore, these results reveal the extent to which measurement error, proxied for by negative features, impacts OLS coefficients. That is, these results compare the same factors with and without negative features. In this way, the only difference between the factors is the presence of negatvie features. Note that the Efficacy factor has only four positive features to use once we exclude the negative features. This could explain the insignificance of these coefficients.

Model	M5	M6	M7	M8
Factors	PCA	EFA	Autoencoder	KPCA
Enjoyment	4.385***	4.498***	0.750	10.231***
	(1.564)	(1.635)	(1.517)	(1.334)
Efficacy	13.905***	8.520***	16.148***	12.829***
	(1.311)	(1.447)	(1.245)	(1.228)
Enjoyment (no neg)	10.112***	9.730***	7.454***	11.645***
	(1.680)	(1.601)	(1.601)	(1.531)
Efficacy (no neg)	-2.162	-0.999	0.323	-3.230**
	(1.380)	(1.371)	(1.303)	(1.319)

Table 5.2: TIMSS OLS Regression Results - Efficacy and Enjoyment with only Positive Features

Only the results for the Enjoyment factor are statistically significant. From the points estimates of the Enjoyment factor, it is evident that all of the coefficients increase in magnitude once the negative features are removed from the factor estimation procedures. That is, the most attenuated point estimates belong to the factors that contain more measurement error.[12] Moreover, this is true for all of the methods of factor estimation, excluding the Autoencoder which is not statistically significant.

Additional information can be gained by analysing the change in the point estimate magnitudes between the factors estimated using only positive features and those using both positive and negative features. The difference is largest for PCA and EFA, and smallest for KPCA.[13] This finding indicates that the PCA and EFA estimates that are derived using negative features (more measurement error) are the most attenuated - these factors are most affected by the measurement error present in the underlying features. Conversely, the smallest difference between point estimates is for the KPCA. This finding indicates that the KPCA outperforms PCA and EFA with regard to reducing the erroneous variation contained within its unidimensional projection. Comparing PCA and EFA, it is evident that the magnitude of the attenuation bias is substantially larger for PCA than it is for EFA. Therefore, among the linearly additive methods, EFA outperforms PCA.

[12] Measurement error here is proxied for by negative features.
[13] Again ignoring the Autoencoder as its point estimated in Model M7 is not statistically significant.

Stellenbosch University https://scholar.sun.ac.za

6. Conclusion

This chapter analyses the performance of several methods of dimension reduction with the regard to their relative ability to reduce the magnitude of erroneously measured variation in their unidimensional projections. Four methods are analysed, exploratory factor analysis, principal component analysis, kernel principal component analysis, and a neural network autoencoder. The analysis makes use of the theory of attenuation bias and OLS regression to draw inference about the relative magnitude of erroneously measured variation captured in the estimated factors. Factors are estimated using these four methods and measured variables derived from the grade 9 South African Trends in International Mathematics and Science Study dataset. In addition to the real analysis, a simulation-based analysis is conducted.

The analysis yields several findings. First, the results of the real analysis indicate that differences in performance do exist between the analysed methods. The findings indicate that non-linear methods (KPCA and AE) perform better when there is substantial measurement error in the features used in the factor estimation procedure. Second, the results of the simulation-based analysis indicate that EFA is the best performing method in the presence of measurement error. Furthermore, the superiority of EFA increases as the magnitude of measurement error grows. Therefore, there is incongruence between the findings of the real and simulated analysis.

This finding, that the simulated results differ from those of the real analysis, supports the need for further research. A major shortcoming of the real analysis in this chapter is its parochial use of only one dataset and a small number of factors. Future work can improve upon the work presented in this chapter in several ways. First, the real analysis can be expanded to include more data and more modelled factors. Second, additional methods of dimension reduction can be analysed. Third, the hyper-parameters for the relevant algorithms can be tuned. This will require further work to be done to find an appropriate means of hyper-parameter tuning given that what is being fit is based on an unobservable construct. Finally, decomposition methods can be investigated. The reliance on the theory of attenuation bias is a novel approach, but it is limiting. Ideally, an approach that can better quantify the magnitude of erroneously measured variation is required.

Stellenbosch University https://scholar.sun.ac.za

7. Appendix A: Simulation Results

The idea underlying the use of a simulation-based analysis is to gain insights in a controlled setting that can be used to aid inference of the real analysis. While the main focus of this chapter is on the real analysis, and this simulation is supplementary, the results of this analysis are themselves valuable. The simulated data is designed in such a way as to mimic real data that is commonly encountered in scale-response surveys. Therefore, the data is generated along a five-point scale, one that replicates observed data that would be obtained when using a Likert-scale response design.

Similarly to the real analysis, the purpose of this analysis is to simulate measured features and use these features to estimate factors using four methods - PCA, EFA, KPCA, and Autoencoder. In addition to the features used as inputs into the procedures of factor estimation, a reading performance measure is also simulated. Thereafter, estimated factors are used in OLS regression as explanatory variables to infer the relative measurement error retained in each unidimensional projection (again relying on the theory of attenuation bias). The simulated reading performance measure is used as the outcome variable in these regressions.

To create the features required to estimate factors and OLS regressions, it is necessary to simulate N 5-point features and an appropriately correlated reading performance measure. This is done by first creating a single normal distribution from which all simulated measures are based. This first step ensures two important requirements are satisfied. First, the used features will be normally distributed. Second, they will be correlated with one another, thereby simulating the concepts of a single underlying latent construct. This initial normal distribution can be thought of as the latent construct to which unrelated (noisy) variation will be added. There is no explicit theoretical construct being used or targeted here, so it will simply be referred to as latent construct X. The central idea underlying the design of each simulated feature is that each reflects X and some magnitude of noise. This simulated noise is designed to mimic measurement error in measured features. Each feature is simulated as

$$x_i = X + e \tag{14}$$

$$x_i = \mathcal{N}(500, 100) + \mathcal{N}(0, j) \tag{15}$$

Where $e = \mathcal{N}(0, j)$ is the simulated measurement error and underlying latent construct is simulated as $X = \mathcal{N}(500, 100)$. j is adjusted to analyse the performance of each method of factor estimation as the simulated measurement error increases. Simply, j is the magnitude of simulated measurement error and it is adjusted to alter the magnitude of measurement error in each simulated feature. Importantly, each simulated feature has a non-zero measure of j. This design is intended to ensure that each feature reflects part of the underlying latent construct while not perfectly reflecting it, and to ensure some variation between features. For the features that are considered to be noisy measures, substantially larger values of j are used compared to those that are not noisy features ($e \neq 0, \forall i$). Another important component of e is the zero mean, which is a typical assumption in studies of measurement error. Furthermore, as the change in j is exogenous, and is entirely uncorrelated with x but is feature-specific, the classical error in variables assumption is valid and the theory of attenuation

Stellenbosch University https://scholar.sun.ac.za

bias can be relied upon for inference.

	Scenario
1	N = 5 | 1/5 is measured with error - 20%
2	N = 7 | 2/7 are measured with error - 29%
3	N = 7 | 3/7 are measured with error - 43%
4	N = 6 | 3/6 are measured with error - 50%

Table 7.1: Four Simulation Scenarios

Four simulation scenarios are analysed to generate a sufficiently broad set of results. The four differ with the proportion of noisy features used to estimate the factors. Table 8.1 outlines the relative number of noisy and non-noisy features used in each scenario. Moreover, in each scenario, the measure of j in each of the noisy simulated features is increased iteratively over 11 values. That is, the analysis looks at both the effect of differing proportions of noisy features and the effects these features have as the magnitude of noise increases, as $\Delta j > 0$.

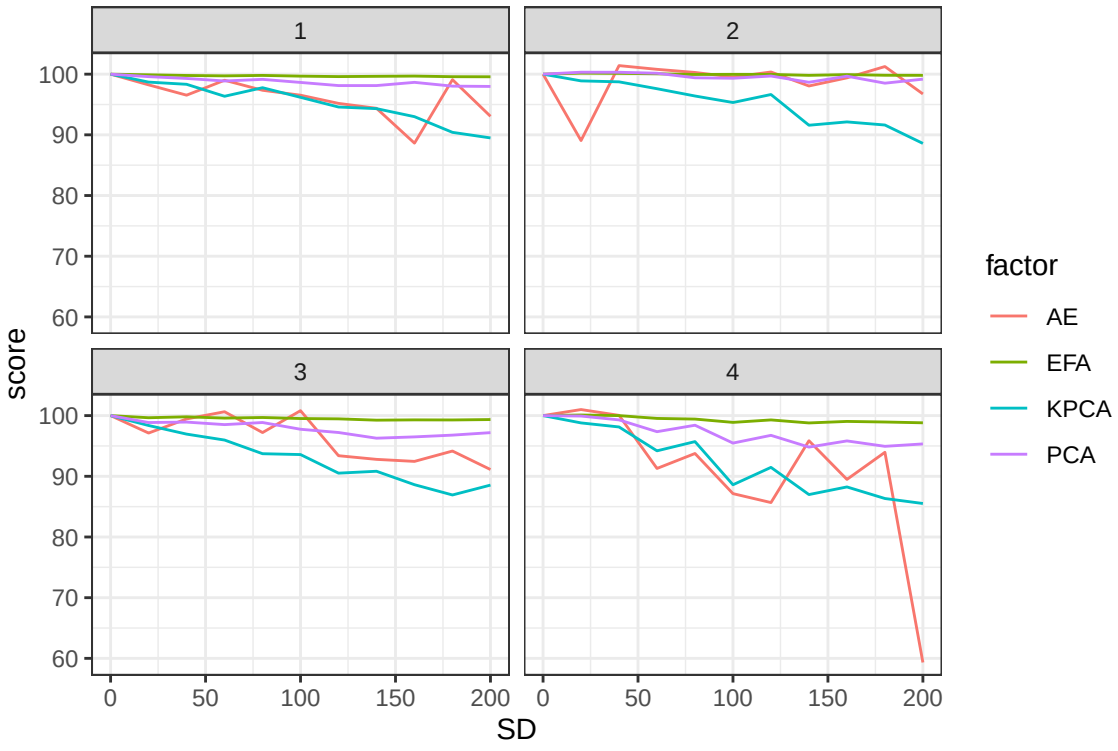


Figure 7.1: Simulation-based OLS coefficient plots

Stellenbosch University https://scholar.sun.ac.za

The results for the OLS coefficients estimated using the simulated data are presented in the four plots in Figure 8.1. The results are presented as follows: each scenario is contained within one graph, each graph contains the results for the estimated OLS coefficients obtained using each of the four methods of factor estimation. Each point of each plot shows the OLS coefficient at an iteratively higher measure of j - more simulated measurement error. Therefore, these results show the impact of increasing the simulated measurement error in a fixed number of features within each plot, and the effects of changing the number of erroneously measured features across each plot. Each initial coefficient is indexed to 100, which is then the base value to which coefficients obtained at subsequent iterations are indexed. This base value is that with the lowest measure of simulated measurement error. The iterative increases in the measurement error are done through the standard deviation of e, j. Each iteration is a % increase over the base value. For example, the point "50" on the x-axis is the point at which the standard deviation of e is 50% larger than it is in the initial value, which is indexed to equal 100. This approach is useful as it enables inference based on the coefficient attenuation in percentage terms.

From the plots in Figure 8.1, it is evident that in the case in which there is more total noisy variation in the input features (scenario 4), all coefficients are more attenuated than they are in the case with the least total noisy variation (scenario 1). It is also evident that KPCA is consistently the worst performer, while EFA is consistently the best performer. A point worth noting here is that the use of KPCA and the Autoencoder here is imperfect in that it is not possible to perfectly tune the hyper-parameters for these estimation techniques in the real analysis, and the hyper-parameters are left un-tuned in this simulation. Conversely, PCA and EFA do not require parameter tuning. The main finding from the results in Figure 8.1 is that EFA achieves the best performance at all levels of j and in all four scenarios.

Stellenbosch University https://scholar.sun.ac.za

8. Appendix B: Factor Correlations

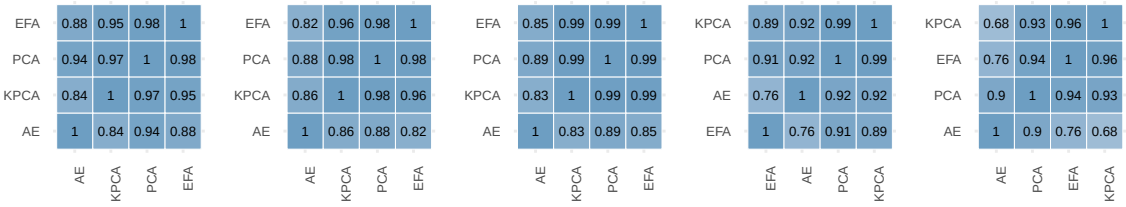


Figure 8.1: TIMSS Factor Correlation Plots - in order of: Belonging, Rejection, Enjoyment, Efficacy, Motivation

From Figure 8.1, it is evident that most of the factors are highly correlated with one another. It is expected that these estimated factors will be highly correlated, they are measuring the same construct using the same inputs. For the purpose of this chapter, too high a correlation value and they are too similar to compare, too low a value and factors are too dissimilar for comparison to be valuable. There exists no precedent regarding what constitutes an ideal correlation coefficient in this context. Therefore, these presented correlation values serve to improve the inference of the chapter's findings rather than inform the strategy. However, values between 0.7 and 0.9 can be considered ideal, as these are sufficiently similar to allow meaningful comparison without being identical.

Stellenbosch University https://scholar.sun.ac.za

9. Appendix C: Response Items Underlying Employed Features

9.1. TIMSS Latent Constructs

	Belonging
1	I like being in school
2	I feel safe when I am at school
3	I feel like I belong at this school
4	Teachers at my school are fair to me
5	I am proud to go to this school

Table 9.1: TIMSS Belonging Latent Construct - Questions

	Rejection
1	Made fun of my clothes and apprearance
2	I am left out of games
3	Spread lies about me
4	Stole something from me
5	Physically hurt me
6	Made me do things I didn't want to do
7	Embarrasing posts on social media
8	Shared embarrassing information or photos of me
9	Threatened me

Table 9.2: TIMSS Rejection Latent Construct - Questions

	Enjoyment
1	I enjoy learning mathematics
2	I wish I did not have to study mathematics (Negative)
3	Mathematics is boring (Negative)
4	I learn many interesting things in mathematics
5	I like mathematics
6	I like any schoolwork that involves numbers
7	I like to solve mathematics problems
8	I look forward to mathematics class
9	Mathematics is one of my favorite subjects

Table 9.3: TIMSS Enjoyment Latent Construct - Questions

Stellenbosch University https://scholar.sun.ac.za

	Efficacy
1	I usually do well in mathematics
2	Mathematics is more difficult for me than for many of my classmates (Negative)
3	Mathematics is not one of my strengths (Negative)
4	I learn things quickly in mathematics
5	Mathematics makes me nervous (Negative)
6	I am good at working out difficult mathematics problems
7	My teacher tells me I am good at mathematics
8	Mathematics is harder for me than any other subject (Negative)
9	Mathematics makes me confused (Negative)

Table 9.4: TIMSS Efficacy Latent Construct - Questions

	Motivation
1	I think learning mathematics will help me in my daily life
2	I need mathematics to learn other school subjects
3	I need to do well in mathematics to get into the college or university of my choice
4	I need to do well in mathematics to get the job I want
5	I would like a job that involves using mathematics
6	It is important to learn about mathematics to get ahead in the world
7	Learning mathematics will give me more job opportunities when I am an adult
8	My parents think that it is important that I do well in mathematics
9	It is important to do well in mathematics

Table 9.5: TIMSS Motivation Latent Construct - Questions

Stellenbosch University https://scholar.sun.ac.za

Chapter 4: Investigating the Determinants of Reading Ability Using Gradient Boosted Regression

Abstract

This chapter investigates the determinants of reading performance of South African grade 4 students. The approach taken is novel in terms of both the statistical methodology and the theoretical framework. The empirical analysis makes extensive use of gradient boosted regression, a statistical learning technique that enables the analysis of complex nonlinear and interactive relationships. The analysis yields several interesting and policy-relevant findings. First, psychological processes are found to be important predictors of reading performance, specifically negative social interaction and self-efficacy. Moreover, the importance of these measured factors is not consistent over the wealth distribution. Second, the importance of the household as a centre for learning is highlighted. Findings indicate that children with parents that are more willing and able to help with their learning tend to perform better than those without helpful or involved parents. Third, the composition of students within the classroom in terms of relative reading ability is a strong predictor of performance. The remedial requirements of portions of students in the class, likely due to learning backlogs, can negatively impact on the learning process of other students in the class. These findings all lend themselves to possible policy interventions.

Keywords: Factor Modelling, Educational Psychology, Statistical Learning, PIRLS, Gradient Boost

JEL classification I21, I24, C49

1. Introduction

The ability to read is arguably the most important skill in most academic settings. All present and future learning is dependent fundamentally on the ability to appropriately interpret text and internalise its meaning. Put simply, a student must first learn to read before reading to learn. Students that do not possess the level of literacy skill necessary to decode text will be incapable of keeping pace with a curriculum that is based primarily on the sequential acquisition of certain skills. From this assertion, the centrality of reading in the context of educational learning becomes apparent. Therefore, it is a serious concern that a large number of students within the South African education system do not possess adequate literacy ability. This illiteracy is omnipresent specifically in poorer schools. The reason behind this is twofold: first, these students receive an extremely poor quality of education; and second, reading is often not a common practice, nor commonly taught adequately, in poorer households. As a result, reading ability is cultivated neither at school nor at home. Therefore, the link between poverty, inadequate education quality, and limited literacy becomes a binding constraint on the ability of poor students to perform academically. This link is a means through which limited labour market prospects - and in many cases, poverty - can become entrenched. Moreover, due to the

Stellenbosch University https://scholar.sun.ac.za

racial and spatial dimensions of South African poverty, this link also enforces and perpetuates current patterns of inequality.

Of notable concern to education researchers is the inability to adequately model and explain the determinants of reading ability. Such research would have direct policy relevance in that its findings can aid future policy-led attempts to improve South African education processes and outcomes. Most contemporary research in the field of the economics of education centres on the use of linearly additive production functions, which generally assume that some linear combination of observable variables constitutes the primary components of the data generating process that underlies academic ability and performance.[1] It is the broad aim of this chapter to move away from this approach in favour of one that is reliant on methods of supervised statistical learning and theoretical constructs borrowed from the educational psychology literature. The reliance on these two broad components in this research enables unique inference of non-linear and interactive relationships between the determinants of reading ability. Such inference is not possible with more typical linear models. One shortcoming of this approach is that, unlike OLS, for example, estimates are not a direct quantitative measure of the relationship between the variable of interest and the outcome variable. Rather, the methods used in this chapter estimate the functional predictive relationship of each feature with the outcome variable and the relative strength with which individual features predict the specified outcome variable.

The analysis employs an approach that incorporates unobservable latent variables (factors) based on theoretical constructs. While factors do receive considerable attention in this study, their inclusion is not done with an *a priori* intention to emphasise their importance. They are included and analysed as is every other feature. However, as the use and demonstration of the value of theoretical latent constructs and their corresponding estimated factors is a central component of this thesis, their results are highlighted and analysed with greater scrutiny. In addition to the factors, the analysis uses the full PIRLS dataset to estimate patterns between variables in the feature space and the outcome variable of reading performance. Therefore, the research question that underlies the analysis is posed rather broadly - the study seeks to highlight important non-linear and interactive relationships among measured features and estimated factors with regard to the prediction of observed reading performance. The aim of uncovering non-linear and interactive effects is supported by the use of gradient boosted regression (GBM). In addition to gradient boost, exploratory factor analysis is used to estimate the factors.

The paper is structured as follows. First, the theoretical foundations upon which the empirical analysis is based are discussed. Second, the data used is briefly discussed and descriptive statistics are provided. Third, the methodological procedure is outlined. Fourth, empirical results are provided.

[1] This empirical approach relies on the underlying and longstanding economic theory that an observed output is the direct result of some group of inputs. Typically, in this case, an estimation procedure would rely on some production function under the assumption that variables input into the function behave in some specified way to produce an output.

Stellenbosch University https://scholar.sun.ac.za

Fifth, the main policy recommendations are outlined. Finally, concluding thoughts and implications are discussed.

2. Theoretical Framework

The central thesis of this analysis is based fundamentally on the notion that there exist two main spheres of influence in the life of a school student - the home sphere and the school sphere (or academic sphere). These two arenas constitute the main environments within which the behavioural characteristics and personality traits of a student are formed. Moreover, these two environments are dynamic arenas in which a student learns and faces numerous challenges and ecological transitions that emerge sequentially as they progress into and through school (Ladd, Birch & Buhs, 1999). These challenges result in certain outcomes within a paradigm of adaptation and maladjustment. That is, either students will adapt within the school sphere and experience little difficulty in coping with social and scholastic challenges, or they will not adapt and will be negatively affected by these challenges. The remainder of the current section briefly describes the important constructs that make up the theoretical framework of the analysis.

A student is endowed with a certain measure of cultural capital within the home sphere. The concept of cultural capital is defined broadly as a measure of familiarity with a dominant culture within society (Sullivan, 2009). In the current study, cultural capital is defined as a measure of familiarity and comfort with the social and scholastic codes that are common in the school sphere. Therefore, cultural capital is made up primarily of social and cognitive maturity, both of which determine the ability of a student to adapt to social and scholastic challenges. A student with a high measure of cultural capital will be comfortable with the codes used in the school sphere and is more likely to adapt well. Cultural capital is a function of the social norms that are present within the home sphere of each individual student. It is with this endowment of cultural capital that a student enters the school sphere, which is the first major ecological transition. Therefore, a large component of cultural capital is what is termed entry factors (Taylor & Machida, 1996).

Entry factors are attributes that operate and emerge prior to the initial entrance into the school sphere. Examples of entry factors are: race, gender, parent's education, pre-school education, and socioeconomic status (SES). Moreover, unobservable entry factors such as family values, behavioural traits, and experiences are also important components of cultural capital. The nature of these entry factors significantly determines the degree to which a student experiences either adaptation or maladjustment upon entering the school sphere. However, and of equal importance, these factors continue to influence the adaptive success of a student after the initial entrance into the school sphere. A student with a low measure of cultural capital (inappropriate entry factors) will experience negative outcomes from the social discontinuities which are realised upon entry into the school sphere This will result in maladjustment. Conversely, a student that has a high measure of cultural capital (appropriate entry

Stellenbosch University https://scholar.sun.ac.za

factors) will not experience negative outcomes from the abrupt social discontinuities realised upon entry. Rather, they will integrate into the school sphere with minimal social and cognitive disruption. It is important to note that most of these concepts - cultural capital, the adaptation and maladjustment paradigm, and entry factors - are measures along some continuum, it is not that students either have or don't have it, they have some measure of it.

Upon entry, the mix of cognitive and social maturity interact in such a way that it determines the outcome of the adaptation-maladjustment paradigm. Within the broad measure of social maturity are individual behavioural styles. It is these behavioural characteristics that influence the manner in which a student reacts to the sequence of social and scholastic challenges. Individual differences in behavioural characteristics largely determine the outcomes experienced by each student as they interact with others in the school sphere. A student may have certain orientations that cause them to move toward or away from social interactions (Caspi, Elder, & Blem, 1987). The degree to which a student adapts within the social domain of the school sphere is termed the sense of belonging in the current analysis; this is the first of six important latent constructs, *(Belonging)*. A student can have either a strong or a weak sense of belonging, which reflects the outcome of being either accepted or rejected by the social unit in the school sphere.

2.1. Sense of Belonging and Social Rejection

Acceptance by the social unit is linked with positive social and scholastic outcomes. Upon entry, the relationships that students form with teachers and peers yield various social provisions that facilitate adaptation (Ladd, Kochenderfer, & Coleman, 1997). Participation within social units offers students a number of supports, such as assistance and security – both of which help with the adaptation to the unique stressors present in the school sphere. Ladd (1990) finds that students adapt better in the early stages of school when they have a friend, develop large social networks, and have relationships with their teachers. In addition, research indicates that there is also a positive relationship between social acceptance and scholastic adaptation (Birch & Ladd, 1996; 1997).

Conversely, rejection by the social unit *(Rejection)* is associated with maladjustment in both the social and scholastic domain of the school sphere (Ladd, Kochenderfer, & Coleman, 1997). Rejected students are more likely to remain friendless and not form relationships with teachers. Moreover, in addition to the lack of relationships, rejected students may experience conflict with fellow students and teachers. In this way, rejected students will have purely negative interactions with the social unit in the school sphere. Therefore, while rejected students will not have access to the above-mentioned resources that facilitate the adaptation to school sphere stressors, they are also more likely to experience direct stress from conflict with the social unit. In short, rejection results in more than just the lack of access to relationships that reduce stress; the lack of relationships may in itself engender stress within the rejected student, which further compounds maladjustment.

Stellenbosch University https://scholar.sun.ac.za

While a strong sense of belonging reflects the lack of resistance from the social unit, the construct of social rejection captures the exact opposite. However, while they are two related concepts, they are not on a single continuum. A sense of belonging can be engendered by positive interactions within the social unit and the broader school environment. Rejection is a phenomenon whereby the social unit actively prevents the entry of specific students. Therefore, social rejection is modelled as its own factor in addition to the sense of belonging. Bullying is an extreme form of rejection by the school sphere's social unit. This outcome could result if the cultural capital endowment - specifically social maturity - of the student does not align with that required in the school sphere.

2.2. Motivation, Engagement, and Enjoyment

While the above-described latent construct of sense of belonging depends on interactions with the social unit, there are components of cultural capital that are innate to each individual student and that are not immediately determined by social interactions at school. A prominent component of cultural capital which is determined primarily outside of the school sphere is intrinsic motivation *(Motivation)* . The current study employs the Self-Determination Theory of Motivation (SDT) (Deci & Ryan, 1980; 1985; Deci, Koestner, & Ryan, 2001). SDT defines motivation as the extent to which an individual actively engages in activities and procedures purely to derive utility (Deci, Vallerand, Pelletier, & Ryan, 1991:327). Here it is assumed that social and scholastic adaptation and performance results in utility. However, while the above definition does outline how the motivation construct enters the model to be used in the current analysis, it does not explain the mechanisms and drivers that are behind the formation of motivation. A useful link can be found in Ryan and Deci (2000), who note that the outcome of intrinsic motivation is the manifestation of psychological needs. That is, a student will experience certain psychological needs and will derive utility from fulfilling these needs. Motivation is a measure of the desire with which they actively attempt to fulfil these needs and subsequently realise utility. Moreover, SDT does not differentiate between the relative engagement with reading activities in the school or home sphere. To gain a deeper understanding of different types of engagement with reading, an additional factor called *Engagement* is estimated. The Engagement factor incorporates features that pertain to inputs into the learning process from both the teacher and the student.[2]

The above enables a theoretical link between intrinsic motivation and cultural capital. That is, within the home sphere, before entry into the school sphere, students develop certain psychological needs. Of consequence for this study is the nature of these needs, specifically, whether or not the student experiences the desire to fulfil needs related to social and scholastic success in the school sphere. Put simply, the study assumes that individual students that have psychological needs related to scholastic achievement are more likely to be motivated to achieve positive scholastic outcomes. The emphasis on needs related to scholastic success rather than social success is due simply to the fact that social

[2]See Appendix E for an outline of the individual features used in the creation of each factor.

Stellenbosch University https://scholar.sun.ac.za

interactions depend fundamentally on the actions of other students, whereas motivation to learn to read is more likely to originate within individual students. Moreover, individual motivation to read can manifest in the act of reading, an act that does not require inputs other than a book. Therefore, motivation is included in the theoretical framework primarily as an entry factor.

However, innate motivation is not disconnected from factors within the school sphere. Ryan and Deci (2000) note that the inherent propensity of motivation is fundamentally dependent on supportive conditions. From this perspective, intrinsic motivation can be understood as an underlying latent propensity which is shaped by factors that elicit and sustain the desires innate within an individual. Therefore, while the existence and emergence of intrinsic motivation is an outcome of supportive conditions at home, its lasting propensity is partly dependent on the presence of supportive conditions within the school sphere. A typical supportive condition is the presence of relationships with fellow students and teachers which, as previously mentioned, facilitate adaptation in the presence of challenges. Therefore, again making a link to a previously mentioned concept, a student's positive sense of belonging will act as a supportive condition and largely determine the continued propensity of the entry factor of intrinsic motivation. Conversely, students that have a low sense of belonging will not have these supportive conditions in place. In this case, even if they enter with a high measure of motivation, the propensity of their motivation will likely diminish as they progress through sequential social and scholastic challenges.

A factor that is related to motivation is *Enjoyment*. This is the degree to which the student derives direct utility from reading. This factor is not given a theoretical outline as is done for the others due simply to the uncomplicated manner with which it is incorporated into the framework. It is assumed that students that more enjoy reading will likely perform better as they have gained practice while deriving utility from reading. The inclusion of enjoyment as a factor distinct from either motivation and engagement is premised upon the notion that it is based on only direct utility from reading, unlike motivation and engagement, which are based more on an indirect derivation of utility from reading.

2.3. Self-Efficacy

While motivation is defined as the desire to read, self-efficacy *(Efficacy)* is a self-perceived measure of reading ability (Bandura, 1997). Schunk (2003) notes that efficacy influences achievement behaviours such as persistence, effort, and choice of task. Therefore, efficacy and motivation are fundamentally linked in that individuals with high efficacy are more likely to partake in activities that improve their chances of achieving some desired outcome. In addition, similarly to the construct of motivation, efficacy is partially an entry factor, the continued propensity of which depends on factors within the school sphere. That is, efficacy is a relative measure that depends not only on the innate ability of the individual, but also on the individual's ability relative to that of others in the class. Therefore, unlike the construct of motivation, efficacy is formed and maintained predominantly in the school sphere.

Stellenbosch University https://scholar.sun.ac.za

3. Methodology

The methodological procedure employed in this chapter is aligned with the research purpose of uncovering non-linear and interactive relationships among measured features and observed reading performance. The sample is analysed in several segments to gain deeper insights into distinct components of the data generating process, which in South Africa is not one distinct phenomenon (Spaull, 2013). The data is disaggregated in two stages. First, the data is separated into four individual sets based on the four different survey respondents.[3] Second, the features found to be the strongest predictors of reading performance in the first stage are combined into one dataset, which is then disaggregated by socioeconomic status and again analysed using gradient boosted regression.

To capture non-linear and interaction effects, gradient boosted regression (GBM) is used as the primary method of modelling in this chapter.[4] In total, 16 regressions are fit, each using a different set of observations. To aid the interpretation of the regression results, models are fit using both the training and testing data. Therefore, each set of observations has two associated regressions, one for the test set and one for the training set.[5] The gradient boosted regressions are estimated using a squared error loss function (MSE). Hyperparameters are tuned using training data and a large grid-search that tests 500 specified parameter combinations to minimise a squared error loss function. Hyperparameters that are selected by the grid-search are subsequently used to fit a regression using the test data to identify possible over- or under-fitting. If any issues are found, the process is repeated and alternative hyperparameters are selected. Regression results are interpreted using variable importance measures (Relative Feature Importance and Permutation Feature Importance) and partial dependence functions. A formal description of these methods of interpretation is given in Appendix D.

The gradient boost algorithm is based on the sequential construction of weak learning predictive models that improve upon the error of the previous learner.[6] While these weak learners need not be regression trees as gradient boost will work with almost any predictive model, it is standard that regressions trees are used. In this way, gradient boost is similar to other ensemble-based methods such as random forest and bagging. However, two fundamental differences make gradient boost unique. First, it uses an ensemble of weak learners. Random forest and bagging models rely on an ensemble in which each individual predictive model is relatively complex and there are fewer limitations on the

[3] The four respondents include: the student themself, the student's parent, the student's teacher, and the student's school head. The teacher dataset was treated differently to the rest due to its large number of features relative to observations. This problem was overcome by first training a regression using all observations and, based on the results of this first regression, removing all but the top 65 most important features (65 strongest predictors). These 65 features were then used in the above-described process.

[4] Regression is the used terminology when the outcome variable is continuous. A binary outcome would necessitate gradient boost classification.

[5] There are two separate regressions run on each of the following sets of observations: Home (parent), School (head), Student, Teacher, SES1, SES2, SES3, and total.

[6] A weak learner is a model that has a predictive accuracy slightly greater than random chance.

Stellenbosch University https://scholar.sun.ac.za

predictive power of each individual learner. Secondly, gradient boost builds this ensemble sequentially. Random forest and bagging models build an ensemble of independent learners, gradient boost trains each new learner to minimise the error of the previous learner. In this way, the final prediction is that of the terminal learner rather than an average of all the learners in the ensemble.

Gradient boost is, as the name suggests, based on a gradient descent algorithm. The algorithm works by first initializing a constant from which the first set of error residuals are calculated. From these errors, a loss function is derived. Gradient boost then minimises this loss function by step-wise iterations along the steepest surface of the loss function as efficiently as possible to reach the global minimum. This is the fundamental premise of gradient boost, to tweak parameters iteratively to minimise the loss function. The gradient boost regression algorithm used is based on that originally proposed by Friedman (2001).[7] The gradient boost procedure and algorithm used in this chapter are as follows.

First, tuned hyperparameters are selected from, in this case, the results of a 500 point grid-search. Chosen hyperparameters are: a loss function (MSE in this case), the number of iterations or trees T, the depth of each tree K, the learning rate λ, and the subsampling rate p. The first step in the process is to initialize the function $\hat{f}(\mathbf{x}_i)$ with a constant

$$\hat{f}(\mathbf{x}_i) = \underset{\rho}{\operatorname{argmin}} \Sigma_{i=1}^{N} L(y_i, \rho) \tag{1}$$

where ρ are the predicted values, $\mathbf{x}_i$ are the input features and y_i is the outcome variable - reading performance score. This constant is simply the mean value of the outcome variable. Next, the negative gradient is computed as

$$z_i = -\frac{\partial L(y_i, \rho)}{\partial f(x_i)}\Big|_{f(x_i)=\hat{f}(x_i)} \tag{2}$$

The z_i value is analogous to the residual term of an OLS regression when the mean squared error loss function is used in gradient boost. It is generally referred to as the pseudo-residual, as it can be derived from any loss function and will not necessarily always correspond to an OLS residual. The following step randomly selects a subsample of $p \times N$ to which a regression tree with K terminal nodes is fit, $g(\mathbf{x}) = E(z|\mathbf{x})$. The algorithm fits trees sequentially; each tree learns from the previous tree by updating and minimising the pseudo-residual from the previous tree. For each iteration $k \in K$, the optimal terminal node prediction is computed as

$$\rho_k = \underset{\rho}{\operatorname{argmin}} \Sigma_{\mathbf{x}_i \in S_k} L(y_i, \hat{f}(\mathbf{x}_i)) + \rho) \tag{3}$$

where S_i is the set of $\mathbf{x}s$ that define the terminal node K. After each iteration, $\hat{f}(\mathbf{x})$ is updated as follows, where $k(\mathbf{x})$ indicates the index of the terminal node into which an observation with features

[7] It is a slightly adapted version coded into the r package "GBM" (Ridgeway, 2020).

Stellenbosch University https://scholar.sun.ac.za

$\mathbf{x}$ would fall.

$$\hat{f}(\mathbf{x}_i) \leftarrow \hat{f}(\mathbf{x}_i) + \lambda \rho_{k(\mathbf{x})} \tag{4}$$

In addition to gradient boost, the analysis in this chapter makes use of exploratory factor analysis (EFA) to estimate factors associated with the studied latent constructs. As discussed in Chapter 3, EFA is a correlation-focused approach in which factors represent the common variance among a group of features. Variance not explain is defined as feature-specific (idiosyncratic) variation. The EFA algorithm used in this chapter estimates each feature as a weighted sum of shared and idiosyncratic variance. The value of this shared variance is then extracted and used as the factor score that represents the unobservable latent construct. Equation 5 describes this procedure.

$$x_i = \Sigma_{k=1}^{M} \beta_i . f_k + e_i \tag{5}$$

β_i is the factor loading of feature x_i onto each of the M factors f_k. e_i is the idiosyncratic variance that is not correlated with the common factors f_k. From the above (assume $M = 1$), the central idea of EFA is apparent. EFA assumes that the observable variance of each feature is a weighted linear combination of variance that is shared with other features and that part that is not shared (idiosyncratic). The strategy used in this chapter is to restrict the number of factors to one (M=1), and use the N observable features to estimate the single factor, and repeat this for all 6 studied latent constructs. EFA is the prefered method for factor estimation in this chapter due to its ability to reduce the magnitude of noise contained in its unidimensional projections, as was shown in Chapter 3 of this thesis.

4. Data

The Progress in International Reading Literacy Study (PIRLS) 2016 dataset is used for the empirical analysis of this chapter. The dataset contains a rich selection of self-reported variables derived from individual survey questionnaires given to students, parents (guardians), teachers, and school heads (principals). The first section of the empirical analysis disaggregates the data by the four respondent groups - students, parents, teachers, and school heads - and analyses each group individually. The second section of the analysis combines the data associated with each of the four respondents into one dataset. The analysis in section 2 disaggregates the combined data by socioeconomic status (SES) and looks at each SES group individually. The SES grouping is done using quintiles based on school wealth, not student wealth, which are divided into three main groups.[8] The first SES group contains the first 3 quintiles, the bottom 60% of the SES distribution. The second SES group contains only

[8]The SES distribution is based on school wealth rather than household or individual wealth. This school wealth measure (SES) is based on the average wealth of the students at the school. Therefore, it is not necessarily the case that a student from a high SES school will be from a household that is as wealthy as those of their classmates.

Stellenbosch University https://scholar.sun.ac.za

those in quintile 4, while the third SES group contains only those in the fifth quintile, the top 20% of the SES distribution.[9]

In addition to the response-based variables, PIRLS contains several measures of reading performance based on tests given to the sampled students. Each measure of reading performance pertains to some aspect of reading ability. Moreover, for each measure, there are five plausible values. These are values that are imputed as not every sampled student completes every one of the reading assessments. In practice, it is typical that methods of multiple imputation are used to factor in the increased sampling variation that arises from the use of plausible values. However, due to the ensembled nature of gradient boost, and lack of estimated standard errors, this increased sampling variation is not an issue and only the first plausible value for the overall measure of reading performance is used as an outcome variable.[10]

The total sample is 12810. However, a large proportion of the sample contains variable non-response. Features for which over 30% of the observations are missing are excluded from the analysis. An analysis of this non-response revealed no distinct patterns. Therefore, concerns of non-random non-response can be ignored. The final sample size used to fit each model varies considerably. However, most are fit using 2000-3000 observations; this includes a combined training and testing set. A split of 70-30 is used, 70% of the sample is used to train the model while 30% is used to test its performance. The final sample size used to train each algorithm is noted in the results section.

Figure 4.1 plots the distribution of reading performance of each of the three SES groupings split by student gender, where the figure to the right is female and that to the left is male. The most notable aspect of the distributions is the substantially better performance by students in the top SES group, the top 20% most wealthy schools. The distribution for SES group 3 is notably different from the other two, indicating that there exists a clear performance advantage for students enrolled at wealthier schools. A second notable aspect of Figure 4.1 is the degree to which performance differs by gender across the three SES groups. Female performance appears to be more defined by SES grouping, and wealthy female students make up the largest proportion of those that achieve high performance scores.

[9] See Appendix A for large clustered correlation heatmaps of the entire feature space for each individual disaggregated dataset.

[10] Analysis was performed on two of the remaining four plausible values and the results were found to be almost identical.

Stellenbosch University https://scholar.sun.ac.za

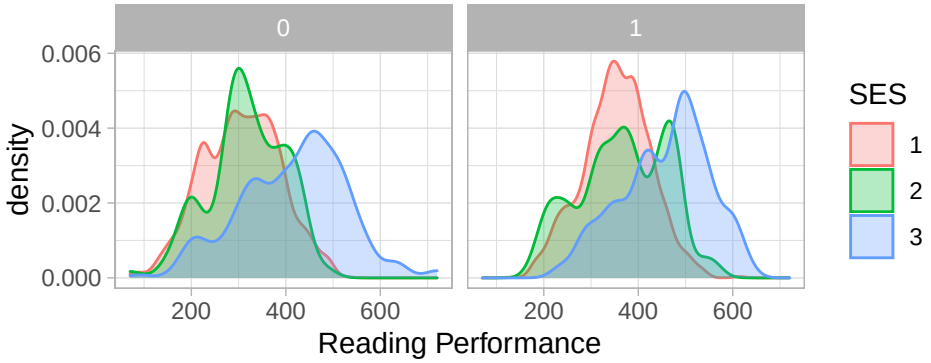


Figure 4.1: Distribution of Reading Performance SES Group and gender, where 0 = Male and 1 = Female

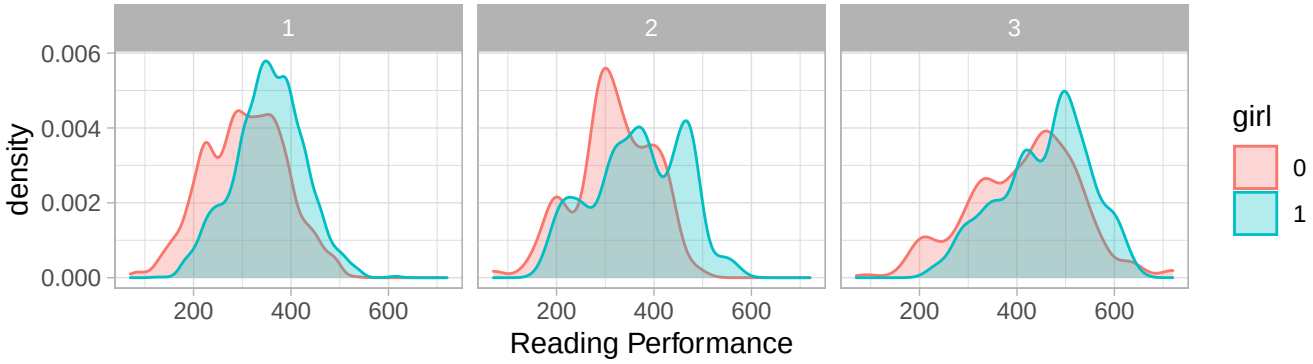


Figure 4.2: Distribution of Reading Performance by Gender and SES group (SES group 1, 2, and 3)

Figure 4.2 shows the distribution of reading performance scores disaggregated by gender and split by SES group. It is clear that students that attend schools in the third SES group perform substantially better than do those in the other two groups - the bottom 80% of the school wealth distribution. Another notable aspect of the distributions is that the female advantage appears to be relatively consistent across all three SES groups. That is, girls do better than boys regardless of average docioeconomic status of the students at the schol (SES grouping). It also appears that the distributions become increasingly flat at higher levels of average school wealth. This indicates that most students in low SES schools perform equally poorly, while those in wealthier schools have a superior but more varied performance.

The contour plot shown in Figure 4.3 visualizes the measure of each factor among sampled students. Each line in the plot represents an individual observation, with each point corresponding to the score (measured value) for that particular factor for that particular individual.[11] It is evident that the

[11] Note, the contour plot is scaled. Therefore, each point is to be interpreted relative to the others.

Stellenbosch University https://scholar.sun.ac.za

majority of students have a relatively low sense of Belonging and Motivation for, and Enjoyment of, reading. Conversely, the factors of Efficacy and Rejection are not notably skewed in any particular direction - it is equally common to see both high and low values of these two factors.

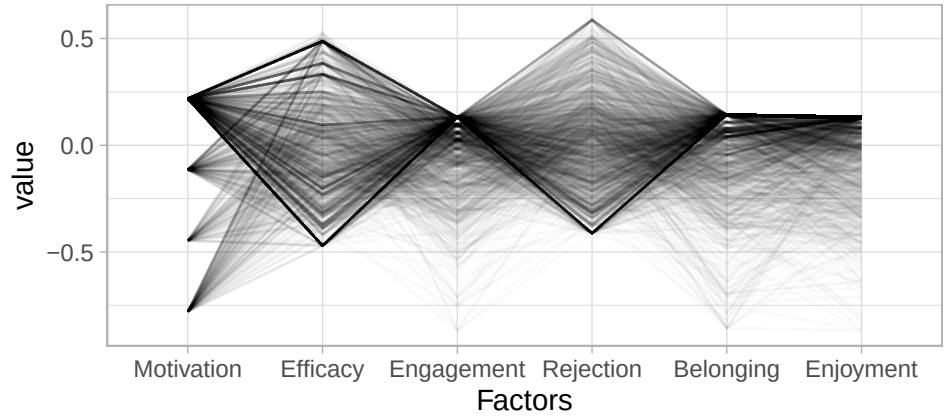


Figure 4.3: Contour Plot - Prevalence of Each Factor

Figure 4.4 shows the bivariate relationship between reading performance and each of the six included factors split by SES group. Each plot includes a bivariate OLS regression line with a 95% confidence interval in addition to a bivariate scatter. From each of the six plots, the SES-based differences in the relationship between performance and each factor can be analysed. The plots reveal that, for certain factors, the relationship between them and reading performance is unique to each SES group, while for others the relationship is almost identical while mean performance scores differ by group. The factor of Belonging has a consistent relationship with reading performance for all three SES groups, only the constant differs. That is, measured reading performance increases in Belonging at a constant rate for all SES groups, but students in SES group 3 consistently achieve higher scores across the entire range of values for the Belonging factor. The Motivation factor has a weak negative relationship with performance among those in wealthy schools and a weak positive relationship for those in poorer schools.

For Efficacy, Rejection, Engagement, and Enjoyment, the bivariate relationship between the factor scores and reading performance does differ by SES group. This difference is most substantial for the Efficacy factor. The relationship between positive self-efficacy and reading performance is much stronger for those in wealthy schools. For students in wealthy, performance increases at a higher rate with Efficacy than is the case for those in poorer schools. Moreover, at low levels of self-efficacy, there is no difference between SES groups. This finding could indicate that students in wealthy schools that have low self-efficacy perform no better than do their counterparts in poorer schools with equally low self-efficacy. A bold interpretation is that low self-efficacy negates the positive relationship

Stellenbosch University https://scholar.sun.ac.za

between wealth and reading performance.[12] Enjoyment has a slightly stronger positive relationship with reading performance among those in SES groups 1 and 2 when compared to those in SES group 3. A similar pattern is found for the Rejection factor, where there exists a stronger negative relationship between the factor score and reading performance for those in SES groups 1 and 2 compared to those in SES group 3.

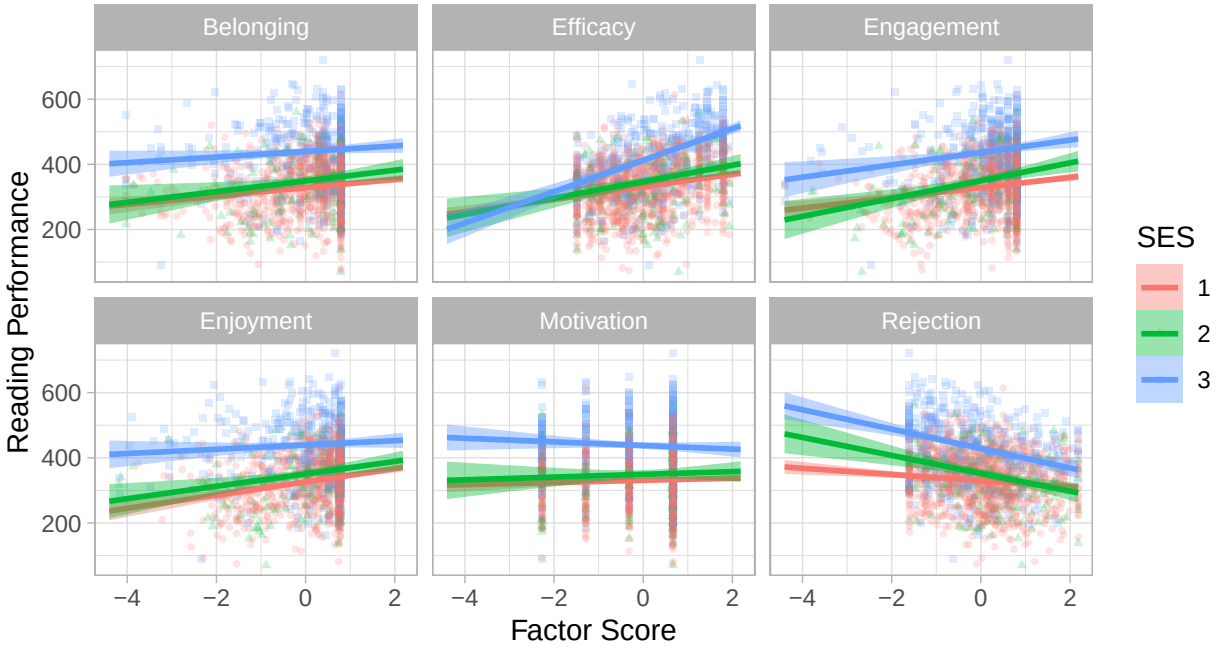


Figure 4.4: Scatter Plot With Fitted Linear Regression Line - by Factor and SES group

5. Results and Discussion

The empirical results section is divided into Section 1 and Section 2. Section 1 contains the results for the individual datasets based on the respondent (Home (parent), School (school head), Teacher, and Student), each of which includes a sample of observations captured within that dataset. Therefore, there are four individual datasets analysed in Section 1. The "Home" dataset contains responses from the students' parent, the "School" dataset contains responses from the students' school head, the "Teacher" dataset contains responses from the students' teacher, and the "Student" dataset contains self-reported responses from the student. Section 2 contains a single dataset that combines the top 10 most important features from the four individual datasets in Section 1. The analysis in Section 2 is sub-divided into 3 socioeconomic groupings, while the total combined sample is also analysed. The value added from the Section 2 results is that it analyses only those features that most strongly

[12] Without proper evidence of causality, such claims are purely speculative.

Stellenbosch University https://scholar.sun.ac.za

predict reading performance and it performs a disaggregated analysis based on socioeconomic status.

The analyses in both Sections 1 and 2 use variable importance plots to aid the interpretation of the results, which employ measures of relative feature importance and permutation feature importance (Friedman, 2001; Breiman, 2001).[13] The variable importance plots include the results of regressions fit using both the testing and training feature sets. While this is atypical, it is a novel means of gaining some measure of statistical significance. The intuition is this: if a feature is found to be a strong predictor of reading performance for both the testing and training set, the strength of its predictive power is robust. Moreover, analysing the results of both feature sets enables consideration of possible over- or under-fitting when interpreting the results of the regressions. The variable importance plots in Figures 5.1 and 5.6 are similar to OLS coefficient plots. Each plotted point corresponds to the importance score of a particular variable. Therefore, each variable has two points, one for the Permutation importance (Permute) score and one for the Relative Importance score (Importance). The further to the right the point, the larger the value of the estimated Permutation or Relative importance measure.

Additionally, the distance between the two points, the difference between the estimate obtained using the testing data and that obtained using the training data, serves as a pseudo measure of statistical significance. Importantly, variable importance plots do not show the direction of the relationship, they only show its strength. In order to analyse the direction of relationships, partial dependence plots are used. Each figure includes four plotted points which are coded as follows in the plot legend: Permutation feature importance using the training set (Train.P) and testing set (Test.P), and relative feature importance using the training set (Train.I) and the testing set (Test.I). Moreover, Appendix F contains a full variable list that outlines the survey questions from which each feature is derived.[14]

Furthermore, two- and three-dimensional partial dependence plots are used to more closely analyse the functional relationship between individual features and reading performance. Two-dimensional partial dependence plots provide information about the relationship between individual features and model predictions and show how model predictions partially depend on the values of individual features (Friedman, 2001). That is, they graphically reveal how model predictions change as the value of a specified input feature changes.[15] Three-dimensional partial dependence plots are theoretically identical to their two-dimensional counterparts. However, three-dimensional plots provide additional information between the values of two specified input features, and how the interaction between these two values influences model predictions.

[13] See Appendix D for a more detailed outline of variable importance measures.
[14] It is worthwhile referring to this list as the feature names presented in the results are essentially codenames that are not necessarily fully informative.
[15] See Appendix D for a more detailed outline of partial dependence plots.

Stellenbosch University https://scholar.sun.ac.za

5.1. Section 1 Results - Individual Datasets

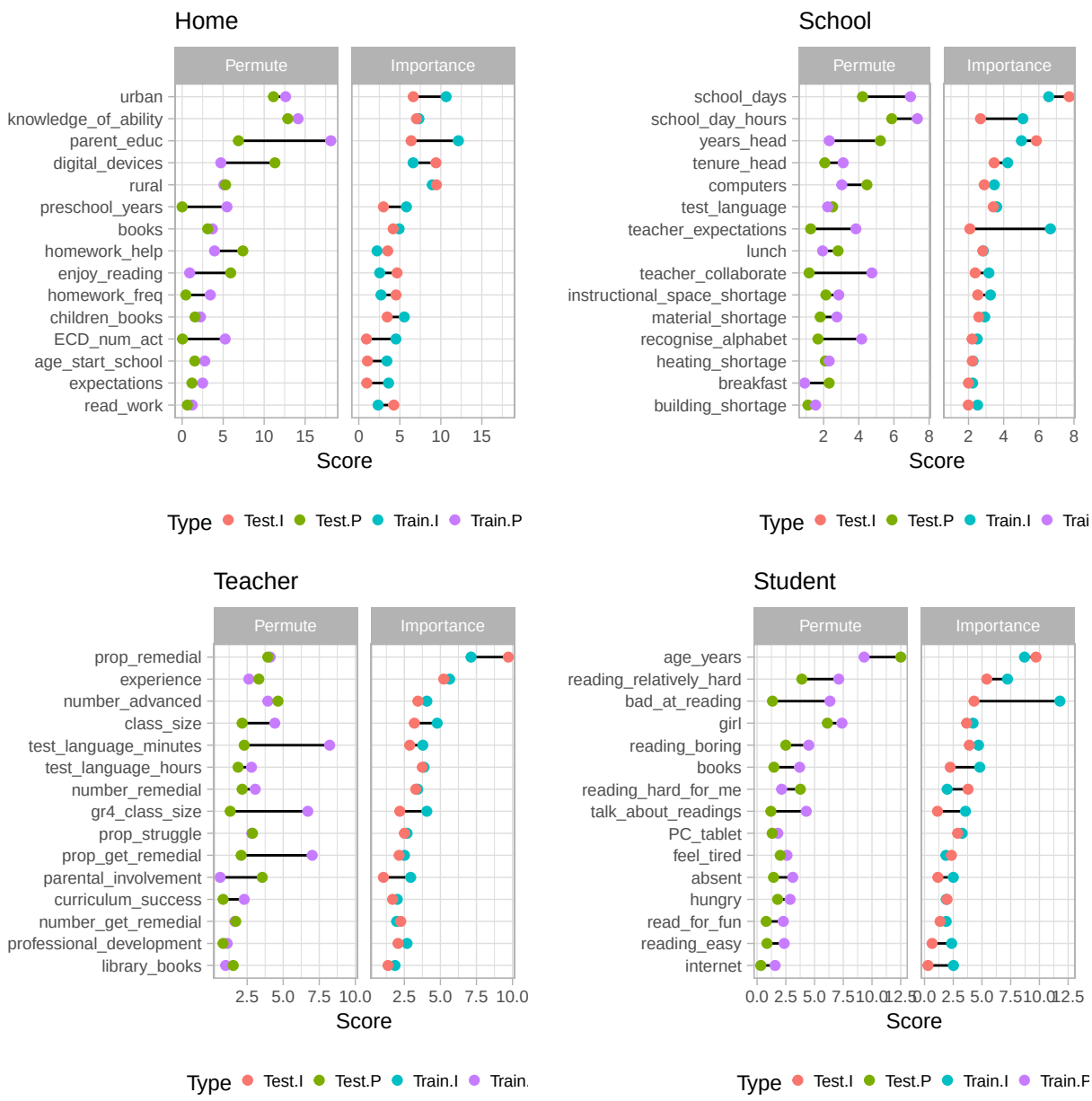


Figure 5.1: Section 1 Feature Importance Plots - Relative Feature Importance and Permutation Feature Importance

5.1.1. Home dataset

The top left variable importance plot shows the results of the gradient boosted regression fit using the Home dataset, which contains responses from the parents or guardians of the sampled students. The strongest predictors of reading performance from this dataset are the location of the household

Stellenbosch University https://scholar.sun.ac.za

(urban/ rural), the parents' knowledge of their child's ability, the level of education of the parent, and the number of digital devices in the household. It is evident that household demographics and wealth are the strongest predictors of the reading performance of the children in the household. Moreover, the importance of the number of books in the household (books) and the degree to which the parent enjoys reading (enjoy_reading) indicate that a cultivated culture of reading in the household can positively impact on the reading ability of the children within the home. Among the weakest predictors of reading performance are perceptions of school safety, perceptions of the degree to which the school helps the students succeed and whether or not the student was born in South Africa. These findings highlight the importance of entry factors and the home sphere as a centre of learning. It is evident that household characteristics, and features that proxy for a culture of reading in the household are strong predictors of reading performance.

Of the findings from the Home dataset, several have policy relevance. Most prominent among these is the finding that the parents' knowledge of their child's ability is a strong predictor of reading performance. This finding reveals that there may exist possible benefits to taking actions designed to improve the ability of parents to identify shortcomings in their child's educational progress and work to overcome them. This knowledge can be used to create and implement programs designed to assist the child at home to improve literacy. Shifting some of the burden from under-resourced schools into the household may be a cost-effective means of improving the literacy ability of students. In some cases, such a method may prove challanging given the limited litracy ability of parent and others in the household. Another policy-relevant finding is that students that receive more homework, and those with parents that more frequently help with homework, perform better. This finding, coupled with that of the parents' knowledge of child's ability, could indicate that students benefit greatly when they are frequently given homework and their parents are willing and able to assist with the homework.

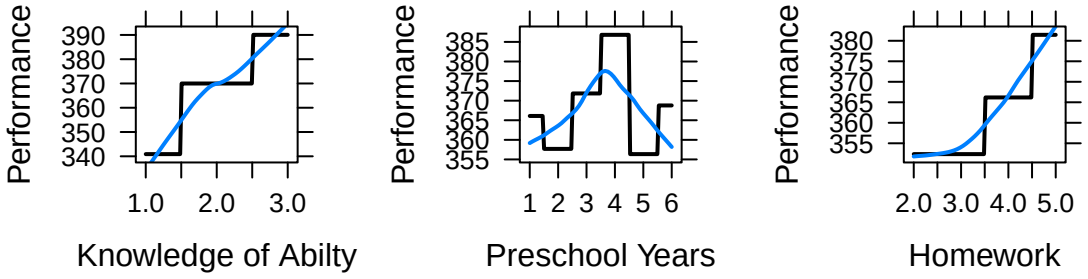


Figure 5.2: Partial Dependence Plots - Select Features From the Home Dataset

The results of the three partial dependence plots in Figure 5.2 show the predictive relationships between three included features and reading performance. Both homework frequency and parents' knowledge of their child's ability have strictly increasing relationships with reading performance. That is, a

Stellenbosch University https://scholar.sun.ac.za

students' reading performance is always predicted to be higher at higher values of these two features. The plot for years of preschooling shows an inverted "U" shaped relationship, indicating that students with very few or very many years of preschooling tend to perform worse than do those with moderate, more typical, years of preschooling. However, this finding is likely reflecting the nature of the type of students that receive extreme years of preschooling rather than the effect of too much or too little preschooling on reading ability. The findings from the partial dependence plots do not add much to the already noted policy relevance of the findings from the Home dataset.

5.1.2. School Dataset

The top right variable importance plot in Figure 5.1 shows the results of a gradient boosted regression fit using the School dataset, which contains responses from the students' school head. The most important features are those that reflect the experience of the school head (years_head, tenure_head) as well as those that contain information about the length of the typical school day and year (school_days, school_day_hours). Furthermore, several features that proxy for the resources of the school are important predictors of reading performance, such as the number of computers and school infrastructure.

Again taking the perspective of policy relevance, the only feature of interest from the findings plotted in figure 5.2 is the measure of the degree to which teachers in the school collaborate with one another. An important value that can be derived by having teachers, specifically teachers from various grades, collaborate is that it could foster a better understanding of the level of ability of students as they move sequentially to higher grades. Students could benefit in terms of a more tailored approach to their education if each teacher has a better understanding of the level required by the teachers of the grade above, and the teachers in the grade above have a better understanding of the ability of the students coming up from the grade below.

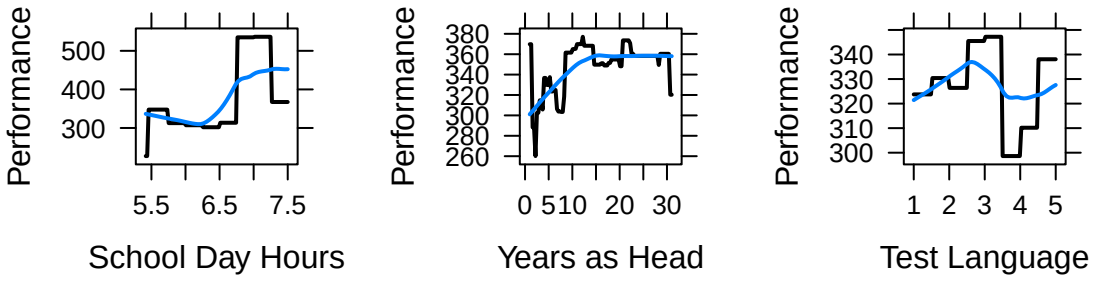


Figure 5.3: Partial Dependence Plots - Select Features From the School Dataset

From the plots in Figure 5.3, it is evident that the experience of the school head, the number of hours in a typical school day, and the frequency with which the test language is taught in class are

Stellenbosch University https://scholar.sun.ac.za

all mostly positively related to the predicted reading performance of students. The experience of the school head has a positive relationship with student performance until a point, about 15 years, after which the gains to additional experience diminish. For the number of hours in the typical school day, the influence of hours is negligible at lower values and is only positively related with predicted reading performance at higher values, indicating that more schooling hours above a certain threshold could benefit reading performance. However, in reality, there is very little variation in the length of school day between most schools. This finding is likely due to measurement error and incorrect response by school heads.

5.1.3. Teacher Dataset

The bottom left variable importance plot in Figure 5.1 shows the results of a gradient boosted regression fit using the Teacher dataset, which contains responses from the students' teacher. A substantial number of the most important features are those that relate to the composition of the students in the classroom, specifically, the relative number of students that require some form of teaching that is different from what is typical (remedial or advanced teaching). While this finding is possibly capturing the poor performance of those that need remedial teaching, it could also be indicating that the teaching process is hindered by heterogeneous ability in the classroom. This finding indicates that it may be difficult for a teacher to adequately teach several students of differing ability. Additionally, features that pertain to the frequency with which the test language is taught in the classroom are also found to be strong predictors of reading performance. The extent to which parents are involved with their child's education is also found to be an important feature. This finding is corroborated by those of regressions fit to the other datasets. That is, parental input into the learning process is a strong predictor of reading performance. These findings again highlight the importance of the home sphere. Moreover, a possible channel through which heterogenous classrooms influence reading performance is through the social fractures that result. That is, relationships between teachers and individual students are unlikely to form. Moreover, this may increase the social fractionalization among the student in the classroom, leading to increased feelings of rejection.

Both of these findings - the importance of classroom composition and parental input - have policy relevance. There exists potential value in further subdividing classes based on ability, or possibly taking students out of less necessary lessons and rather placing them in remedial reading lessons. Moreover, the finding that parental involvement is a strong predictor of success, coupled with the earlier-mentioned findings that parental knowledge of their child's ability and helping with homework are also strong predictors of success, further indicates that there exists possible value in increasing the capacity and ability of parents to aid in their child's learning process. Taken in combination, these two findings indicate that shifting the burden of remedial reading lessons to the household by aiding parents and enabling them to adequately assist their children may be productive. That is, there exists possible value in cultivating centres of learning within the household.

Stellenbosch University https://scholar.sun.ac.za

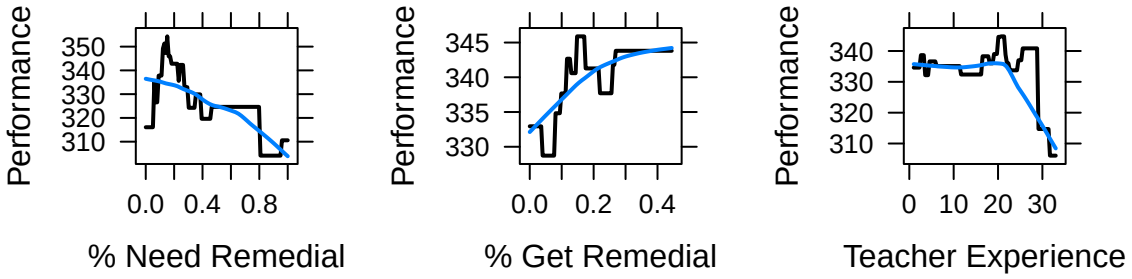


Figure 5.4: Partial Dependence Plots - Select Features From the Teacher Dataset

5.1.4. Student Dataset

The bottom left variable importance plot in Figure 5.1 shows the results of a gradient boosted regression fit using the student dataset, which contains responses from the student. Demographic and home background characteristics such as age, gender, the number of books at home, internet connection, hunger and absenteeism are found to be important predictors of reading performance. In addition to these characteristics, opinions towards reading are also found to be strong predictors. Features that reveal perceptions of one's ability, aptitude, and preferences regarding reading are found to be important predictors. The importance of latent constructs is made evident by the relative importance of the features: reading_relatively_hard (Efficacy) and reading_boring (Enjoyment of mathematics). An interesting finding regarding the opinion-based features is that negative opinions appear to be much stronger predictors than positive opinions. From the variable importance plot, features that relate to feelings of low self-efficacy and low interest are stronger predictors than features that relate to the enjoyment of, and positive self-efficacy for, reading. This finding indicates that psychological processes do play an important part in the learning process and are fundamental components of the data generating process underlying observable reading ability.

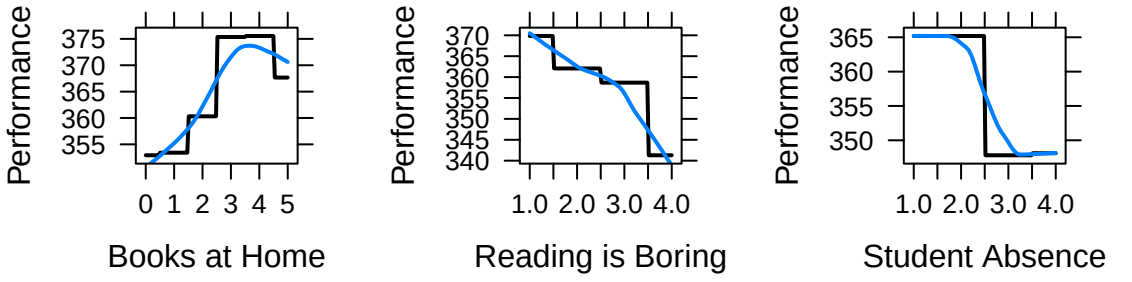


Figure 5.5: Partial Dependence Plots - Select Features From the Student Dataset

Stellenbosch University https://scholar.sun.ac.za

5.2. Section 2 Results - Combined Dataset

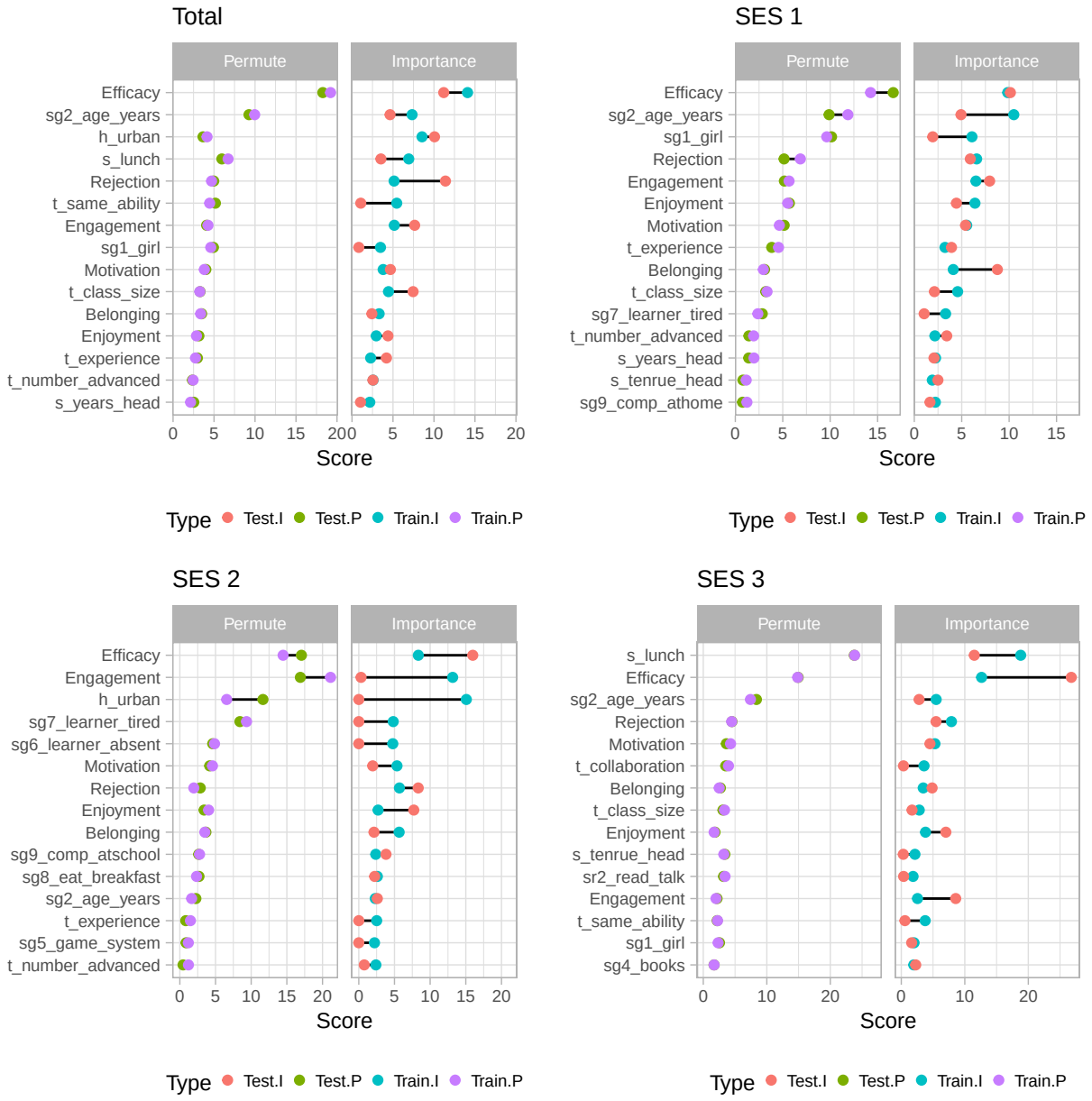


Figure 5.6: Section 2 Feature Importance Plots - Relative Feature Importance and Permutation Feature Importance

Figure 5.6 contains the results of four separate gradient boosted regressions, with one fit using the entire combined dataset, and three fit using one of each of the SES groups. Moreover, the data used to fit these regressions includes the estimated factors for the studied latent constructs - Efficacy, Motivation, Belonging, Rejection, Enjoyment, and Engagement. From all four regressions, self-efficacy

Stellenbosch University https://scholar.sun.ac.za

is consistently a very strong predictor of reading performance.[16] Moreover, the age of the student, social rejection, and the level of engagement with reading are also consistently strong predictors across all four regressions. Classroom composition is also found to be a strong predictor by all four. This finding further highlights the importance of this classroom characteristic and adds additional weight to the need for policy implementation that targets it. Other features that are found to be important in all four regressions are the experience of both the teacher and the school head, as well as the number of students in the classroom. Moreover, these findings indicate that psychological processes are important components of reading performance.

Several inconsistent findings are worth noting. The gender of the student is a substantially stronger predictor among SES group 1 than it is among the other two groups. Moreover, individual factors such as the urban/rural location of the household are not found to be important among the poorest students. This is likely due to the relative demographic homogeneity of the majority of these students. For the poorest schools, opinion-based and psychological features are the strongest predictors, marginally more so than for the other SES groups. Furthermore, for SES groups 1 and 2, the frequency with which students experience tiredness at school is an important predictor, which is not the case among those in SES group 3. This finding further suggests that household factors and characteristics are very important, in addition to school-level factors. There may be value in policies directed at effecting changes in the household rather than an exclusive focus on the school environment.

A notable finding that warrants more discussion is the relative importance of Motivation and Engagement across the SES groupings. It is useful here to recall that the Motivation factor reveals a self-motivated engagement with reading while the Engagement factor reveals more teacher- and classroom-motivated engagement. From Figure 5.6, it is evident that Engagement is a stronger predictor of reading performance among SES groups 1 and 2 than is Motivation. For those in SES group 3, Engagement is a relatively weak predictor. This indicates that teacher input into student engagement with reading is more important for poorer students while wealthier students are more reliant on their own input (or effort) for higher reading performance. The findings for SES group 2 are relatively unreliable, as made evident by the large spread between the estimates derived from the testing and training data. However, most of the features found to be strong predictors by this data are similar to those fit by the other regressions - self-efficacy and proxy features for household and school wealth.

From the results of the regressions fit using only students in SES group 3, the most important feature is whether or not the student receives lunch at school, this relationships is negative (results not shown). This finding highlights the relative difference in wealth within the top wealth quintile and the effects of this relative difference on reading performance. This is a characteristic of South African inequality, the top wealth quintile of schools is the most heterogeneous in terms of wealth and living standard

[16] it is difficuilt to assign causality here. It could be the case that these students have high self-efficacy because they have performed well academically in the past.

Stellenbosch University https://scholar.sun.ac.za

(Armstrong at al., 2008). The feature for lunch receipt can be interpreted as a proxy for school wealth, as most wealthy schools (except for a small number of boarding schools) would not typically provide lunch to their students, while some poorer schools might have lunch feeding schemes. Therefore, even within the top quintile of wealthy schools, wealth differentials remain the strongest predictors of reading performance.

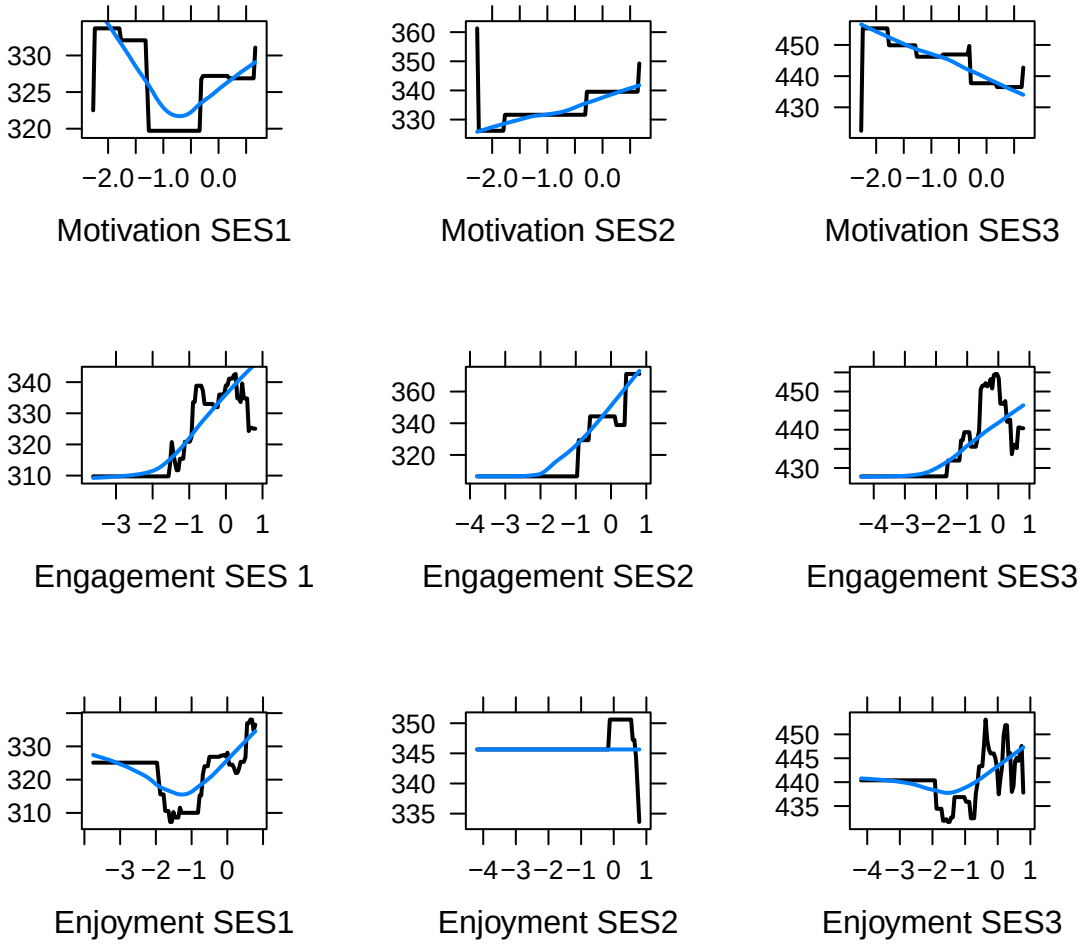


Figure 5.7: Factor Partial Dependence Plots - Motivation, Engagement, and Enjoyment

The remainder of this subsection focuses on the results of only the estimated factors associated with the six latent constructs. To aid the interpretation of the results, Loess regression functions are plotted over the partial dependence plots in each figure. Figure 5.7 shows partial dependence plots for the factors of Motivation, Engagement, and Enjoyment on reading performance by SES group. Motivation provides extremely inconsistent predictions of predicted reading performance across the three SES groups. For SES group 3, reading performance is strictly decreasing in Motivation,

Stellenbosch University https://scholar.sun.ac.za

for SES group 2 it is strictly increasing, and for SES group 1 it follows an inverted U shape and has a relatively weak relationship. The predicted relationship between Engagement and reading performance is consistent across all three SES groups, with performance increasing in Engagement. However, at higher levels of Engagement, performance appears to decline marginally. This may be the result of the scarcity of high levels of measured Engagement (See Figure 4.3). The findings for the impact of engagement indicate that taking actions designed to nudge students toward more active engagement with reading material could be productive. The Enjoyment factor has a relatively weak relationship with reading performance across all three SES groups while being marginally positively related to reading performance at higher values.

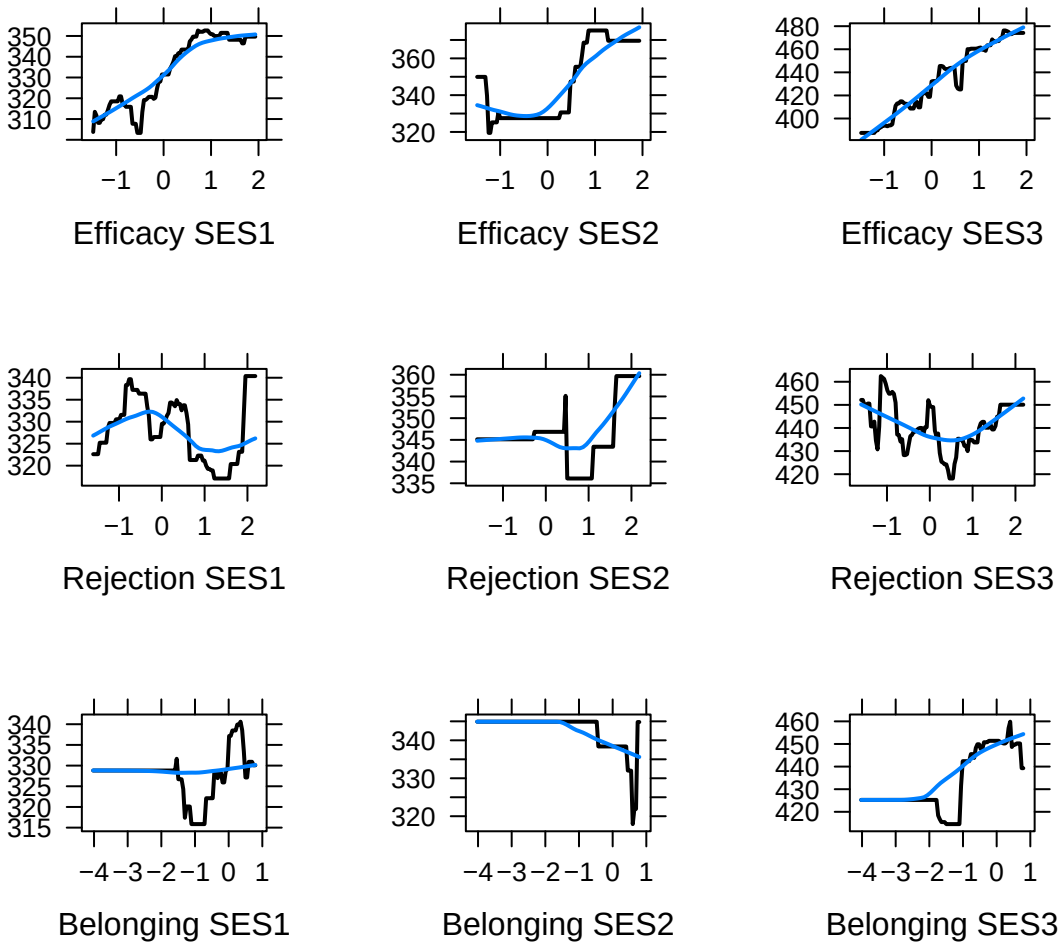


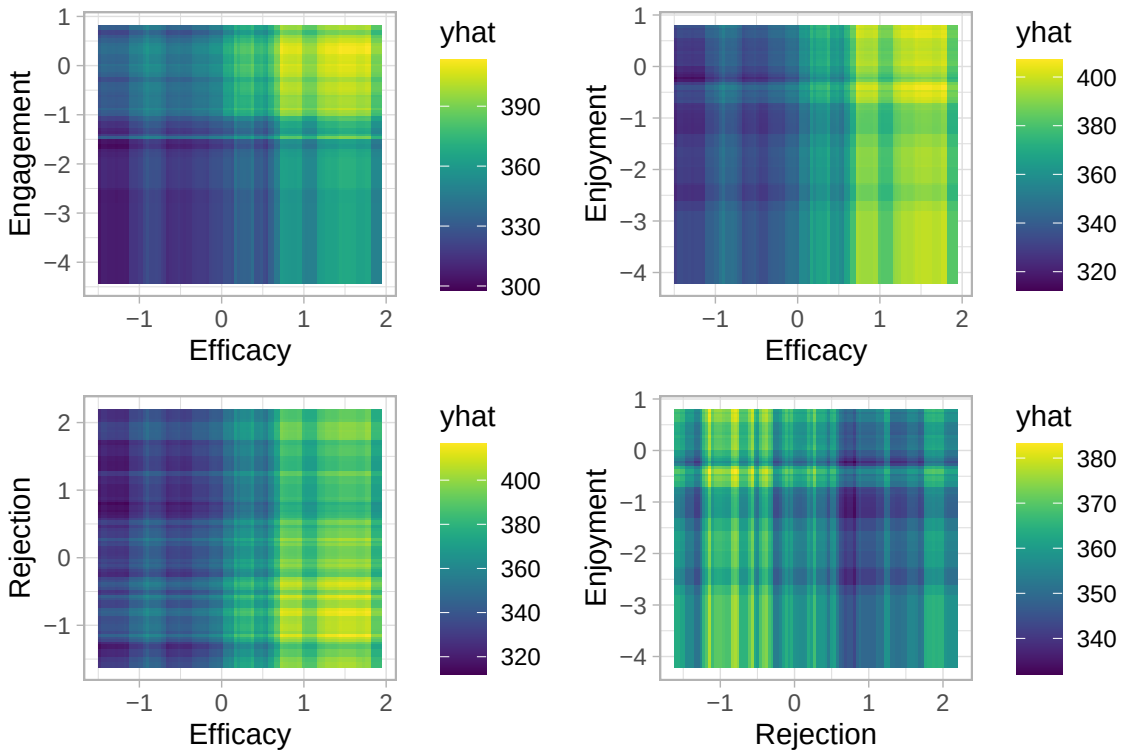
Figure 5.8: Factor Partial Dependence Plots - Efficacy, Rejection, and Belonging

Figure 5.8 shows partial dependence plots for the factors of Efficacy, Rejection, and Belonging with

Stellenbosch University https://scholar.sun.ac.za

reading performance by SES group. Predicted reading performance is mostly increasing in Efficacy, and this relationship is consistent across all three SES groups. Looking at the scale on the left of each figure for the Efficacy factor, it is evident that the gradient is steepest for SES group 3. This finding is corroborated by that shown in Figure 4.4. The Rejection factor has a relatively weak but consistent relationship with predicted reading performance across the three SES groups. However, it appears that predicted reading performance tends to be higher at higher levels of Rejection. The relationship between social standing and academic performance is a complex phenomenon. For example, it could be the case that students that do very well academically are seen as uncool. More simply, and in line with the specific variables used to create the Rejection factor, those that perform well may choose not to spend time with classmates as they prefer to focus on their studies. There are just two examples, either could be the case here. It is not possible to do more than speculate given the findings in Figure 5.8.

The Belonging factor also has an inconsistent relationship with predicted reading performance across the three SES groups. For SES group 1, there is no relationship. For SES group 2, higher levels of Belonging predict lower reading performance. For SES group 3, higher levels of Belonging predict higher levels of reading performance.



Stellenbosch University https://scholar.sun.ac.za

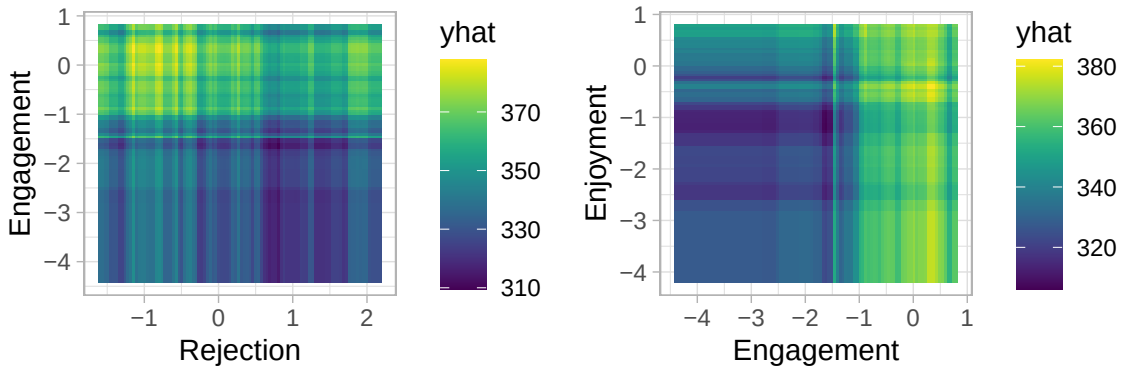


Figure 5.9: Interacted Partial Dependence Plots: top left - Engagement and Efficacy, top right - Enjoyment and Efficacy, middle left - Rejection and Efficacy, middle right - Enjoyment and Rejection, bottom left - Engagement and Rejection, bottom right - Enjoyment and Engagement

Figure 5.9 shows two-way interaction partial dependence plots (3D plots). Six individual plots are included, and all interactions between the factors of Enjoyment, Engagement, Efficacy, and Rejection are analysed. The plots are created using the entire sample, they are not split up by SES groups. As the entire sample is used, only relationships that are relatively consistent between the SES groups are analysed (see Figures 5.7 and 5.8). Therefore, the factors for Belonging and Motivation are excluded due to their inconsistent relationships with reading performance across the three SES groups. These plots are interpreted as three-dimensional figures, with the third dimension represented by the colour contours. Starting from the top left plot, and moving sequentially rightward from there, it is evident that predicted reading performance is substantially higher for students that have high levels of both Engagement and Efficacy. The predictive strength of the Efficacy factor is further highlighted by its interaction with Enjoyment. Predicted performance is high at higher levels of Efficacy regardless of the level of the Enjoyment factor. Interacting Efficacy and Rejection reveals that higher levels of Rejection offset the positive relationship between Efficacy and performance. That is, even among those with high levels of Efficacy, predicted reading performance is lower for those with higher measured levels of the Rejection factor. Interacting Rejection with Enjoyment further highlights the predictive power of Rejection. At all levels of Enjoyment, predicted reading performance is higher for students with lower levels of Rejection. This finding is further corroborated by interacting Rejection with Engagement. At higher levels of Rejection, predicted reading performance is lower, while it is highest at low levels of Rejection and high levels of Engagement. And finally, interacting Enjoyment and Engagement reveals that predicted reading performance is higher at higher levels of Engagement given any level of Enjoyment.

Stellenbosch University https://scholar.sun.ac.za

6. Policy Implications

The findings of this chapter lend themselves to several potential policy interventions that may be productive means of improving the reading ability of South African primary school students. The overarching theme of the ideas generated from this study relates to the need for increased importance to be given to the home sphere. The cultivation of a culture of reading in the household has the potential to greatly improve literacy rates in poorer areas. While this means of policy intervention is complex, novel methods such as "nudges" could prove useful (Thaler & Sunstein, 2009). The list of policy implications is limited to six main points, each of which can individually add value, but would be best implemented in combination with one another. Moreover, while they are six unique points, they do overlap with one another to a certain extent.

First, empower parents and provide the necessary resources to enable them to assist their child's education in the household. The findings in this chapter indicate that increased parental involvement is positively related to reading performance. This is in line with the broader findings throughout the chapter that a cultivated culture of reading in the household can positively impact on the reading ability. Therefore, a policy that targets increase parental assistance while simultaneously increasing the degree to which parents are capable of helping would be productive. Such a policy would need to affect parents by incerasing both their willingness and ability to help

Second, in conjunction with the policies targeting increased parental willingness and ability to help, the frequency with which students receive homework should be increased. This additional homework should be designed to enable maximum parental input. One possible practical implementation of this would be to send additional resources home with the student to give to their parents that can be used by the parent to assist with the homework.

Third, shift part of the remedial teaching burden to the household. In line with the first two points, taking steps to increase the amount of remedial teaching that is done in the household will greatly reduce the burden that is currently borne by teachers. Such a policy would require a similar practical application to the one described in the second point.

Fourth, take steps to create a more homogenous classroom composition based on relative ability. If possible, classrooms should be divided further based on reading ability. If the resources are not available, classroom division could be based on ability in a particular sunject rather than grade. Additionally, while extreme, it could also be productive to remove students from less necessary classes and rather place them in remedial reading classes. There is a relatively large literature that proposes that heterogeniety in the classroom is academically beneficial (Taylor, Muller, & Vinjevold, 2003). Therefore, more work is likely needed in this area. However, the findings in this chapter indicate that heterogenous classroom may not be a good idea.

Stellenbosch University https://scholar.sun.ac.za

Fifth, foster increased collaboration between teachers of different grades within the school. Such collaboration would improve the ability of teachers to target problem areas and address them before they persist and ultimately become binding constraints to further progress. Of all the proposed policy measures, this one will likely prove to be the most cost-effective.

Sixth, take steps to reduce the prevalence of negative social interactions with the classroom and school. The findings of this chapter highlight the importance of social rejection for reading performace. While causality and direction of effect are difficult to ascertain with certainty, its importance is evident. Therefore, programs designed to alleviate the negative effects of phenomena such as bullying could create an environment that is more condusive to teaching, learning, and academic success.

7. Conclusion

This chapter investigated the determinants of reading performance of South African grade 4 students with specific emphasis given to measured psychological factors. The approach taken is novel in terms of both the statistical methodology and the theoretical framework of the study. The empirical analysis makes extensive use of gradient boosted regression, a statistical learning technique that enables the analysis of complex non-linear and interactive relationships. To aid the precision of interpretation, the data was disaggregated and analysed in two stages. First, the data was separated into four individual sets based on the identity of the four different survey respondents, student, teacher, school head, and parent. Second, the features found to be the strongest predictors of reading performance in the first stage were combined into one dataset which was then disaggregated by socioeconomic status and again analysed using gradient boosted regression in the second stage.

The analysis yields several interesting and policy-relevant findings. First, psychological processes are important predictors of reading performance. Moreover, the importance of these measured factors is not consistent over the wealth distribution. Second, the importance of the household as a centre for learning is highlighted. Findings indicate that children with parents that are more willing and able to help and participate with their child's education perform better than those with parents that are less willing or less able. Third, this study finds that it is the composition of the students within the classroom that strongly predicts reading performance, specifically, composition in terms of the relative homogeneity of reading ability among the students in the class.

The approach taken in this chapter is novel and makes several contributions to the economics of education literature. First, it borrows extensively from the psychology literature. The inclusion of measured psychological processes is validated by the significant predictive power of these features. Moreover, even if used in a more general OLS-based analysis, controlling for these components of the data generating process will improve the fit and precision of estimated models. Second, this chapter makes extensive use of gradient boosted regression with tuned hyperparameters. While these mod-

Stellenbosch University https://scholar.sun.ac.za

els are uncommon in contemporary research in the field, when they are used, they are often used inappropriately and hyperparameters are left un-tuned. This chapter has highlighted the value of using these methods in research pertaining to the economics and psychology of education. Finally, this chapter has introduced a novel means of feature interpretation by combining the results of variable importance measures estimated using both the testing and training datasets. This approach enables interpretation that controls for possible over- or under-fitting and provides a pseudo measure of statistical significance.

Future work in this direction can make one of several improvements. First, the outcome variable of interest can be altered. Rather than using only reading performance, a measure of relative reading performance can be used. Second, the analysis can, given a sufficiently large training set, split the analysis further by gender to look at possible feature relationship heterogeneity based on gender. Third, to better accommodate the use of measured psychological processes and the effects of characteristics of the home sphere, the use of panel data would greatly improve the reliability of estimates and add increased significance to the findings of research on this specific topic.

To take stock, this chapter has shown one possibly productive way forward. It has demonstrated the value of using methods borrowed from the statistical learning and educational psychology literatures respectively. It has also introduced new methods of interpretation to the literature. The recent expansion in the availability of data on human behaviour, opinions, and preferences as well as recent advances in computational power ensure that this is an exciting time for research in this field. Researchers must take advantage of this by employing new techniques and borrowing extensively from several research frontiers rather than allowing their work to be constrained by contemporary academic norms.

Stellenbosch University https://scholar.sun.ac.za

8. Appendix A: Feature Space

8.1. Section 1 - Home Feature Space

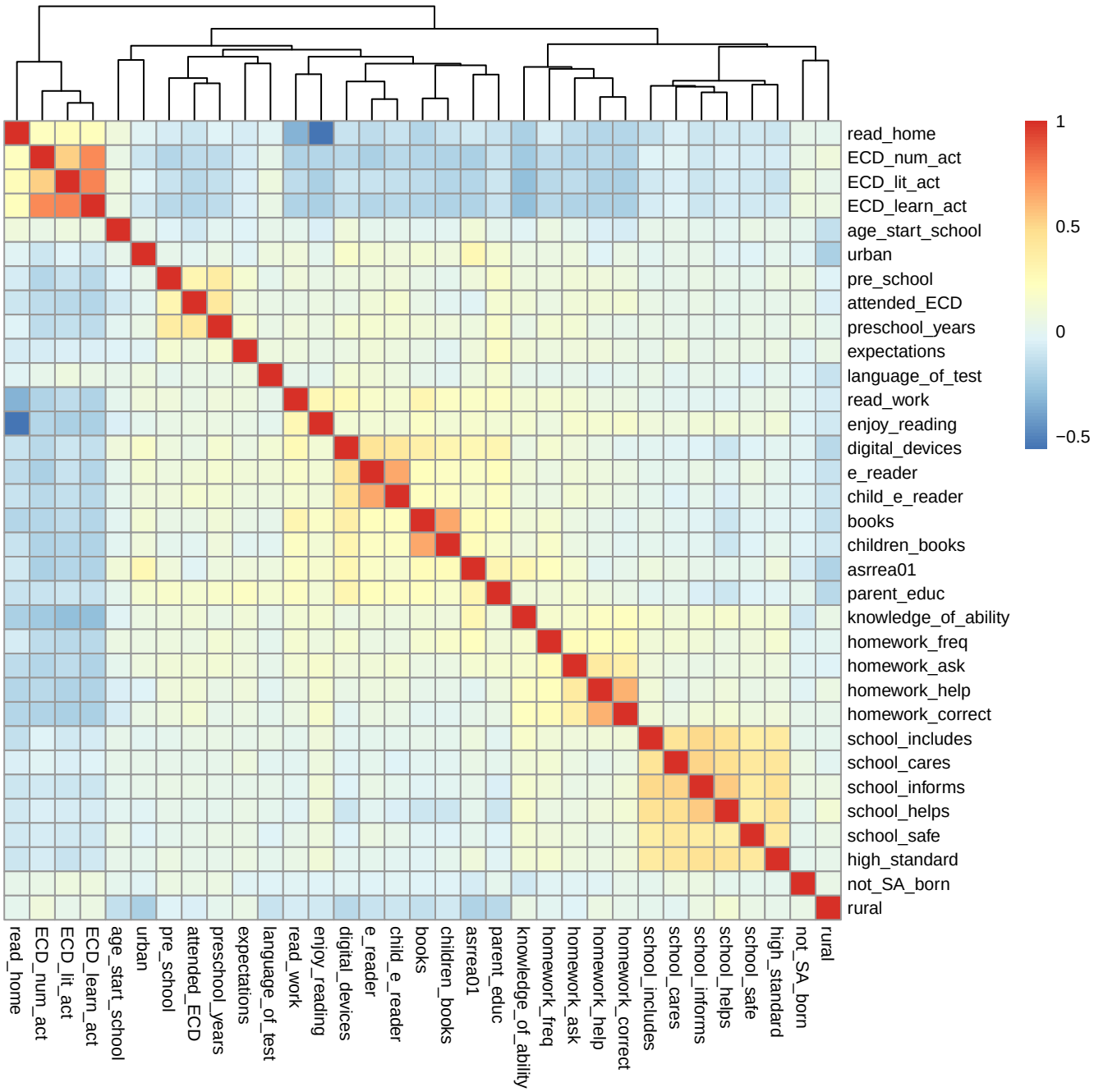


Figure 8.1: Feature Space Correlation Heatmap - Home Feature Space

Stellenbosch University https://scholar.sun.ac.za

8.2. Section 1 - School Feature Space

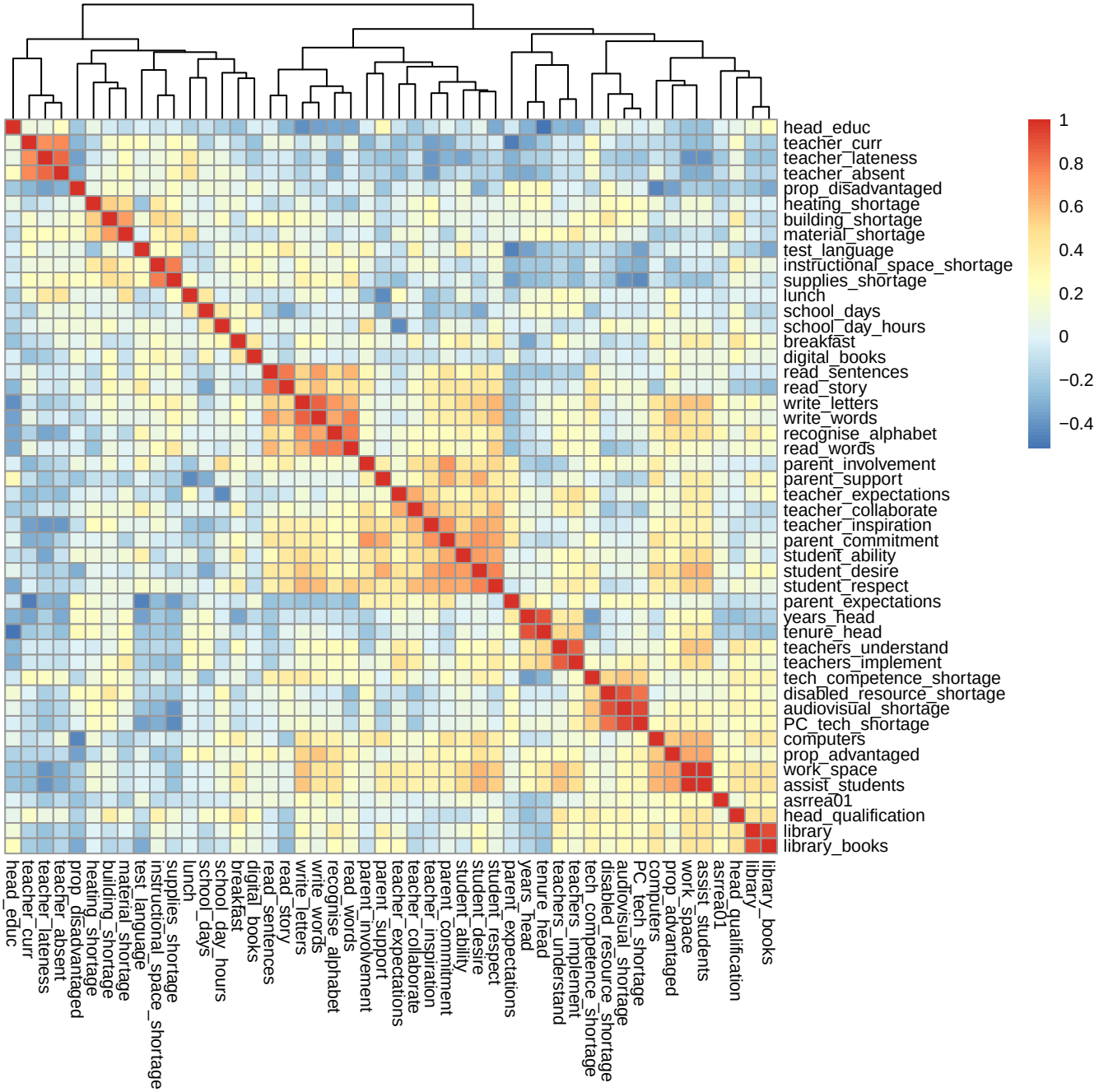


Figure 8.2: Feature Space Correlation Heatmap - School Feature Space

Stellenbosch University https://scholar.sun.ac.za

8.3. Section 1 - Teacher Feature Space

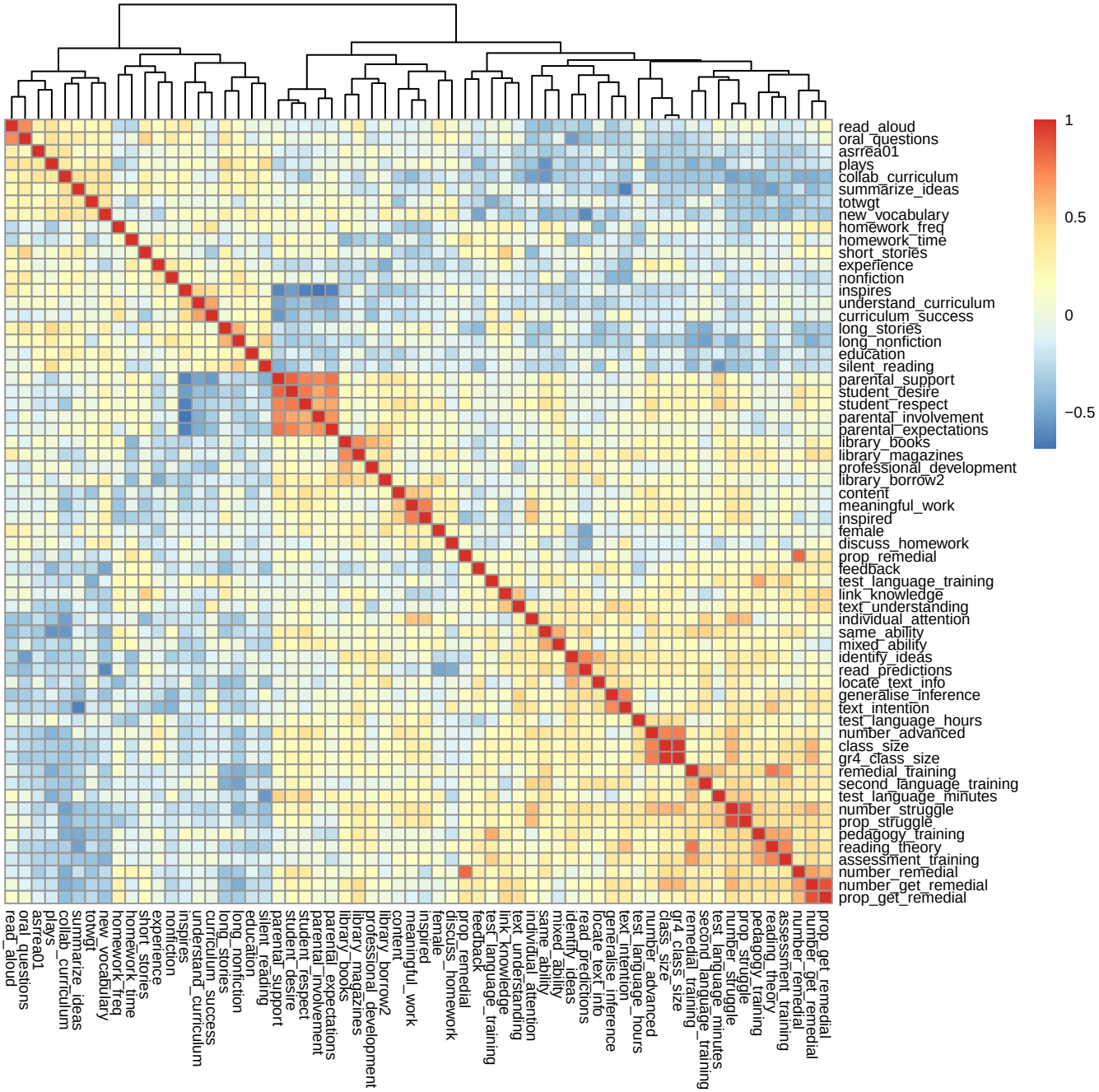


Figure 8.3: Feature Space Correlation Heatmap - Teacher Feature Space

Stellenbosch University https://scholar.sun.ac.za

8.4. Section 1 - Student Feature Space

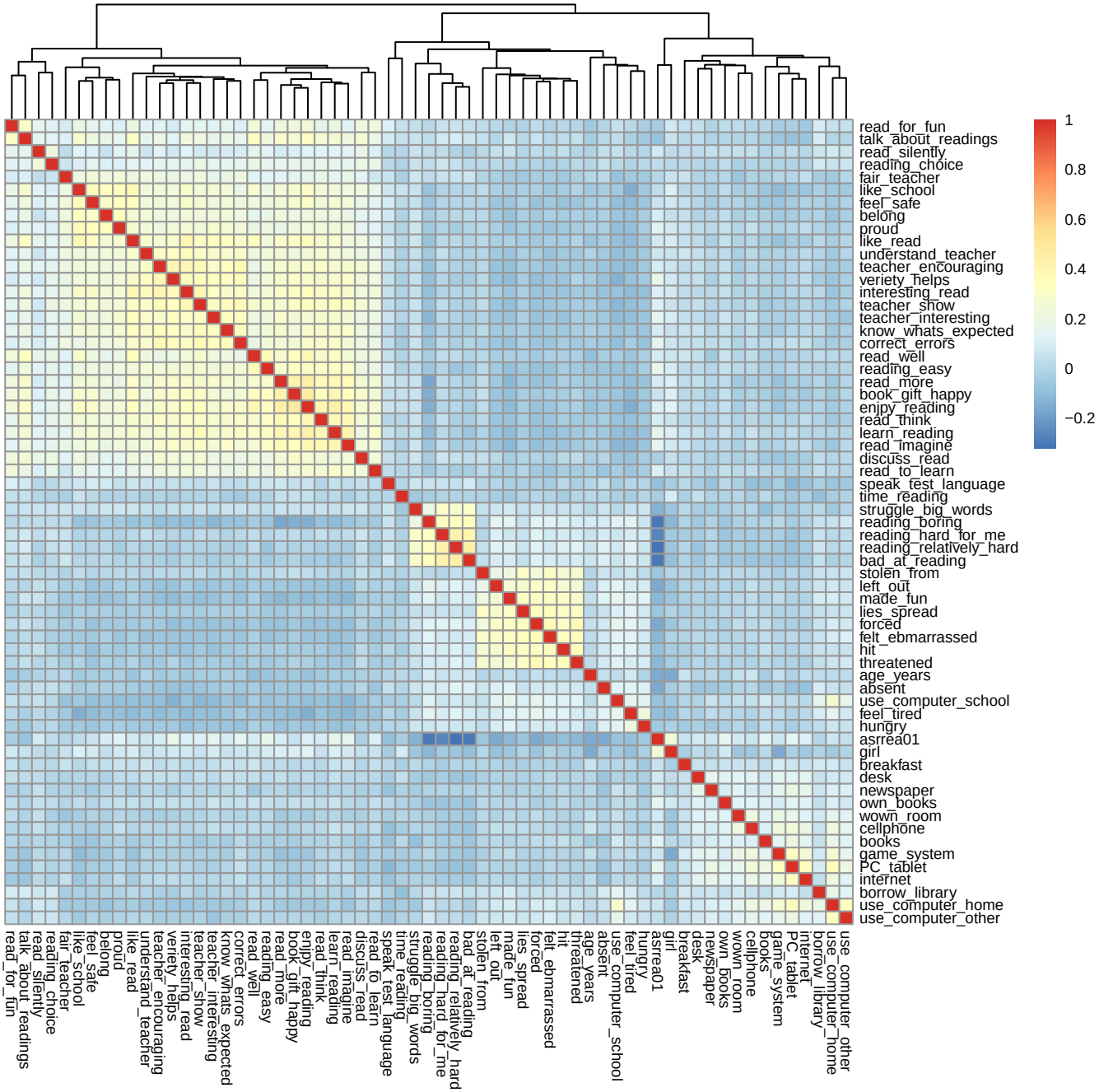


Figure 8.4: Feature Space Correlation Heatmap - Student Feature Space

Stellenbosch University https://scholar.sun.ac.za

8.5. Section 2 - Combined Feature Space Total

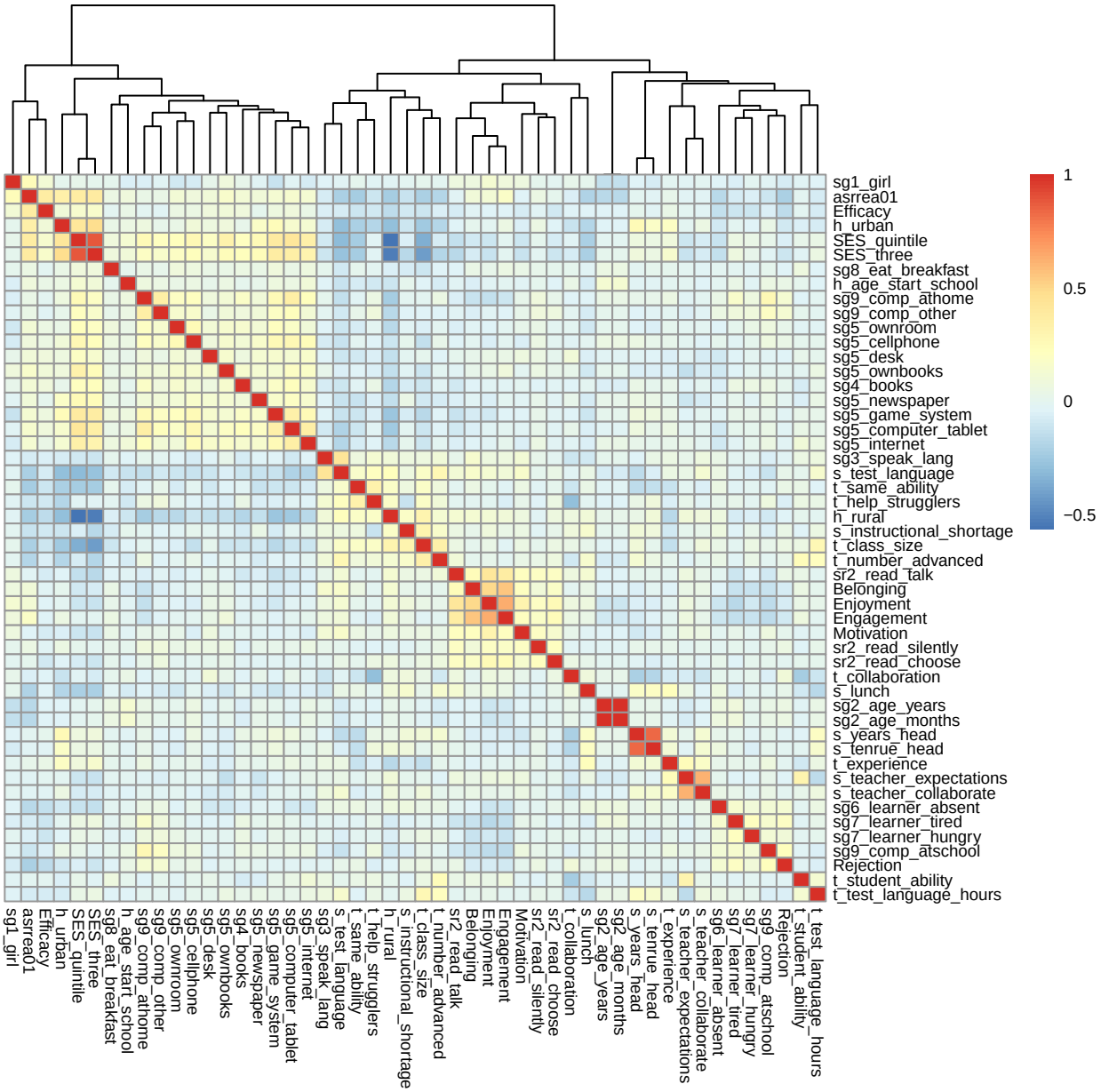


Figure 8.5: Feature Space Correlation Heatmap - Combined Feature Space (Total)

Stellenbosch University https://scholar.sun.ac.za

8.6. Section 2 - Combined Feature Space SES 1

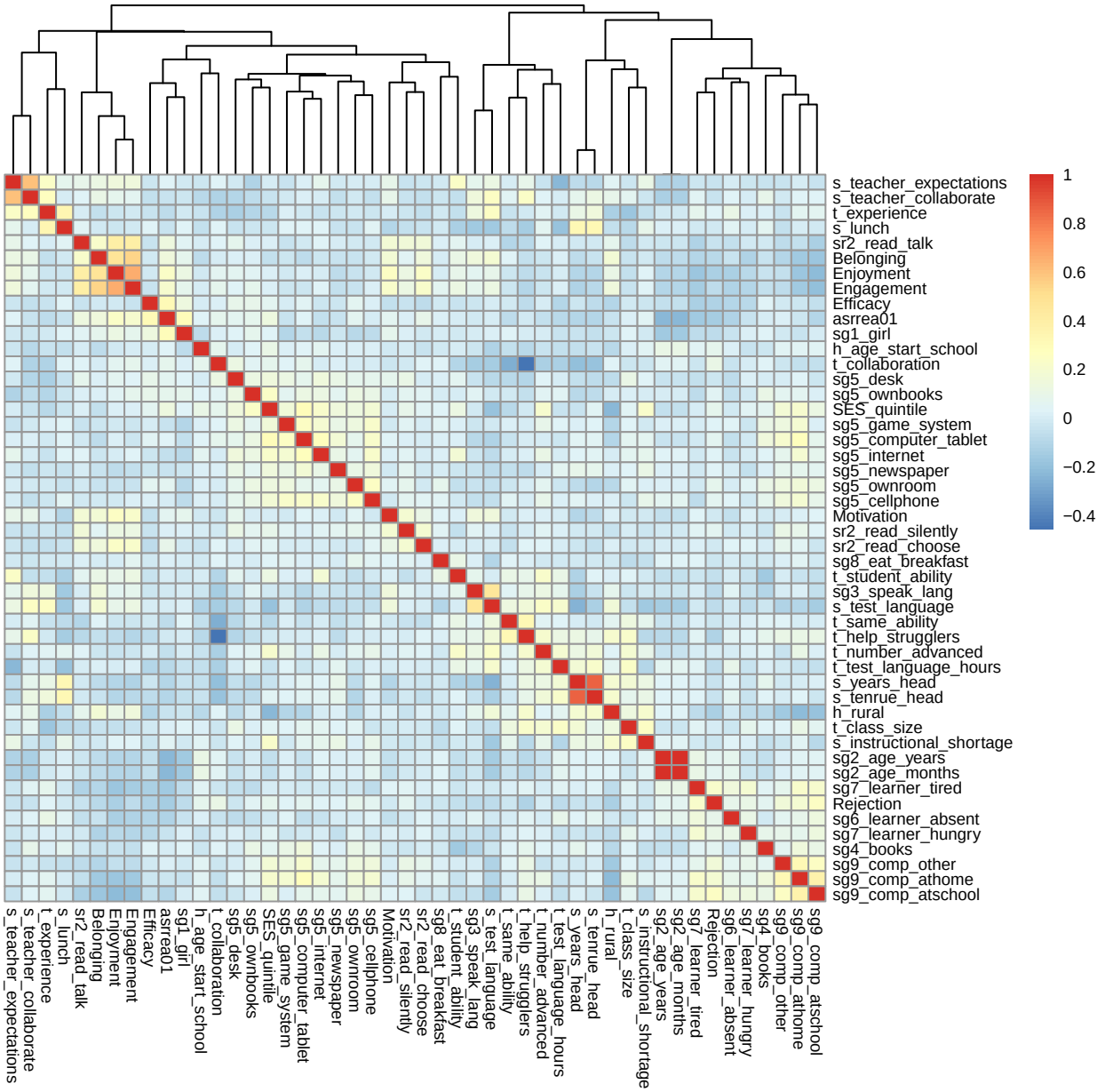


Figure 8.6: Feature Space Correlation Heatmap - Combined Feature Space (SES 1)

Stellenbosch University https://scholar.sun.ac.za

8.7. Section 2 - Combined Feature Space SES 2

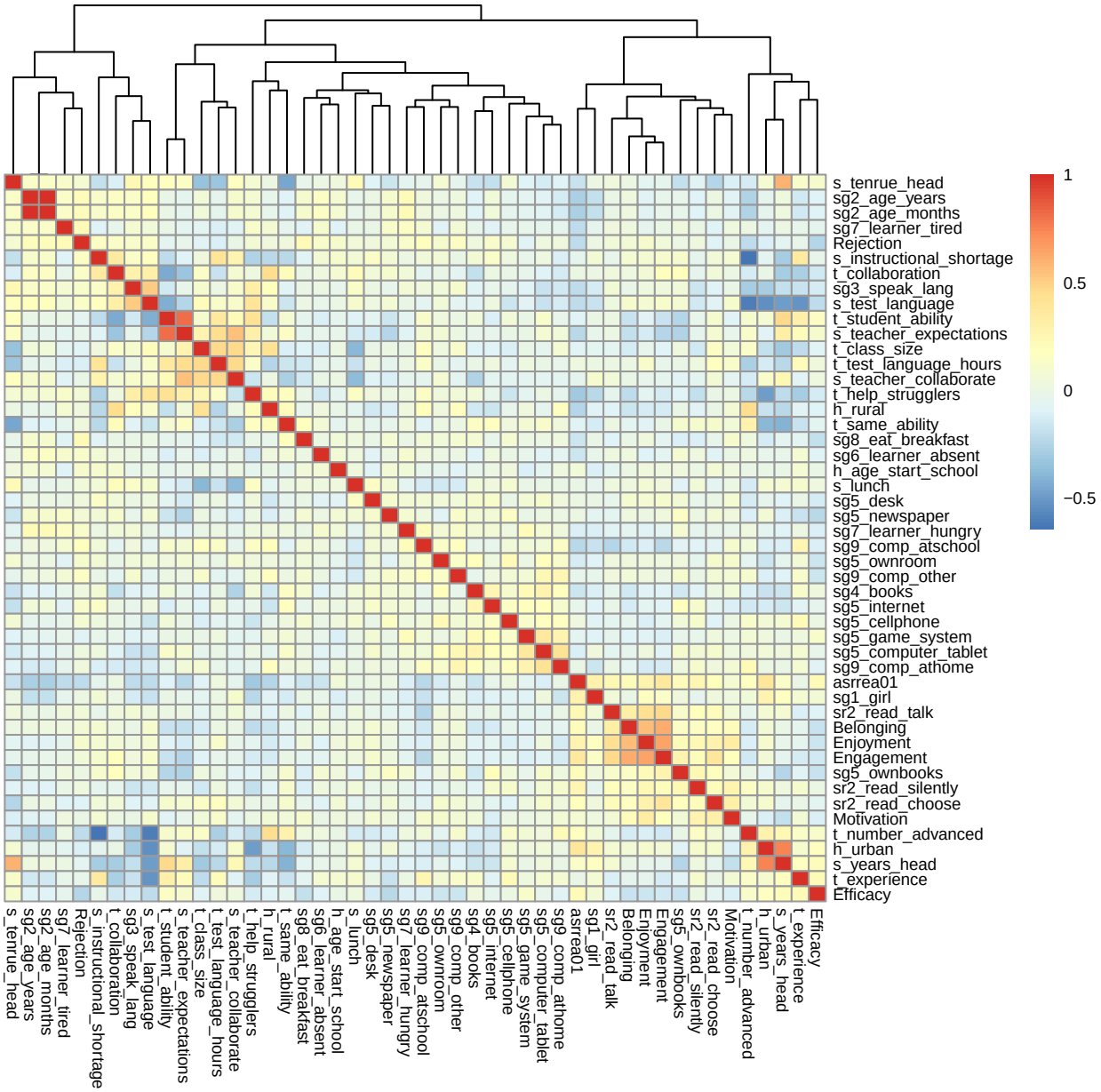


Figure 8.7: Feature Space Correlation Heatmap - Combined Feature Space (SES 2)

Stellenbosch University https://scholar.sun.ac.za

8.8. Section 2 - Combined Feature Space SES 3

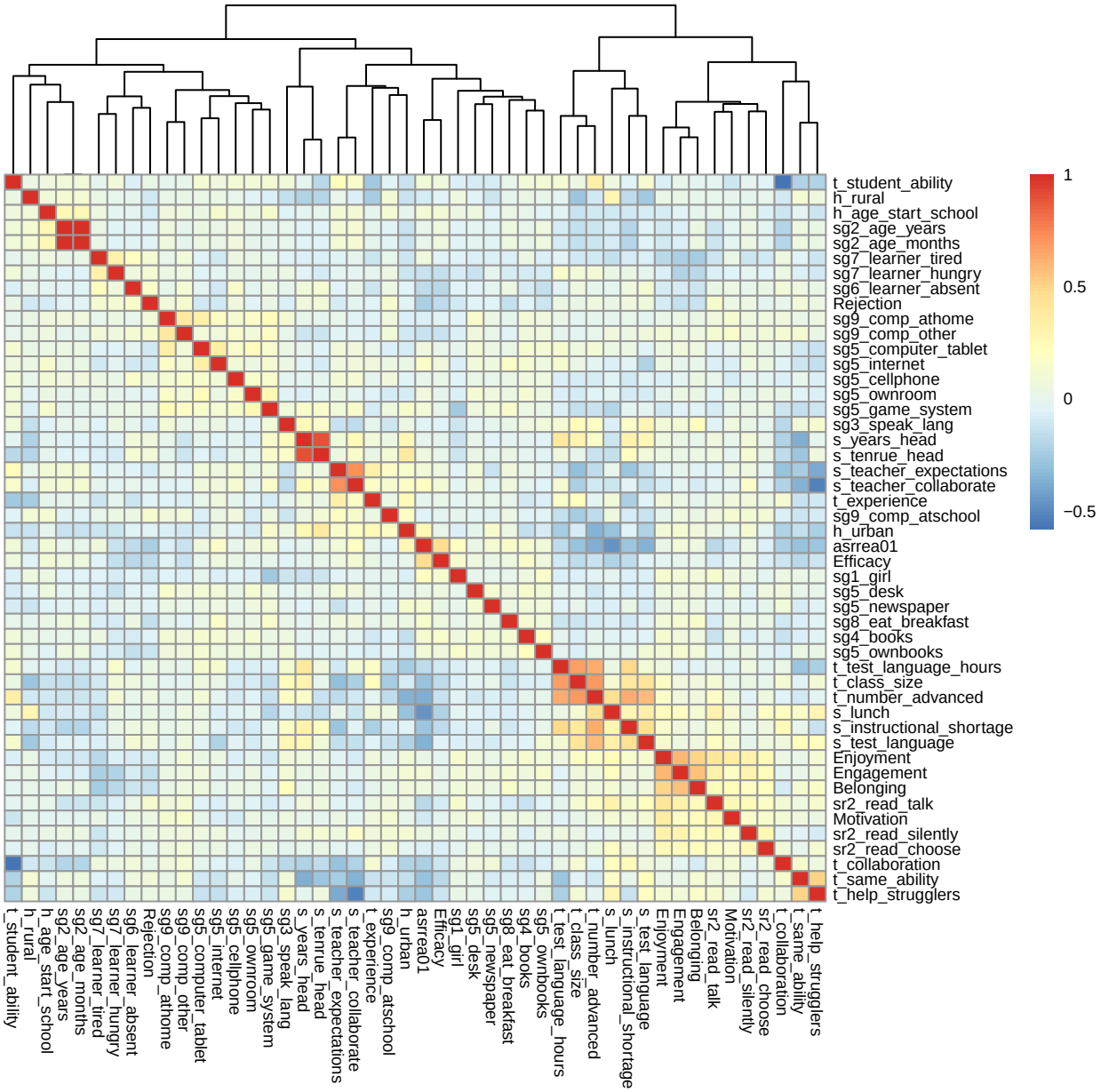


Figure 8.8: Feature Space Correlation Heatmap - Combined Feature Space (SES 3)

Stellenbosch University https://scholar.sun.ac.za

9. Appendix B: Section 1 Full Feature Importance Results

9.1. Full Home Feature Results

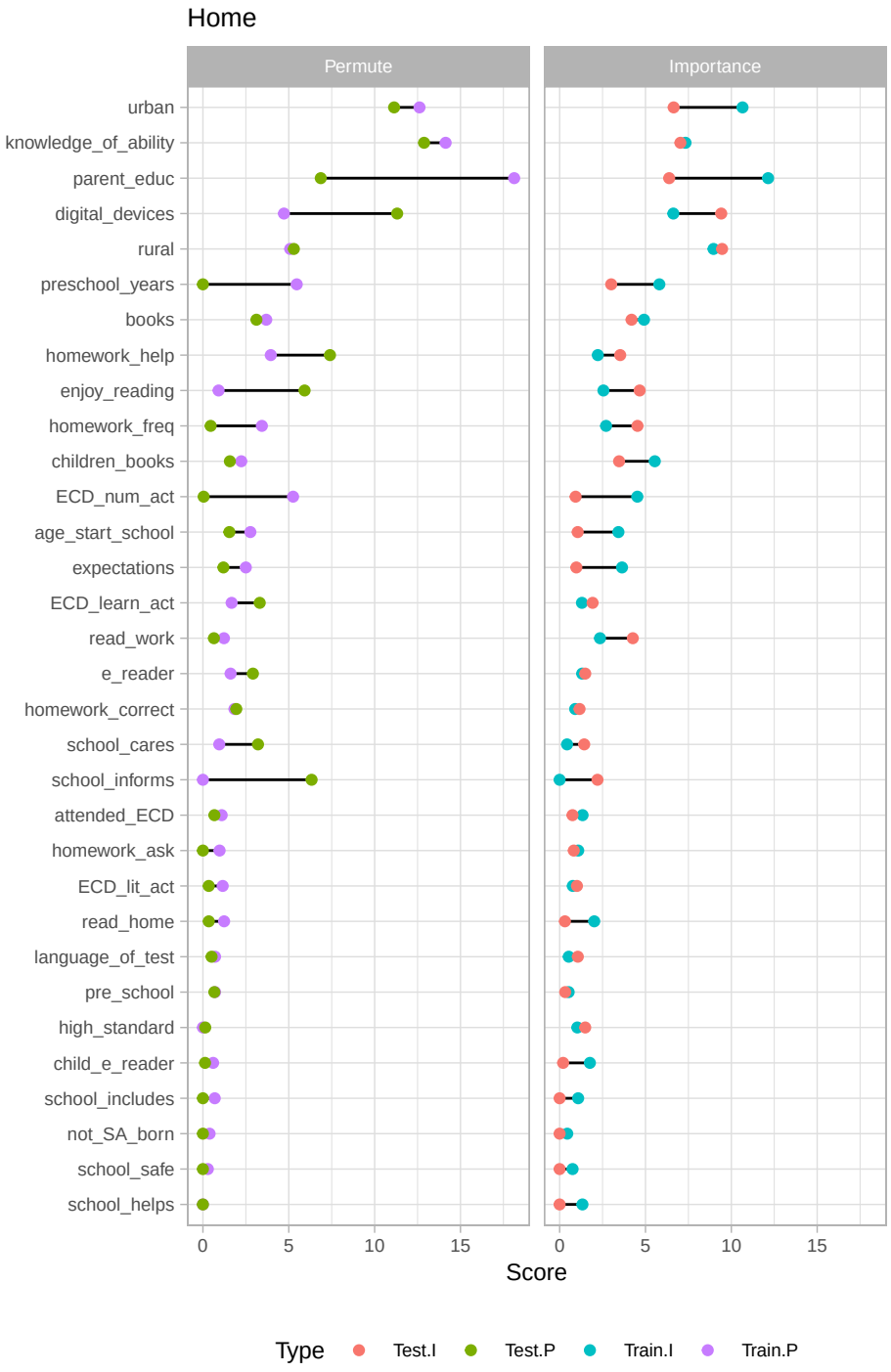


Figure 9.1: Full Feature Importance Plot - Home

Stellenbosch University https://scholar.sun.ac.za

9.2. Full School Feature Results

Figure 9.2: Full Feature Importance Plot - School

Stellenbosch University https://scholar.sun.ac.za

9.3. Full Teacher Feature Results

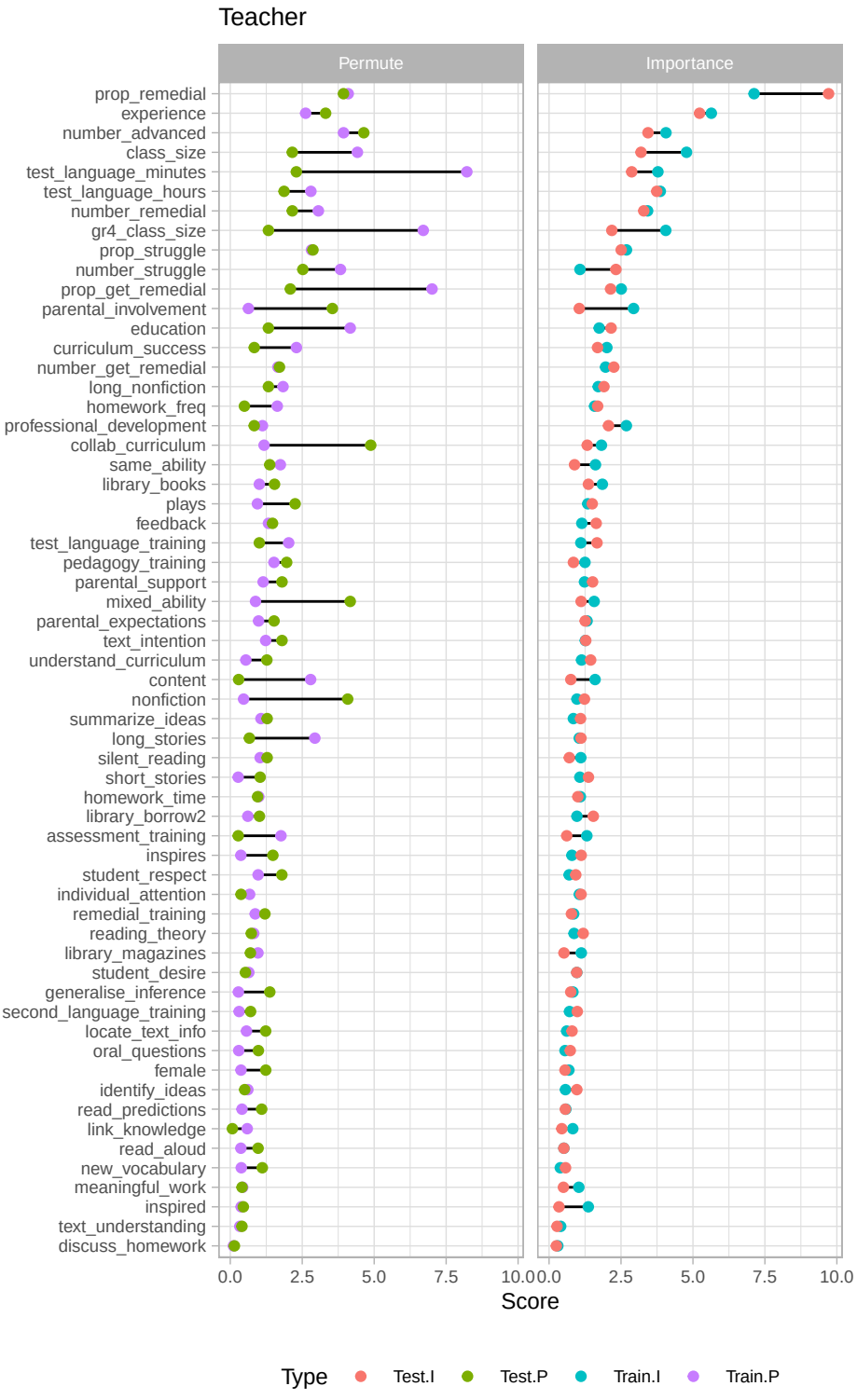


Figure 9.3: Full Feature Importance Plot - Teacher

Stellenbosch University https://scholar.sun.ac.za

9.4. Full Student Feature Results

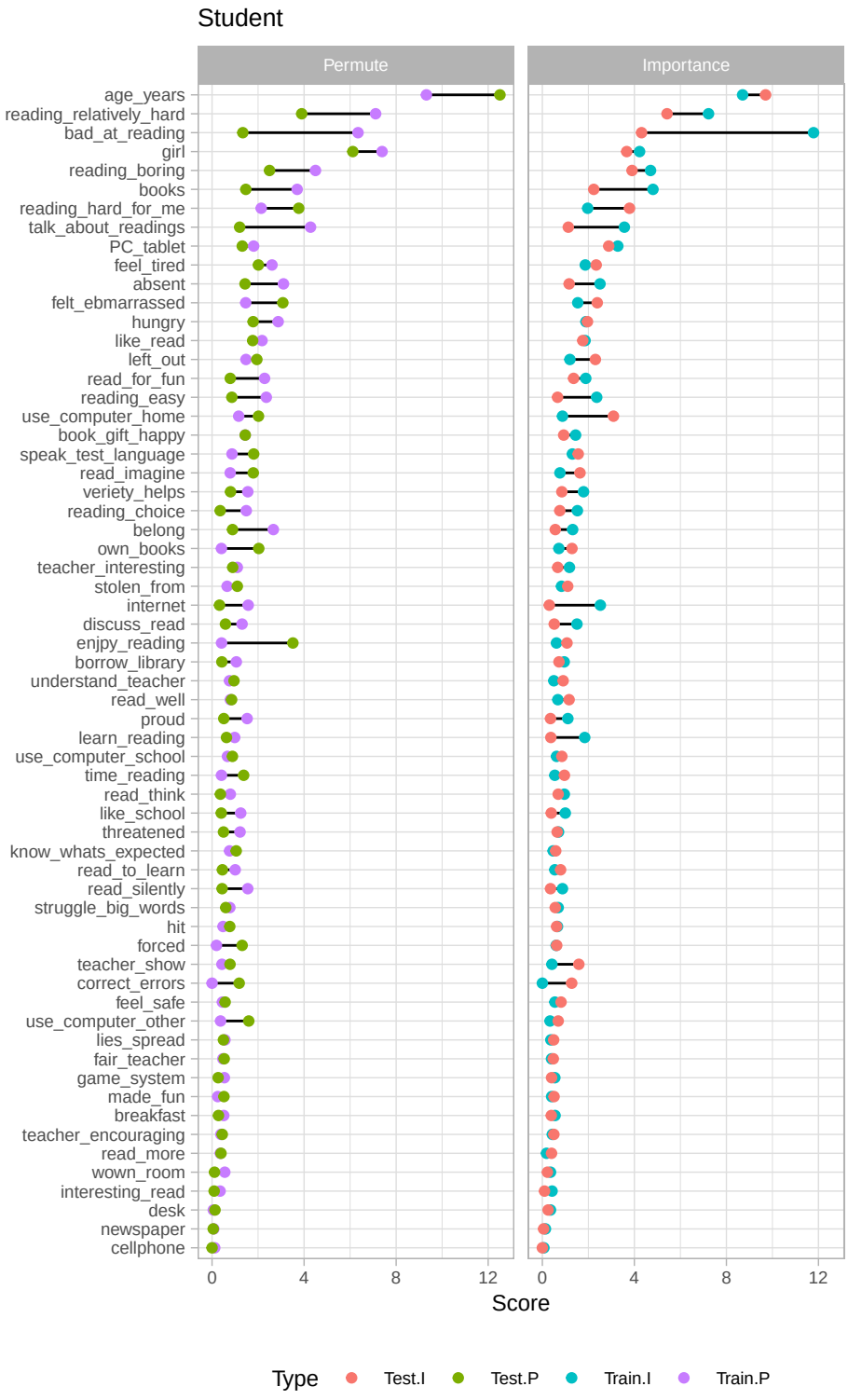


Figure 9.4: Full Feature Importance Plot - Student

Stellenbosch University https://scholar.sun.ac.za

10. Appendix C: Section 2 Full Feature Importance Results

10.1. Full Home Feature Results

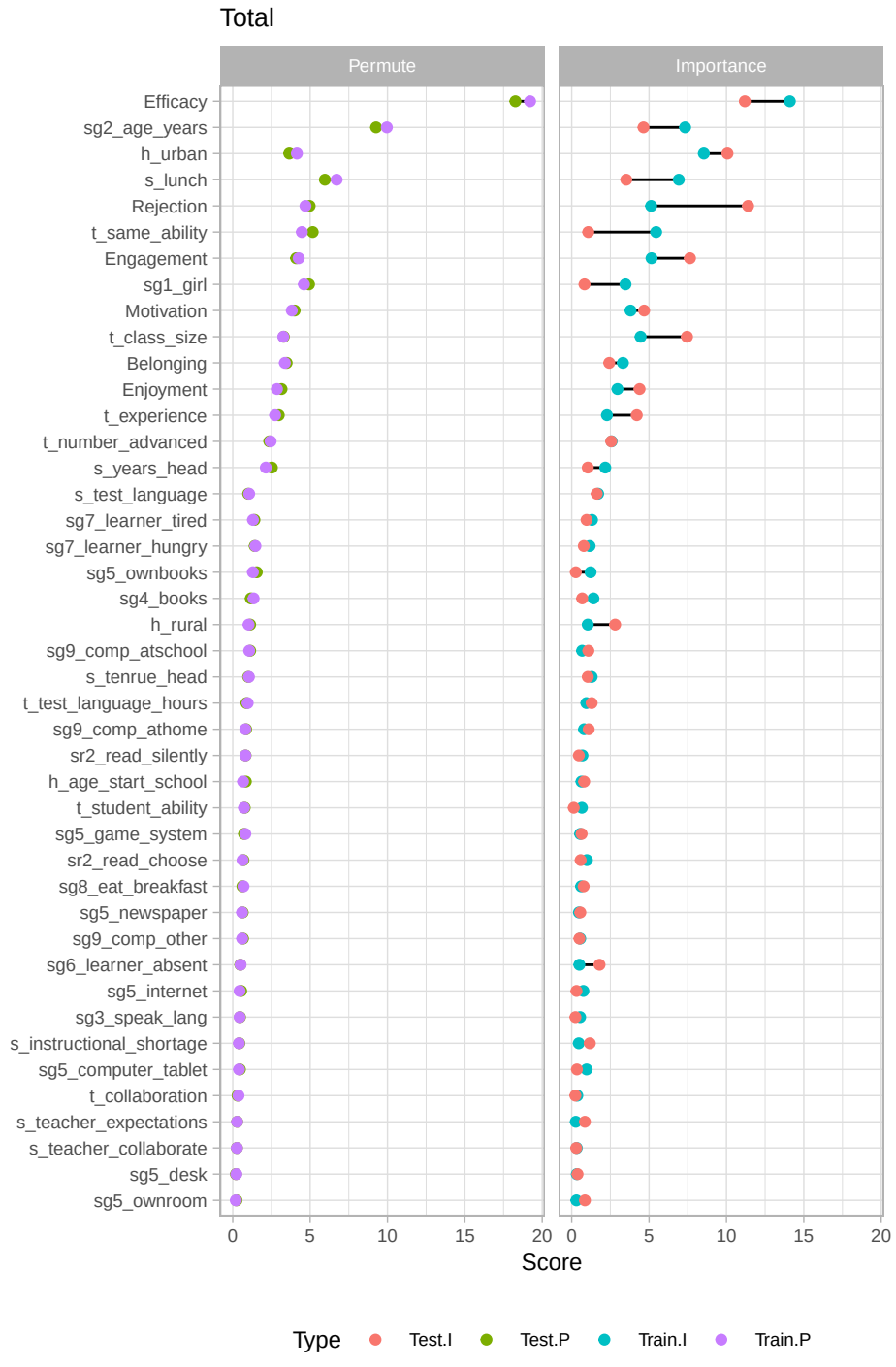


Figure 10.1: Full Feature Importance Plot - Combined Total

Stellenbosch University https://scholar.sun.ac.za

10.2. Full School Feature Results

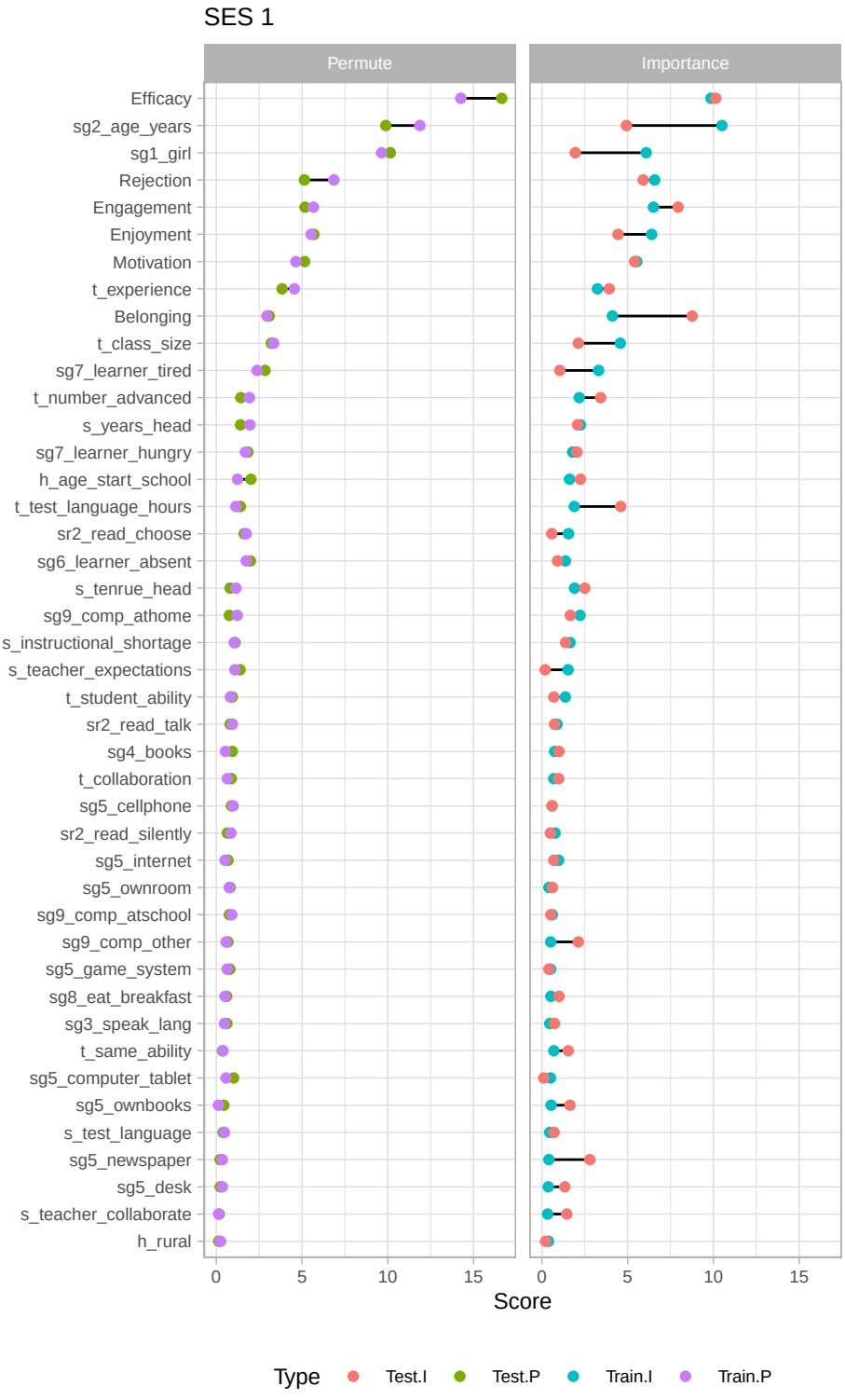


Figure 10.2: Full Feature Importance Plot - Combined SES 1

Stellenbosch University https://scholar.sun.ac.za

10.3. Full Teacher Feature Results

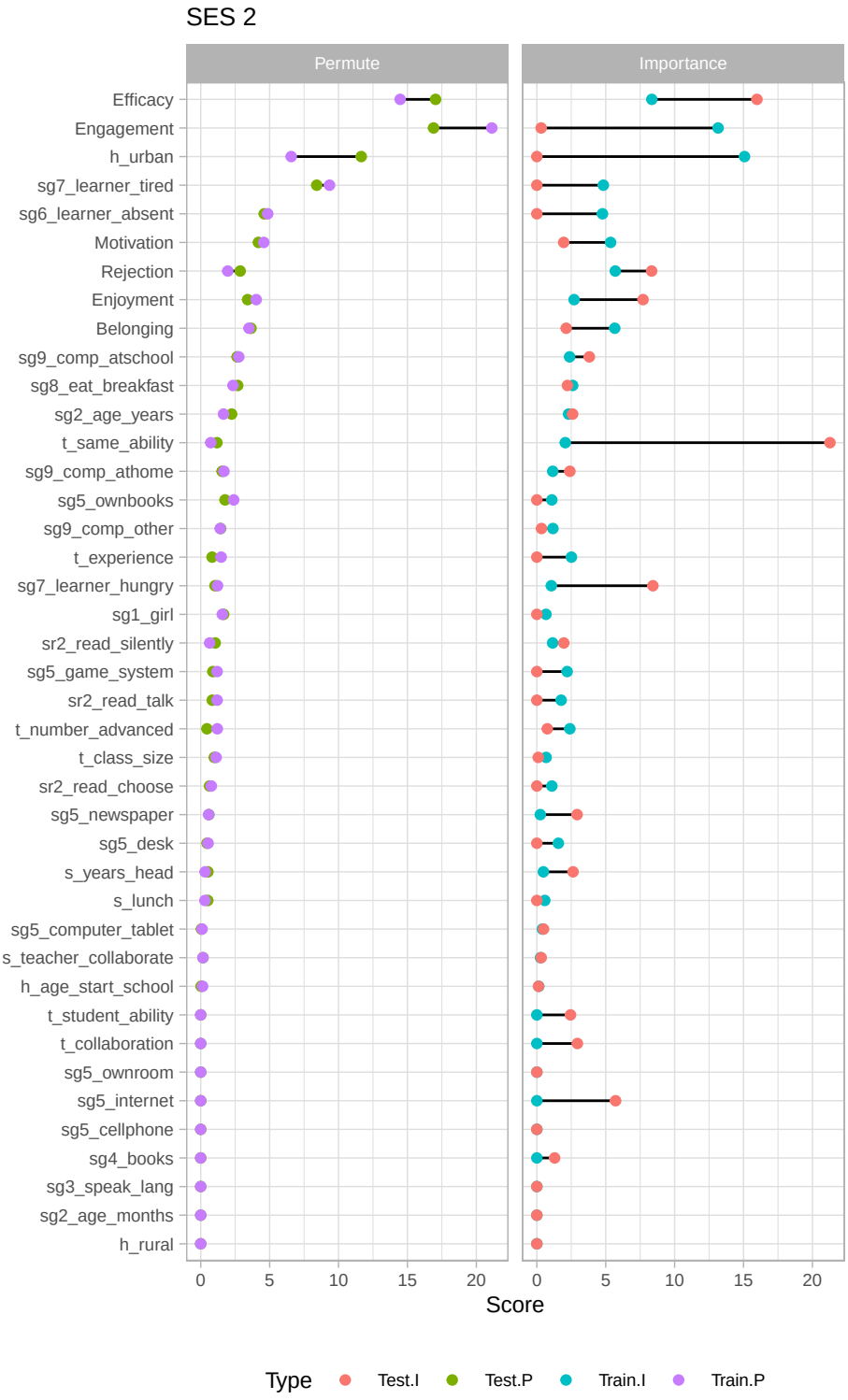


Figure 10.3: Full Feature Importance Plot - Combined SES 2

Stellenbosch University https://scholar.sun.ac.za

10.4. Full Student Feature Results

Figure 10.4: Full Feature Importance Plot - Combined SES 3

Stellenbosch University https://scholar.sun.ac.za

11. Appendix D: Methods of Interpretation

11.1. Relative Feature Importance (Importance)

Relative feature importance is a model inspection technique used for the interpretation of complex machine learning algorithms. It is a measure that captures the influence I_j of each feature $x_j, \forall j \in J$, relative to all others in the training set, on the variation of $\hat{F}(x)$ over the joint distribution. Feature importance is a broad term and can be associated with several specific techniques. The function used in this chapter to estimate relative feature importance is (Friedman, 2001)

$$I_j = (E_x \left[\frac{\partial \hat{F}(x)}{\partial x_j} \right] .var_x[x_j])^{\frac{1}{2}} \tag{6}$$

For approximations produced by decision trees, as used in our gradient boosted regressions, equation 6 does not strictly exist and must be approximated. A common approximation, and one that is used here, is proposed by Breiman et al. (1983) as

$$\hat{I}_j^2(T) = \sum_{t=1}^{J-1} \hat{i}_j^2 1(v_t = j) \tag{7}$$

The summation in equation 7 is over the non-terminal nodes t of the J-terminal node tree T. v_t is the splitting variable for node t and $\hat{i}_j^2$ is the associated empirical improvement in squared-error (reconstruction error) that results from the split. This highlights the basic theoretical notion underlying the measure of relative importance - the degree to which the addition of a new feature to a tree improves the specified loss function. If the addition of feature x_j greatly reduces reconstruction error (improves the loss function), X_j is a relatively important feature in the prediction of $\hat{F}(x)$. When dealing with a collection of trees $[T_m]_1^M$, as is the case in gradient boost, equation 7 can be generalized by taking the average over all trees in sequence as

$$\hat{I}_j^2 = \frac{1}{M} \sum_{m=1}^{M} \hat{I}_j^2(T_m) \tag{8}$$

11.2. Permutation Feature Importance (Permute)

The concept underlying permutation feature importance is similar to that of relative feature importance. The importance of a specific feature is determined by calculating the change in the fit of a models prediction error before and after permuting (corrupting) the values of that specific feature (Fisher at al., 2018). A feature is deemed important if corrupting its values changes the model's prediction error s by a large margin.

Stellenbosch University https://scholar.sun.ac.za

$$i_j = s - \frac{1}{K}\sum_{k=1}^{K} s_{kj} \tag{9}$$

The permutation importance value i_j for an individual feature j is obtained by permuting its values k times, calculating the associated k measures of prediction error s_{kj} and averaging over them. This average prediction error is then subtracted from the error of the original model s which is estimated using non-corrupted values of feature j.

11.3. Partial Dependence Plots

Partial dependence plots (PDPs) show the marginal effects of one or two features on the predicted value estimated by the fit model.[17] Unlike feature importance, a PDP is a tool for visual interpretation and provides a graphical rendering of the value of $\hat{F}(x)$ as a function of its individual arguments. Therefore, PDPs provide a summary of the dependence of $\hat{F}(x)$ on the joint values on each input feature marginalizing over the values of all other features (Friedman, 2001). The function that is plotted in a PDP is

$$\hat{f}_{x_S}(x_S) = E_{x_C}\left[\hat{f}(x_S, x_C)\right] = \int \hat{f}(x_S, x_C) d\mathbb{P}(x_C) \tag{10}$$

Where x_S is a feature vector containing features to be plotted by the PDP and X_C is a feature vector containing the complement features, those not to be plotted in the PDP. The partial dependence function is estimated by marginalizing the gradient boost algorithm over the distribution of the complement feature vector X_C. Therefore, the PDP function plots the relationship between features in X_S and $\hat{f}(x)$. By marginalizing over the complement feature vector, the PDP highlights the relationship between X_S and $\hat{f}(x)$ while factoring in the interactions between X_S and X_C.

To estimate equation 10, $\hat{f}_{x_S}(x_S)$ is calculated as

$$\hat{f}_{x_S}(x_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_S, x_C^{(i)}) \tag{11}$$

From equation 11, it is evident that the partial dependence function is estimated by averaging the marginal effect of x_S over the n observations in the training sample. It is important to note that an assumption of PDPs is that the features in the vectors X_S and x_S should not be highly correlated. In this application, such correlations do exist. Therefore, the interpretation of the PDPs here should not be exact to each point on the function, rather, the broad trend of the function is interpreted. Moreover, Loess regression curves are plotted in addition to the PDPs in order to gain a smoothed visualization of the relationship between X_S and $\hat{f}(x)$. However, the correlation between features

[17] A partial dependence plot is a visualization of the partical dependence function. The terms are used interchangibly.

Stellenbosch University https://scholar.sun.ac.za

need not be an issue in this case when one considers the desired interpretation. Ideally, having non-correlated features would enable the interpretation of *ceteris paribus* relationships. Whan features are correlated, the interpretation needs to consider possible interaction effects that are not identifiable by looking only at the PDP. That is, a partial dependence function of feature x on y may contain some indirect influence of feature z through its interaction with x.

12. Appendix E: Features Used to Estimate the Factors

12.1. PIRLS Latent Constructs

	Motivation
1	How often do you borrow books from the library
2	How much time per day do you spend reading outside of school
3	How often do you read for fun outside of school
4	How often do you read about thing that interest you outside of school

Table 12.1: PIRLS Motivation Latent Construct - Questions

	Enjoyment
1	How often do you talk to people about things that you have read
2	Would you be happy to get a book for a gift
3	Do you think that reading is boring
4	Would you like more time to read
5	Do you enjoy reading
6	Do you learn a lot from reading
7	Do you like to read think that make you think
8	Do you like it when books can help you imagine another world

Table 12.2: PIRLS Enjoyment Latent Construct - Questions

	Efficacy
1	I usually do well in reading
2	Reading is easy for me
3	I have trouble reading stories with difficult words
4	Reading is more difficult for me than it is for my classmates
5	Reading is more difficult than other subjects at school
6	I am not that good at reading

Table 12.3: PIRLS Efficacy Latent Construct - Questions

Stellenbosch University https://scholar.sun.ac.za

	Engagement
1	Do you like what they read about in school
2	Are you interested in the things that the teacher gives you to read
3	Do you know what the teacher expects you to do
4	Do you find your teacher easy to understand
5	Are you interested in what your teacher says
6	Does the teacher encourage you to read out loud
7	Does the teacher allow you to show the class what you have learned
8	How many things does the teacher do to help you read
9	Does the leather tell you how to fix your mistakes

Table 12.4: PIRLS Enjoyment Latent Construct - Questions

	Belonging
1	I like being in school
2	I feel safe when I am at school
3	I feel like I belong at this school
4	Teachers at my school are fair to me
5	I am proud to go to this school

Table 12.5: PIRLS Belonging Latent Construct - Questions

	Rejection
1	Made fun of my clothes and appearance
2	I am left out of games
3	Spread lies about me
4	Stole something from me
5	Physically hurt me
6	Made me do things I didn't want to do
7	Shared embarrassing information or photos of me
8	Threatened me

Table 12.6: PIRLS Rejection Latent Construct - Questions

Stellenbosch University https://scholar.sun.ac.za

Chapter 5: Conclusion and Implications

Keywords: Latent Variable Approach, Non-Linear Modelling, Reading Performance, PIRLS

JEL classification L250, L100

1. Introduction

The South African education system is failing the nation's poor. Rather than acting as a catalyst for social mobility, it hinders it. Most students enrolled in South African schools are receiving an education of unacceptable quality. In this way, the system is not fulfilling its mandate to educate South Africa's youth. This failing necessitates action. The work contained in this thesis forms a humble part of this action. While the barriers to successful improvement of the system can seem insurmountable, it must be attempted. In this thesis, work is done in an attempt to improve the knowledge available to both South African policymakers and education researchers.

Most work on the South African education system, especially that with an economics bent, does not stray far from the extant consensus in terms of methodology, findings, and implications. Typical findings relate to the influence of socioeconomic status, and home and school-based resources. Additional studies add nuance to what is already known, trying some slightly new approach to get some slightly different answer. This thesis has two main functions. First, and most importantly, it looks with scrutiny at the methodological aspect of contemporary research. This function is designed to induce increased consideration of data-related issues, and methodological experimentation in order to derive more accurate results. The second function is performed mostly in chapter 4, which identifies functional relationships between variables in the prediction of reading performance using a statistical learning approach. This approach departs from common linear modelling methods to incorporate non-linear techniques and theoretical latent constructs.

2. Chapter 2

This chapter analyses the degree to which the orientation of the wording (positive or negative) of a survey question can influence the variation captured by the variable derived from that question. The results contained in this chapter indicate that homogenous positive and negative features differ with regards to their ability to capture information of the same underlying latent constructs. That is, the

Stellenbosch University https://scholar.sun.ac.za

two distinct orientations of questionnaire response items do not illicit identically aligned responses. Rather, the responses to negatively worded items appear to be noisy reversed reflections of those to positively worded items. This finding is of particular significance in the field of factor modelling, where the accurate estimation of factors and capturing of underlying latent constructs are essential for appropriate inference.

The implications of this chapter are entirely methodological. The findings of the chapter can be used by both researchers and survey construction specialists. The main findings are listed as. First, survey response items with negative wording do not necessarily capture identical but reversed variation that an identical but positively worded response item would. Therefore, when employing a factor remodelling approach, or using methods of dimension reduction for index creation, consideration should be given to the combination of features derived from questions of differing orientation. Finally, increased scrutiny should be applied when interpreting the results of traditional measures of sampling adequacy such as the KMO statistic. The findings of this chapter indicate that these measures of sampling adequacy do not adequately capture incongruous variation among a group of features derived from similar survey response items.

3. Chapter 3

This chapter analyses the performance of several methods of dimension reduction with the regard to their relative ability to reduce the magnitude of erroneously measured variation in their unidimensional projections. Four methods are analysed, exploratory factor analysis, principal component analysis, kernel principal component analysis, and a neural network autoencoder. The analysis makes use of the theory of attenuation bias and OLS regression to draw inferences about the relative magnitude of erroneously measured variation captured in the estimated factors. Factors are estimated using these four methods and measured variables derived from the grade 9 South African Trends in International Mathematics and Science Study dataset. In addition to the real analysis, a simulation-based analysis is conducted.

The analysis yields several findings. First, the results of the real analysis indicate that differences in performance do exist between the analysed methods. The findings indicate that non-linear methods (KPCA and AE) perform better when there is substantial measurement error in the features used in the factor estimation procedure. Second, the results of the simulation-based analysis indicate that EFA is the best performing method in the presence of measurement error. Furthermore, the superiority of EFA increases as the magnitude of measurement error grows. Therefore, there is incongruence between the findings of the real and simulated analysis.

The main value of this chapter derives from the extent to which it reveals the potential for future work in this direction. The findings reveal that different methods of dimension reduction can have

Stellenbosch University https://scholar.sun.ac.za

dramatically different effects on subsequent estimation results. This information is useful mostly for future work done using a factor modelling framework in which efficient methods of dimension reduction are central to estimation results. In addition to the implication for factor modelling, the findings of Chapter 3 also have implications for index creation. It is typical to use PCA and use the first component as the desired index variable. The findings of Chapter 3 indicate that this approach may not be ideal, both empirically and theoretically. First, the components of PCA are simply geometric abstracts that condense total variation based on direction. PCA has no theoretical consideration for underlying drivers as does EFA. Moreover, the findings of Chapter 3 indicate that EFA performs better than does PCA when measurement error is present in the features used in the procedure. Therefore, for index creation, EFA is arguably a more useful method than PCA. Another useful finding of this chapter is that untuned machine learning algorithms are weak algorithms. This finding is relevant given the degree to which un-tuned hyper-parameters are used in education research.

The main implication of Chapter are this: first, The method of dimension reduction used in a factor modelling framework can have serious implications for final derived estimates. Second, the use of EFA in index creation rather than PCA should be explored more. EFA is, as this chapter has shown, likely a superior method both empirically and theoretically. Third, untuned machine learning algorithms should not be used in place of more simple techniques such as OLS. That is, when left untuned, the advantages of machine learning algorithms over more simple methods such as OLS disappear.

4. Chapter 4

This chapter investigates the determinants of reading performance of South African grade 4 students with specific emphasis given to measured psychological factors. The approach taken is novel in terms of both the statistical methodology and the theoretical framework of the study. The empirical analysis makes extensive use of gradient boosted regression, a statistical learning technique that enables the analysis of complex non-linear and interactive relationships. To aid the precision of interpretation, the data is disaggregated and analysed in two stages. First, the data is separated into four individual sets based on the identity of the four different survey respondents. Second, the features found to be the strongest predictors of reading performance in the first stage are combined into one dataset which is then disaggregated by socioeconomic status and again analysed using gradient boosted regression in the second stage.

Stellenbosch University https://scholar.sun.ac.za

5. Thesis Bibliography

Armstrong, P., Lekezwa, B., & Siebrits, K. 2008. Poverty in South Africa: A profile based on recent household surveys. Stellenbosch Economic Working Paper, 4(08), Stellenbosch University.

Bandura, A. 1997. Self-efficacy: The exercise of control. Macmillan. New York. United States.

Birch, S. H., & Ladd, G. W. 1997. The teacher-child relationship and children's early school adjustment. Journal of School Psychology, 35(1): 61-79.

Birch, S. H., & Ladd, G. W. 1998. Children's interpersonal behaviours and the teacher-child relationship. Developmental Psychology, 34(5): 934-946.

Booysen, F., Van Der Berg, S., Burger, R., Von Maltitz, M., & Du Rand, G. 2008. Using an asset index to assess trends in poverty in seven Sub-Saharan African countries. World Development, 36(6): 1113-1130.

Breiman, L. 2001. Random Forests. Machine Learning, 45(1): 5-32.

Breiman, L., Friedman, J. H., Olshen, R., & Stone, C. 1983. Classification and Regressions Trees. CRC press.

Brown, T.A. 2015. Confirmatory factor analysis for applied research. Guilford publications.

Caspi, A., Elder, G.H., & Bem, D.J. 1987. Moving against the world: Life-course patterns of explosive children. Developmental Psychology, 23(2): 308-313.

Cerny, C.A., & Kaiser, H.F. 1977. A study of a measure of sampling adequacy for factor-analytic correlation matrices. Multivariate Behavioral Research, 12(1): 43-47.

Chiavaroli, N. 2017. Negatively-worded multiple choice questions: An avoidable threat to validity. Practical Assessment, Research, and Evaluation, 22(1): 3 - 18.

Cronbach, L.J. 1951. Coefficient alpha and the internal strucuture of tests. Psychometrika, 16(3): 297-334.

Deci, E.L., & Ryan, R.M. 1980. The empirical exploration of intrinsic motivational processes. In Advances in experimental social psychology, 13(2): 39-80.

Deci, E.L., & Ryan, R.M. 1985. The general causality orientations scale: Self-determination in personality. Journal of research in personality, 19(2): 109-134.

Stellenbosch University https://scholar.sun.ac.za

Deci, E.L., Koestner, R., & Ryan, R.M. 2001. Extrinsic rewards and intrinsic motivation in education: Reconsidered once again. Review of educational research, 71(1): 1-27.

Deci, E.L., Vallerand, R.J., Pelletier, L.G., & Ryan, R.M. 1991. Motivation and education: The self-determination perspective. Educational psychologist, 26(3-4): 325-346.

Fisher, A., Rudin, C., & Dominici, F. 2018. Model class reliance: Variable importance measures for any machine learning model class, from the "Rashomon" perspective. arXiv preprint arXiv:1801.01489, 68.

Friedman, J., Hastie, T., & Tibshirani, R. 2001. The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.

Friedman, J.H. 1997. On bias, variance, 0/1—loss, and the curse-of-dimensionality. Data mining and knowledge discovery, 1(1): 55-77.

Friedman, J.H. 2001. Greedy function approximation: a gradient machine. Annals of statistics, 29(5): 1189-1232.

Greene, W.H. 2000. Econometric analysis 4th edition. International edition, New Jersey: Prentice Hall.

Guttman, L. 1945. A basis for analyzing test-retest reliability. Psychometrika, 10(4): 255-282.

Hartley, J. 2014. Some thoughts on Likert-type scales. International journal of clinical and health psychology, 14(1): 83-86.

Hunt, E. 1978. Mechanics of verbal ability. Psychological Review, 85(2): 109 – 130.

Hyslop, D.R., & Imbens, G.W. 2001. Bias from classical and other forms of measurement error. Journal of Business & Economic Statistics, 19(4): 475-481.

Jöreskog, K.G. 1967. Some contributions to maximum likelihood factor analysis. Psychometrika, 32(4): 443-482.

Kaiser, H. 1974. An index of factor simplicity. Psychometrika, 39: 31–36.

Kaiser, H.F., & Rice, J. 1974. Little jiffy, mark IV. Educational and psychological measurement, 34(1): 111-117.

Ladd, G.W., Birch, S.H., & Buhs, E.S. 1999. Children's social and scholastic lives in kindergarten: Related spheres of influence? Child development, 70(6): 1373-1400.

Stellenbosch University https://scholar.sun.ac.za

Ladd, G.W., Kochenderfer, B.J., & Coleman, C.C. 1997. Classroom peer acceptance, friendship, and victimization: Distinct relation systems that contribute uniquely to children's school adjustment?. Child development, 68(6): 1181-1197

MacCallum, R.C. 2009. Factor Analysis, in Millsap & Maydeu-Oliveras. Quantitative Methods in Psychology. 123 - 147.

Marsh, H.W. 1986. Negative item bias in ratings scales for preadolescent children: A cognitive-developmental phenomenon. Developmental Psychology, 22(1): 37 - 49.

McDonald, M.E. 2013. The Nurse Educator's Guide to Assessing Learning Outcomes. (3rd ed.). Burlington, MA: Jones and Bartlett.

McInnes, L., Healy, J. & Melville, J. 2018. Umap: Uniform manifold approximation and projection for dimension reduction.

Mincer, J. 1974. Schooling, Experience, and Earnings. Human Behavior & Social Institutions No. 2.

Ridgeway, G. 2020. Generalized boosted models: A guide to the gbm package [Online]. Available: http://cran.open-source-solution.org/web/packages/gbm/vignettes/gbm.pdf. [2020, September 03]

Schölkopf, B., Smola, A., & Müller, K.R. 1998. Nonlinear component analysis as a kernel eigenvalue problem. Neural computation, 10(5): 1299 - 1319.

Schunk, D.H. 2003. Self-efficacy for reading and writing: Influence of modelling, goal setting, and self evaluation. Reading & Writing Quarterly, 19(2): 159-172.

Spaull, N., & Hoadley, U. 2017. Getting Reading Right In: Jamieson L, Berry L & Lake L (eds) South African Child Gauge 2017. Cape Town, Children's Institute, University of Cape Town.

Spaull, N. 2013. Poverty & privilege: Primary school inequality in South Africa. International Journal of Educational Development, 33(5): 436-447.

Suhr, D.D. 2006. Exploratory or confirmatory factor analysis?.

Sullivan, A. 2009. Academic self-concept, gender and single-sex schooling. British educational research journal, 35(2): 259-288.

Taylor, A., & Machida, S. 1996. Student-teacher relationships of Head Start children. Paper presented at the annual meeting of the American Educational Research Association, New York.

Taylor, N., Muller, J., & Vinjevold, P. 2003. Getting schools working: Research and systemic school reform in South Africa. Pearson South Africa.

Stellenbosch University https://scholar.sun.ac.za

Thaler, R.H., & Sunstein, C.R. 2009. Nudge: Improving decisions about health, wealth, and happiness. Penguin.

Thurstone, L.L. 1947. Multiple factor analysis, Chicago, IL: University of Chicago Press.

Vincent, P. 2011. A Connection Between Score Matching and Denoising Autoencoders. Neural Computation, 23(7): 1661–74

Weijters, B., Baumgartner, H., & Schillewaert, N. 2013. Reversed item bias: An integrative model. Psychological methods, 18(3): 320 – 334.

Wong, N., Rindfleisch, A., & Burroughs, J.E. 2003. Do reverse-worded items confound measures in cross-cultural consumer research? The case of the material values scale. Journal of consumer research, 30(1): 72-91.

Wooldridge, J.M. 2002. Econometric analysis of cross section and panel data MIT press. Cambridge, MA, 108.

Stellenbosch University https://scholar.sun.ac.za

13. Appendix F: Variable list

13.1. Home Dataset Variable List

Variable	Survey Question
not_SA_born	learner was not born in South Africa yes/ no
attended_ECD	child attended early childhood education (children under 3 years) yes/no
pre_school	child attended pre-school
preschool_years	number of years child spent in pre-school
age_start_school	age of child when they started first grade of primary school
knowledge_of_ability	knowledge (perception) of child's abilities in literacy tasks at entering school
homework_freq	how often does the learner receive homework
homework_ask	how often does someone ask if the learner has done their homework
homework_help	how often does someone help the learner with their homework
homework_correct	how often does someone review and correct the learner's homework
read_work	how much time does the parent spend reading for themselves at work in a typical day
books	how many books does the parent report having at home
children_books	how many childrens books does the parent report having at home
e_reader	here is an electronic device at home for reading yes/ no
child_e_reader	there is an electronic device at home for the child to read
digital_devices	how many digital devices are available in the home
parent_educ	highest level of parental education within the home
expectations	how far do the parents expect their child to go in education
rural	live in a rural area
urban	live in a urban area
ECD_lit_act	index of literacy activities in early childhood
ECD_num_act	index of numeracy activities in early childhood
ECD_learn_act	index of learning (literacy and numeracy) activities in early childhood
school_includes	parent believes that the child's school does a good job of including them in education
school_safe	parent believes that the child's school provides a safe environment
school_cares	parent believes that the child's school cares about child's progress
school_informs	parent believes that the child's school informs them of child's progress
high_standard	parent believes that the child's school promotes high academic standards
school_helps	parent believes that the child's school does good job helping child become better
read_home	how much time does the parent spend reading for themselves at home in a typical
enjoy_reading	index of how much parents like reading
language_of_test	how often does learner speak the test language at home (according to parent)

Stellenbosch University https://scholar.sun.ac.za

13.2. Student Dataset Variable List

Variable	Survey Question
girl	learner is a girl
age_years	age in years
speak_test_language	how often does learner speak the test language at home
books	how many books does the learner have at home
PC_tablet	yes/ no
desk	yes/ no
wown_room	yes/ no
internet	yes/ no
cellphone	yes/ no
game_system	yes/ no
own_books	yes/ no
newspaper	yes/ no
absent	how often was learner absent from school
feel_tired	how often did the learner feel tired when arriving at school
hungry	how often did the learner feel hungry when arriving at school
breakfast	how often did the learner eat breakfast on a school days
use_computer_home	how often does the learner use a computer or tablet at home
use_computer_school	how often does the learner use a computer or tablet at school
use_computer_other	how often does the learner use a computer or tablet at another place
like_school	learner likes being in school
feel_safe	learner feels safe at school
belong	learner feels like they belong at the school
fair_teacher	learner things teachers are fair at the school
proud	learner is proud to be at the school
made_fun	learner is made fun of
left_out	learner is left out of games
lies_spread	learner had lies spread about them
stolen_from	learner had things stolen from them
hit	learner was hit or hurt at school
forced	learner was forced to something they didn't like
felt_ebmarrassed	learner felt embarrassed
threatened	learner was threatened at school
read_silently	how often are learners allowed to read silently on their own
reading_choice	how often are learners allowed to read things they choose themselves
discuss_read	how often do teachers ask learners to talk about what they have read
borrow_library	how often does the learner borrow books from a library
time_reading	how much time per day does learner spend reading outside of school
read_for_fun	how often does the learner read outside of school for fun

Stellenbosch University https://scholar.sun.ac.za

read_to_learn	how often does the learner read about things they want to learn outside of school
like_read	learner likes what they read about in school
interesting_read	learner thinks the teacher gives them interesting things to read
know_whats_expected	learner knows what teacher expects them to do
understand_teacher	learner finds teacher easy to understand
teacher_interesting	learner finds what teacher says interesting
teacher_encouraging	learner believes teacher encourages them to think about what they read
teacher_show	teacher lets learner show what they have learned
veriety_helps	teacher does a variety of things to help learning
correct_errors	teacher tells learners how to do better when they make a mistake
talk_about_readings	learner likes to talk about they read with other people
book_gift_happy	learner would be happy recieving a book as a present
reading_boring	learner thinks that reading is boring
read_more	learner would like to have more time to read
enjpy_reading	learner enjoys reading
learn_reading	learner learns a lot from reading
read_think	learner likes to read things that make them think
read_imagine	learner likes it when a book helps them imagine other worlds
read_well	learner believes they usually do well in reading
reading_easy	learner believes that reading is easy for them
struggle_big_words	learner has trouble reading stories with difficult words
reading_hard_for_me	learner believes that reading is harder for them than most of their classmates
reading_relatively_hard	learner thinks that reading is harder than other subjects
bad_at_reading	learner think that they are just not good at reading

Stellenbosch University https://scholar.sun.ac.za

13.3. Teacher Dataset Variable List

Variable	Survey Question
experience	years of teaching experience
female	teacher is female yes/ no
education	teacher's level of education
reading_theory	reading theory formed part of formal trainingt
pedagogy_training	pedagogy of reading formed part of formal trainingt
remedial_training	remedial reading formed part of formal trainingt
second_language_training	second language learning formed part of formal trainingt
assessment_training	assessment methods in reading formed part of formal trainingt
test_language_training	language of test formed part of formal trainingt
professional_development	hours spent in professional development dealing with teaching reading over the
parental_involvement	teacher's perception of parental involvement in school activities
parental_expectations	teacher's perception of parental expectations for student achievement
parental_support	teacher's perception of parental support for student achievement
student_desire	teacher's perception of students' desire to do well in school
student_respect	teacher's perception of students' respect for classmates who excel academically
understand_curriculum	teacher's perception of teachers' understanding of curricular goals
curriculum_success	teacher's perception of teachers' success in implementing curriculum
inspires	teacher's perception of teachers' ability to inspire students
collab_curriculum	how often does the teacher work with teachers from other schools on the curricul
class_size	number of students in the class
gr4_class_size	number of students in the class that are in grade 4
number_struggle	the number of grade 4 learners in the class that have difficulty understanding
prop_struggle	the proportion of grade 4 learners in the class that have difficulty understanding
number_remedial	the number of grade 4 learners that need remedial instruction in reading
prop_remedial	the proportio of grade 4 learners that need remedial instruction in reading
number_get_remedial	the number of grade 4 learners that receive remedial instruction in reading
prop_get_remedial	the proption of grade 4 learners that receive remedial instruction in reading
number_advanced	the number of grade 4 learners that are advanced instruction readers
test_language_hours	hours spent in instruction in test language
test_language_minutes	minutes spent in reading instruction
same_ability	how often does teacher instruct class in same
mixed_ability	how often does teacher instruct class in mixed
individual_attention	how often does teacher instruct class individually for reading

Stellenbosch University https://scholar.sun.ac.za

library_books	how many books are in the classroom library
library_magazines	how many magazines are in the classroom library
homework_freq	how often does the teacher assign reading as part of homework
homework_time	how much time does the teacher expect learners to spend on reading homework
content	how often does the teacher feel content with their profession as a teacher
meaningful_work	how often does the teacher find their work full of meaning and purpose
inspired	how often does the teacher feel that their work inspires them
short_stories	how often does the teacher use short stories for reading instruction
long_stories	how often does the teacher use long stories for reading instruction
plays	how often does the teacher use plays for reading instruction
nonfiction	how often does the teacher use non
long_nonfiction	how often does the teacher use long non
read_aloud	how often does the teacher read aloud to students during reading instruction
silent_reading	how often does the teacher ask students to read silently during reading instruct
new_vocabulary	how often does the teacher teach new vocabulary systematically during reading in
summarize_ideas	how often does the teacher teach students how to summarize main ideas during reading
link_knowledge	how often does the teacher link new content to prior knowledge
text_understanding	how often does the teacher encourage development understanding of the text
feedback	how often does the teacher give individual feedback to students
locate_text_info	how often does the teacher ask students to locate information within the text
identify_ideas	how often does the teacher ask students to identify main ideas of what they have
read_predictions	how often does the teacher ask students to make predictions about what will happen
generalise_inference	how often does the teacher ask students to generalise and draw inference
text_intention	how often does the teacher ask students to determine perspective/intention of author
oral_questions	how often does the teacher ask students to answer oral questions about what they read
library_borrow2	How often are students sent to other library to borrow books?
discuss_homework	how often does the teacher discuss the homework in class

Stellenbosch University https://scholar.sun.ac.za

13.4. School Dataset Variable List

Variable	Survey Question
prop_disadvantaged	proportion of student population from economically disadvantaged background
prop_advantaged	proportion of student population from economically advantaged background
breakfast	does school provide free breakfast to students
lunch	does school provide free lunch to students
school_days	number of calendar days in the school year
school_day_hours	number of instructional hours in a school day
work_space	Is there a space where students can do schoolwork before or after school yes/no
assist_students	Is there sometime to assist students with schoolwork before or after school Yes/no
library	school has a library yes/no
library_books	How many library books are in the library?
digital_books	students have access to digital books yes/no
computers	total number of computers for grade 4 use
instructional_space_shortage	shortage of instructional materials
supplies_shortage	shortage of supplies
building_shortage	shortage of buildings and grounds
heating_shortage	shortage of heating and lighting
material_shortage	shortage of instructional space
tech_competence_shortage	shortage of technologically competent staff
audiovisual_shortage	shortage of audio and visual resources
PC_tech_shortage	shortage of computer technology for instruction
disabled_resource_shortage	shortage of resources for students with disabilities
teacher_lateness	To what degree is teacher lateness a problem?
teacher_absent	To what degree is teacher absenteeism a problem
teacher_curr	To what degree is teacher failure to complete curriculum a problem?
recognise_alphabet	proportion of students that recognised letters of alphabet when beginning school
read_words	proportion of students that could read some words when beginning school
read_sentences	proportion of students that could read sentences when beginning school
read_story	proportion of students that could read stories when beginning school
write_letters	proportion of students that could write letters when beginning school
write_words	proportion of students that could write words when beginning school
years_head	number of years in role of school head
tenure_head	number of years in role of school head at this school
head_educ	highest level of formal education of school head

Stellenbosch University https://scholar.sun.ac.za

head_qualification	school head has qualification in school leadership yes/no
test_language	proportion of student population that speak the language of the test
teachers_understand	How does school head characterise teachers' understanding of curricular goals
teachers_implement	How does school head characterise teachers' success in implementing curriculum
teacher_expectations	How does school head characterise teachers' expecations for student success
teacher_collaborate	How does school head characterise teachers' ability to work together to improve
teacher_inspiration	How does school head characterise teachers' ability to inspire students
parent_involvement	How does school head characterise parent involvement in school activitie
parent_commitment	How does school head characterise parent commitment to learner school readiness
parent_expectations	How does school head characterise parent expecations for student achievement
parent_support	How does school head characterise parent support for student achievement
student_desire	How does school head characterise students' desire to do well
student_respect	How does school head characterise students' respect for classmates who excel
student_ability	How does school head characterise students' ability to reach academic goals