Evaluating and extending statistical methods for estimating the construct-level predictive validity of selection tests

Lazaro Mwakesi Mwandigha, MSc.

PhD

University of York

Health Sciences

December, 2017

Abstract

Background:

In the thesis two medical selection challenges were addressed using the United Kingdom Clinical Aptitude Test (UKCAT) and Professional and Linguistic Assessments Board (PLAB) test in the selection of undergraduate medical school entrants and International Medical Graduates (IMGs) in the UK as motivating examples. Firstly, methods for correcting for bias in the estimate of predictive validity due to range restriction (particularly Multiple Imputation (MI) and Full Information Maximum Likelihood (FIML)) were evaluated for the *predictive validity, single hurdle concurrent* and *multiple hurdle* validity designs under varying degrees of strictness in selection. For MI, the impact of the composition of the imputation model was also investigated. Secondly, the Number Needed to Reject (NNR), a concept analogous to Number Needed to Treat was evaluated with its uncertainty tackled as a missing data and resampling problem.

Methods:

The performance of MI and FIML was tested through Monte Carlo simulations and validated using PLAB data. The uncertainty about NNR was estimated by use of MI and case resampling bootstrap using UKCAT data.

Results:

Generally, MI and FIML were found to be equivalent in performance and superior to other methods of correcting for range restriction bias for selection ratios of $\leq 20\%$ only in instances where data were multivariate normal. The inclusion of highly predictive variables in the imputation model increased the precision of MI. The percentile bootstrap confidence intervals contained reliable estimates for NNR.

Conclusion

MI and FIML are viable alternatives for tackling bias in the estimate of predictive validity for direct range restricted data that satisfies the assumption of multivariate normality. Caution should be taken to avoid their application in instances where the assumption of multivariate normality is violated. A combination of imputations and *case resampling bootstrap* is recommended for estimating uncertainty about NNR when data are incomplete and (un)clustered.

Contents

List of Figures

List of Tables

Acknowledgements

"Iron sharpens iron, So one man sharpens another" Proverbs 27:17.

I owe a debt of gratitude to many who have sharpened my personality and wit during my PhD journey, without whom this labour of work would not be possible. I am grateful to the UKCAT Board, and specifically, *Rachel Greatrix*, for her help administering my stipend. I am grateful for the support of *Hull York Medical School*, and specifically *Professor Trevor Sheldon*, who provided the funds for two years of my tuition fees at the *University of York*, and to *Oliver Short* who helped me with administering the finances. I owe a vote of thanks to the *Health Sciences* and *International Student Support Team* who made my student life memorable. I am grateful to *Drs Paul Tiffin* and *Adetayo Kasim* who welcomed me to the UK and have provided supervision throughout my studies at both Durham and York. They, together with *Professor Catherine Hewitt* and *Dr Jan Böhnke* have offered me invaluable academic mentorship, pastoral support and constructive criticism in my second and third year of my PhD at the *University of York*. I have greatly cherished their readiness and availability to offer great feedback. I thank my father, mother and siblings who have been supportive of me in my studies. Special thanks goes out to my lovely wife, *Maria Cristina Mwandigha*, a real gem and dependable ally in this life. Lastly, certainly not least, I am grateful for the divine providence that has seen me journey from the rough neighbourhood of Nairobi, Kenya to study and work alongside the brightest minds in statistical methodology and medical education.

Author’s declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, university. All sources are acknowledged as references. From the thesis, chapter 5 and 6 are intended to be published as statistical methodology papers. The working titles for the papers are listed under subsection entitled “working scientific publication (from PhD thesis)”. During my research for this thesis, I collaborated with other researchers within the area of medical education selection. My contribution was the provision of statistical support in terms of data management, statistical modelling, interpretation and critical appraisal of the written reports. This led to submission of papers numbered 2 to 5 listed under subsection entitled “related scientific publication”. The first paper on the list is my first author publication which is based on a multilevel mediation analysis conducted on the same data used in chapter 6 of this thesis.

Working scientific publication (from PhD thesis)

1. *Dealing with attenuation in the Pearson correlation coefficient due to range restriction as a missing data problem: A Monte Carlo simulation and validation with the aid of a contrived example using PLAB data* (based on chapter 5 of this thesis)
2. *Estimating uncertainty in the estimate of Number Needed to Reject (NNR) and proof of concept for “Peer Competition Rescaling (PCR)”* (based on chapter 6 of this thesis)

Related scientific publications

1. Mwandigha LM, Tiffin PA, Paton LW, Kasim AS, Böhnke JR. *What is the effect of secondary (high) schooling on subsequent medical school performance? A national, UK-based, cohort study*. BMJ Open 2018;8:e020291.

2. Finn G, Mwandigha LM, Paton LW, Tiffin PA. *The ability of ’non-cognitive’ traits to predict undergraduate performance in medical schools: a national linkage study*. BMC Medical Education 2018 18:93.

3. Tiffin PA, Mwandigha LM, Paton LW, Hesselgreaves H, McLachlan JC, Finn GM, Kasim AS. *Predictive validity of the UKCAT for medical school undergraduate performance: a national prospective cohort study*. BMC medicine. 2016 Sep 17;14(1):140.

4. Tiffin PA, Paton LW, Mwandigha LM, McLachlan JC, Illing J. *Predicting fitness to practise events in international medical graduates who registered as UK doctors via theProfessional and Linguistic Assessments Board (PLAB) system: a national cohort study*. BMC medicine. 2017 Mar 20;15(1):66.

5. Paton LW, Tiffin PA, Smith D, Dowell JS, Mwandigha LM. *Predictors of Fitness to Practise Declarations in UK Medical Undergraduates at Provisional Registration*. BMC Medical Education 2018 18:68.

Part I.

Preface

1. Background

Worldwide competition for medical school places is generally fierce. In the United Kingdom (UK), it is approximated that there are 11 applicants per place available to study medicine (Medical School Council, 2014). Most applicants are academically (and often non-academically) highly achieving and thus it is often difficult to discriminate between a relatively homogeneous group of highly achieving individuals in a high stakes situation. This has led to the implementation of more structured and formal selection processes in the medical professions (including dentistry and veterinary science). Within the selection process there are competing agendas- a need to defend high stakes decisions (and notably the rejection of very able candidates at times), a desire to identify those most suitable for a medical career and also to widen access to traditionally under-represented groups. This has led to the development and implementation of *aptitude tests*. The first of these, was the *Scholastic Aptitude Test for Medical Schools* developed by physician and psychologist F.A Moss (hence sometimes referred to as the Moss test), was implemented almost a century ago for none of these listed reasons- ironically the aim of the Moss test was to address the high drop-out rates from USA medical schools (McGaghie, 2002) (see section 2.1.2 for further details). Nevertheless such aptitude tests, intended to evaluate the characteristics deemed to predict a successful career in medicine, are now widely implemented.

Ideally selection tests should be *reliable* (i.e. scores should be internally consistent and reproducible). Item Response Theory (IRT) has extended this concept of reliability to *test information* - the accuracy to which a psychometric instrument is able to discriminate between testees across differing trait or ability levels (usually denoted θ). Test information is thus an *epistemo-*

logical issue (the way we know things)- the extent we can gain knowledge of a construct via a measurement process. Tests should also be *valid*, in that they should be measuring the intended construct. This is a usually a more complex and challenging concept than reliability and raises numerous questions. Is the construct being measured uni or multidimensional? Can the construct be operationalised and defined easily (e.g. *professionalism* versus *verbal reasoning*? Is the construct being directly (and accurately) measured by the instrument or is it a proxy? Thus the concept of validity raises crucial *ontological* issues (what things are). Moreover in selection science we have additional psychometric and statistical challenges. These are outlined in greater detail later but consist of issues such as *range restriction*- i.e. the inability to observe an outcome in those candidates who are not selected. This gives rise to an attenuation in the degree of correlation (association referred to as predictive validity) between the selection test score (also known as selection measure or predictor) and the outcome (also known as outcome measure or criterion variable) of interest. This can easily lead to the erroneous conclusion that the selection test is invalid. Indeed it was this effect that led Sir George Smart, at the General Medical Council (GMC) conference on Methods of Examination and Assessment in 1973 to state:

"As predictors of future performance examinations were not highly successful, as was shown by the low correlation of A level GCE grades with subsequent performance in medical school" (General Medical Council, 1973; McManus, Dewberry, Nicholson, Dowell, et al., 2013)

This thesis is thus primarily statistical in focus, evaluating a number of methods for correcting for these, and other attenuating effects, in selection tests. The case of the Professional and Linguistic Assessments Board (PLAB) test, the main route by which International Medical Graduates (IMGs) demonstrate they have the required skills and knowledge to work in the UK, is used as a motivating example. Several competing statistical approaches are evaluated and methods are also developed and extended to support selectors in understanding the effectiveness (or otherwise) of their tests. In this sense the aim is to understand construct-level predictive validity- an important term introduced by *McManus, Dewberry, Nicholson, Dowell, et al. (2013)* to indicate an ideal situation whereby the *true* (i.e. free of bias and measurement error) underlying relationship between a selection test score and a target criterion could be estimated. This construct-level predictive validity may also be viewed as estimate arrived at by correcting the predictive validity estimate for the effects of bias and measurement error.

Further, this thesis explores other statistical challenges related to selection, one being the potential capability of using aptitude tests as screening tests to detect potentially poor candidates during the medical school selection process. The term "poor candidates" in this specific context and motivating example in the thesis refers to *those applicants who present a high risk of failing at least one exam at first sitting during undergraduate medical school training*. The concept of *Number Needed to Reject (NNR)* (analogous to *Number Needed to Treat* within biomedical health research), introduced by *Tiffin, Paton, et al. (2017)* for the selection context, estimates the number of good candidates that would be rejected in order to get rid of one poor candidate during the selection process. The term "good candidates" in this specific context and motivating example in the thesis refers to *those applicants who are at a very low risk of a specified adverse outcome (i.e. failing at least one year at medical school).* If the aptitude test is a good screening test, then it is expected to be characterised by a low NNR. Notably, from a statistical standpoint, the introduction of the concept of NNR was not accompanied by precision estimates (standard errors and confidence intervals). To address this limitation, several methodological approaches are developed in the thesis with recommendations for when they are most appropriate.

The other statistical challenge explored in the thesis has to do with the problem of statistical "nationalisation" of local medical school outcomes. Oftentimes, at the point of entry, applicants to medical school are selected on the basis of their scores on a national predictor like an aptitude test. However, after graduation, selection for national opportunities, is done (to a large extent) based on local medical school outcomes. For example, in the UK, the Educational Performance Measure (EPM) as a medical school measure based on a ranking of a student within their medical school (Medical School Council, 2017a). However, this measure is partly used to select candidates for the national *Foundation Programme* (Foundation Programme, 2017). Likewise, attempts to establish the validity of predictors for medical selection, the outcomes used are often local measures such as medical school performance in *knowledge* and *skills*-based exams.

The use of local measures of performance as proxies for national ones may lead to bias and incorrect inference. To correct for this, an exploration of potential methodology for adjusting local measures in order to "nationalise" them is worthwhile. In the thesis, this is referred to as *"Peer Competition Rescaling (PCR)"*, a term coined by *Tiffin and Paton (2017).* The statistical potential for the usefulness of PCR and future research implications are detailed in the thesis. For both the NNR and PCR, the United Kingdom Clinical Aptitude Test (UKCAT), the most widely used aptitude test for medical selection in the UK, is used as a motivating example.

Although the project is primarily statistical in focus it is important to contextualise the work. Thus, in this background section there is a brief overview of the main current selection approaches used in medical education, with a more expanded section on the use of aptitude testing.

1.1. Medical selection in the UK context

Historically, selection of medical school students in the UK (as elsewhere) has been primarily based on educational achievement. Specifically, in the UK, students generally sit their General Certificate of Secondary Education (GCSE) at around the age of 16 years. Many go on to study for their General Certificate of Education Advanced Level (A-level) exams which they take at around the age of 18 (Wright and Bradley, 2010). Thus, for most school leavers from England and Wales (Scotland has a different examination system based around Scottish Highers) applying for medical school, they will not have obtained their A-level results at the time of application. Thus, in most cases, provisional offers are made on the basis of their GCSE results and predicted A-level grades. Applicants must then achieve the required A-level grades, specified by the medical school at the time of provisional offer, in order to matriculate. Overall, selection based on prior educational achievment has been a relatively effective approach as research has consistently observed that this factor predicts undergraduate performance, scores on any subsequent postgraduate or licensing examinations and career progression (Benbassat and Baumal, 2007). In recent times, however, student selection based on A-level grades has been complicated by the fact that an increasing number of students obtain top grades at A-level which makes it difficult for medical schools to discriminate applicants due to their relatively homogeneous score distribution. It has also been found that selection based on A-level performance favours applicants from selective schools. Moreover, many school leavers apply to medical school with other education qualifications including International Baccalaureate and Scottish Highers. This is further complicated by the fact that medical schools also admit entrants who may not be hailing directly from school (Ferguson, James, and Madeley, 2002; James, Yates, and Nicholson, 2010; McManus, Powis, et al., 2005).

Consequently, efforts have been directed towards adjusting for A-level bias in order to compensate for the fact that A-level performance may reflect resource deprivation, demographic, social and quality of secondary school differences rather than an applicant's true ability. Unfortunately applicants from non-selective secondary schools have been found to be underrepresented in UK medical schools compared to their representation in society even after such adjustments (Wright and Bradley, 2010). In fact, it was recently highlighted that 80% of those studying medicine in the UK applied from only 20% of the countrys secondary schools a vast majority of which are selective (Medical School Council, 2014). Although these applicants from selective secondary schools are much more likely to be selected for medical school, they tend to underperform compared to their counterparts from non-selective schools (McManus, Powis, et al., 2005). In light of these circumstances, other evidence-based and fair selection procedures that may be used singly or in combination with A-level have been considered (Benbassat and Baumal, 2007).

In the UK, this search led to the adoption of UKCAT in 26 UK medical schools in 2006 for applicants starting their studies in 2007. Since then, a majority of UK medical schools have required applicants to sit for the UKCAT as tool for selection for entry since the academic year 2007-2008. The UKCAT was developed by a consortium of medical schools along with the commercial testing company Pearson VUE, a global computer-based testing company that is part of Pearson plc. The UKCAT is delivered worldwide through Pearson VUE's high street test centres. The UKCAT is required for admission into UK medical and dentistry schools from applicants in the UK, EU and a majority of other countries outside the EU. Exemptions on geographical grounds may be considered in exceptional cases if an applicant lives or is educated in a country where UKCAT is not offered and it is not possible for the said applicant to sit for UKCAT in a neighbouring country. Applicants are required to sit for UKCAT once in each test cycle. Rejected applicants to UK medical and dentistry schools in a given UCAS cycle who wish to reapply in subsequent UCAS cycles are required to retake the test. Applicants can take UKCAT at their convenience any time between July and October every year (specific dates in those months may vary from year to year), the standard duration of the UKCAT is 120 minutes although applicants with special needs such as *dyslexia, dyspraxia, dysgraphia, dysorthographia, attention deficit disorder* or *working memory deficit* may be given an additional 30 minutes to complete the test by registering for the UKCATSEN (SEN stands for Special Educational Needs). Applicants from the UK and EU may be eligible for bursaries under certain specified circumstances (Adam, Bore, Childs, et al., 2015; Lynch et al., 2009; UKCAT, 2015; Wright and Bradley, 2010).

The UKCAT assesses a range of mental abilities identified by medical and dental schools as important. There is no curriculum or science content as the test examines aptitude and hence cannot be revised for. Nevertheless, prospective applicants can boost their performance through practice. The test consists of five subtests, four of which test for *cognitive skills* and one which tests for *situated-cognitive skills*. The four cognitive based subtests are *Quantitative Reasoning (QR), Decision Analysis (DA), Verbal Reasoning(VR)* and *Abstract Reasoning (AR)* while situated-cognitive based subtest is the *Situation Judgment Tests (SJTs)*. *QR* assesses an applicants ability to critically evaluate information presented in numerical form, *DA* assesses the ability to make sound decisions and judgements using complex information, *VR* assesses the ability to critically evaluate information that is presented in a written form, *AR* assesses the use of convergent and divergent thinking to infer relationships from information while *SJTs* measures the capacity to understand real world situations and to identify critical factors and appropriate behaviour in dealing with them. Each of the subtests is in a multiple-choice format and is timed separately. Each of the cognitive subtest have their raw score converted to a scale score that ranges from 300 to 900. Therefore total scale score for all of the cognitive subtests range from 1,200 to 3,600. The non-cognitive SJTs raw scores are expressed in one of four bands (band 1=highest, band 4=lowest) each with a unique interpretation of performance. The results of the test are valid for the particular UCAS admission cycle. The UKCAT is used by universities depending on the selection procedure in place. Some universities consider the total score of the cogntive subtests, others consider the individual subtests and may even set a threshold for each subtest, some use UKCAT in addition to *prior educational achievement, personal statements* and *interview performance*, others only use UKCAT to discriminate applicants who have scored equally at some point in their selection process while a small number of universities use UKCAT to widen access by considering applicants who do not stand a chance of progressing through their selection process but have performed well in the UKCAT (Adam, Dowell, and Greatrix, 2011; UKCAT, 2017). During inception of UKCAT, concerns were raised regarding the lack of information of its predictive validity, the perception that an additional unnecessary hurdle was being introduced for entry into medical and dental school. It was also claimed that the registration fee was prohibitive for poor students even though bursaries were available for eligible students in the UK and EU (Wright and Bradley, 2010)(costs in 2017 range between £65 and £115 depending on the time of the year one is scheduled to sit the UKCAT test and whether the test is to be taken in or outside the EU (UKCAT Consortium, 2017)).

Evidence for the performance of the UKCAT has been mixed although largely positive. Most studies have found the UKCAT to be (modestly) predictive of undergraduate medical school performance. For instance, a study conducted by the University of Aberdeen and University of Dundee medical schools showed that there was no correlation between UKCAT and Year 1 performance. This finding was the first study that examined the predictive validity of UKCAT. It was noted that since the study was based on two medical schools, it would prove difficult to make generalisations over the entire set of medical schools which adopted the use of UKCAT in the UK. In addition, the study was conducted during a period at which non-cognitive parts of the UKCAT were still going through development hence some predictive aspects of UKCAT may have been lacking as a result. Nevertheless, it was concerning that the findings failed to find moderate predictive validity demonstrated by similar selection tools (Lynch et al., 2009). A later study established that UKCAT has predictive validity for Year 1 and Year 2 of medical school (in all but one knowledge based examinations). No predictive validity for OSCE was detected which may be indicative of lower reliability of Objective Structured Clinical Examination (OSCE) or that UKCAT is more useful in predicting performance in pre-clinical years. In addition, UKCAT has been found to be less sensitive to school type attended (Wright and Bradley, 2010). Other findings have shown evidence indicative of the UKCAT's *DA* subtest and UKCAT overall total scores predictive validity for Year 1 and Year 2 overall exam score. In recent years, there have been notable studies that have looked into the performance and predictive validity of the UKCAT (Husbands, Mathieson, et al., 2014; James, Yates, and Nicholson, 2010; Lala, Wood, and Baker, 2013; McManus, Dewberry, Nicholson, and Dowell, 2013; McManus, Dewberry, Nicholson, Dowell, et al., 2013; Sartania et al., 2014). A summary of the findings from the studies may be viewed in Table 8.2 in the Technical Appendices.

1.2. Towards understanding construct-level predictive validity, uncertainty for Number Needed to Reject (NNR) and Peer Competition Rescaling (PCR)

Predictive validity estimated as a correlation coefficient between selection test score (predictor) and outcome (criterion) measure is widespread but fraught with challenges. Problems arise due to the fact that whilst the selection test is administered on an entire pool of applicants, validation of the selection test (predictor) by use of predictive validity is applied to only those who entered medical school. This is because the outcome (criterion) is only observed for the entrants. Consequently, the entrants have high but less variable scores on selection test scores (predictor) than the entire pool of applicants, a concept referred to as range restriction. This range restriction causes the estimated correlation coefficient to be artificially deflated (attenuated or downward bias) and may lead to the conclusion that there is modest or at worst no association between selection and outcome measures (McManus, Dewberry, Nicholson, Dowell, et al., 2013). The deflation (downward bias) of the estimated correlation may also be exacerbated by measurement error, a problem that stems from imperfect reliability (Fisher, 2014; Neter et al., 1996). There have been attempts to deal with attenuated correlation starting with Spearman over a century ago whose work on the subject formed a foundation for further research on methods for dealing with the problem. This led to the contribution of knowledge regarding how to adjust for attenuation, computation of associated standard errors and confidence intervals for the dis-attenuated correlation coefficient (Bedeian, Day, and Kelloway, 1997; Behseta et al., 2009; Charles, 2005; Mendoza and Mumford, 1987a). A detailed exploration of these methods together with a treatment of downward bias in associations estimated by β regression coefficients are covered in chapter 3.

Recent proposals from continuing research for dealing with downward bias in the correlation coefficient due to range restriction have been treating the non-existent outcome (criterion) scores of the rejected (non-selected) applicants as a special case of missing data (Mendoza, Bard, et al., 2004; Pfaffel, Kollmayer, et al., 2016; Pfaffel, Schober, and Spiel, 2016; Pfaffel and Spiel, 2016; Wiberg and Sundström, 2009). Figure 1.1 is a schematic representation of the missing data paradigm for estimating construct-level predictive validity that makes use of the selection context in the UK.

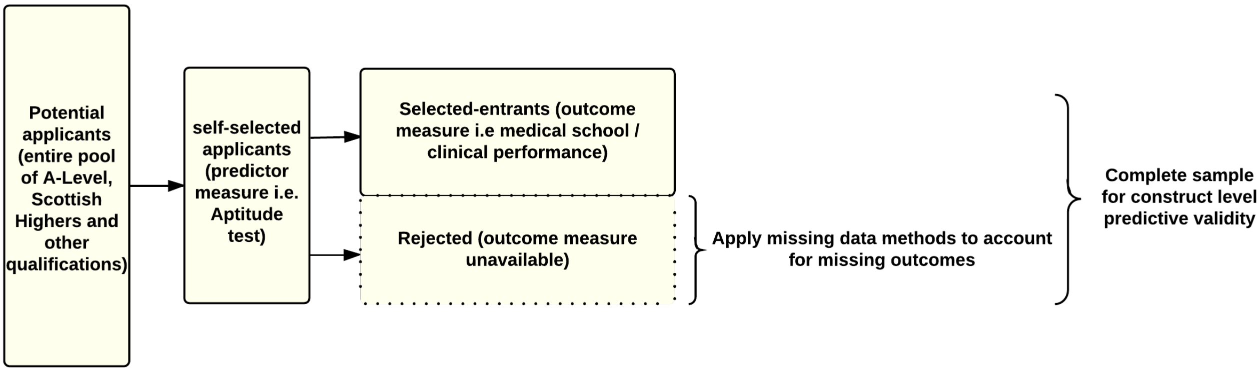

Potential applicants (entire pool of A-Level, Scottish Highers and other qualifications)
self-selected applicants (predictor measure i.e. Aptitude test)
Selected-entrants (outcome measure i.e medical school / clinical performance)
Rejected (outcome measure unavailable)
Apply missing data methods to account for missing outcomes
Complete sample for construct level predictive validity

Figure 1.1.: *Missing data approach to the UKCAT's construct-level predictive validity* [1]

Dealing with the effects of range restriction using missing data handling methods will lead to construct-level predictive validity (association, in this thesis correlation or regression coefficient, between the selection test score (predictor) and outcome (criterion) measure of interest that is free of the effects of range restriction and/ or measurement error). Generally, the missing data handling methods in the literature may be classified into five broad categories. These are *deletion methods, filling-in (imputation) methods, weighing of observations, ignoring of the missing data process of the data* and the more complex *joint modelling of the missing data and measurement(observed) data process*. The validity of these methods in estimating the construct-level predictive validity of selection tests under different selection designs and degrees of strictness of the selection process are explored in chapter 4. The idea therefore will be to extend the statistical methodologies that exist in the literature for estimating construct-level predictive validity by viewing it as a special case of a missing data problem. In addition, viewing uncertainty for NNR and exploring the viability of PCR as other special cases of missing data problems within the selection context, presents an opportunity to apply ,in a novel way, existing statistical methodology for handling missing data.

1.3. Aims

1. To extend statistical methodology for the more accurate appraisal of construct-level predictive validity in selection tests, with the aid of a contrived example using the Professional and Linguistic Assessments Board (PLAB) data.
2. Developing approaches to “nationalising” local outcome measures via “Peer Competition Rescaling (PCR)” using UKCAT as a motivating example.
3. Given aims (1) and (2) above, develop approaches to estimating uncertainty for the estimates of the Number Needed to Reject (NNR) for the UKCAT as a motivating example.

[1] Figure 1.1 and all other conceptual diagrams that do not show results of statistical analyses are constructed in Lucidchart (Lucid Software Inc, 2018)

1.4. Objectives

1. Conduct reviews of the existing literature regarding

    a) The predictors of undergraduate medical school performance used in the selection of medical school entrants.

    b) The predictive validity and possibly construct-level predictive validity of the various aptitude tests used in medical selection internationally (restricted to cognitive tests of fluid intelligence and/or semantic knowledge).

    c) Statistical methodologies for adjusting for attenuation observed in the association between predictors (selection test) and outcomes (criterion) in the selection context.

2. To compare and contrast a variety of approaches for the establishment of construct-level predictive validity in selection tests by use of simulated data.

3. To apply these methods in a contrived example using real-world data in order to establish their potential for predicting performance in medical school applicants, given the challenges of "missing data" especially related to unobserved outcomes in unsuccessful applicants.

4. To apply existing statistical methodology in a novel way in the estimation of the uncertainty for Number Needed to Reject (NNR) and Peer Competition Rescaling (PCR) by viewing them as special cases of "missing data modelling" within the selection context.

1.5. Chapter summary

In this preface, the statistical focus and scope of this thesis have been introduced. In particular, the statistical challenges of the estimation of construct-level predictive validity, NNR and PCR are covered. In the next chapter, the main methods used in the selection of medical school entrants worldwide are briefly summarised. In addition, findings of literature review conducted to meet objectives 1(a) and 1(b) outlined in section 1.4 will be discussed.

Part II.

Literature review

2. Predictors of medical school performance used in medical selection and construct-level predictive validity of aptitude tests

In this thesis chapter, findings of literature reviews conducted in line with objectives 1(a) and 1(b) of the objectives listed in section 1.4 are presented. The findings for objective 1(a) relating to predictors of undergraduate medical school performance used to select medical school entrants were notably similar to findings from a recent systematic review conducted by *Patterson, Knight, et al. (2016)* entitled *How effective are selection methods in medical education? A systematic review*. In their publication, *Patterson et al.* summarised their findings into two categories, namely *short-listing* and *final-stage* selection methods. Under the short-listing methods, were *aptitude tests, prior academic attainment, personal statements, references, Situation Judgment Tests (SJTs)s, personality assessment and Emotional Intelligence (EI)* while final-stage methods consisted mainly of the different forms of *interviews* such as *traditional, structured* and *Multiple Mini Interviews (MMIs).* Therefore, the findings of the literature review for objective 1(a) in this thesis chapter are presented using the same categorisation in section 2.1 and 2.2 respectively. In addition, findings of the literature search for objective 1(b) relating to the construct-level predictive validity of the various aptitude tests used in medical selection are presented in section 2.3.

For the literature review related to objective 1(a), two sources, *Google Scholar* (Google Scholar, 2017) and *Web of Science (previously known as Web of Knowledge)* (Clarivate Analytics, 2017) were selected. The "PICO" methodology (Centre for Reviews and Dissemination (CRD), 2008) was used to formulate the search terms as demonstrated in Table 2.1.

Acronym letter	**Description**	**Contextualisation**
P	Population, problem, patient	Medical school students
I	Intervention	Predictors
C	Comparison, control, comparator	-
O	Outcome	Success in medical school

Table 2.1.: *The PICO methodology contextualised in the review of literature on predictors of medical school performance*

Figure 2.1 shows the search term, the inclusion and exclusion process for the review conducted for objective 1(a). In all, 97 papers were considered to be relevant for the review. The predictors identified in the review are summarised in following sections 2.1 and 2.2.

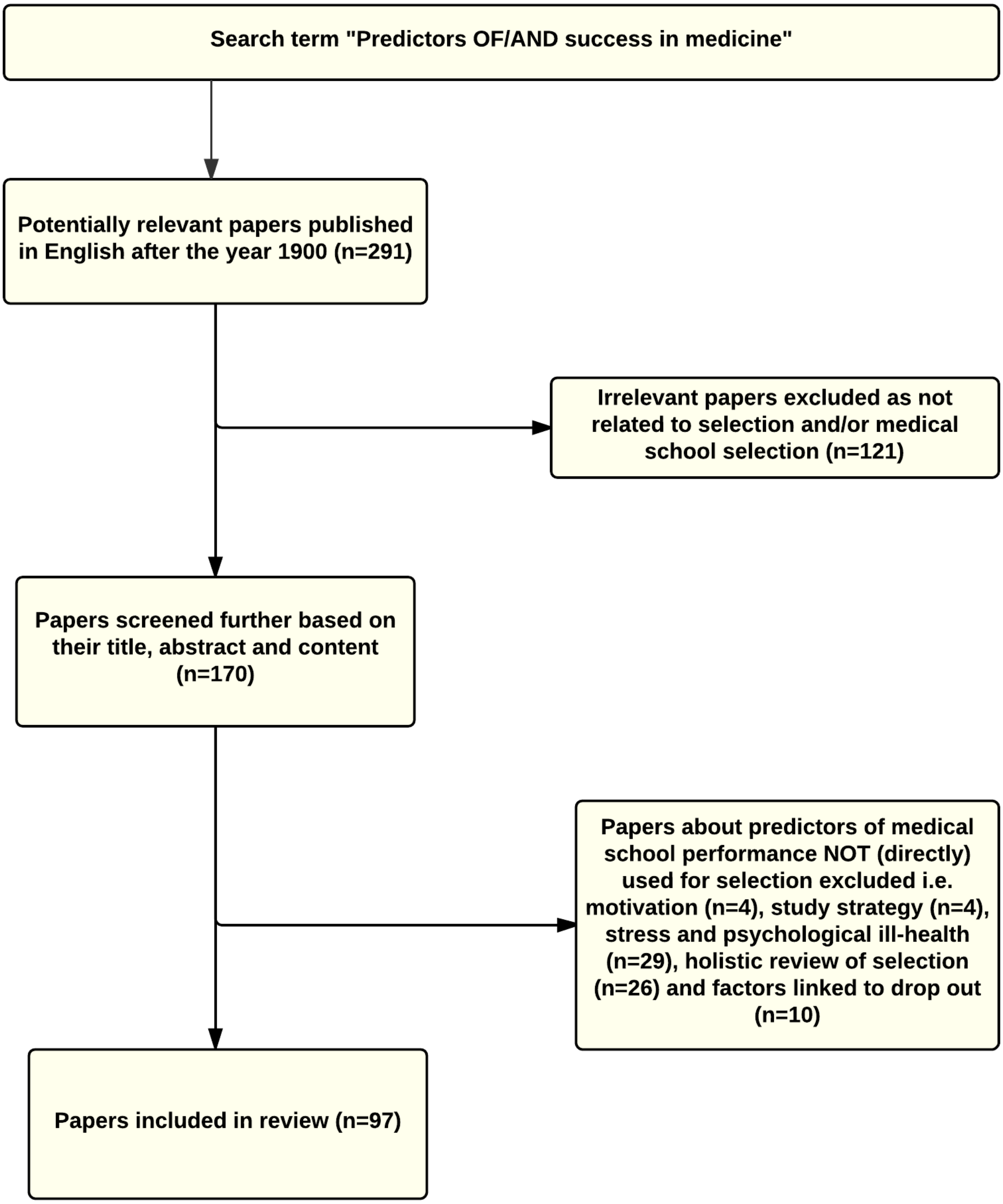

Search term "Predictors OF/AND success in medicine"
Potentially relevant papers published in English after the year 1900 (n=291)
Irrelevant papers excluded as not related to selection and/or medical school selection (n=121)
Papers screened further based on their title, abstract and content (n=170)
Papers about predictors of medical school performance NOT (directly) used for selection excluded i.e. motivation (n=4), study strategy (n=4), stress and psychological ill-health (n=29), holistic review of selection (n=26) and factors linked to drop out (n=10)
Papers included in review (n=97)

Figure 2.1.: *Flow chart of papers included in the review of predictors of undergraduate medical school performance used in selection of medical school entrants*

2.1. Predictors of undergraduate medical school performance used in the short-listing stage of selection

2.1.1. Prior educational achievement

Examinations deemed to be a reliable measure of academic (or scholastic) ability of applicants are in use across many countries. In Australia, educational performance at secondary school has been shown to have predictive validity for medical school outcomes (Mercer and Puddey, 2011). This is collaborated by evidence from a different study in an Australian university that showed that an integrated selection procedure for medical school that utilised prior educational achievement was predictive of learning outcomes (Simpson et al., 2014). This is also supported by evidence from a South Korean study that reported that pre-admission GPA was a reliable predictor during medical school and postgraduate clinical performance (Kim, Chang, et al., 2016). An early study into the effect of prior educational achievement in the USA found that almost without exception, high achieving pre-medical students tend to be high achieving medical school students (Buehler and Trainer, 1962).

In the UK, General Certificate of Education Advanced Level (A-level) and General Certificate of Secondary Education (GCSE) have been demonstrated to have a strong and persistent but diminishing predictive validity for all the undergraduate and postgraduate medical school examinations. A different study found that GCSE had greater predictive validity than A-level for undergraduate and postgraduate medical school examinations with incremental predictive validity for clinical and post-graduate performance (McManus, Woolf, et al., 2013). In addition, the predictive validity of A-level and GCSE grades were found to be higher than that of aptitude tests (McManus, Dewberry, Nicholson, Dowell, et al., 2013) although the observed differences in ability of medical school entrants based on this exam disappeared or reduced during medical school (Thiele, Pope, et al., 2016). In India and the Czech republic, prior educational performance has also been found to be a good predictor of overall medical school performance (Gupta, Nagpal, and Dhaliwal, 2013; Štuka et al., 2012). A Saudi Arabian study found that final high school exam scores were more predictive of medical school pre-clinical performance than aptitude tests (Al Alwan et al., 2013). A different study in Saudi Arabia found that high school scores were not predictive of medical school clinical performance (Salem et al., 2016). A study at Leicester-Warwick medical schools concluded that graduate entry students performed as well as their undergraduates counterparts in final examinations despite having lower General Certificate of Education Advanced Level (A-level) grades and a shorter four year accelerated course (Shehmar et al., 2010). A New Zealand study found that the pre-admission GPA had predictive validity for year two and three academic performances and was further predictive of whether or not a medical student would earn a "Distinction" rather than a "Pass" in year four (Shulruf et al., 2012).

The emerging evidence suggests that prior educational achievement is highly predictive of medical school performance and subsequent career outcomes, for this reason, prior educational achievement is an important predictor used in the medical school selection process. While this is defensible, caution has been advised by those who argue that prior educational achievement should be used in the selection process in a way that mitigates the barriers that exist against under-represented and disadvantaged groups (Patterson, Knight, et al., 2016).

2.1.2. Aptitude tests

Aptitude tests for medical school selection were first used in the USA in the late 1920s at a time in which concerns about medical school attrition rates of 5% to 50% were rife. Prior to this time, selection to medical schools in the USA had been based on undergraduate education and sometimes on a high school diploma together with biographical information and recommendation letters. The first aptitude test developed and implemented by physician and psychologist F.A Moss, the *Scholastic Aptitude Test for Medical Schools*, was administered between 1928 and 1946. By 1946, the aptitude test was credited with reducing the attrition rate to 7% based on selection of students purely on academic grounds. This aptitude test later evolved into the *Professional School Aptitude Test and Medical College Admission Test* and was administered between 1946 and 1962 (McGaghie, 2002). Thereafter, it later evolved into the Medical College Admissions Test (MCAT) which was sponsored and administered by the Association of American Medical Colleges (AAMC). Table 2.2 shows the timeline and evolution of the MCAT (now used in the USA and Canada) for medical school selection (Eskander, Shandling, and Hanson, 2013).

Since the adoption of the MCAT in the USA, aptitude tests have come into widespread use across the world for medical school selection. For example, in the UK increase of medical applicants with top A-level grades made it difficult for medical schools to conduct selection. In addition, A-level performance had been shown to be biased in favour of female applicants and those from selective secondary schools. This led to the adoption the BioMedical Admissions Test (BMAT) and United Kingdom Clinical Aptitude Test (UKCAT), as selection tools in 2003 and 2006 respectively (Emery and Bell, 2009; Ferguson, James, and Madeley, 2002; James, Yates, and Nicholson, 2010; McManus, Powis, et al., 2005). Similar concerns were rife in Ireland about the fairness of selecting medical school applicants based on Irish Leaving Certificate Examination (or equivalent). In 2009 an aptitude test for selection, the Health Professions Admission Test-Ireland (HPAT-Ireland), was adopted (O'Flynn, Fitzgerald, and Mills, 2013). In Australia the Graduate Medical School Admissions Test (GAMSAT) was adopted as a selection tool into graduate-entry programs in medicine or dentistry in 1996 (Coates, 2008) while the Undergraduate Medical and Health Sciences Admission Test (UMAT) has been in use in New Zealand and Australia since 2003 for selection of applicants into medicine, dentistry and health science degree programs at undergraduate level (Poole et al., 2012; Puddey and Mercer, 2013). Other examples of aptitude tests used for medical school selection include the *Medical and Dental Colleges Entrance Tests* in Pakistan, *MCAT* in Austria, *HAM-Nat* in Germany, *Saudi Aptitude Test* in Saudi Arabia, *Konkoor* in Iran and *Eignungstest für das Medizinstudium in der Schweiz (EMS)* or Aptitude test for medical studies in Switzerland (Abbiati, Baroffio, and Gerbase, 2016; Farrokhi-Khajeh-Pasha et al., 2012; Habersack et al., 2015; Hissbach, Klusmann, and Hampe, 2011; Khan, Mukhtar, and Tabasum, 2014; Khan, Tabasum, and Mukhtar, 2013).

	Aptitude test type				
	Scholastic Aptitude Test for Medical Schools	Professional School Aptitude Test and Medical College Admission Test	MCAT	“New” MCAT	“Current” MCAT
Test years	1928 to 1946	1946 to 1962	1962 to 1977	1977 to 1991	1991 to 2002
Sub-tests					
1.	Visual memory	Verbal ability	Verbal ability	Science knowledge	Verbal reasoning
2	Memory for content	Quantitative ability	Quantitative ability	Science problems	Biological sciences
3	Scientific vocabulary	Science achievement	Science achievement	Skills analysis:reading	Physical sciences
4	Scientific definitions	Understanding modern society	General information	Skills analysis: quantitative	Writing sample
5	Understanding of written materials				
6	Premedical information				
7	Logical reasoning				
Score range	0-250, 0-275, 0-300, 0-385	200-800	200-800	1-15	1-13, 1-15, J to T for writing sample
Question type	True-False and Multiple choice	Multiple choice	Multiple choice	Multiple choice	Multiple choice and writing sample that elicits constructed response

Table 2.2.: *The evolution of the first developed aptitude test, the Medical College Admissions Test (MCAT)*

The factors that influence the performance of aptitude tests have been studied. In Pakistan, emerging evidence suggests that female and highly economically and academically developed districts score higher in *Medical and Dental Colleges Entrance Tests* (Khan, Tabasum, and Mukhtar, 2013). In the UK, the BMAT has been shown to have predictive validity for the pre-clinical years of medical school (Emery and Bell, 2009). It has been shown that the socio-demographic factors that predict A-level performance also predict UKCAT performance although compared to A-level, male and native English speakers fair much better in UKCAT suggesting that the UKCAT may be less sensitive to gender and secondary school type (Tiffin, McLachlan, et al., 2014). A study that examined the use of GAMSAT over a ten year period found it to have high reliability and slightly sensitive to age, gender, level and discipline of previous academic study and language background of the applicant (Mercer, Crotty, et al., 2015). In Austria, it has been shown that male applicants tend to score higher than female applicants on the Austrian version of the MCAT upon completing secondary education (Habersack et al., 2015). In terms of the predictive validity of aptitude tests for medical school performance, the UMAT has less predictive power compared to the GPA although the UMAT has incremental predictive validity when used together with GPA (Poole et al., 2012). A different study on UMAT found that its predictive validity varied between schools, across medical school years, and within sections of the UMAT and socio-economic strata (Edwards, Friedman, and Pearce, 2013; Puddey and Mercer, 2013). The *HAM-Nat* has been found to be predictive of medical school performance independent of high school performance although this predictive power is higher in females than in males (Hissbach, Klusmann, and Hampe, 2011).

In the USA, the MCAT has been found to have weak to moderate predictive validity for multiple choice based medical school assessments. The predictive validity of the MCAT was found to be stronger in early years of medical school with no predictive power in the clinical years (Saguil et al., 2015) while another study found that both undergraduate GPA and total MCAT scores were strong predictors in medical school performance throughout medical school (Dunleavy et al., 2013). Another study on the predictive validity of MCAT found that it explained twice as much variance in performance at medical school than undergraduate GPA (Julian, 2005).

An early study of the predictive validity of the UKCAT at two Scottish universities found no predictive power in year one of medical school (Lynch et al., 2009). This same conclusion was made by a different study from England which reported that the UKCAT did not independently predict medical school performance (Yates and James, 2013). A different Scottish university found that the UKCAT had modest but diminishing predictive power of performance throughout the undergraduate medical school above that of school science achievements or pre-admission interview (Sartania et al., 2014). A different Scottish study found that the predictive validity of UKCAT was stronger in the latter years of medical school (Husbands, Mathieson, et al., 2014). A study in Australia did not find evidence for the predictive validity of the GAMSAT (Groves, Gordon, and Ryan, 2007) while other studies reported that GAMSAT and GPA used in combination were predictive of medical school performance in year one (Coates, 2008; Puddey and Mercer, 2014). The *Saudi Aptitude Test* has been reported to have predictive validity for medical school performance alongside the pre-admission high school and achievement tests (Al Alwan et al., 2013). In Pakistan, the *Medical and Dental Colleges Entrance Tests* has been found to be reliable and predictive of performance in the pre-clinical years of medical school (Khan, Mukhtar, and Tabasum, 2014). A different study in the UK established that critical thinking aptitude tests had incremental predictive validity for final psychology exams outcomes (O'Hare and McGuinness, 2015). In Iran, the *Konkoor* alone, even with combination with high school GPA, has poor and diminishing predictive validity for performance throughout medical school (Farrokhi-Khajeh-Pasha et al., 2012). The HPAT-Ireland has been reported to be a good predictor of performance at year two medical school examination and Objective Structured Clinical Examination (OSCE) (Kelly, Regan, et al., 2013).

Overall, aptitude tests, have been evidenced to have some incremental predictive validity for medical school performance over and above other predictors of medical school performance, one of the concerns regarding the use of aptitude tests as selection tools is the impact of coaching, practicing and re-sitting on aptitude test performance. It has been demonstrated that those applicants who are coached, practice or re-sit the UMAT tend to have higher scores. It is not completely understood whether this improvement represents just an artificially induced increase in performance or an increase in ability and therefore increased competence. If the improvement represents an artificially induced increase in performance, it may be indicative of “coaching effects” rather than improved true ability (Laurence et al., 2013; Puddey, Mercer, et al., 2014). There is also emerging evidence that medical school entrants who were coached for the UMAT have inferior academic outcomes throughout medical school (Griffin, Yeomans, and Wilson, 2013). The other concern is the reliability and validity of aptitude tests may vary depending on how they are used in the selection process, this is a major concern as there is no consensus on how medical schools should use aptitude tests in the selection process (Patterson, Knight, et al., 2016). To illustrate, consider the use of the UKCAT in the medical school selection process, some universities consider the total score of the cogntive subtests, others consider the individual subtests and may even set a threshold for each subtest. Some use the UKCAT in addition to *prior educational achievement, personal statements* and *interview performance*, others only use the UKCAT to discriminate applicants who have equal scores at some point in their selection process. A small number of universities use the UKCAT to widen access by considering applicants who do not stand a chance of progressing through their selection process but have performed well in the UKCAT (Adam, Dowell, and Greatrix, 2011).

2.1.3. Situation Judgment Tests (SJTs) and Emotional Intelligence (EI)

SJTss are test of non-academic attributes designed to assess a candidates’s judgement regarding scenarios encountered in specific scenarios (Patterson, Cousans, et al., 2017). They include tests such as the *Clinical Problem Solving Test (CPST)* which has been shown to be predictive of performance at medical school (Ahmed, Rhydderch, and Matthews, 2012; Koczwara et al., 2012). In the UK, studies have demonstrated that the SJTss components of the Personal Qualities Assessment (PQA) tests together with UKCAT have significant associations with medical school performance and later professional behaviours of medical students (Adam, Bore, Childs, et al., 2015; Adam, Bore, McKendree, et al., 2012). The SJTs has been shown to be have significant association with supervisors ratings of integrity and Multiple Mini Interviews (MMIs) suggesting that carefully designed SJTs may augment more costly MMIs (Husbands, Rodgerson, et al., 2015). This has been confirmed by another study from Australia which found that SJTs and MMIs scores were both predictive of all end of postgraduate general practice training and were complimentary in predicting end of training scores (Patterson, Rowett, et al., 2016). A study on the SJTs component of the UKCAT concluded it has the potential to diversify medical school intake since it is less sensitive to socio- economic status compared to cognitive tests (Lievens et al., 2016). Other researchers have remained unconvinced that SJTs measure what they are purported to measure and that there is no justifiable evidence for them to be used for medical school selection (Harris, Walsh, and Lammy, 2015). Unlike aptitude tests, the emerging evidence does not suggest that applicant performance on SJTs may be enhanced significantly by coaching (Taylor et al., 2016) although research has demonstrated that mode of administration may affect the properties of SJTs, with video-based SJTs having greater operational validity compared to paper-based SJTs. In addition, different response instructions may affect their validity (Patterson, Knight, et al., 2016).

Emotional Intelligence (EI) refers to the awareness and ability of a person to respond to his/her emotions and those of other people (Cherry et al., 2014). EI may thus be used as a measure of interpersonal and communication skills which together may be viewed as *interpersonal aptitude* (Carr, 2009). In addition, EI may contribute to *professionalism, compassionate* and *empathic patient care* (Cherry, 2014). Research has demonstrated that EI is associated with good social skills, good academic performance and empathy towards patients. A Malaysian study found that medical school students who possessed higher EI performed better in both first and fifth year of continuous assessments under consideration. In the fifth year, however, EI explained a small variance of the performance (Chew, Zain, and Hassan, 2013). A study at the University of Bali involving Psychology students demonstrated that EI is positively associated with academic achievement with higher associations reported for men than women. It was also found that EI had predictive validity for scholastic performance over and above cognitive ability and personality variables (Lanciano and Curci, 2014). Another Caribbean study also found that younger and male medical school students possessed higher EI than females (Sa et al., 2014). It has been shown that EI may differ by gender and nationality. A Japanese study found stronger and significant positive effects for male and non-Japanese medical school students although over the long term, female and Japanese students showed greater increase in EI (Abe et al., 2013). An international study across four countries, the UK, Australia, Ireland and Hong Kong, involving first year radiography students evaluated EI differences based on gender, age and culture. In keeping with other studies, it was found that across the four countries, male radiography students had significantly higher EI scores than their female counterparts. There was no evidence of association between age and EI suggesting that EI remains stable over a person's lifetime. The Hong Kong EI student's scores differed from other scores of students from the UK, Australia and Ireland, a finding which is consistent with what other researchers have concluded regarding EI differences from western and Eastern cultures (McNulty et al., 2016). The cumulative GPA in Saudi Arabia has been shown to be a predictor of EI (Naeem et al., 2014) although another study from Canada found no association between EI and selection measures such as weighted GPA, personal statements and interview scores (Leddy et al., 2011). Another Canadian study concluded that EI scores derived from *Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT)* for those selected through interviews did not have predictive validity for future medical school performance (Humphrey-Murto et al., 2014). A separate Canadian study also did not find evidence supporting the suitability of the use of EI scores for potential pre-screening tool for MMIs (Yen et al., 2011). The research on the suitability and applicability of EI in medical school selection and training has been mixed. There is some acknowledgement that there is some association between EI and medical school performance and that EI may vary by gender, age and socio-economic status (Mankus, Boden, and Thompson, 2016). Future research will inform and shape both the debate and medical school policies with respect to EI (Cherry, 2014).

2.1.4. Personality tests

Personality tests are *non-cognitive* based tests designed to evaluate a candidate's non-academic attributes of character and psychological make up (MacKenzie, Dowell, et al., 2017). It is known that a doctor's personality may influence aspects of care such as patient satisfaction, adherence to treatments and health outcomes. For this reason, the use of personality testing for selection has received growing interest in recent years. For example, the *Five Factor Model (FFM)* is widely used across the world and measures five attributes, namely, *conscientiousness* (which constitutes self-discipline, persistence and striving for achievement), *extraversion* (which constitutes sociability, positive affect and energetic behaviour), *agreeableness* (which constitutes altruistic affective and collaborative behaviour), *neuroticism* (constitutes anxiety, fearfulness, and insecurity in relationship) and *openness to experience* (constitutes active imagination, preference for variety and intellectual curiosity) which when taken together reflect individual differences in social, emotional and behavioural patterns (Costa et al., 2014).

Other non-knowledge workplace-based simulation tests have also become of interest. The *Five Factor Model (FFM)* may be thought of as a *personality test.* The Persian version of the non-cognitive PQA in Iran and *Five Factor Model (FFM)* in Netherlands have been shown to be useful for selecting medical students with desired non-academic attributes (Nedjat et al., 2013; Schripsema et al., 2016). Studies in Japan and Taiwan on the *Libertarian-Dual-Communitarian moral orientations (Mojac)* and *Narcissism, Aloofness, Confidence, and Empathy (NACE)* components of the PQA found them to be internally consistent for use in selection with the caveat that cultural considerations of the tests would have to be made (Fukui et al., 2014; Tsou et al., 2013). In Sri-Lanka, the *Non-Cognitive Questionnaire (NQ)*, when combined with prior educational achievement has been reported to have utility for early detection of academic strugglers in medical school (Ranasinghe, Ellawela, and Gunatilake, 2012). This has been further supported by a different study in Netherlands which concluded that non-cognitve tests on their own are not sufficient to select the best academically performing students at medical school because cognitive skills are also needed to succeed in medical school (Lucieer et al., 2016) while a study in Switzerland found that there was no compelling evidence that cognitive tests advantage or preclude applicants with more desired personality traits for medical school selection (Abbiati, Baroffio, and Gerbase, 2016).

Those opposed to use of the *personality tests* in selection argue that the association between personality traits like conscientiousness with learning outcomes, for example, may change in direction (from enhancing to inhibiting) when the context changes. It has been demonstrated that clinical knowledge and skills can be separated and predicted by different prior learning patterns and personality traits. Therefore personality tests may need to be designed in a way that is specific to the learning outcome under consideration (Ferguson, Semper, et al., 2014). There are concerns that widespread adoption of personality tests in selection may lead to homogenisation of medical school students and thus lead to an increase of biases. In addition, it has also been suggested that personality traits develop until the age of thirty and beyond thus it would be unwise to select medical school applicants based on personality traits in their late teens or early twenties (Mushtaq and Ratneswaran, 2016; Wilson et al., 2013).

2.1.5. Personal statements and recommendation letters

There is little (if any) evidence that personal statements (or motivational statements) distinguish selected from non-selected medical school applicants in terms of motivation (Wouters et al., 2014). Proponents of the use of personal statements argue that applicants typically include in personal statements more information than what is required for selection. The extra information included in the personal statements may add value by providing applicants with a platform to tell their own story and hence encourage them to gain more information about medical school programs. This may thus help them make an informed decision about applying to medical school. Overall, personal statements lack the reliability and validity associated with other selection measures (predictors of medical school performance) although they remain widely used in medical school selection. Opponents of the use of personal statements argue that the extra information included by applicants may cloud the judgement of the individuals making selection decisions (Patterson, Knight, et al., 2016). A meta-analysis on the predictive power of recommendation letters, involving both medical and non-medical students, revealed that overall recommendation letters have a weak but positive association with predictors of performance in post-secondary education (Kuncel, Kochevar, and Ones, 2014). The study also found that recommendation letters also provide incremental validity for degree attainment. For the non-medical students, the predictors of performance used were Grade Point Average (GPA) while the performance outcomes of interest were performance ratings, degree attainment and research productivity. For the medical students, the predictors of performance used were the GPA and internship. It was also noted that the effect of recommendation letters may be moderated by the format of the letter.

2.1.6. Use of background (contextual) data

The medical profession has, for decades, been associated with the elite, those of high social economic status, who traditionally have had high matriculation rates in medical schools (Carlisle, Gardner, and Liu, 1998). In response, there have been policies and programmes developed to specifically address this by increasing the number of under-represented groups in medical schools. These policies, referred to as *Affirmative Action* in the USA (Davidson and Lewis, 1997; Garces and Mickey-Pabello, 2015; Lakhan, 2003) or *Widening Partcipation (Access)* in the UK (British Medical Association, 2017; General Medical Council, 2017a; Medic Portal, 2017) are increasingly being used during the medical school selection process. These policies typically aim to address gender, racial, ethnic and socio-economic imbalances in the medical school matriculation rates. For example, in the UK, it is has been established that 80% of medical school entrants come from 20% of the secondary schools (Medical School Council, 2014). Most of these secondary schools are selective or grammar schools which are located in affluent and highly resourced areas. In response, some medicals schools have started to incorporate the use of applicant's background information into the selection process so as to redress the trend. For example, King's College London, have an *Extended Medical Degree Course* developed to target applicants studying at A-level at non-selective secondary schools in England (King's College London, 2017), the University of Birmingham has two programmes, one called *Access to Birmingham (A2B)* developed for student applicants to medical school who hail from families or communities in the West-Midlands who have little or no experience with Higher Education (University of Birmingham, 2017a). The other is called *Routes to the Professions* developed for applicants from under-represented groups to access courses that lead to careers in four key professions including Medicine and Dentistry (University of Birmingham, 2017b). The University of Manchester's *Manchester Access Programme* targets under-represented applicants to medical school from low socio-economic areas with no history of participation in Higher Education (University of Manchester, 2017).

The factoring in of background (contextual) data in the selection of medical school entrants worldwide is growing in popularity. The background (contextual) data considered have included *age, degree history, rurality, gender, age, ethinicity* and *nationality* among others. A study in Ireland concluded that medical school students who enter medical school through the four year *Graduate Entry Programme (GEP)* perform better than the undergraduate five or six year *Direct Entry Programme (DEP)* students in the common final two years of medical school exams for the two groups. There were no significant differences in performance between those medical school entrants with *science* and *non-science* background within the GEP or those with EU or non-EU nationality (Byrne et al., 2014). A similar conclusion was reported by a Chiropractic college in the USA which found that degree holding students, as well as female and native English speaking students, performed better in year one exams (Green, Johnson, and McCarthy, 2003). A South African study found that there were differences between medical school entrants from urban and rural areas in the perception of all but one aspect of the first and final year of university life. Medical school students from rural areas found the medical school environment, language of instruction, technology, finances and personal difficulties more challenging than their urban counterparts. There was ,however, no difference in the perception of academic content of the medical school curriculum (Diab et al., 2015). An exploratory study at a university in South East England found that ethnic minority physiotherapy students faired worse in final clinical placement scores with no significant differences observed in scores based on gender and age (Naylor, Norris, and Williams, 2014). A Dutch study found that both ethnicity and social background were independent predictors of medical school selection. Dutch applicants stood a higher chance of selection compared to Surinamese/Antillean, Turkish/Moroccan/African and Asian applicants. The performance in pre-university GPA partly explained this disparity in all but the Surinamese/Antillean applicant group. Sociodemographic variables partly accounted for the difference found for the Asian applicants only. First generation immigrants were more likely not to be selected for medical school compared to non-first generation immigrants (Stegers-Jager, Steyerberg, Lucieer, et al., 2015). Related studies found that the selection disparity persisted in medical school performance with Dutch students performing better than the Surinamese/Antillean and Asian students in pre-clinical courses. Dutch students were more likely to earn higher marks than all non-Dutch counterparts in clinical courses even after adjusting for age, gender, pre-university GPA and socio-demographic variables (Stegers-Jager, Steyerberg, Cohen-Schotanus, et al., 2012; Stegers-Jager, Themmen, et al., 2015).

Similar conclusions were made from a meta-analysis of UK medical school students, the results suggested that non-white medical students performed less well compared to white students across different medical schools, different types of exams, and in undergraduate and postgraduate medical courses (Woolf, Potts, and McManus, 2011). In the USA, it is known that historically, under-represented minorities earn lower scores compared to their white counterparts in GPA and MCAT (Veloski et al., 2000). This has led to under-represented minorities entry scores being lowered at selection for medical school. This trend (of lower academic achievement under-represented minorities) has been shown to persist for some medical school performance measures but disappears in clinical years when performance between ethnic groups is equivalent. It has also been demonstrated that medical school student selection in the UK has historically highly favoured female and affluent applicants (based on area of residence and social-economic status of parents). Applicants from affluent areas were much more likely to apply for medical school and to be selected. Such applicants were also much more likely to attend selective and grammar secondary schools. The rates of attendance in these secondary school types is highest in Scotland followed by England, Wales and Northern Ireland (Houston, Osborne, and Rimmer, 2015; McManus, Dewberry, Nicholson, and Dowell, 2013; Simmenroth-Nayda and Görlich, 2015; Steven et al., 2016). It is known however that medical school entrants from low performing state and comprehensive schools are more likely to achieve the highest degree classifications despite low chance of selection (Thiele, Singleton, et al., 2015).

In Turkey, a study found that female and financially well off medical school students stood a higher chance of academic success (Ogenler and Selvi, 2014). A study from University College London (UCL) confirmed that ethnic minority medical students under-perform academically in final year examinations compared to non-ethnic minority applicants. This negative association of performance with ethnicity was not mediated by age, socio-economic group, sex, schooling, parents education, language, personality, study habits, or motivation (Woolf, McManus, et al., 2013). A Dutch university found that medical school applicants who had parents who were physicians accounted for close to half of all applications. A rate much higher than reported in the literature. Although those who had parents who were physicians were much more likely to apply, there were no demonstrated accrued advantages in their performance in MMIs and interview scores, preparation for the admission test, or in receiving or accepting a place at medical school (Jerant et al., 2015). Furthermore, a different study showed that those medical school entrants who had at least one parent who was a physician, male and aged between 19 and 21 years (inclusive) were less likely to complete year one successfully (Stegers-Jager, Themmen, et al., 2015).

In summary, the emerging evidence does not suggest that in themselves *background (contextual) data* are predictive of medical performance but are increasingly used in the medical school selection process to increase matriculation rates for groups who have traditionally been underrepresented and disadvantaged at selection. Proponents argue that the non-traditional medical students may contribute towards the education of their peers by challenging the prevailing medical culture. It is also argued that in future a diverse workforce of medical practitioners will be better able to engage with patients from a diverse background, thus improving the overall quality of healthcare. Opponents argue that there are no valid or reliable methods for quantifying the accrued benefits of factoring in *background (contextual) data* in the medical school selection process (Patterson, Knight, et al., 2016). In fact, there have been several legal challenges in the USA to affirmative action such as *Regents of University of California v. Bakke (1978)*, *Grutter v. Bollinger (2003)* and *Fisher v. University of Texas at Austin (2012-2013)* (Valarie Blake, 2012).

2.2. Predictors of undergraduate medical school performance used in final stage of selection

2.2.1. Interviews

Interviews are tests of non-academic attributes used by medical schools in a variety of ways during the selection process. For example, they may used to *gather information* about an applicant, make the decision to *accept or reject* an applicant, obtain *non-cognitive data* and *verify information* provided in the application and to *select* particular applicants. Interview formats may be *one-on-one*, *group* with one interviewer and several interviewees, *panel* with many interviewers for each applicant or a *combination* of any of the formats mentioned (Edwards, Johnson, and Molidor, 1996). *Multiple Mini Interviews (MMIs)* are an improvement on the traditional (panel) interview process which use brief, sequential interviews with structured tasks and independent assessments within each interview (Pau, Chen, et al., 2016). There have been a number of studies focussed on the usefulness of interviews in selection and their utility in predicting medical school performance. The results of these studies have been mixed but largely positive. A study in Taiwan called into question the ability of interviews to help medical schools to gather non-cognitive data from applicants when it was found that selected and non-selected interviewees did not differ significantly in their non-cognitive attributes (Fan et al., 2010). Another study found that the proportion of male medical students enrolled was significantly increased when interviews were removed from the selection process (Wilkinson, Casey, and Eley, 2014). In contrast, a similar study found that compared to the rejected, using a 12 station MMIs, the selected applicants performed better in Canadian National Licensing Examinations (Eva, Reiter, et al., 2012). Selecting applicants for interviews is a complicated process and it has been suggested that this process may be simplified by using an aptitude test as a screening test such as the UKCAT (Turner and Nicholson, 2011) which may have the added advantage of increasing the enrolment of ethnicities under-represented in medicine (Terregino, McConnell, and Reiter, 2015). On the other hand some studies have hinted that interviews should be independently used to add value to the selection process over and above that provided by aptitude tests such as the MCAT (VanSusteren et al., 1999). A Swedish study found that there was little difference in performance between medical school students selected through academic merit and those selected through interviews in the clinical years when evaluations are primarily conducted through written exams. Those selected through interviews, however, tended to perform better in the fourth year of medical school when evaluations are hinged on interpersonal and communication skills (Dahlin et al., 2012; Gutowski et al., 2010; Oluwasanjo, Wasser, and Alweis, 2015). Another study in the USA determined that the predictive validity of interviews generally increased throughout medical school unlike the for the case of MCAT and pre-admission GPA (Elam and Johnson, 1992). This positive finding has been echoed to some extent by different studies in the UK which reported that MMIs scores had the most consistent predictive validity for medical school performance in the early years of medical school (Husbands and Dowell, 2013) and that the MMIs approach is reliable, acceptable and feasible (Pau, Jeevaratnam, et al., 2013).

Research has demonstrated that the MMIs process may be improved by the use of several groups of interviewers (examiners) whose role would be to score medical school applicants independently. This would lead to each applicant receiving a weighted and fairer score from the MMIs. For the purpose of quality control, those examiners with wildly different scores may be identified for further training (Till, Myford, and Dowell, 2012). It has also been demonstrated that internet based MMIs (that is iMMIs) conducted online via Skype, for example, are also as reliable as the in person MMIs with both iMMIs and MMIs producing comparably similar, acceptable and reliable results. Compared to MMIs, internet based MMIs enables selection to be done by medical schools with reduced resources (Tiller et al., 2013). The advent of the several station MMIs signalled the transformation of the traditional interview format. There are several reliable and cost efficient MMIs designs (Knorr and Hissbach, 2014). The MMIs is very flexible since the test characteristics for the several stations may be structured to enable selection of the most suitable candidates for medical school. A study conducted in Canada, determined that, stations could be categorised into three groups. The first group was *Situation Judgment Tests (SJTs)* where applicants were asked to imagine what they would do in specific situations. The second group was *Behavioural Interview (BI)* where applicants were asked to recall what they did in experienced situations. The third group was *Free Form (FF)* where the interview was unstructured. The examiner was given a brief explanation of the intent of the interview without further guidance on how it should be conducted. The results suggest that structuring of the MMIs stations has value, although that value is gained only through the use of BI stations (Eva and Macala, 2014). This has however been contradicted by a different study in Japan which concluded that SJTs and BI structured MMIs are equally reliable. These may be used together although maximal utility may be gained when used independently in different stations (Yoshimura et al., 2015). The MMIs have been shown to have predictive validity for medical school success in Ireland devoid of the influence of gender, age and socio-economic status, although those medical students from non-EU countries that did not speak English as a native language tended to have significantly lower MMIs scores (Kelly, Dowell, et al., 2014; Oliver et al., 2014). A medical school in the USA found that selection based on MMIs did not disfavour under-represented and ethnic minority applicants. Applicants from lower socio-economic status were more likely to be invited to an Multiple Mini Interviews (MMIs) and recommended for acceptance even though they had lower scores (Yoshimura et al., 2015). This is in line with an Australian study which concluded that applicants from lower social-economic status were less likely to be disadvantaged by selection based on interviews unlike other selection measures in which female applicants from lower social-economic status were particularly disadvantaged (Griffin and Hu, 2015).

Overall, interviews are the widely used in the medical selection process. Emerging evidence suggests that traditional format interviews are too lacking in reliability and validity to justify their use. However, the MMIs offers sufficient evidence of reliability and validity required for the medical school selection process. Medical schools differ widely in terms of length, panel composition, structure, content and scoring methods for interviews. These differences may affect the reliability and validity of interviews. In addition, it has also been demonstrated that interviewee coaching may significantly affect performance at interviews (Patterson, Knight, et al., 2016).

2.3. Review of construct-level predictive validity of aptitude tests for undergraduate medical school performance

This section of the thesis is devoted to objective 1(b) outlined in section 1.4. The goal will be the determination of predictive validity and possibly construct-level predictive validity of the aptitude tests used for medical school selection worldwide. To recap, predictive validity is concerned with the association computed between a selection measure or predictor (aptitude test in this case) and an outcome or criterion (in this case medical school exams or future medical licensing exams). As already mentioned in section 1.2 in the background chapter, within the selection context, the estimate of predictive validity is typically biased downwards in the presence of range restriction and/or measurement error. The correction for the effect of range restriction and measurement error results in construct-level predictive validity. Note that this only applies to predictive validity estimated by a correlation coefficient rather than a regression coefficient. The reason for this will be clearly demonstrated in great detail in chapter 3, for now, take that to be case.

Therefore the task at hand will be to determine instances from the literature where the predictive validity of aptitude tests are computed for medical school (or related) performance. Where the correlation coefficients are presented, corrections for range restriction or / and measurement error will constitute construct-level predictive validity. Since the construct-level predictive validity of aptitude tests for medical school performance is part of the focus of this thesis, it will be of interest to determine the extent to which the construct-level predictive validity of aptitude tests for medical school performance are reported in the literature. This will thus make it possible to ascertain the research gap that this thesis will attempt to fill with respect to construct-level predictive validity.

The review of the literature was conducted within two sources, *Google Scholar* (Google Scholar, 2017) and *Web of Science (previously known as Web of Knowledge)* (Clarivate Analytics, 2017) as summarised in Figure 2.2. Altogether, 24 papers were considered to be relevant. The categorisation of the results with regards to the predictive validity and construct-level predictive validity from the review can be found in Table 2.3. It was observed that in all of the papers reviewed, predictive validity was estimated in terms of correlation coefficients, regression coefficients and Odds Ratios.

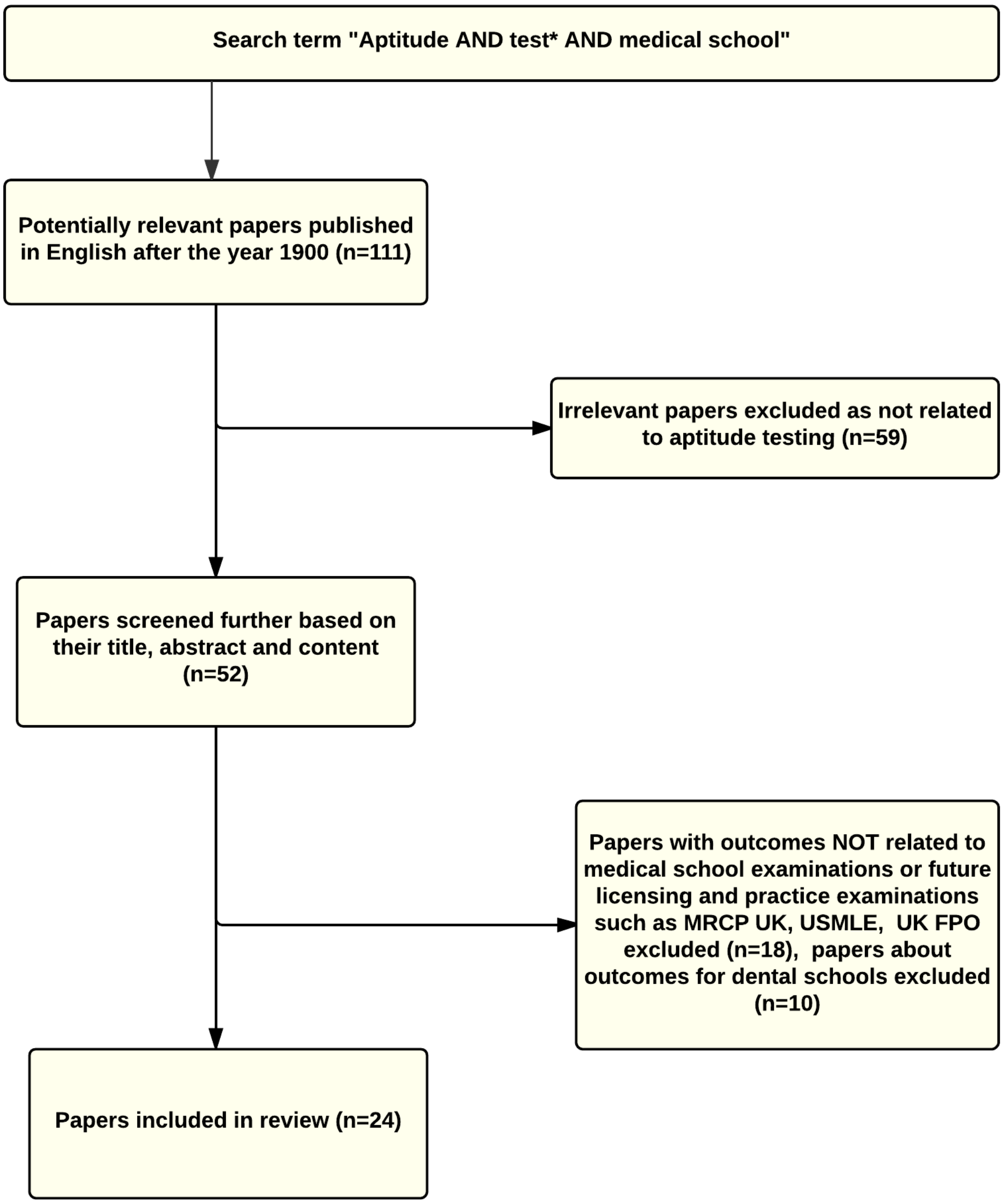

Search term "Aptitude AND test* AND medical school"
Potentially relevant papers published in English after the year 1900 (n=111)
Irrelevant papers excluded as not related to aptitude testing (n=59)
Papers screened further based on their title, abstract and content (n=52)
Papers with outcomes NOT related to medical school examinations or future licensing and practice examinations such as MRCP UK, USMLE, UK FPO excluded (n=18), papers about outcomes for dental schools excluded (n=10)
Papers included in review (n=24)

Figure 2.2.: *Flow chart of papers included in the review of construct-level predictive validity of aptitude testing for medical school outcomes. Note that the medical school related outcomes included in the flowchart in abbreviated form are Membership of the Royal Colleges of Physicians of the United Kingdom (MRCP UK), United States Medical Licensing Examination (USMLE) and United Kingdom Foundation Programme (UK FPO)*

As may be observed from the Table 2.3, estimation of the predictive validity of aptitude tests for medical school (and related) performance by means of correlation coefficients is prevalent. However, the reported correlation coefficients are biased estimates of predictive validity due to range restriction and /or measurement error. Furthermore, it was observed that attempts to correct the estimate of predictive validity for the effects of range restriction and/or measurement error are extremely rare ($\frac{4}{19} * 100 \approx 21$ % of the time when required). This means that most of estimated predictive validity estimates reported as correlation coefficients are highly likely to be biased downwards. This therefore results in underestimation of the predictive effects of aptitude tests for medical school (related) performance in most studies. This carries over to meta-analytic studies as pooled estimates of these underestimated correlation coefficients are computed.

Predictive validity			Construct-level predictive validity		
Estimated by	r_{po}	19	Corrected for	Range restriction	3
	β_{po}	3		Censoring, range restriction and measurement error	1
	Odds Ratio	1		Censoring	1
	Odds Ratio and β_{po}	1			
	Total	24		Total	5

Table 2.3.: *Summary of figures relating to predictive validity and construct-level predictive validity from the review of the literature. Note that the acronym subscript "po" represents "predictor outcome", with r and β denoting correlation and regression coefficients respectively*

.

Table 2.3 is a concise summary of the review conducted, the full details of the review for the 24 papers are included in Table 2.4 which spans 10 pages. The last column of the Table includes notes on the necessity and computation of construct-level predictive validity. To make it easy for the reader to identify the studies where construct-level predictive validity estimates are computed, the text in the last column corresponding to those studies are presented in bold.

Aptitude test (predictor/selection test)	Country	Study	Outcome (criterion)	Association	Results	Construct-level predictive validity
Aptitude Test (APT)	Saudi Arabia	*Al-Rukban et al. (2010)*	End year Grade Point Average (GPA) for students in year one to four	Correlation coefficient	r= (0.11, 0.19) with p values (0.07, 0.09) for those students with GPA (>= 3, <3) respectively	Not computed
		Albishri, Aly, and Alnemary (2012)	Performance at end of year six	Correlation coefficient	r=0.24, p value < 0.0001	Not-computed
		Al Alwan et al. (2013)	Performance in year one and two	Correlation coefficient	r= (0.80, 0.81), r= (0.79, 0.78) and r=(0.74, 0.66) for year one and two for 2007/8, 2008/9 and 2009/10 cohorts respectively. All p values > 0.05	Not-computed
Aptitude Test (APT)	Saudi Arabia	*Alhadlaq et al. (2015)*	Performance in first and second year	Regression coefficient	β=-0.019, p value > 0.010	N/A

Aptitude test (predictor/selection test)	Country	Study	Outcome (criterion)	Association	Results	Construct-level predictive validity
Test for Medical Studies (TMS)	Germany	*Kadmon and Kadmon (2016)*	Performance in pre-clinical years	Regression coefficient and Odds Ratio	β=(0.442, 0.446) for academic performance, OR (0.890, 0.853) for continuity of studies	N/A
AH5 (Group Test for High Grade Intelligence)	UK	*McManus, Woolf, et al. (2013)*	Basic Medical Sciences (BMS) for year one and two and MRCP UK	Correlation coefficient	r= 0.050±0.042 (-0.033,0.131), r= 0.120±0.062 (0.002,0.249) and r= 0.189±0.072 (0.034,0.321) for BMS outcome, MRCP UK Part I (written) and Part II (clinical) respectively	**Adjusted for censoring**

Aptitude test (predictor/selection test)	Country	Study	Outcome (criterion)	Association	Results	Construct-level predictive validity
UKCAT and AH5	UK	*McManus, Dewberry, Nicholson, Dowell, et al. (2013)*	All undergraduate, All postgraduate and All MRCP UK exams	Meta-analysis of correlation coefficients	r=0.181, n=4 studies, CI (0.055,0.302). r=0.226, n=3 studies, CI (0.108, 0.339). r=0.243, n=5 studies, CI (0.090, 0.385) for all undergraduate exams, all MRCP UK exams and postgraduate exams respectively	**Adjusted for censoring, reliability and range restriction**
UKCAT	UK	*MacKenzie, Cleland, et al. (2016)*	Total EPM and UK FPO scores	Correlation coefficients	r=(0.193, 0.253) with total EPM and UK FPO scores respectively. P values <0.05	Not computed
UKCAT	UK	*Lynch et al. (2009)*	Year one knowledge score and performance score (OSCE)	Correlation coefficient	r=(0.062, 0.036) with p values=(0.291, 0.668) for knowledge and performance scores respectively	Not computed

Aptitude test (predictor/selection test)	Country	Study	Outcome (criterion)	Association	Results	Construct-level predictive validity
		Tiffin, Mwandigha, et al. (2016)	Skills and Knowledge based exams for years one to five	Regression coefficient	Adjusted significant total UKCAT effect, β= (0.15, 0.25) and β= (0.10, 0.20) for Knowledge and skills exams for years one to five	**Adjusted for range restriction (but not needed)**
Graduate Medical School Admissions Test (GAMSAT)	Australia	*Sladek et al. (2016)*	Years one to four exams	Odds Ratios (OR)	OR 1.09 with 95% CI (1.04-1.14) for years one to two	N/A
UKCAT	UK	*Sartania et al. (2014)*	Total EPM	Regression coefficient	β= 0.216, p value 0.003 and β= 0.212, p value 0.005 for unadjusted and adjusted total UKCAT effect	N/A
		Yates and James (2013)	Weighted average of knowledge and skills exams, phase one to three	Correlation coefficient	Significant r=0.173, 0.242 and 0.193 for phase one, two and three respectively	Not computed

Aptitude test (predictor/selection test)	Country	Study	Outcome (criterion)	Association	Results	Construct-level predictive validity
(Trial) Geneva Aptitude Test	Switzerland	*Cerutti, Bernheim, and Van Gessel (2013)*	Year one exam	Correlation coefficient	r=0.47	Not computed
Undergraduate Medical and Health Sciences Admission Test (UMAT)	Australia and New Zealand	*Edwards, Friedman, and Pearce (2013)*	Years one, two and three exams at three institutions identified by A, B and C	Correlation coefficient	Total UMAT scores had significant r= (0.34, 0.41, 0.41) for years one to three at institution A, significant r=(0.26, 0.19) for years one and two in institution B and r=0.32 for year one at institution C	**Adjusted for range restriction**
MCAT	USA	*Casey et al. (2016)*	USMLE (Part I and II)	Correlation coefficient	r= 0.39 and 0.35 with USMLE Part I and II respectively with p values <0.001	Not computed

Aptitude test (predictor/selection test)	Country	Study	Outcome (criterion)	Association	Results	Construct-level predictive validity
UKCAT	UK	*Husbands, Mathieson, et al. (2014)*	Year four written and and OSCE for year four and five	Correlation coefficient	Total UKCAT scores had r=0.24, 0.36 and 0.29 respectively for years four written, years four and five for OSCE exams with all p values <0.05 (Aberdeen University). Total UKCAT scores had r=0.34 (p values <0.05), and r=0.15 (p values > 0.05) for four written and OSCE exams respectively (Dundee University)	Not computed

Aptitude test (predictor/selection test)	Country	Study	Outcome (criterion)	Association	Results	Construct-level predictive validity
UKCAT	UK	*McManus, Dewberry, Nicholson, and Dowell (2013)*	Skills and knowledge exams in year one	Correlation coefficient	Significant effect for UKCAT subtests AR, DM, QR, VR and Total UKCAT scores with r= (0.053,0.052), (0.056,0.077), (0.044,0.079), (0.028,0.177) and (0.75,160) for skills and knowledge exams respectively	Not computed
MCAT	USA	*Donnon, Paolucci, and Violato (2007)*	Preclinical Basic Science and USMLE Part I	Correlation coefficient	0.43 and 0.66 Preclinical Basic Science and USMLE Part I respectively	**Adjusted for range restriction**

Aptitude test (predictor/selection test)	Country	Study	Outcome (criterion)	Association	Results	Construct-level predictive validity
Health Professions Admission Test-Ireland (HPAT-Ireland)	Ireland	*Kelly, Regan, et al. (2013)*	Clinical, Communication and Total OSCE for year one and two. Total Multiple Choice Questions (MCQ) clinical exams for year one	Correlation coefficient	Non-significant (p value > 0.05) with r= 0.07, 0.13, 0.18 and 0.09 for clinical OSCE, communication OSCE, Total OSCE and MCQ clinical exams for year one. Significant(p value < 0.05) with r= 0.29, 0.21 and 0.28 for clinical OSCE, communication OSCE and Total OSCE for year two	Not computed

Aptitude test (predictor/selection test)	Country	Study	Outcome (criterion)	Association	Results	Construct-level predictive validity
BMAT	UK	*Emery and Bell (2009)*	Year one and two exams	Correlation coefficient	BMAT subtests aptitude and skills component had significant r= 0.13 and r=0.22 for years one and two respectively while BMAT subtest scientific knowledge had significant r= 0.26 and r=0.25 for years one and two respectively	Not computed
UKCAT	UK	*Yates and James (2010)*	Average for preclinical theme courses A (the cell), B (person), C (community), D (personal and professional development) and E (OSCE)	Correlation coefficient	r= 0.211 (p value=0.003), r= 0.126 (p value 0.078), r= 0.232 (p value 0.001), r=-0.085 (p value 0.237) and r=-0.014 (p value 0.849) theme course A, B, C D and E	Not computed

Aptitude test (predictor/selection test)	Country	Study	Outcome (criterion)	Association	Results	Construct-level predictive validity
Health Professions Admission Test-Ireland (HPAT-Ireland)	Ireland	*Kelly and O'Flynn (2017)*	OSCE for year one, two, overall OSCE and assessments for year one	Correlation coefficient	Significant r=(-0.25 and -0.27) for overall OSCE and overall assessment for year one. Significant r=(0.29, 0.21 and 0.28) for communication, clinical and year two total OSCE	Not computed
MCAT	USA	*Siu and Reiter (2009)*	Performance in year one and two	Meta-analysis correlation coefficients	Physical sciences (PS), Biological Sciences (BS), Writing Sciences (WS) and Verbal Reasoning (VR) have r=0.23, 0.32, -0.13 and 0.19 for year one respectively. Similarly r=0.36, 0.39, 0.07 and 0.27 for year two.	Not computed

Table 2.4.: *Summary of papers from the review of the construct-level predictive validity of aptitude tests used in medical school selection world wide*

2.4. Chapter summary

In this thesis chapter, results relating to the objectives 1(a) and 1(b) outlined in section 1.4 were discussed. From the literature review conducted, the predictors used for medical school selection were categorised into short-listing and final stage selection methods. Under each of the category, the individual predictors were covered in great detail. Notably, under the short-listing selection methods, aptitude tests have been demonstrated to play a major role in medical school selection. This may also be seen from the studies presented following the literature review relating to objective 1(b). There is a lot of research that has sought to quantify the predictive effects of aptitude tests for medical school (related) performance through the estimation of correlation coefficients. Unfortunately, these predictive effects are, for the most part, underestimated (downward biased) in about 4 out of 5 such research studies.

In the next chapter, several methods available in the literature to deal with the downward bias in the predictive validity in line with objective 1(c) outlined in section 1.4 will be discussed. The effects of range restriction and/or measurement error that cause this bias will not only be shown from literature but also empirically through Monte Carlo simulation. Furthermore, it will be demonstrated from literature, and also empirically through Monte Carlo simulation, that this downward bias in the predictive validity is only a problem that affects correlation coefficients rather than regression coefficients. This downward bias in the computed correlation coefficients will also be explored for a variety of selection validity designs. This will thus lay the foundation for work in chapter 5 relating to objective 2 outlined in section 1.4.

3. Dealing with bias in the estimation of predictive validity

3.1. Background and scope

So far in this thesis, the statistical scope of the thesis has been presented in chapter 1 and the review of the literature for objectives 1(a) and 1(b) outlined in section 1.4 presented in chapter 2. In this third chapter, the thesis will focus primarily on objective 1(c) outlined in section 1.4 which relates to the statistical methodologies for adjusting for the attenuation (downward bias) in the association (hereafter referred to as predictive validity) between predictors (selection test) and outcomes (criterion) in the selection context. Statistically speaking, these associations may be estimated by means of correlation coefficients, regression coefficients and odds ratios as was observed in Table 2.4. As was mentioned in the chapters 1 and 2, attenuation (downward bias) in the predictive validity between the predictor, selection (aptitude) test, and the outcome (criterion) may be due to range restriction and measurement error (Burt, 1943; Fisher, 2014; General Medical Council, 1973; McManus, Dewberry, Nicholson, Dowell, et al., 2013; Neter et al., 1996). However, it is known that range restriction causes attenuation (downward bias) in the predictive validity within the selection context only when it is estimated by correlation coefficient rather than regression coefficients and odds ratios (Bengt O. Muthén and Asparouhov, 2016, pp 443-445; Fife, Mendoza, and Terry, 2013). The statistical focus of this thesis is range restriction (and not measurement error). Therefore, this chapter will from this point deal primarily with the effect of range restriction on the correlation coefficient and the statistical methodologies that are available to correct the computed correlation coefficients to achieve construct-level predictive validity. The effect of measurement error on correlation coefficients is however acknowledged (and will be covered briefly in this chapter) as a contributing factor to attenuation (downward bias) of predictive validity. A detailed coverage on a variety of statistical methods for dealing with measurement error to achieve construct-level predictive validity are presented in section 8.1 of the Technical Appendices for the interested reader.

The effects of range restriction and measurement error on the predictive validity estimated by correlation coefficients are clearly documented in literature. Their effects can also be demonstrated empirically by means of Monte Carlo simulations. At this point, the effect of measurement error on the computed correlation is demonstrated first. Assume that a predictor (selection test) is denoted by x and that an outcome (criterion) is denoted by y. Assume further that any observed value for the predictor (selection test), x, can be decomposed to its true unobserved score t_x and some amount of measurement error ε_x as shown in equation 3.1.1. It follows that an observed score of the outcome (criterion), y, can also be decomposed to its true unobserved score t_y and some amount of measurement error ε_y as shown on 3.1.2. It is also assumed that the measurement error of one predictor (selection test), say ε_x, is uncorrelated with true score of another predictor (selection test), say t_y, and that ε_x and ε_y are independent (Charles, 2005).

$$x = t_x + \varepsilon_x \tag{3.1.1}$$

likewise,

$$y = t_y + \varepsilon_y \tag{3.1.2}$$

Note that $E(x) = t_x, \quad E(y) = t_y$ since $\varepsilon_x \sim N(0, \sigma^2_{\varepsilon_x}), \quad \varepsilon_y \sim N(0, \sigma^2_{\varepsilon_y})$ and

$$\sigma^2_x = \sigma^2_{t_x} + \sigma^2_{\varepsilon_x} + 2\sigma_{t_x\varepsilon_x} \tag{3.1.3}$$

Under the strict assumption of independence between true value t_x and measurement error ε_x,

$\sigma_{t_x\varepsilon_x} = 0$, therefore

$$\sigma_x^2 = \sigma_{t_x}^2 + \sigma_{\varepsilon_x}^2 \tag{3.1.4}$$

similarly

$$\sigma_y^2 = \sigma_{t_y}^2 + \sigma_{\varepsilon_y}^2 \tag{3.1.5}$$

If the variables, x and y, are continuous and a linear relationship assumed, then *Pearson product- moment correlation coefficient* or simply *Pearson correlation coefficient* may be computed by the formula

$$r_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{3.1.6}$$

where r_{xy} , σ_{xy}, σ_x and σ_y represent the observed correlation, covariance between x and y and standard deviation for x and y respectively. To demonstrate the effect of measurement error on the Pearson correlation coefficient, assume data devoid of measurement error generated from a standardised bivariate normal distribution shown in equation 3.1.7. The covariance is randomly selected from a uniform distribution. $\sigma_{t_x t_y} \sim U[0.1, 0.9]$

$$\begin{pmatrix} t_x \\ t_y \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{t_x t_y} \\ \sigma_{t_x t_y} & 1 \end{pmatrix}\right] \tag{3.1.7}$$

To introduce data affected with measurement error, consider the equations 3.1.1 and 3.1.2 where it is assumed that there are varying but increasing levels of measurement error, ε_x and ε_y, such that $\varepsilon_x \sim N(0, \sigma_{\varepsilon_x}^2)$, $\varepsilon_y \sim N(0, \sigma_{\varepsilon_y}^2)$ have pairs of $(\sigma_{\varepsilon_x}^2, \sigma_{\varepsilon_y}^2) = \{(0.2, 0.2), (0.4, 0.4), (0.6, 0.6), (0.8, 0.8)\}$. To assess the effect of measurement error on the computed Pearson correlation coefficient in equation 3.1.6, interest will be the bias estimated as the difference between r_{xy} and $r_{t_x t_y}$. Note that $r_{t_x t_y}$ denotes the Pearson correlation coefficient between the perfectly measured scores of the predictor (selection test) and outcome (criterion). The term r_{xy} denotes the

Pearson correlation coefficient between the imperfectly measured scores of the predictor (selection test) and outcome (criterion). Since the bias is a random variable, it is recommended that it be estimated over several samples (chosen arbitrarily to be 5,000 samples in this case) by computing its expected (average) value. Figure 3.1 shows the expected (average) bias from the simulations, it was observed that as documented in literature (Muchinsky, 1996; Van Iddekinge and Ployhart, 2008; Viswesvaran et al., 2014; Wang, 2010), measurement error does induce attenuation (downward bias) on the Pearson correlation coefficient. The magnitude of this attenuation (downward bias) worsens with increasing magnitude of the measurement error between the predictor (selection test) and outcome (criterion). Several statistical methodologies have been developed to counteract the effects of measurement error, these methodologies are detailed in section 8.1 of the Technical Appendices for the interested reader.

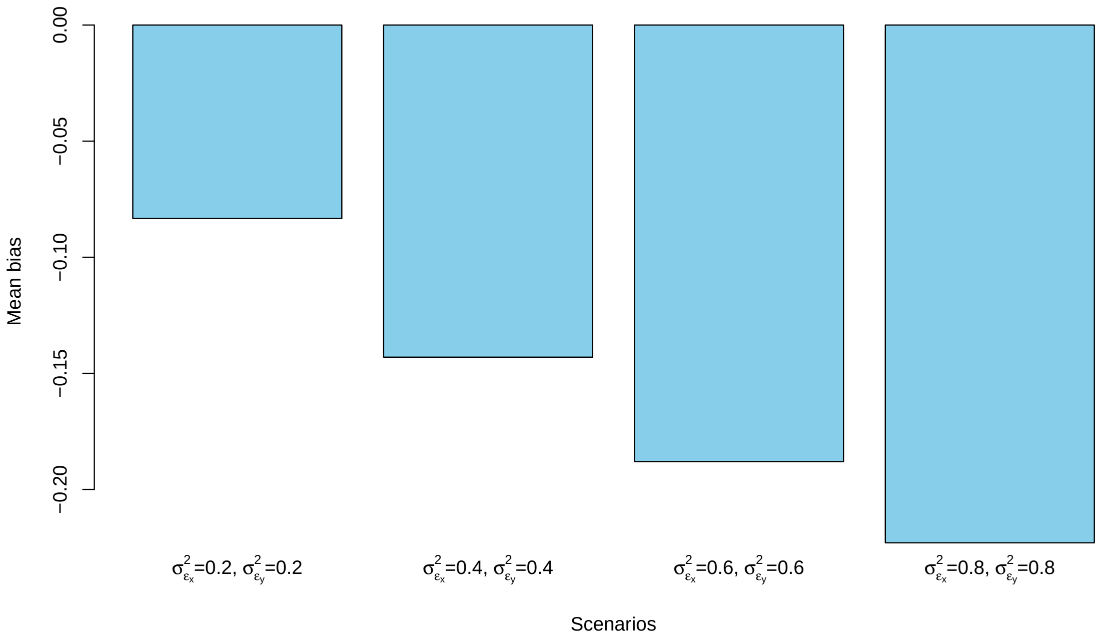


Figure 3.1.: *The attenuating effect of measurement error on the Pearson correlation coefficient between a simulated predictor (selection test), x, and a simulated outcome (criterion), y.*

With regard to range restriction, it is possible to show its effects on the Pearson correlation and the regression coefficients, the two estimands used to measure predictive validity as reported in Table 2.4. For the Pearson correlation coefficient, data were simulated from the standardised bivariate normal distribution in equation 3.1.8.

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & r^u_{xy} \\ r^u_{xy} & 1 \end{pmatrix}\right] \tag{3.1.8}$$

The true correlation between the predictor (selection test), x, and the outcome (criterion) y is denoted by r^u_{xy} with the superscript *"u"* indicating that the correlation is unrestricted. This is the Pearson correlation coefficient that is untainted by range restriction and is randomly drawn from the uniform distribution $r^u_{xy} \sim U[0.1, 0.9]$. To determine the effect of range restriction, the selection of entrants (which introduces range restriction) is then simulated by assuming the data consists of all applicants with scores on the predictor (selection test), x. Based on their scores, selection ratios of 20%, 40%, 60% and 80% (alternatively 0.2, 0.4, 0.6 and 0.8) are simulated. This means that for example for selection ratio 0.2, only those applicants in the top 20 % of the score distribution for predictor (selection test) x are selected. The Pearson correlation coefficient based on the selected (top 20 % applicants) is then computed between scores on their predictor (selection test), x, and scores on their corresponding outcome (criterion) y.

To assess the effect of range restriction on the computed Pearson correlation coefficient, two performance measures, bias and precision are used. Bias measures how much the computed Pearson correlation coefficient, r^r_{xy}, (superscript *"r"* indicates correlation is restricted due to selection) deviates from the known true correlation coefficient, r^u_{xy} which is unaffected by range restriction. On the other hand, precision measures the uncertainty associated with the estimate of r^r_{xy}. Since both bias and precision are random variables, it is recommended that they be evaluated over many samples (chosen arbitrarily to be 5,000 in this case). Therefore equation 3.1.8 was used to generate 5,000 samples, for each sample, selection ratios of 0.2, 0.4, 0.6 and 0.8 were simulated. The bias and precision across all of the samples for each of the selection ratios were evaluated by mean bias and Root Mean Square Error (RMSE) using the formulas in equation 3.1.9 and 3.1.10 respectively. Note that n_{sims} denotes the number of simulated samples and that smaller values of $\widehat{Meanbias}$ and $\widehat{RMSE}$ are preferred. This is because they signify little (if any) deviation from the known true correlation and little (if any) uncertainty associated with the estimate of the correlation coefficient.

$$\widehat{Meanbias} = \frac{1}{n_{sims}} \sum_{i=1}^{n_{sims}} (r^r_{xy_{(i)}} - r^u_{xy_{(i)}}) \tag{3.1.9}$$

$$\widehat{RMSE} = \sqrt{\frac{1}{n_{sims}} \sum_{i=1}^{n_{sims}} (r^r_{xy_{(i)}} - r^u_{xy_{(i)}})^2} \tag{3.1.10}$$

Figures 3.2 and 3.3 show the results of the simulations of the effect of range restriction on the Pearson correlation coefficient. As may be observed in the Figures, the stricter the selection process (lower the selection ratios), the higher the attenuation (downward bias) of the computed Pearson correlation coefficient. This is accompanied by a loss of precision which worsens with strictness of the selection process.

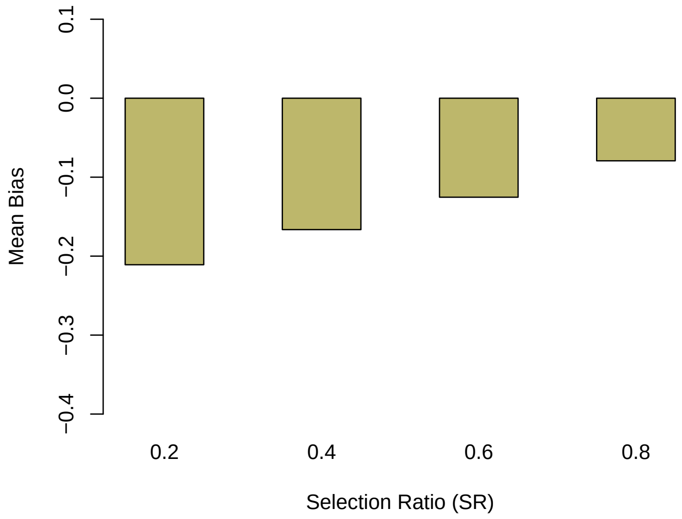


Figure 3.2.: *The attenuating (downward bias) effect of range restriction on the Pearson correlation coefficient between a simulated predictor (selection test), x, and a simulated outcome (criterion), y.*

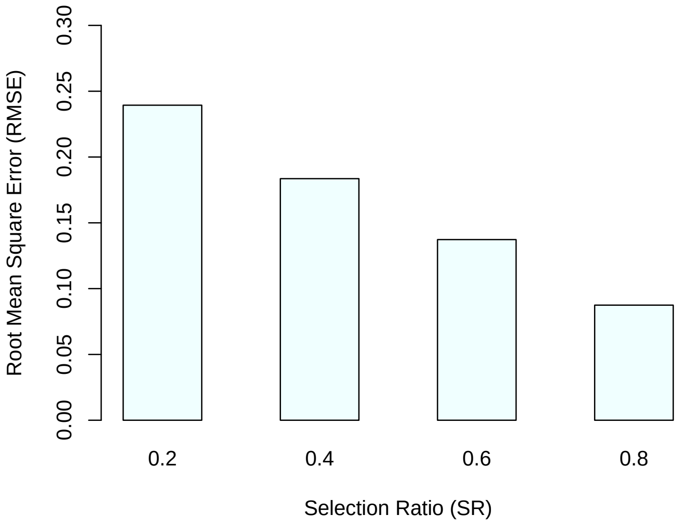


Figure 3.3.: *The effect of range restriction on the precision of the estimate for the Pearson correlation coefficient between a simulated predictor (selection test), x, and a simulated outcome (criterion), y.*

For regression coefficients, the effect of range restriction may be empirically demonstrated using Monte Carlo simulation. Assume that there are two predictors (potential selection tests) z and x but selection is based only on z. This then means that z induces direct range restriction and that x is another predictor (correlated somewhat with z but not used for selection). Further, as before, y is an outcome (criterion) of interest. Based on these assumptions, data were generated from the standardised trivariate normal distribution in equation 3.1.11. The elements r^u_{xy}, r^u_{zy} and r^u_{zx} are randomly drawn from the uniform distribution $r^u_{xy}, r^u_{zy} \sim U[0.1, 0.9]$ and $r^u_{zx} \sim U[0.1, 0.2]$. The subscript *"u"* indicate that the correlation coefficients are unrestricted. The direct predictor, z, and predictor, x, are assumed to have low (to modest) positive correlation (covariance) so as to prevent the generated data from suffering from inter-correlation (multicollinearity) problems.

$$\begin{pmatrix} z \\ x \\ y \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & r^u_{zx} & r^u_{zy} \\ r^u_{zx} & 1 & r^u_{xy} \\ r^u_{zy} & r^u_{xy} & 1 \end{pmatrix} \right] \qquad (3.1.11)$$

To assess the effect of range restriction on the predictive validity estimated by regression coefficients, the performance measures of bias and precision will be used under two modelling scenarios. In the first scenario, interest will constitute regressing the outcome (criterion), y, on the predictor, x, to obtain $\beta^r_{y|x}$. In the second scenario, interest will constitute controlling for both the direct predictor (selection test), z, and the predictor, x, to obtain $\beta^r_{y|x,z}$. Note that the restricted regression coefficients $\beta^r_{y|x}$ and $\beta^r_{y|x,z}$ will be evaluated against their unrestricted counterparts, that is $\beta^u_{y|x}$ and $\beta^u_{y|x,z}$. Therefore the subject of interest will be the evaluation of whether the effect of range restriction on the predictive validity estimated by regression coefficients is potentially moderated by the choice of predictors in the regression model. Since both bias and precision are random variables, it is recommended that they be evaluated over many samples (chosen arbitrarily to be 5,000 in this case). To induce range restriction, the selection process is simulated by applying a selection ratio of 0.2, 0.4, 0.6 and 0.8.

Figure 3.4 shows the results of the simulations of the effect of range restriction on the regression coefficient, as may be observed in the Figure, exclusion of the direct predictor (selection test), z, from the regression model of y on x, results in $\beta_{y|x}$ that is amplified (biased upwards) and imprecise. The magnitude of the bias and imprecision worsens with increase in the strictness of the selection process (decrease in selection ratios). When both the direct and indirect predictor (selection tests) are included in the model, this results in $\beta_{y|x,z}$ that has negligible bias and is highly precise. The reason for this is that, even in the presence of range restriction due to selection based on z, the assumption underlying regression is that the marginal distribution of $y|z$ is unaffected even when the distribution of z is curtailed by range restriction. This curtailment of the distribution of z would however lead to a biased (Pearson) correlation coefficient,

r_{yz}, because the underlying assumption of the estimand includes bivariate normality of y and z (whose distribution is affected by range restriction) (Bengt O. Muthén and Asparouhov, 2016, pp 443-445).

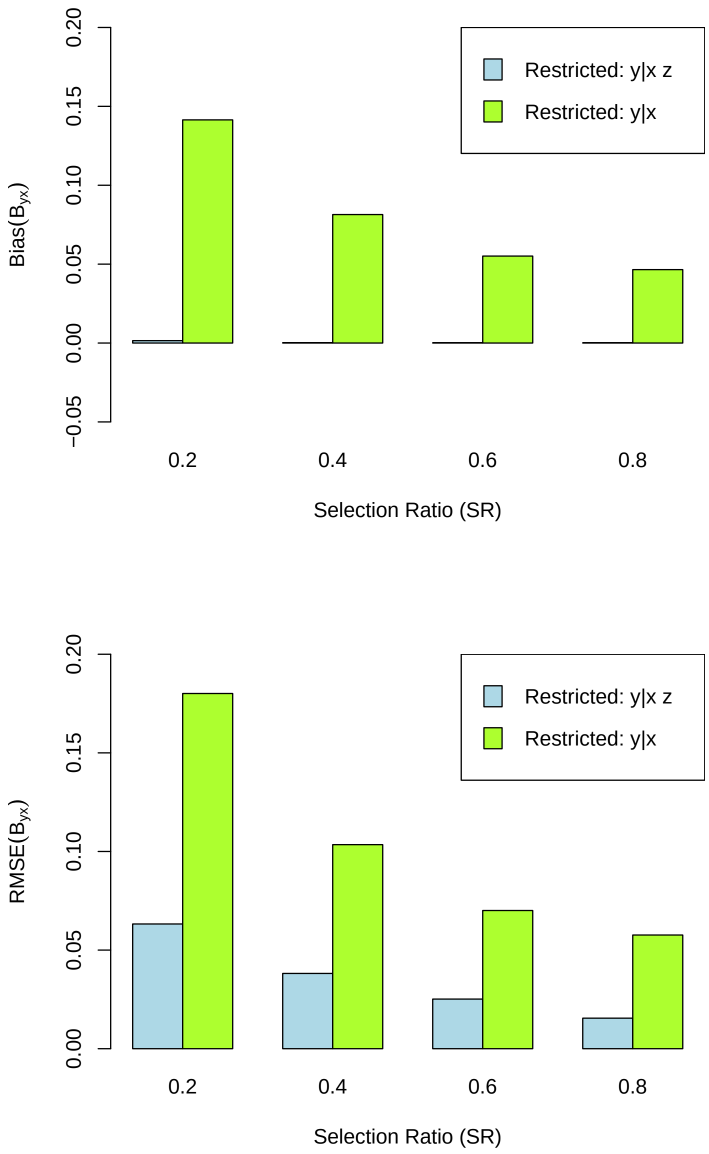


Figure 3.4.: *Mean bias and RMSE for regression coefficient for indirect range restriction induced by selection on variable z for selection ratio 0.2, 0.4, 0.6 and 0.8.*

The results of the simulations confirm what is known in the literature. That is, the effect of range restriction on predictive validity is problematic in instances where it is estimated by Pearson correlation coefficients. When predictive validity is to be estimated by regression coefficients, the effect of range restriction is problematic only when the direct predictor (selection test) is excluded from the regression model (Behseta et al., 2009; Dunlap and Cureton, 1930; Huitema and Stein, 1993; Mendoza and Mumford, 1987a). These findings were summarised by *Mendoza, Bard, et al. (2004)* in their paper *"Criterion-Related Validity in Multiple-Hurdle Designs: Estimation and Bias"* in which the effects of range restriction on both the Pearson correlation and regression coefficient were evaluated under three selection designs. The designs were the *single hurdle concurrent validity, predictive validity* and *two hurdle validity* selection designs. These designs are described next with the aid of Figures 3.5, 3.6 and 3.7.

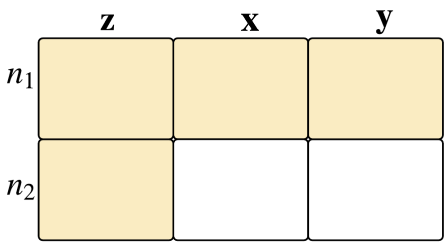


Figure 3.5.: *Conceptual diagram of data structure for the single hurdle concurrent validity design showing observed data in (orange) and missing data (white) for the predictors and criterion variables*

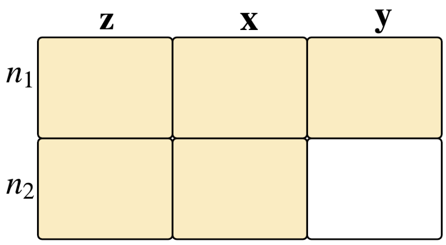


Figure 3.6.: *Conceptual diagram of data structure for the predictive validity design showing observed data in (orange) and missing data (white) for the direct predictor and criterion variables*

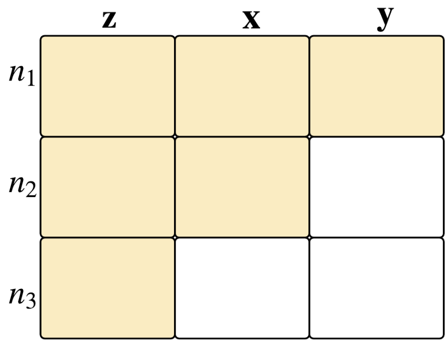


Figure 3.7.: *Conceptual diagram of data structure for the two hurdle validity design showing observed data in (orange) and missing data (white) for the direct predictors and criterion variables*

For the *single hurdle concurrent validity* selection design, take the the total number of applicants to be $(n_1 + n_2)$. The entrants, n_1, are selected on the basis of a predictor (selection test) z which has full information available for all the $(n_1 + n_2)$ applicants. Partial information is available only for n_1 entrants for predictor, x, and outcome (criterion) y. For the *predictive validity* selection design, selection is also based on direct predictor (selection test), z, which has full information available for the direct predictor, z, and predictor, x, for all the applicants taken to be $(n_1 + n_2)$ in total. Partial information is available only for the n_1 entrants for the outcome (criterion) y. This is a result of applicants sitting for two selection tests z and x but selection being based only on z. Lastly, for the *two hurdle validity* selection design, the selection is based on a two stage process, firstly based on the predictor (selection test) z, and secondly at a subsequent time point based on another predictor (selection test) x. For this design, take the total number of applicants to be $(n_1 + n_2 + n_3)$ with full information available only for the first predictor (selection test) z.

The bias related to predictive validity estimated by (bivariate) Pearson correlation and regression coefficients for the three selection designs is presented in Tables 3.1 and 3.2. The results affirm the results of the Monte Carlo simulation that range restriction always leads to biased estimates of predictive validity if it is estimated by a bivariate Pearson correlation coefficient. When predictive validity is estimated by a regression coefficient, range restriction leads to bias only when the predictor (selection test) on which the selection is conducted is omitted from the regression model. Therefore, with respect to predictive validity and construct-level predictive validity, for the rest of this thesis, focus will be devoted to the (bivariate) Pearson correlation coefficient.

Association	Regression (β) coefficient	Bivariate correlation coefficient	Sample size used
y with (x,z)	unbiased	biased	n_1
y with z	unbiased	biased	n_1
y with x	biased	biased	n_1
x with z	unbiased	biased	n_1
x with z	**unbiased**	**unbiased**	$\mathbf{n_1 + n_2}$

Table 3.1.: *Bias in regression and bivariate Pearson correlation coefficients for both the concurrent single hurdle concurrent validity design and predictive validity selection design. Note that the last appearing in bold only applies to the predictive validity selection design*

Association	Regression (β) coefficient	Bivariate correlation coefficient	Sample size used
y with (x, z)	unbiased	biased	n_1
y with z	biased	biased	n_1
y with x	biased	biased	n_1
x with z	biased	biased	n_1
x with z	unbiased	biased	$n_1 + n_2$

Table 3.2.: *Bias in regression and bivariate Pearson correlation coefficients for the two hurdle validity selection design* [1]

The contents of this thesis chapter so far lay the foundation for conducting a literature review in accordance to objective 1(c) found in section 1.4. This literature review conducted within two sources, *Google Scholar* (Google Scholar, 2017) and *Web of Science (previously known as Web of Knowledge)* (Clarivate Analytics, 2017) is summarised in Figure 3.8 which includes the inclusion and exclusion criteria. The focus of this thesis chapter is range restriction and

[1] Tables adapted from *"Criterion-Related Validity in Multiple-Hurdle Designs: Estimation and Bias"* by *Mendoza, Bard, et al. (2004)*

its effect on predictive validity estimated by (Pearson) correlation coefficients. Therefore, only those results of the review related to range restriction and statistical methods for correction for the effect of range restriction are discussed next (section 3.2). For the results related to attenuated correlation coefficients due to measurement error and statistical methods for correction for the effect of measurement error are discussed in section 8.1 of the Technical Appendices for the interested reader.

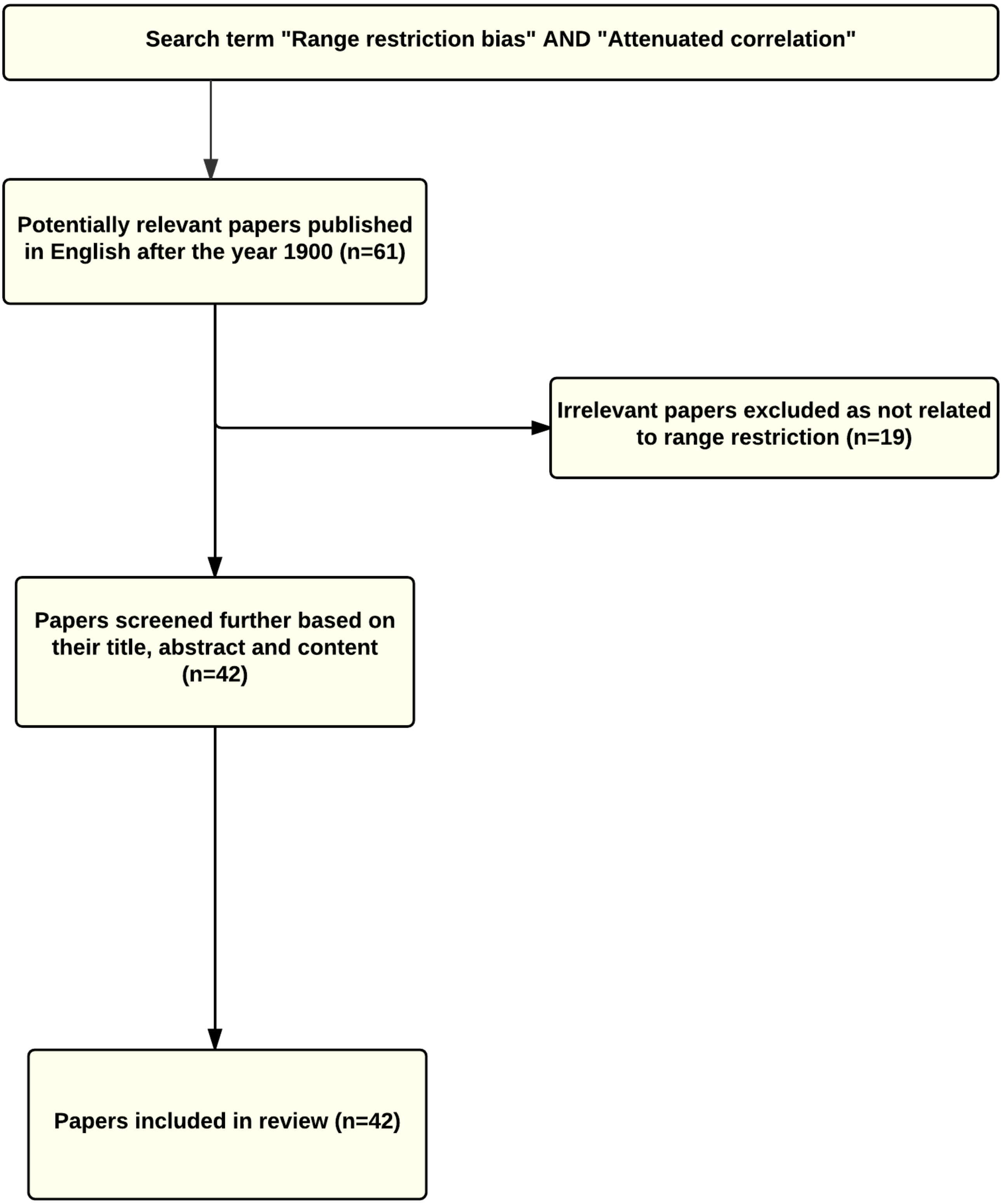

Search term "Range restriction bias" AND "Attenuated correlation"
Potentially relevant papers published in English after the year 1900 (n=61)
Irrelevant papers excluded as not related to range restriction (n=19)
Papers screened further based on their title, abstract and content (n=42)
Papers included in review (n=42)

Figure 3.8.: *Flow chart of papers included in the review of the effects of range restriction on the (Pearson) correlation coefficient*

3.2. Statistical methods for correcting for the bias in the (Pearson) correlation coefficient due to range restriction

Before the presentation of the results for the literature review in Figure 3.8, it is important to emphasise the notation that will be used for the formulas and their underlying assumptions. Variables x and z denote the predictors (selection tests). Instances where either or both are used in the selection process will be explicitly defined. The notation for the outcome (criterion) of interest is y. It is assumed that x, z and y are observed (manifest), continuous and that interest is the bivariate correlation coefficient between either one of the predictor (or selection test), x or z, and the outcome (criterion) y. Further it is assumed that the variables to be used for the computation of the bivariate correlation coefficient are bivariate normal, linearly related and homoscedastic (variables have common error variance). These underlying assumptions support the computation of the Pearson correlation coefficient (Held and Foley, 1994; InfluentialPoints, 2017; Laerd Statistics, 2013a,b).

Next, the notations for the formulas to be presented are explained. For the purpose of demonstration, assume that interest is the estimation of the predictive validity by Pearson correlation coefficient between the predictor (selection test), x, and outcome (criterion), y, in a selection context, r^r_{xy}. The superscript “r” indicates that correlation coefficient is restricted. This means that it suffers from the effects of range restriction. It has been shown both from literature and empirically in this thesis chapter that r^r_{xy} would be attenuated (biased downwards). Attempts to correct for this downward bias would result in construct-level predictive validity which is denoted by r^c_{xy}. The superscript “c” indicates that a correction for the effects of range restriction has been applied on r^r_{xy} (Huffcutt, Culbertson, and Weyhrauch, 2014; Hunter, Schmidt, and Le, 2006; Le and Schmidt, 2006). The assumptions underlying r^c_{xy} are linear relationship between the predictor (selection test), x and outcome (criterion), y, and that both variables are homoscedastic (Culpepper, 2015; Gross and Fleischman, 1983; Gross and Fleischman, 1987; Holmes, 1990). The extent to which the violations of these assumptions may adversely affect inference depend on the severity of range restriction observed (Greener and Osburn, 1979, 1980).

Although the computation of the standard error for the estimate of the construct-level predictive validity, which is here denoted $SE(r^c_{xy})$, is beyond the scope of this thesis, some statistical methods for its computation are presented briefly for the interested reader. One of the simplest and earliest methods shown by equation 3.2.1 involve making use of the restricted sample size (restricted here means affected by range restriction, that is sample after selection) used in the computation of r^r_{xy} denoted by n_r. Other proposals have included formulas that utilise both the attenuated correlation, r^r_{xy}, and the correlation corrected for range restriction, r^c_{xy}, such as equations 3.2.2, 3.2.3 and 3.2.4 (Duan and Dunlap, 1997). Note that $U_x = \frac{1}{u_x}$ where $u_x = sd_x/SD_x$ represents the ratio of the standard deviation of x from the restricted sample to that of the standard deviation of x from the unrestricted sample x (unrestricted here means not affected by range restriction, that is sample before selection). In recent times, the *Fisher's Z transformation* (Mendoza, 1993), resampling techniques such as the *jackknife* and *bootstrap* (see sections 8.3.3 and 8.4 of the Technical Appendices for detailed explanation of these two resampling techniques) have emerged as promising alternative methods for the computation of $SE(r^c_{xy})$(Allen and Dunbar, 1990; Li, Chan, and Cui, 2011; Padilla and Veprinsky, 2014; Rogers, 1976).

$$SE(r^c_{xy}) = \frac{1-(r^r_{xy})^2}{\sqrt{n_r - 1}} \tag{3.2.1}$$

$$SE(r^c_{xy}) = \frac{1-(r^r_{xy})^2}{\sqrt{n_r - 2}} \frac{r^c_{xy}(1-(r^c_{xy})^2)}{r^r_{xy}(1-(r^r_{xy})^2)} \tag{3.2.2}$$

$$SE(r^c_{xy}) = \frac{r^c_{xy}(1-(r^c_{xy})^2)}{r^r_{xy}(1-(r^r_{xy})^2)(n_r - 1)} \tag{3.2.3}$$

$$SE(r^c_{xy}) = \frac{U_x(1-(r^r_{xy})^2)}{\sqrt{n_r(1-(r^r_{xy})^2 + U_x^2(r^r_{xy})^2)^3}} \tag{3.2.4}$$

At this point, care must be taken to differentiate between two kinds of range restriction, Direct Range Restriction (DRR) and Indirect Range Restriction (IRR). Suppose that at point of application to a particular medical school in the UK, applicants have their (predictor) scores on (predicted) A-level grade and the UKCAT considered for selection. Note that the UKCAT score considered at point of selection for an applicant may be derived from any of the four cognitive based UKCAT sub-tests or sum of scores of the four sub-tests as described in section 1.1. Further, suppose that the medical school is interested in examining the association between the applicants' UKCAT scores and their corresponding future performance in *knowledge-based* exams at the end of the first year medical school training (i.e. the UKCAT's predictive validity for first year *knowledge-based* exams). If the applicants are selected exclusively based on their UKCAT scores then the *knowledge-based* exams scores would not be observed for those applicants below a specified UKCAT score threshold. In this case, the predictor whose predictive validity is to be examined, plays a direct role in determining who are selected (entrants) into the medical school. Therefore, the range restriction induced by the UKCAT score is said to be *direct* thus the term used is DRR. On the other hand, if selection into the medical school was exclusively based on applicants (predicted) A-level grade results, those selected (entrants) would not be determined by the UKCAT score, the predictor whose predictive validity is to be examined. In this case, the range restriction induced with respect to the UKCAT score would not be direct, it can therefore be said to be *indirect*, thus the use of the term for range restriction is IRR. In practice however, range restriction is never singly *direct* or *indirect* but rather a combination of the two as selection is often a multi-hurdle process involving multiple predictors (as depicted by the *two hurdle validity selection design* in Figure 3.7). For example, in the UK, this may be implemented by selecting applicants based on their (predicted) A-level grade results and then further whittling down the numbers of potential entrants by examining their UKCAT and Multiple Mini Interviews (MMIs) scores. In the next section, different statistical methods that correct predictive validity estimates to achieve construct-level predictive validity are discussed for different selection validity designs.

3.2.1. Thorndike Case I formula

As already demonstrated in Figure 3.2, selection introduces range restriction which leads to attenuated (downward bias) of the predictive validity estimated by the Pearson correlation coefficient. This is due to the fact that whilst the predictor (selection test) is administered on an entire pool of applicants, validation of the selection test by use of predictive validity is applied to only those who have been selected. This is because the outcome (criterion) measure is only observed for the entrants rather than all applicants. When selection is based on a predictor (selection test), x, and y is the outcome (criterion). The correction for the effects of range restriction to achieve construct-level predictive validity denoted by r^c_{xy} may be done using the *Thorndike Case I formula* shown in equation 3.2.5 as proposed by *Thorndike (1949).* Note that $U_y = \frac{1}{u_y}$ where $u_y = sd_y/SD_y$ represents the ratio of the standard deviation of the restricted sample y (restricted here means affected by range restriction, that is sample after selection) to that of the unrestricted sample y (unrestricted here means not affected by range restriction, that is sample before selection).

$$r^c_{xy} = \sqrt{(1 - u^2_y(1 - (r^r_{xy})^2))} \tag{3.2.5}$$

Note that equation 3.2.5 makes use of the quantity u_y which is not possible to obtain. This is because its computation would need the unrestricted sample for the outcome (criterion) y. The outcome (criterion) scores are only observed for those selected (that is the restricted sample) and can never be observed for the rejected applicants (Duan and Dunlap, 1997; Saupe and Eimers, 2010). For this reason, the *Thorndike Case I formula* has no practical application when selection (as is often the case) is based on the predictor (selection test). However, in instances where selection is based on the outcome (criterion), *Thorndike Case I formula* may be used. For example, suppose there is interest in the estimation of the predictive validity of *year one knowledge based examinations* for *year two knowledge based examinations* for undergraduate medical school entrants. If the data for *year two knowledge based examinations* and corresponding *year one knowledge based examinations* were made available only for those entrants who were classified as having passed *year two knowledge based examinations*, then the estimate of predictive validity would be biased downwards. However, If data for *year two knowledge based examinations* for those who had also failed was to be obtained, then the term $u_y = sd_y/SD_y$ would be easily obtained thus facilitating the use of *Thorndike Case I formula*.

3.2.2. Thorndike Case II formula

The *Thorndike Case II formula* also proposed by *Thorndike (1949)*, was intended to be used to compensate for the drawbacks of *Thorndike Case I formula*. Since the term $u_y = sd_y/SD_y$ is not possible to determine, an alternative $u_x = sd_x/SD_x$ is used. This means that for the *Thorndike Case II formula*, the predictor (selection test) is of crucial importance. Note that the problem encountered when attempting to make use of *Thorndike Case I formula* disappears completely as information on the unrestricted sample (before selection) and restricted sample (after selection) for the predictor (selection test) is always available through all the applicants and entrants respectively. The *Thorndike Case II formula* shown in equation 3.2.6 is thus used to obtain construct-level predictive validity, r^c_{xy}, where selection is based on a predictor (selection test), x, and y is the outcome (criterion). Note that $U_x = \frac{1}{u_x}$ which is $u_x = sd_x/SD_x$, this represents the ratio of the standard deviation of the restricted sample x to that of the unrestricted sample x (Alliger, 1987; Pfaffel, Kollmayer, et al., 2016; Saupe and Eimers, 2010; Wiberg and Sundström, 2009).

$$r^c_{xy} = \frac{U_x r^r_{xy}}{\sqrt{1 + (U^2_x - 1)(r^r_{xy})^2}} \tag{3.2.6}$$

3.2.3. Thorndike Case III formula

The *Thorndike case I* and *Thorndike case II formulas* are proposed for use in instances of Direct Range Restriction (DRR), an example of DRR would be where interest is the estimation of predictive validity, when selection is based on the predictor (selection test), x, and the outcome (criterion) of interest is y. Thus the predictor (selection test) upon which selection is based, is featured in the computation of the bivariate correlation coefficient. However, consider a scenario in which selection is based on a predictor (selection test), z, while interest is the estimation of predictive validity of x another predictor (correlated with z) for the outcome (criterion) of interest y. This example would constitute an instance of Indirect Range Restriction (IRR), the use of *Thorndike case II formula* to achieve construct-level predictive validity, r^c_{xy}, in this instance would lead to an under correction. That is r^c_{xy} would still be attenuated (downward biased) (Schmidt, Oh, and Le, 2006; Sjöberg et al., 2012). Therefore, the *Thorndike case III formula* also proposed by *Thorndike (1949)* as shown in 3.2.7 is used to achieve construct-level predictive validity, r^c_{xy}, in situations of IRR. Note that $U_z = \frac{1}{u_z}$ where $u_z = sd_z/SD_z$ is the ratio between standard deviation of the restricted sample z to that of the unrestricted sample z as a result of explicit selection on z (Alliger, 1987; Li, Chan, and Cui, 2011; Pfaffel, Kollmayer, et al., 2016; Saupe and Eimers, 2010).

$$r^c_{xy} = \frac{r^r_{xy} + r^r_{xz} r^r_{yz} (U_z^2 - 1)}{\sqrt{(1 + (U_z^2 - 1)(r^r_{xz})^2)}\sqrt{(1 + (U_z^2 - 1)(r^r_{yz})^2)}} \tag{3.2.7}$$

3.2.4. Method of Hunter, Schmidt, & Le (Thorndike Case IV formula)

A practical challenge in the implementation of the *Thorndike Case I*, *Thorndike Case II* and *Thorndike Case III* formulas is that standard deviation from the unrestricted sample required to use them may not be available (Alexander, Alliger, and Hanges, 1984; Alexander, Hanges, and Alliger, 1985; Fife, Mendoza, and Terry, 2013; Held and Foley, 1994). For this reason, *Sackett and Yang (2000)* proposed an expanded classification system in which additional range restriction scenarios were evaluated, they comprised of (a) whether selection is based on either predictor (selection tests), x or z, or outcome (criterion) y (b) whether the required standard deviation from the unrestricted sample for the selection variable is known (c) whether the predictor (selection test) z if involved in the selection process is measured or not. From these scenarios, the one which has seen a lot of research focus is the case of IRR where the predictor (selection test) z is unmeasured but the standard error of x from the unrestricted sample is known (or unknown but can be accurately estimated in some way) (Li, 2015). This scenario is an extension of *Thorndike Case III* and is thus sometimes refereed to as *Thorndike Case IV*. A method for dealing with range restriction in this scenario was proposed by *Hunter, Schmidt, and Le (2006)*. The proposed method was more realistic as it may also correct the bias in predictive validity due to range restriction in instances where z is not a single predictor (selection test) but a set of predictors (selection tests) used in combination to make selection decisions.

The method is explained by aid of Figure 3.9, assume that selection is based on a predictor (selection test) z (denoted as construct named as suitability S) which is composite of several unobserved predictors and that r^r_{yz} is not allowed to vary independently from r^r_{xy}. Instead S is allowed to induce range restriction on true scores of x, t_x, but exhibit no direct restriction on true scores of y, t_y, except through t_x (Li, Chan, and Cui, 2011).

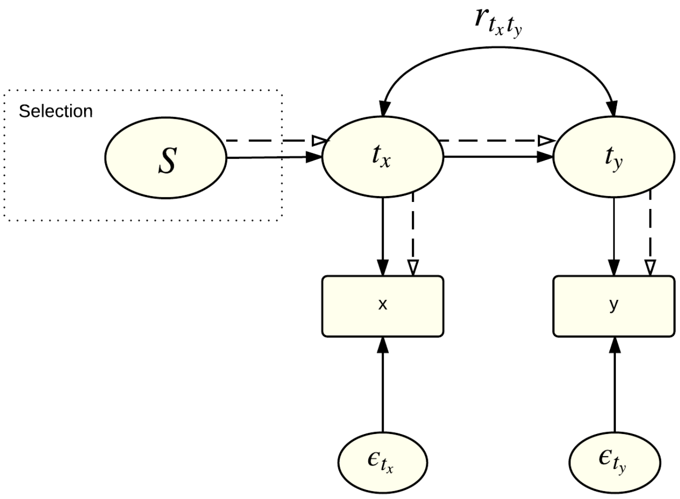


Figure 3.9.: *Model of method of Hunter, Schmidt, & Le (Case IV) for correcting range restriction, unbroken arrow denotes direction of structural relationship and broken arrows denotes direction of range restriction* [2]

The correction accuracy of this method depends on correct sequencing, first for correcting for the effects of measurement error then followed by correcting for the effects of range restriction (Hunter, Schmidt, and Le, 2006; Le and Schmidt, 2006). The formula of *Hunter, Schmidt, & Le* for correcting for the effect of range restriction is shown in equation 3.2.11 and is implemented

[2] Figure adapted from *"Correcting for Indirect Range Restriction in Meta-Analysis: Testing a New Meta-Analytic Procedure (Le and Schmidt, 2006)"* and constructed in LucidChart ®, *For more details, visit : www.lucidchart.com*

in three steps. The first step involves equation 3.2.8 where the term r_{xx_u} is obtained from 3.2.9. Recall that the subscript "u" and "r" denote unrestricted and restricted samples respectively. The terms r_{xx_u}, r_{xx_r} and u_x are the reliability estimates of the predictor (selection test) x from unrestricted sample, reliability estimates of the predictor (selection test) x from restricted sample and ratio of the standard deviation of the restricted sample x to that of the unrestricted sample x respectively. The second step involves correcting for the effect of measurement error using equation 3.2.10. In third step and final step, the effect of range restriction is addressed by applying the formula 3.2.11.

$$u_t = \sqrt{\frac{u_x^2 - (1 - r_{xx_u})}{r_{xx_u}}} \tag{3.2.8}$$

$$r_{xx_u} = 1 - u_x^2(1 - r_{xx_r}) \tag{3.2.9}$$

$$r_{t_x t_y} = \frac{r_{xy}^r}{\sqrt{r_{xx_r} r_{yy_r}}} \tag{3.2.10}$$

$$r_{xy}^c = \frac{U_t r_{t_x t_y}}{\sqrt{U_t^2 r_{t_x t_y}^2 - r_{t_x t_y}^2 + 1}} \tag{3.2.11}$$

The major challenge of the method is the computation of u_t (needed in $U_t = \frac{1}{u_t}$ for equation 3.2.11 is unobserved unlike u_x) is highly dependent on the value of r_{xx_u}. To circumvent this challenge, one may work backwards, instead of incorporating the values u_x and r_{xx_u} to estimate values of u_t, one may iteratively examine until convergence the appropriateness of different "plausible u_t values". This would involve examining how close, based on some criteria, they can reproduce the original u_x distribution when combined with the r_{xx_u} distribution. This approach is logically appropriate because u_x results from u_t and r_{xx_u} not the other way around as seemingly suggested by the formula.

3.2.5. Pearson Lawley formula

The *Thorndike Case I* and *Thorndike Case II formulas* estimate a bivariate relationship under DRR selection scenario, the *Thorndike Case III formula* and the method of *Hunter, Schmidt, & Le (Thorndike Case IV formula)* estimate a bivariate relationship under IRR selection scenario although *Thorndike Case IV formula* is able to tackle the effects of range restriction when a composite of predictors are used in the selection process. The *Pearson Lawley formula* is an extension of the *Thorndike Case IV formula* in that it can tackle the effects of range restriction when a composite of predictors are used in the selection process and/or when multiple outcomes are of interest (Allen and Dunbar, 1989; Held and Foley, 1994). Therefore the *Pearson Lawley formula* is multivariate in nature. To demonstrate how the *Pearson Lawley formula* works, assume a simple univariate selection scenario where interest is a single outcome (criterion), y, with selection based on a predictor (selection test), z. The estimation of predictive validity of x for y may be obtained by the *Pearson Lawley formula* expressed in equation 3.2.12. Note that $U_z = \frac{1}{u_z}$ where $u_z = sd_z/SD_z$ represents the ratio of the standard deviation of the restricted sample z to that of the unrestricted sample z. The *Pearson Lawley formula* may be extended and generalised for the multivariate case where the formula requires an estimate of the unrestricted covariance matrix of the predictors used for selection and restricted covariance matrix to estimate construct-level predictive validity (Dunbar and Linn, 1991; Fife, Hunter, and Mendoza, 2016).

$$r^c_{xy} = \frac{r^r_{xy}[U^2_z - 1]r^r_{xz}r^r_{yz}}{\sqrt{(1+[U^2_z-1](r^r_{zx})^2)(1+[U^2_z-1](r^r_{yz})^2)}} \tag{3.2.12}$$

3.2.6. Missing data approaches

A detailed coverage of missing data frameworks, mechanisms, patterns and handling methods will be presented in chapter 4. For now, two missing data handling methods for tackling the effects of range restriction when estimating predictive validity are discussed.

3.2.6.1. Full Information Maximum Likelihood (FIML)

Maximum Likelihood (ML) parameter estimation involves constructing a probability function of the data given the parameter(s) of interest. This probability function is what is referred to as a *likelihood (function)*. Given the likelihood (function), the task at hand becomes the determination of the parameter values that maximise the likelihood (function). These parameter values are referred to as Maximum Likelihood Estimates (MLE) (National Institute of Standards and Technology, 2003; S Purcell, 2007). Generally in the presence of missing data, ML estimation will be biased except in non-selection contexts under Missing At Random (MAR) (this is explained in section 4.5.4, for now take it as given) (Enders, 2001b; Molenberghs and Kenward, 2007). Within the context of selection, assume for the sake of simplicity, a DRR scenario (although it is easy to generalise to scenarios of IRR or selection based on multiple predictors) where n applicants, are subjected to a selection process based on a predictor (selection test), x, and the outcome (criterion) of interest is y for those who will be selected (entrants). As only the selected (entrants) will have outcome data available, to obtain an unbiased of r_{xy}, which is corrected for the effect of range restriction denoted by r^c_{xy}, the FIML is implemented as follows.

Assume the DRR selection context described, further assume MAR and that the data follow a multivariate normal distribution with mean, μ, covariance Σ and $d = \{x, y\}$ denoting the data. Firstly, the case wise observed data likelihood is obtained by maximising the function shown in equation 3.2.13. Note that K is a constant that depends on the number of complete data points, whereas μ, and Σ represent the parameter estimates for the mean and covariance matrices for the data, d, that are complete. Secondly, the casewise likelihood functions are accumulated across the entire sample and maximised based on equation 3.2.14 (Little, Jorgensen, et al., 2013). From the description, it may be seen that parameter estimation is accomplished by making use of all available data. Firstly, equation 3.2.13 makes use of all available data from all the n applicants. All the n cases are considered but a case i contributes to the parameter estimation for which there are complete data. Secondly, for 3.2.14, the use of all available data is accomplished by estimating the parameters of y (which has missing values for those rejected applicants) by incorporating information from x for the n applicants. The probable values for the missing values of y are implied by the available x through r^r_{xy} (Dong and Peng, 2013). Note that although the missing values y are not imputed, the borrowing of information from the observed portion of x for y is akin to replacing the missing data y with $E(y|x)$ (Enders and Bandalos, 2001).

$$LogL_i = K - \frac{1}{2}log|\Sigma| - \frac{1}{2}(d_i - \mu)^i\Sigma^{-1}(d_i - \mu) \tag{3.2.13}$$

$$Log(\mu,\Sigma) = \sum_i^n LogL_i \tag{3.2.14}$$

In the end, the computed Pearson correlation coefficient between x and y corrected for the effects of range restriction, denoted by r^c_{xy}, constitutes construct-level predictive validity.

3.2.6.2. Multiple Imputation (MI)

MI is slowly gaining prominence as a means of dealing with the attenuation (downward bias) in the predictive validity estimated by Pearson correlation coefficients within the selection contexts. For a simple DRR selection context where selection is based on a predictor (selection test), x, and the outcome of interest is y. Range restriction occurs as a result of missing data points of y for the rejected applicants. MI may be used to remedy this by generating a set of plausible values for the missing y so as to constitute complete data devoid of range restriction (refer to Figure 1.1 for a conceptual representation of this). This would thus facilitate estimation of construct-level predictive validity (Mendoza, Bard, et al., 2004; Pfaffel, Kollmayer, et al., 2016; Pfaffel, Schober, and Spiel, 2016; Pfaffel and Spiel, 2016; Wiberg and Sundström, 2009). Note that several estimates for the construct-level predictive validity will be obtained equal to the number of imputations conducted. To obtain a single estimate of the construct-level predictive validity, these estimates are pooled by obtaining their arithmetic mean. The associated standard error involves taking into account the within and between imputation variability as demonstrated in section 4.5.7. There are several methods by which missing data for y may be imputed, these methods with their pros and cons will be discussed in section 4.6.

To empirically illustrate the effect of range restriction and the feasibility of the MI as a means of determining construct-level predictive validity, assume a DRR selection scenario where the predictor (selection test) is x and the outcome (criterion) is y. For the scenario, data were drawn from the bivariate normal distribution shown in equation 3.2.15 with the unrestricted correlation arbitrarily taken to be r^u_{xy}=0.8. The simulated data are shown in Figure 3.10. On the top panel, before selection, range restriction is absent and the entire range of values for x and y are displayed. On the bottom panel, the selection ratio is arbitrarily set at 0.8 (top 80% of scores for x are selected), the red points illustrate the range of x values rejected which leads to their corresponding y values being unobserved. Computing the Pearson correlation coefficient between x and y values from the data on the bottom panel will definitely lead to downward bias.

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & r^u_{xy} \\ r^u_{xy} & 1 \end{pmatrix}\right] \tag{3.2.15}$$

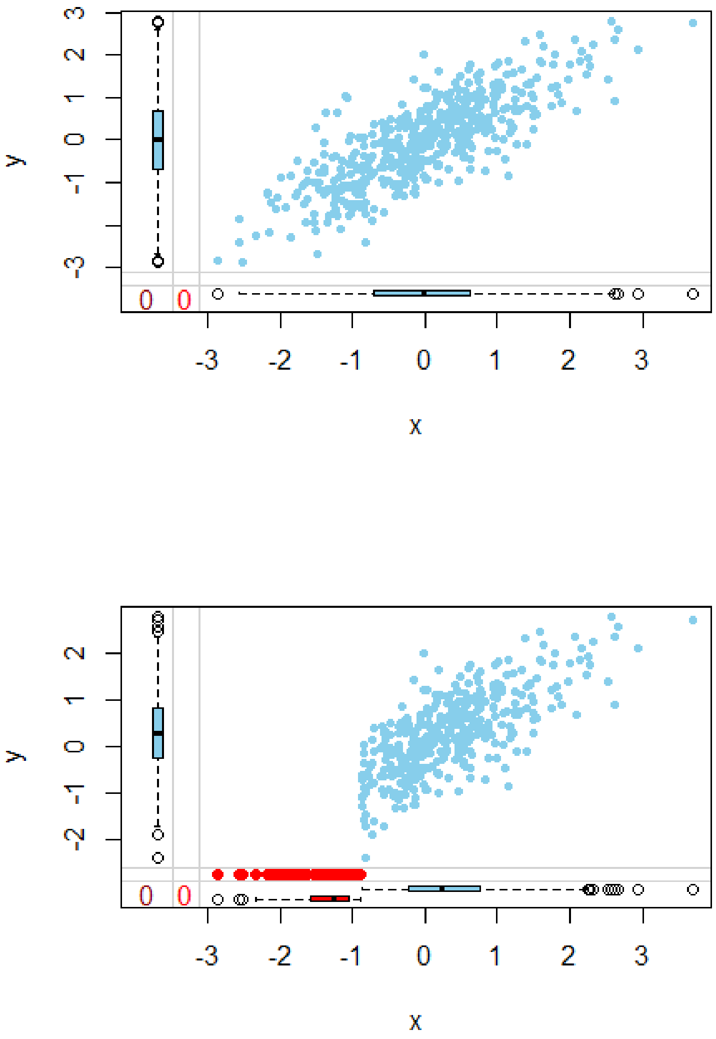


Figure 3.10.: *The effect of range restriction on the simulated predictor (selection test) x and outcome (criterion) y for arbitrary* $r_{xy}^{u} = 0.8$ *and selection ratio=0.8*

To correct for this bias, MI may be used to “reconstitute” the full data depicted in the top panel of Figure 3.10. This would involve imputing plausible values for the missing *y*. Figure 3.11 demonstrates how through the imputation of the outcome (criterion) *y* based on the predictor (selection test) *x*, the original full data for the 5 Monte Carlo simulated samples are “reconstituted”. The distribution of the *x* and imputed *y* seem plausible and similar to the original distribution of *x* and *y* depicted in Figure 3.10 with range restriction absent. This clearly demonstrates that tackling range restriction as a missing data problem is potentially useful in dealing with its effects on Pearson correlation coefficients.

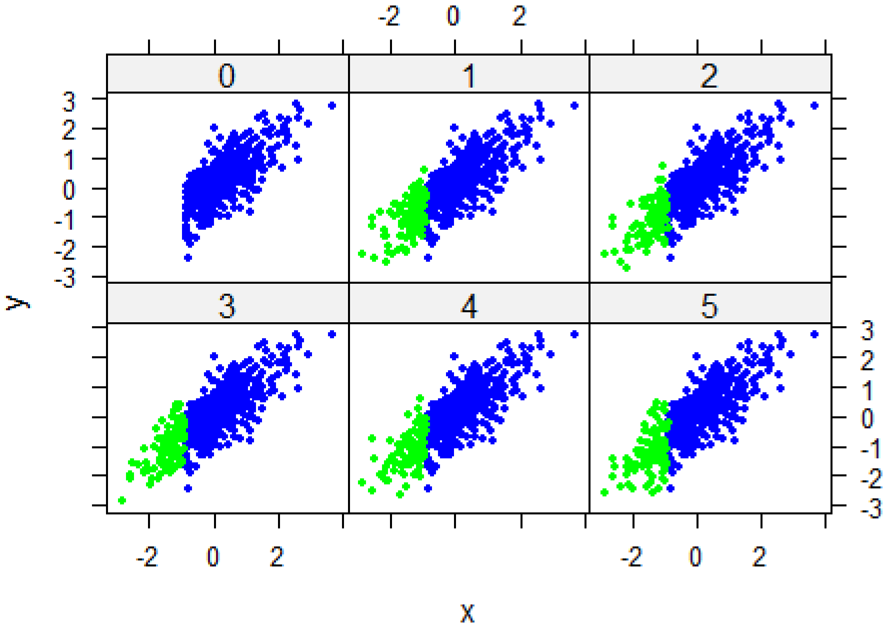


Figure 3.11.: *MI for the simulated y outcome based on the simulated selection test x with* r^u_{xy}*=0.8 and selection ratio arbitrarily set at 0.8. For the five Monte Carlo samples, the original x and y pairs are plotted in blue while the rejected x and imputed y pair plotted in green*

3.3. Chapter summary

In this thesis chapter, it was demonstrated from literature and by Monte Carlo simulations that range restriction has an adverse effect of inducing downward bias in predictive validity estimated by (Pearson) correlation rather than regression coefficients. The impact of measurement error in contributing to this downward bias in the estimate of predictive validity was also noted from literature (and demonstrated by Monte Carlo simulations). This downward bias was shown to be independent of the selection validity design considered and was most acute when selection process was strict. In addition, the several statistical methodologies for tackling the downward bias in the estimate of predictive validity were reviewed with those based on missing data handling techniques being of prime importance for the scope of the thesis.

In the next chapter, missing data frameworks, mechanisms, patterns and handling methods will be discussed. This will be done with the view of determining which of the missing data handling methods are most promising in the handling of the effects of downward bias due to range restriction under different selection validity designs. This will then lay a foundation for work relating to objective 2 outlined in section 1.4 that will involve using Monte Carlo simulations to compare the statistical methodologies (that may be) used to tackle the downward bias in the estimate of predictive validity.

4. Missing data patterns, frameworks, mechanisms and handling methods

In section 1.2, the focus of the thesis was introduced and the potential of missing data handling methods for estimating construct-level predictive validity highlighted (refer to Figure 1.1 for a conceptual representation). In section 3.2, statistical methods for estimating construct-level predictive validity were highlighted with those that make use of missing data handling methods covered in section 3.2.6. The treatment of construct-level predictive validity as a missing data problem has slowly been gaining prominence since the turn of the century. This is due to several construct-level predictive validity publications that have made use of missing data handling methods (Fife, Hunter, and Mendoza, 2016; Mendoza, Bard, et al., 2004; Pfaffel, Kollmayer, et al., 2016; Pfaffel, Schober, and Spiel, 2016; Pfaffel and Spiel, 2016; Wiberg and Sundström, 2009). Although missing data handling methods have been under development since the pioneering work of *Rubin (1976)*, it is only after the turn of the 21st century that computer hardware and software development have enabled the wide scale implementation and use of these methods (within the selection context) (Bengt O. Muthén and Asparouhov, 2016, pp 443-445; Molenberghs and Kenward, 2007; Schafer and Graham, 2002). In chapter 5, the performance of statistical methods for estimating construct-level predictive validity will be compared. The statistical methods based on formulas will be compared to those based on missing data handling methods. The comparison will be done under a variety of selection ratios. Methodologically an attempt will be made to extend this comparison to different selection validity designs and *auxillary variables* (to be defined later in section 4.5.7). For now, this thesis chapter will explore the missing data frameworks, mechanisms, patterns and handling methods. By the end of this chapter, those missing data handling methods that will be found to be viable for estimating construct-level predictive validity will form basis for further work in chapter 5.

4.1. Missing data patterns

In order to understand what missing data handling methods may be applicable within the selection context, it is important to first review the different types of *missing data patterns*. These are discussed primarily based on *Schafer and Graham (2002)*'s *"Missing data: our view of the state of the art"*. Assume as before that, the predictor (selection test) is denoted by x and that the outcome (criterion) is denoted by y. Further assume that in some instances, the criterion $\boldsymbol{y}$ is longitudinal with measurements taken over 3 time points such that $\mathbf{y} = \{y_1, y_2, y_3\}$ as shown in Figure 4.1. Thus the criterion is defined in both multivariate and univariate terms so as to facilitate the demonstration of the three missing data patterns as follows.

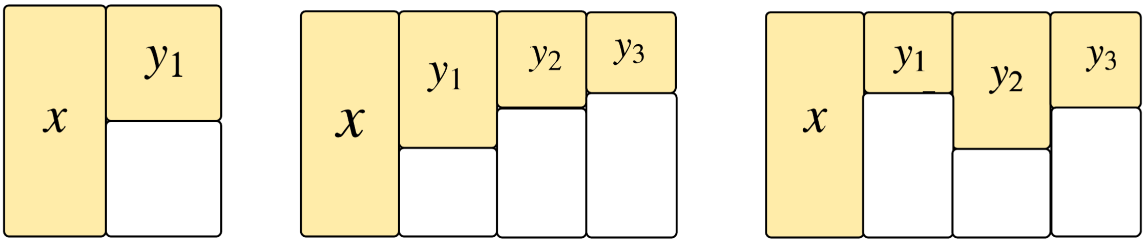


Figure 4.1.: *The three missing data patterns: univariate pattern of missingness on the extreme left, monotone pattern of missingness in the middle and arbitrary pattern of missingness on the extreme right. The yellow highlighted cells represent observed data while the white unhighlighted cells represent missing data.* [1]

For the univariate criterion y_1, selection is based on the predictor x and criterion observed only for those selected (entrants). The criterion values for y_1 are unobserved, that is missing, for those not selected (rejected). The missing data pattern under this scenario is referred to as *univariate missing data pattern*. For the multivariate or longitudinal criterion, the values for

[1] Figure adapted from *"Missing data: our view of the state of the art (Schafer and Graham, 2002)"* and constructed in LucidChart ®, *For more details, visit : www.lucidchart.com*

y are also observed only for those selected (entrants), however, if among the selected, only a fraction have **y** reported at each subsequent time point then the missing data pattern will either be *monotone* or *arbitrary missing data pattern*. As shown in Figure 4.1, *monotone missing data pattern* is created when an entrant has criterion values reported up to a particular time point with no value(s) reported at subsequent time point(s). For the *arbitrary missing data pattern* (also known as *intermittent missing data pattern*), the reporting of criterion values occurs randomly across the time points for those selected thus creating a non-monotone pattern of missingness such as the one shown.

4.2. Missing data frameworks

In section 4.1, the three different possible types of missing data patterns in a selection context were introduced. All of the three may be modelled under the Selection Model (SeM) framework. The missing data patterns for the multivariate or/and longitudinal data, i.e. *monotone* and *arbitrary patterns of missingness*, are modelled under the Pattern Mixture Model (PMM) and Shared Parameter Model (SPM) frameworks (Molenberghs and Kenward, 2007). Since the focus of this thesis is construct-level predictive validity estimated by a bivariate Pearson correlation coefficient between a single predictor (selection test) and a single outcome (criterion), the missing data pattern that will be of importance is the *univariate missing data pattern*. For this reason, only the SeM framework will be covered in great detail. For the interested reader, further references that deal with the treatment of *monotone* and *arbitrary missing data patterns* for the multivariate and/or longitudinal data are provided for the PMM framework (Hedeker and Gibbons, 1997; Lin, McCulloch, and Rosenheck, 2004; Little, 1993, 1994; Little and Wang, 1996; Molenberghs, Michiels, et al., 1998) and the SPM framework (Enders, 2011; Gao, 2004; Ibrahim and Molenberghs, 2009; Roy, 2003) respectively. A brief introduction to these frameworks can be found in section 4.4.

Selection Model (SeM) framework		
Missing Completely At Random (MCAR)	Missing At Random (MAR)	Missing Not At Random (MNAR)
Listwise deletion Pairwise deletion Last Observation Carried Forward (LOCF)	Direct likelihood Full Information Maximum Likelihood (FIML) Multiple Imputation (MI) Weighing observations	Joint Modelling Sensitivity analysis
Pattern Mixture Model (PMM) framework		
Shared Parameter Model (SPM) framework		

Table 4.1.: *Layout of missing data framework, mechanisms and handling methods adapted from Molenberghs and Kenward (2007)'s book on "Missing data in clinical studies".*

Table 4.1 shows the summary of the different missing data frameworks, mechanisms and handling methods. For the Selection Model (SeM) framework, the missing data mechanisms (described in great detail in sections 4.4.1 to 4.4.3) are presented in order of complexity (left to right). The methods that appear under each of the missing data mechanisms are presented in random order. In the following sections, the different mechanisms and handling methods that apply in the selection context for the *univariate missing data pattern* are explored. Before that however, a further definition of notations is in order.

4.3. Distribution of missingness

So far, the missing data patterns and frameworks have been covered in section 4.1 and 4.2, at this point, the notation that describes the distribution of missingness in data is defined. This is important for one to identify where and how missingness occurs in the data. To accomplish this, use of an indicator variable, r, is employed. In the selection context, with regard to construct-level predictive validity, interest is in the missing values of the outcome (criterion) for those not selected. Thus r can be defined as a binary term that assumes a value of "1" when y is observed (y^{obs}) and a value of "0" when y is missing (y^{miss}) as shown in equation 4.3.1. In addition, just as the predictor x is assumed to influence y, it is assumed that another predictor m is said to influence r. This forms the foundation for the discussion of the missing data mechanisms assuming an *univariate missing data pattern*.

$$r_i = \begin{cases} 1, & \text{if } y_i \textit{ is observed} \\ 0, & \textit{otherwise} \end{cases} \tag{4.3.1}$$

4.4. Missing data mechanisms

The missing data mechanisms are discussed assuming an *univariate missing data pattern*. This means that the SeM framework is of prime importance. Making use of notation described in section 4.3, the differences between the frameworks are identified. The frameworks differ in the way the joint distribution of the complete data, $f(y,r|x,m)$, is factorised as shown in equations 4.4.1, 4.4.2 and 4.4.3 for the SeM, PMM and SPM framework respectively.

$$f(y,r|x,m) = f(y|x)f(r|y,m) \tag{4.4.1}$$

$$f(y,r|x,m) = f(r|m)f(y|r,x) \tag{4.4.2}$$

$$f(y,r|x,m,b) = f(y|x,b)f(r|m,b)f(b) \tag{4.4.3}$$

It is possible to demonstrate that SPM framework may be derived based on either the SeM and PMM framework as expressed on equations 4.4.4 and 4.4.5 respectively. Under conditional independence, that is y and r are independent given b, then equations 4.4.4 and 4.4.5 simplify to equation 4.4.3 (Molenberghs and Kenward, 2007).

$$f(y,r|x,m,b) = f(y|x,b)f(r|y,m,b)f(b) \tag{4.4.4}$$

$$f(y,r|x,m,b) = f(y|r,x,b)f(r|m,b)f(b) \tag{4.4.5}$$

For the SeM framework in equation 4.4.1, the joint distribution of the complete data, $f(y, r|x, m)$, is factored into a marginal density of the *measurement* and *missingness process*. The term, $f(r|y, m)$, that defines the missingness process may be thought of as a description of an observation units's selection mechanism into the data. Thus the term Selection Model (SeM) introduced by *Heckman (1977).* For the PMM framework in equation 4.4.2, a different model is fitted for each pattern of missing data present. The observed data is taken to be a mixture of these missing data patterns weighted by the probability of each missing value. For the SPM framework in equation 4.4.3, it is assumed that joint distribution of the data can be factored into terms joined together with latent or random effect structure denoted by b (Molenberghs and Kenward, 2007). Given this understanding, the missing data mechanisms will now be discussed assuming a SeM framework whilst making use of the term, $f(r|y, m)$, that defines the missing data process.

4.4.1. Missing Completely At Random (MCAR)

For the Missing Completely At Random (MCAR) mechanism, the missing values of y, denoted by $r_i = 1$, are influenced only by a predictor m that is unrelated to x as conceptually demonstrated in Figure 4.2.

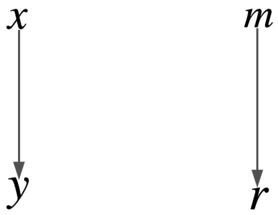


Figure 4.2.: *Conceptual representation of Missing Completely At Random (MCAR) for a univariate missing data pattern. Note that the arrows do not imply a causal relationship but simply a relationship.* [2]

In a selection context, selection is based on predictor (selection test) x which in turn leads to the outcome (criterion) y to be observed only for those selected. It thus inconceivable that MCAR would apply in a selection context when a selection test is used to determine entrants. MCAR

[2] Figure adapted from *"Missing data: our view of the state of the art (Schafer and Graham, 2002)"* and constructed in LucidChart ®, *For more details, visit : www.lucidchart.com*

would only apply in a selection context when the missing values for y occur by design. For example, in the UK medical selection context, this may occur if all applications to a particular medical school are arranged in random order, and subsequently every n^{th} application is selected for an offer regardless of the applicant's score on the UKCAT or (predicted) A-level grade. Subsequently, missing criterion values (*knowledge* and *skills*-based exam outcomes in the undergraduate medical school) for those rejected will be MCAR since they would be inherently missing by design rather than due to their scores on the selection test (the UKCAT or (predicted) A-level grade). For the MCAR, the probability of missingness $p(r|y,x,m)$ is defined by equation 4.4.6 (Bengt O. Muthén and Asparouhov, 2016, pp 443-445; Molenberghs and Kenward, 2007; Schafer and Graham, 2002).

$$p(r|y,x,m) = p(r|m) \tag{4.4.6}$$

4.4.2. Missing At Random (MAR)

For the Missing At Random (MAR) mechanism, the missing values of y, denoted by $r_i = 1$, are influenced by x and another predictor m that is unrelated to x as conceptually demonstrated in Figure 4.3.

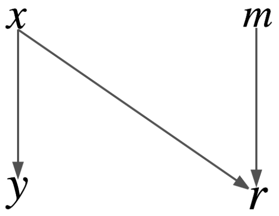


Figure 4.3.: *Conceptual representation of Missing At Random (MAR) for a univariate missing data pattern. Note that the arrows do not imply a causal relationship but simply a relationship.* [3]

In a selection context, MAR is more plausible. This is because selection is based on a predictor (selection test), x, which influences missingness in the outcome y as only those applicants that

[3] Figure adapted from *"Missing data: our view of the state of the art (Schafer and Graham, 2002)"* and constructed in LucidChart ®, *For more details, visit : www.lucidchart.com*

are selected have observed values for y. MAR applies in practice in the UK medical selection context when undergraduate medical schools makes offers to medical school applicants based on their scores on the UKCAT or/and A-level grades. Thus an applicant's score on any or both of these tests has an influence on whether they would be selected for medical school and subsequently whether their criterion values (*knowledge* and *skills*-based exam outcomes in undergraduate medical school) are observed or missing. The probability of missingness $p(r|y,m,x)$ is thus defined by equation 4.4.7 (Bengt O. Muthén and Asparouhov, 2016, pp 443-445; Molenberghs and Kenward, 2007; Schafer and Graham, 2002).

$$p(r|y,m,x) = p(r|m,x) \tag{4.4.7}$$

4.4.3. Missing Not At Random (MNAR)

For the Missing Not At Random (MNAR) mechanism, the missing values of y, denoted by $r_i = 1$, are influenced by x, another predictor m that is unrelated to x and y as conceptually demonstrated in Figure 4.4. Note that MNAR mechanism is equivalent to the MAR mechanism with the added restriction that missingness in y is influenced by y itself. Note that given this definition, the term $p(r|y,m,x)$ can not be simplified further.

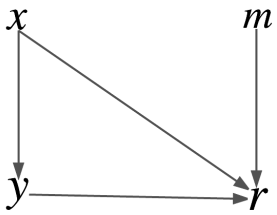


Figure 4.4.: *Conceptual representation of Missing Not At Random (MNAR) for a univariate missing data pattern. Note that the arrows do not imply a causal relationship but simply a relationship.* [4]

In a selection context, consider an hypothetical example in which selection data is made available in which entrants were enrolled into medical school based on a predictor (selection test)

[4] Figure adapted from *"Missing data: our view of the state of the art (Schafer and Graham, 2002)"* and constructed in LucidChart ®, *For more details, visit : www.lucidchart.com*

such as the UKCAT with outcome (criterion) *knowledge*-based year one exams scores being of interest and observed as expected only for those selected. This constitutes an hypothetical example in which MAR applies as defined in section 4.4.2. Now suppose that it was known that values for *knowledge*-based year one exams were self reported by the entrants themselves. If it is also later discovered that those who had a failing score of $<40\%$ decided not to report their score. Then the missing values for *knowledge*-based year one exams depend on both the predictor (selection test) UKCAT and outcome (criterion) *knowledge*-based year one exams which is characteristic of MNAR (Bengt O. Muthén and Asparouhov, 2016, pp 443-445; Molenberghs and Kenward, 2007; Schafer and Graham, 2002).

4.5. Missing data handling methods

So far, the missing data patterns, missing data frameworks and mechanisms have been discussed and contextualised within the selection process in sections 4.1, 4.2 and 4.4. Next, the different methods of handling missing data are presented, they comprise of *Deletion methods, Last Observation Carried Forward (LOCF), Weighing of observations, Ignorability, Full Information Maximum Likelihood (FIML), joint modelling of missing and observed data* and *Imputation methods*. Each of these will be presented in turn and contextualised within the selection context. Detailed explanations will be provided for those methods that are useful in handling *univariate missing data patterns* under the SeM. Where applicable information regarding the applicability of the method under each of the *missing data mechanisms* will be provided.

4.5.1. Deletion methods

4.5.1.1. Listwise Deletion (LD)

This method is also known as *case deletion* or *complete case (CC)* (Baraldi and Enders, 2010; Molenberghs and Kenward, 2007). To demonstrate how it is implemented, consider the predictive validity selection design conceptualised in Figure 3.6. Selection is based on predictor (selection test) z with the outcome (criterion) of interest being y. At the point of selection, there are $n_1 + n_2$ applicants with n_1 being selected for whom the outcome is y observed. For *LD*, the n_2 observations or cases with missing values for y are deleted (Eekhout et al., 2012; Enders, 2010). This means that the predictive validity estimated by the bivariate Pearson correlation coefficient, makes use of n_1 observations (or cases). The resulting Pearson correlation coefficient, r^r_{zy}, computed after *LD* is attenuated due to the effects of range restriction as demonstrated in chapter 3. For this reason, *LD* is never recommended as a missing data handling method in the selection context.

4.5.1.2. Pairwise Deletion (PD)

Also known as *Available Case (AC)* or *Pairwise Inclusion* (Baraldi and Enders, 2010; Molenberghs and Kenward, 2007). *PD* makes more use of data compared to *LD*. To demonstrate how it is implemented, consider the *two hurdle selection validity design* conceptualised in Figure 3.7. Selection is based first on predictor (selection test), z then subsequently on predictor (selection test), x. At point of initial selection, the total number of applicants is $n_1 + n_2 + n_3$ with $n_1 + n_2$ selected based on z. The bivariate Pearson correlation coefficient between z and x makes use of $n_1 + n_2$ observations. At the second point of selection based on x, n_1 out of the $n_1 + n_2$ are selected. The bivariate Pearson correlation coefficient between x and y makes use of n_1 observations. This highlights a key feature of *PD*, deletion of observations is done only in those variables required for the analysis at hand. This results in differing number of observations being used in our example, $n_1 + n_2$ and n_1 for the analyses considered (Eekhout et al., 2012; Enders, 2010). Note that under *LD*, only n_1 observations would have been used for both analyses. The resulting Pearson correlation coefficients, r^r_{zx} and r^r_{xy} , computed after *PD* are attenuated due to the effects of range restriction as demonstrated in chapter 3. For this reason, *PD* is never recommended as a missing data handling method in the selection context.

4.5.2. Last Observation Carried Forward (LOCF)

This method may be applied in settings where missingness is due to attrition, that is, in longitudinal or prospective cohort studies. It may be applied to both *monotone pattern of missingness* and *arbitrary (intermittent) pattern of missingness*. For each case in the data, it involves substituting a missing observation (value) for a variable at a particular time point with the observation (value) observed for that variable at a preceding time point hence the name *LOCF* (Lachin, 2016; Molenberghs and Kenward, 2007; Molnar, Hutton, and Fergusson, 2008). Generally the method is easy to implement but makes the unrealistic assumption that a case's profile changes up to the last observation and that the measurements flatten or remain unchanged from the moment of missingness onwards or during the period of unobserved values for *arbitrary or intermittent pattern of missingness*. *LOCF* tends to be biased even under the MCAR with the direction of the bias differing on a case to case basis (Kenward and Molenberghs, 2009; Molenberghs, Thijs, et al., 2004; Saha and Jones, 2009). In a selection context, where selection is based on a predictor x with interest being the outcome (criterion) y. Following selection, there would be no values observed for the outcome y for those rejected. Note that this applies even when several measurements for y are taken at different time points in a longitudinal study following selection. Therefore it would be practically impossible to implement LOCF within a selection context simply because there would be no values among the rejected to carry forward.

4.5.3. Weighting observations

The bias introduced into the analysis as a result of missing data may be counteracted by assigning weights to the observed data. This may be done such that the observed data infuses more information into the analysis to compensate for the information that is lost as a result of missing data. The *weighing of observations* is a method normally applied in the longitudinal context under a *monotone pattern of missingness*. It is implemented by first computing the probability of missingess for each observation within a case. This is accomplished by fitting a logistic or probit regression where the outcome assigned a value of "1" or "0" ("missing" or "not missing") for each observation within a case. The predictors in the model include the outcome (1 or 0) for a previous time point for a particular case (Molenberghs and Kenward, 2007; Molenberghs and Verbeke, 2006; Schafer and Graham, 2002). Next the weights are computed as the product of probabilities of not missing up to particular time point for a case. The weights are introduced into the model as *Inverse Probability Weights (IPWs)* obtained by taking the reciprocal of the product of probabilities. Note that this implies that for observations within a case, the lower the probabilities of not missing, the higher the resulting weight and the more information the observations infuse into the analysis. An instance in which *Weighting observations* is implemented is in the Weighted Generalized Estimating Equations (WGEE) model (Chen, Yi, and Cook, 2010; Preisser, Lohman, and Rathouz, 2002; Silva, Colosimo, and Demarqui, 2015). In a selection context, subject to selection based on a predictor x, the outcome y are not observed for those not selected with a probability of 1. This means that the probabilities of not missing would be 0 and the *IPWs* would be undefined. For this reason, *weighting of observations* in a selection context would be impossible to implement regardless of whether the pattern of missingness is *univariate* or *monotone*.

4.5.4. Direct likelihood

To demonstrate how the *direct likelihood* works in handling missing data under SeM framework, assume that the *measurement* and *missingness process* introduced in section 4.4 have parameters represented by Φ and φ respectively. Now the joint distribution of the full data may be expressed as in equation 4.5.1. Since the full data are never actually observed for the outcome in the selection context, in terms of the observed data equation 4.5.1 may be re-written as equation 4.5.2. Assuming the parameters Φ and φ are disjoint, the likelihood equation 4.5.2 may be written out as a product containing the two terms in the equation. Under MAR mechanism, if the interest is the *measurement process*, then the term $f(y^{obs}|x,\Phi)$ is used while term $f(r^{obs}|y,m,\varphi)$ ignored during the construction of the likelihood hence the term *direct likelihood* or *ignorability* (Molenberghs and Kenward, 2007; Molenberghs and Verbeke, 2006).

$$f(y,r|x,m,\Phi,\varphi) = f(y|x,\Phi)f(r|y,m,\varphi) \tag{4.5.1}$$

$$f(y^{obs}, r|x, m, \Phi, \varphi) = f(y^{obs}|x, \Phi) f(r|y^{obs}, m, \varphi) \tag{4.5.2}$$

It has been shown that when the *measurement* and *missingness process* are disjoint, under a MAR mechanism, *direct likelihood* generates parameter estimates that are unbiased (Beunckens, Molenberghs, and Kenward, 2005; Kadengye, Ceulemans, and Van den Noortgate, 2014; Kadengye, Cools, et al., 2012). However in the selection context, the *measurement* and *missingness process* are not disjoint as those applicants below a particular threshold score in the selection test, x, have a missing measurement for y. Therefore, in the selection context, the implementation of *direct likelihood* or *ignorability* results in bivariate Pearson correlation coefficients estimates that are attenuated (biased downwards).

4.5.5. Joint modelling of observed and missing data

For the *direct likelihood* method, the *missingness process* is ignored but the *measurement process* is modelled. Unlike the *direct likelihood* method, the *measurement* and *missingness process* may be modelled jointly. For data with multivariate or longitudinal outcomes, this may be accomplished under a SPM framework in which the joint distribution of the data is factorised into terms under girded by latent or random effects structure as shown in equation 4.4.3 (Creemers et al., 2010; Gad and Darwish, 2013; Molenberghs and Kenward, 2007; Vonesh, Greene, and Schluchter, 2006). With respect to construct-level predictive validity, the joint modelling of *measurement* and *missingness process* is of special interest. This entails handling of *univariate missing data pattern* assuming the SeM framework under the MNAR mechanism. This is discussed next with aid of the *Heckman model* developed by *Heckman (1977)* to deal with missingness in outcomes due to *self-selection* and *non-random selection* (Briggs, 2004; Van de Ven and Van Praag, 1981; Vance, 2009).

In the *Heckman model*, an unbiased β regression coefficient is obtained by regressing an outcome on a predictor whilst accounting for the missing mechanism for the outcome in the same regression model (Gross, 1990; Heckman, 1977). In a selection context assume that selection is based on a predictor (selection test), z, but interest is the modelling of the $\beta_{y|x}$ where x is another predictor of interest and y is the outcome (criterion). As was clearly demonstrated in chapter 3, from literature and also from simulation, $\beta_{y|x}$ would be biased. A simple solution to the problem would be to estimate $\beta_{y|x,z}$ rather than $\beta_{y|x}$. This is because range restriction has no effect on $\beta_{y|x,z}$ hence its estimate is always unbiased as was clearly shown in chapter 3. Another solution that may be explored is the joint modelling of the observed and missing data for y using the *Heckman model* in three steps.

The first step involves fitting the *selection model* shown in equation 4.5.3 which models the selection process. If an applicant, i, is selected, then s_i takes a value of “1” and “0” otherwise. In the second step, the relationship between y and x are modelled by estimating $\beta_{y|x}$ using the *prediction model* shown in equation 4.5.4. In the model, when an applicant, i, is rejected, y_i^* is not observed and is then assigned a value of “0”. The error terms , ε_{s_i} and ε_{y_i}, from the two models are assumed to follow the bivariate normal distribution in equation 4.5.5. Finally, the third step involves correcting the bias in $\beta_{y|x}$ using the model in equation 4.5.6 to obtain $\beta^c_{y|x}$. In the model, the the term $\rho\sigma_{\varepsilon_{s_i}}\sigma_{\varepsilon_{y_i}}$ corrects for selection bias (range restriction). When $\rho = 0$, there is no bias in the $\beta_{y|x}$ whereas when $\rho < 0$ and $\rho > 0$, $\beta_{y|x}$ will be biased downwards and upwards respectively. The corresponding bias in the intercept, β_0, will swing in the opposite direction, that is, upwards and downwards respectively. The size of the bias depends on the magnitude of the ρ, the relative variance of the error $\frac{\sigma_{\varepsilon_{s_i}}}{\sigma_{\varepsilon_{y_i}}}$, and the strictness of the selection process.

$$s_i = \alpha_0 + \alpha * z_i + \varepsilon_{s_i} \tag{4.5.3}$$

$$y_i^* = \begin{cases} \beta_0 + \beta_{y|x} * x_i + \varepsilon_{y_i}, & \text{if } \; s_i = 1 \\ 0, & \text{if } \; s_i = 0 \end{cases} \tag{4.5.4}$$

$$\begin{pmatrix} \varepsilon_{s_i} \\ \varepsilon_{y_i} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{\varepsilon_{s_i}} & \rho\sigma_{\varepsilon_{s_i}}\sigma_{\varepsilon_{y_i}} \\ \rho\sigma_{\varepsilon_{s_i}}\sigma_{\varepsilon_{y_i}} & \sigma^2_{\varepsilon_{y_i}} \end{pmatrix}\right] \tag{4.5.5}$$

$$\begin{aligned} E(y_i^* \text{ } is \text{ } observed) &= \beta_0^c + \beta_{y|x}^c * x_i + E(\varepsilon_{y_i}|s_i = 1) \\ &= \beta_0^c + \beta_{y|x}^c * x_i + \rho\sigma_{\varepsilon_{y_i}}\sigma_{\varepsilon_{s_i}}\frac{\phi(z_i\alpha)}{\Phi(z_i\alpha)} \end{aligned} \tag{4.5.6}$$

The *Heckman model*, as constituted, works well in correcting for *self-selection* and *non-random selection*. However, in selection settings, the usefulness of the *Heckman model* in estimating construct-level predictive validity in regression models is susceptible to the following challenges. First, when x and z, are not independent then the model will fail to converge due to multicollinearity problems (Toomet, Henningsen, et al., 2008). Second, if selection is based on a single selection test, z, the *selection model* shown in equation 4.5.3 is not estimable as the relationship betwen s_i and z becomes deterministic. Nevertheless, the *Heckman model* may prove useful when interest is $\beta_{y|x}$ and the information for the predictor upon which selection was conducted is missing but information is available on another predictor which is known to influence selection. That predictor may then incorporated in the *selection model* shown in equation 4.5.3. The selection model can also accommodate multiple predictors known to influence selection (Kennet-Cohen and Bronner, 1998).

4.5.6. Full Information Maximum Likelihood (FIML)

The usefulness of FIML in handling the missing values of the outcome y to enable the estimation of construct-level predictive validity in a selection context can be found in section 3.2.6.1. Generally, FIML replaces the missing data for a particular variable by the conditional expectation of the observed data for the variable given other observed variables data. Parameters estimated through FIML have been demonstrated to be unbiased and efficient under a MAR mechanism (Enders, 2001a,c; Enders and Bandalos, 2001; Lindé, 2005; Wothke, 2000).

4.5.7. Multiple Imputation (MI)

MI enables the analysis of incomplete data in three steps, the *imputation (also called filling in stage)*, the *analysis stage* and the *pooling stage* conceptually shown in Figure 4.5. *Rubin* (Rubin, 1976, 1996) demonstrated that if imputations are done properly, then the statistical inference obtained from the data will be valid.

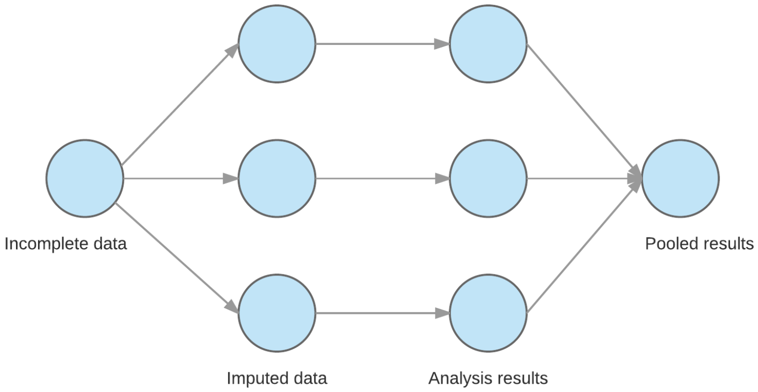


Figure 4.5.: *Conceptual representation of the stages of Multiple Imputation (MI).* [5]

The three steps for MI are explained next, assume that for the data containing missing values for some variables, each missing value for a variable is replaced with a set of P (say) plausible values. The replacement is done multiple times (P times) to reflect the uncertainty about the true values for the missing data. This would constitute the *imputation (also called filling in stage)*. In the *analysis stage*, each of the $p = 1,2,...,P$ data sets are analysed separately to obtain P parameter estimates, denoted by μ^p where "p" indicates the imputed data from which the parameter is estimated . In the *pooling stage*, first the MI estimate of the parameter is computed as a simple arithmetic mean of the estimates as shown in equation 4.5.7.

$$\hat{\mu}^* = \frac{\sum_{p=1}^{P} \hat{\mu}^p}{P} \tag{4.5.7}$$

[5] Figure adapted from *"Multiple Imputation" by Stef Van Buuren (2017)* and constructed in LucidChart ®, *For more details, visit : www.lucidchart.com*

Secondly, a measure of precision (uncertainty) for μ^* may be obtained by formula in equation 4.5.10 where B and W are the *between-imputation* and *within-imputation* variability computed in equation 4.5.8 and 4.5.9 respectively (Molenberghs and Kenward, 2007; Rubin and Schenker, 1991; Schafer, 1999; Schafer and Olsen, 1998; Sterne et al., 2009).

$$B = \frac{\sum_{p=1}^{P}(\hat{\mu}^p - \hat{\mu}^*)(\hat{\mu}^p - \hat{\mu}^*)'}{P-1} \tag{4.5.8}$$

$$W = \frac{\sum_{p=1}^{P} V^p}{P} \tag{4.5.9}$$

$$V = W + \left(\frac{P+1}{P}\right)B \tag{4.5.10}$$

Note that the MI is generally conducted while utilising two models, the *imputation model* and the *substantive model*. The *imputation model* is the model used to produce the imputations while the *substantive model* is the model used for latter analysis. In other words, the *substantive model* is the one that would have been fit in the first place had the data been complete. For best results, it is recommended that all variables that are to be included in the *substantive model* be included in the *imputation model* as well. In addition, other variables that are known to be correlated with variables containing missingness, should be included in the *imputation model* even when they are to be excluded from the *substantive model*. This is because, these variables referred to as *auxiliary variables* have been shown to be capable of substantially improving imputation thus reducing bias and increasing efficiency (Graham, 2003; Newgard and Haukoos, 2007). In the selection context, MI is slowly gaining prominence as a means of dealing with the attenuation (downward bias) in the predictive validity estimated by Pearson correlation coefficients (Mendoza, Bard, et al., 2004; Pfaffel, Kollmayer, et al., 2016; Pfaffel, Schober, and Spiel, 2016; Pfaffel and Spiel, 2016; Wiberg and Sundström, 2009). A detailed explanation of the usefulness of MI in the selection context is found in section 3.2.6.2.

4.6. Forms of imputations

In this section, the forms of imputations will be covered with a view of determining which form yields the best results in the selection context. Take x and z, to be predictors and y the outcome (criterion). Selection may induce either (both) DRR or (and) IRR. Situations where each applies, will be explicitly defined. Given this information, the forms of imputations that may apply in a selection context are broadly classified as

4.6.1. Imputing unconditional means

Assume a DRR scenario where selection is based on a predictor (selection test), x, with y the outcome (criterion). The missing values for y (y^{miss}) for those rejected may be imputed by the arithmetic mean of the observed values of y (y^{obs}). This form of imputation is easy to implement, preserves the mean of y but is problematic as it distorts its the variance and quantiles. In turn, the covariances and the bivariate Pearson correlation coefficient between x and y are attenuated. This form of imputation is referred to *imputing unconditional means* as each missing value of y (y^{miss}) is imputed without using information from the predictor (selection test) x. Instead of the mean, the median or the mode of y^{obs} may also be used in the imputation. This form of imputation is a *single imputation* method and therefore does not account for the uncertainty associated with the imputation process. (Donders et al., 2006; Haukoos and Newgard, 2007; Heijden et al., 2006; Zhang, 2016).

4.6.2. Imputing from unconditional distributions

Assume an IRR scenario where selection is based on a predictor (selection test),z, with x being a predictor and y an outcome (criterion). *Imputing from unconditional distributions* involves replacing the missing values for y (y^{miss}) for those rejected, with values of y from "similar" profiles based on the predictor x. This may be accomplished by creating classes within the observed values of x with somewhat similar observed values. Subsequently, a random value of y^{obs} is drawn from any of the classes whose corresponding value of x closely matches with the value of x for which y is missing (Durrant et al., 2005). Many variations and modifications of this form of imputation exist such as *random hot-deck imputation*, *deterministic hot-deck imputation*, *cold-deck imputation* and *predicted mean matching* (Andridge and Little, 2010). The most common approach involves imputation following sampling with replacement from the classes containing potential values of y. This form of imputation has the advantage of not requiring parametric assumptions for its implementation. However, it does not yield good results as it distorts the covariance and Pearson correlation coefficients between z, x and y (Andridge and Little, 2010; Haukoos and Newgard, 2007; Kim and Fuller, 2004; Myers, 2011; Schafer and Graham, 2002).

4.6.3. Imputing conditional means

This form of imputation is another instance of *single imputation* method and therefore does not account for the uncertainty associated with the imputation process. Assume a DRR scenario where selection is based on a predictor (selection test), x, with y the outcome (criterion). The missing values for y (y^{miss}) for those rejected may be imputed by the predicted values of y given x (from a regression model of y on x). For this reason, *imputing conditional means* is also known as *regression imputation*. This form of imputation does not yield good results because it overstates the covariance and the Pearson correlation coefficient between x and y. In fact, the The R^2 among the imputed variables is 1.00. If there is no significant association between x and y, this form of imputation reduces to *unconditional mean imputation* (Greenland and Finkle, 1995; Haukoos and Newgard, 2007; Schafer and Graham, 2002; Shao and Wang, 2002).

4.6.4. Imputing from conditional distributions

Assume a DRR scenario where selection is based on a predictor (selection test), x, with y the outcome (criterion). The missing values for y (y^{miss}) for those rejected may be imputed by a predicted value of y given x from the regression of y on x plus a residual error term. Note that this form of imputation amounts to *imputing conditional means* described in section 4.6.3 but with an extra level of sophistication. This form of imputation is also referred to as *stochastic regression imputation* since the residual error term is drawn from a normal distribution with mean zero and variance equal to residual mean square error (Iris Eekhout, 2016). If the imputation model is correctly specified, this method produces (nearly) unbiased parameter estimates (Haukoos and Newgard, 2007; Schafer and Graham, 2002).

4.7. Algorithms for Multiple Imputation (MI)

4.7.1. Expectation Maximisatiom (EM)

To illustrate how the EM algorithm is utilised for imputation, assume as before that in the selection context, that x is a predictor (selection test) and y is the outcome (criterion). Due to selection, outcomes for the rejected applicants will be unobserved, y^{miss} while outcomes for those selected will be observed y^{obs}. Now consider equation 4.4.1, which highlights how the joint distribution of the data is factorised under the Selection Model (SeM) framework. It follows then, equation 4.7.1 applies. Under the MAR mechanism, equation 4.7.2 is obtained since the focus is the observed data. The resulting likelihood may be written as expressed in equation 4.7.3 where Θ is the parameter of interest. Note that based on expected iterations, equation 4.7.4 holds given a flat prior on Θ. The EM algorithm, is then used to find the mode of the posterior distribution. Subsequently, values may be drawn from this posterior distribution and imputed to replace the missing values y^{miss} (Catellier et al., 2005; Honaker, King, Blackwell, et al., 2011; Lin, 2010; Schafer, 1999). Thereafter, imputations may be conducted several times so as to accomplish MI as described in section 4.5.7.

$$p(y,r|x,m) = p(y|x)p(r|y,m) \tag{4.7.1}$$

$$p(y^{obs},r|x,m) = p(y^{obs}|x)p(r|y^{obs},m) \tag{4.7.2}$$

$$L(\Theta|y^{obs}) \propto P(y^{obs}|\Theta) \tag{4.7.3}$$

$$P(y^{obs}|\Theta) = \int P(y|\Theta)dy^{miss} \tag{4.7.4}$$

4.7.2. Markov Chain Monte Carlo (MCMC)

As in section 4.7.1, assume that x is the predictor (selection test) and y the outcome (criterion). Assume the joint distribution of the data is completely defined by $p(y|x,\Theta)$. The posterior distribution may be obtained by Multiple Imputation Chained Equation, also known as FCS (Full Conditional Specification) (MICE) which is based on MCMC algorithm (Azur et al., 2011; White, Royston, and Wood, 2011). The joint distribution of the data is expressed as a product of conditional densities. The Gibbs sampler is used to conduct MCMC imputation as follows. First, a random value for $\Theta^{*(0)}$ is sampled. Then the posterior distribution is explored by generating $\Theta^{*(t)}$ and $y^{*(t+1)}$ where t=0,1,2,3.... in a sequential manner as shown in equation 4.7.5 and 4.7.6. Note that the $y^{*(t+1)}$ is the imputed value at iteration t. At the end of each iteration in equation 4.7.6, the data consists of $\{y^{*(t+1)}, y^{obs}\}$. These are then fed back into equation 4.7.5. This process is done continuously and sequentially until convergence is achieved. Convergence is said to occur when it can be demonstrated based on some measures that sampling is being done from the targeted posterior distribution (Lesaffre and Lawson, 2012; Lin, 2010; Rubin, 2003; Zalewska, Niemiro, and Samoliński, 2010). Thereafter, imputations may be conducted several times so as to accomplish MI as described in section 4.5.7.

$$\Theta^{*(t)} \sim p(\Theta|y^{obs(t)}, x) \tag{4.7.5}$$

$$y^{*(t+1)} \sim p(y^{obs}|x, \Theta^{*(t)}) \tag{4.7.6}$$

4.8. Chapter summary

In this thesis chapter, missing data patterns, frameworks, mechanisms and handling methods were discussed. With regard to construct-level predictive validity estimated by a bivariate Pearson correlation coefficient, it was demonstrated that the *univariate missing data pattern* under the SeM framework was of prime importance. For the missing data mechanisms, it was demonstrated that MCAR is not applicable in the selection context. Although both the MAR and MNAR are applicable in the selection context, MNAR is appropriate for the joint modelling of missing and observed data and thus beyond the scope of this thesis. Under MAR, the missing data handling methods found to be useful in estimating construct-level predictive validity are the FIML and MI. The form of imputation found to be most useful in producing unbiased and efficient parameter estimates is *imputing from a conditional distribution*.

To recap, so far in this thesis, objective 1 outlined in section 1.4 has been met. Objective 1(a) and 1(b) were addressed in chapter 2 while objective 1(c) was addressed in chapter 3. The aim of this chapter was to lay the foundation for work relating to objective 2 and 3 in chapter 5. In that chapter, the performance of the statistical methods for establishing construct-level predictive validity discussed in chapter 3 will be evaluated. The statistical methods based on formulas will be compared to those based on missing data handling methods. This comparison will be conducted by means of a Monte Carlo simulation whose results will be validated with the aid of a contrived example using real-world data.

Part III.

Implementation of proposed methodology

5. Correcting for range restriction bias using missing data handling methods: simulation and validation study

In this chapter, the objectives 2 and 3 outlined in section 1.4 will be addressed. This chapter is thus divided into two broad phases that address the two objectives separately. These are the *testing* and *validation* phases which involve simulations that will make use of both artificial (pseudo-random) and a contrived example using real-world data respectively. The goal will be to evaluate the performance of statistical methods for achieving construct-level predictive validity. The statistical methods based on formulas will be compared to those based on missing data handling methods. The considerations for the design, conduct, analysis and reporting of the simulations in this chapter are inspired by *"The design of simulation studies in medical statistics"* by *Burton et al. (2006)*. These considerations are presented as follows

1. **Objectives**

   Using simulation techniques, evaluate statistical formula based methods for achieving construct-level predictive validity and compare them against those based on missing data handling methods. This will be accomplished in two phases

   a) *Testing phase* which will involve the use of artificial also known as (pseudo) random data drawn from specified distributions and;

   b) *Validation phase* which will involve a contrived example using real-world data for the purpose of determining the performance of the methods on real world data.

5. *Correcting for range restriction bias using missing data handling methods*

   This will help inform discussion on the specific circumstances in which missing data handling methods may optimally correct for range restriction.

2. **Choice of resampling methods**

   Simulations rely on *resampling methods*. These resampling methods are *Monte Carlo, randomisation (permutation) tests, jackknife* and the *bootstrap*. A detailed discussion of these resampling methods can be found in section 8.3 of the Technical Appendices. Randomisation (permutation) tests are optimal for calculation of pvalues given specified hypotheses. Therefore, given the aims in (1) above, randomisation (permutation) tests are unsuitable. Monte carlo, bootstrap and jackknife are optimal for assessing the performance of statistical methods under specified circumstances. These were thus considered suitable for the stated aims in (1) above. From the three methods, Monte Carlo and (modified) bootstrap, were chosen for testing and validation of the methods under investigation respectively. Monte carlo resampling was chosen for the testing phase since, under the method, data required may be simulated from known distributions. A Modified version of the case resampling bootstrap was chosen for the validation phase since bootstrap resampling may be used in an contrived example using real-world data. The real-world data used to accomplish the validation phase was derived from the Professional and Linguistic Assessments Board (PLAB) test and was made available by the General Medical Council (GMC). More details on the various aspects of the data are included in section 5.3.

3. **Data generating mechanism**

   For the simulations, data were generated from multivariate normal distributions. The estimand affected by range restriction is the (Pearson) correlation coefficient. In order to satisfy the restrictions imposed on correlations coefficients, the individual elements of the multivariate normal were drawn from a uniform distribution. In addition, it was assumed that the (Pearson) correlation coefficient was positive. This is because in the UK, the context in which the motivating examples for the thesis are obtained, all predictive validity studies report positive correlations (whether statistically significant or otherwise) between selection tests (e.g. UKCAT, GCSE, A-level and PLAB tests) and criterion (e.g. *knowledge* and *skills*-based) (see (McManus, Dewberry, Nicholson, and Dowell, 2013; McManus, Dewberry, Nicholson, Dowell, et al., 2013; McManus, Powis, et al., 2005; McManus, Woolf, et al., 2013; Tiffin, Mwandigha, et al., 2016; Tiffin, Paton, et al., 2017)). For this reason, in all simulated settings, the correlation coefficients were simulated from distributions with positive support (values). To evaluate, the impact of the magnitude of correlation coefficient (with positive support between zero and one) on construct-level predictive validity two settings were investigated. The first involved adopting four different values of 0.2, 0.4, 0.6 and 0.8 for the correlation coefficients (estimates of predictive validity). These values were chosen in such a way that the entire range of plausible values for the positive correlation coefficients under DRR and IRR would be represented (see Tables 5.3, 5.4 and 5.5 in section 5.1). Second, the average impact of the correlation coefficient (estimates of predictive validity) was subsequently examined (in sections 5.2 and 5.3) for the different selection validity designs (more on that follows next).

4. **Selection scenarios to be investigated**

   For the simulations, four selection proportions were considered. Specifically, the top 20%, 40% , 60% and top 80% of the applicants were selected. These selection proportions may be thought of as selection ratios of 0.2, 0.4, 0.6 and 0.8 respectively. These values were chosen to capture the entire range of possible selection ratios between zero and one. A selection ratio is obtained by computing the fraction of those selected over all applicants. For example, a selection ratio of 0.2 implies that only applicants with predictor (selection test) scores in the top 20% of the distribution are selected. For each selection ratio, the *single hurdle concurrent validity, predictive validity,* and *two hurdle validity* selection designs were considered. These selection designs are conceptually illustrated in Figures 3.5, 3.6 and 3.7 respectively.

5. **Statistical methods evaluated**

   The methods evaluated in the simulation study were the *Thorndike Case II, Thorndike Case III* and *Pearson Lawley*. These methods have been traditionally used in correcting for the effect of range restriction in the (Pearson) correlation coefficients. A detailed overview of these methods is found sections 3.2.2, 3.2.3 and 3.2.5 respectively. The performance of these methods were compared against those based on missing data handling methods which were found to be applicable in the selection context (see section 4.5). These missing data handling methods, used to correct for the effects of range restriction, were Full Information Maximum Likelihood (FIML) and Multiple Imputation (MI). A detailed overview of these methods is found in sections 3.2.6.1 and 3.2.6.2 respectively. MI correction was compared to *Thorndike Case II* and *Thorndike Case III* for the DRR and IRR respectively under the *predictive validity* selection design. MI correction and FIML were compared to *Pearson Lawley* correction under the *single hurdle concurrent validity* and *two hurdle validity* selection designs. For the MI, the Expectation Maximisatiom (EM) and Markov Chain Monte Carlo (MCMC) imputation algorithms were evaluated. The convergence stopping rule for the algorithms were arbitrarily chosen to be tolerance of $1 * 10^{-6}$ (determined to be sufficiently close to zero) and $n_{iter} = 50$ (more than twice the recommended number of iterations after sampling from target distribution has commenced, (Royston and White, 2011, p. 2)) for the EM and MCMC algorithms respectively. This means that for the EM algorithm, convergence was said to occur when there was no more than $1 * 10^{-6}$ difference in the loglikelihood values from any two consecutive iterations. For MCMC algorithm, convergence was said to occur after 50 iterations. A description of these imputation algorithms can be found in section 4.7.1 and 4.7.2 respectively.

6. **Performance measures**

In order to compare the performance of the different methods evaluated, *Mean Bias* and *Root Mean Square Error (RMSE)* were used. Their formulas are expressed in equation 3.1.9 and 3.1.10. These are conveniently presented again in Table 5.1 where the number of simulations are denoted by n_{sims}. In a simulation setting, bias may be defined as the average discrepancy between the simulated estimates and the true (known) estimate. Methods with little or no (mean) bias are preferred. On the other hand, RMSE quantifies the precision of the simulated estimates. Smaller magnitudes of RMSE lead to greater precision of the simulated estimates. Just like in the case of bias, estimators with little RMSE are preferred. Note that the square of RMSE which is MSE is a combination of bias and variance of an estimator, which may be expressed as shown in equation 5.0.1. To objectively compare the methods, formal statistical tests, that is the T-test and F-tests, were used. The (two sample) T-test was developed for the testing of equality of means of two independent random samples and was thus used to compare the equality of bias between any two methods under consideration (Student, 1908). Based on these assumptions, the hypothesis of interest is shown on equation 5.0.2. The subscript k and j denote the first and second method under consideration respectively. The critical values of the T distribution are $t^*_{0.025,df}$ and $t^*_{0.975,df}$. The subscript *df* is the degree of freedom computed for each test assuming non-equality of variance of the two samples (the computation of the *df* depends on whether the variances of the two samples under consideration are assumed to be equal or not, for more details see (Coombs, Algina, and Oltman, 1996; Moser and Stevens, 1992; Ruxton, 2006)). The subscripts 0.025 and 0.975 correspond to the probabilities of the distribution for a two sided hypothesis at specified level of significance (type I error rate, see item 8 below) (Derrick et al., 2017; Witt and McGrain, 1985). The MSE is a measure of precision which may be interpreted as a measure of variance. To test for equality of MSE between the first (k) and second method (j) respectively, equality of variance hypothesis testing was done as shown in equation 5.0.3. Note that this is an Anova type experiment with the critical value of the F distribution being $F^*_{0.95,n_1,n_2}$ where $n_1, n_2 = (n_{sims} - 1)$ (Pfaffel, Kollmayer, et al., 2016; Walther and Moore, 2005).

Criteria	**Formula**
$\widehat{Meanbias}$	$\frac{1}{n_{sims}} \sum_{i=1}^{n_{sims}} (r^r_{xy_{(i)}} - r^u_{xy_{(i)}})$
$\widehat{RMSE}$	$\sqrt{\frac{1}{n_{sims}} \sum_{i=1}^{n_{sims}} (r^r_{xy_{(i)}} - r^u_{xy_{(i)}})^2}$

Table 5.1.: *Performance measures for the different methods evaluated by simulation. Note that n_{sims} denotes the number of simulated samples*

$$\widehat{MSE} = \widehat{Meanbias}^2 + Var \tag{5.0.1}$$

$$\begin{aligned} H_0 &: \widehat{Meanbias}_k = \widehat{Meanbias}_j \\ H_1 &: \widehat{Meanbias}_k \neq \widehat{Meanbias}_j \end{aligned} \tag{5.0.2}$$

$$\begin{aligned} H_0 &: MSE_k = MSE_j \\ H_1 &: MSE_k > MSE_j \end{aligned} \tag{5.0.3}$$

7. **Number of simulations n_{sims} and samples size n**

In order to evaluate the methods, the empirical distribution of the parameters of interest is required. In simulation studies, this involves the generation of many samples which are then used to estimate the empirical distribution of the parameter of interest. The number of samples needed for a simulation study is not a straightforward matter. Several issues need to be considered, these include the specified type I error rate (α), the magnitude of the parameter estimate σ under consideration and the maximum allowable bias in the estimates and standard errors of interest (δ). These are represented by the formula in equation (5.0.4). The term $Z_{(1-\frac{\alpha}{2})}$ is the quantile $(1-\frac{\alpha}{2})$ from a standard normal distribution. Assuming 1% maximum bias for a known Pearson correlation estimate of 0.2345 with standard error 0.0345 and type I error rate of 5%. Using the formula, the number of simulated samples should be at least 832 as shown in equation 5.0.5. To account for power, formula (5.0.6) may be used where power is taken to be equal to one less the quantity of type II error rate $(1-\beta)$. Note that this formula gives equivalent results to formula in equation 5.0.4 when power is set at 50% as shown in equation 5.0.7. Choosing the desired power leads to an appropriate number of simulations required (Burton et al., 2006; Muthén and Muthén, 2002). Another issue that was considered was the Monte Carlo Standard Error (MCSE). This is the uncertainty associated with the simulation results, it has been demonstrated that $n_{sims} \geq 1000$ leads to acceptable levels of MCSE (rule of thumb, less than 5%) (Harding, Tremblay, and Cousineau, 2014; Koehler, Brown, and Haneuse, 2009; Lesaffre and Lawson, 2012; Wehrens, Putter, and Buydens, 2000).

$$n_{sims} = \left(\frac{Z_{(1-\frac{\alpha}{2})}\sigma}{\delta} \right)^2 \tag{5.0.4}$$

$$n_{sims} = 831.50 = \left(\frac{1.960 * 0.0345}{0.0100 * 0.2345} \right)^2 \tag{5.0.5}$$

$$n_{sims} = \left(\frac{[Z_{(1-\frac{\alpha}{2})} + Z_{1-\beta}]\sigma}{\delta} \right)^2 \tag{5.0.6}$$

$$n_{sims} = 831.50 = \left(\frac{[1.960 + 0.0000] * 0.0345}{0.0100 * 0.2345} \right)^2 \tag{5.0.7}$$

For the simulations, n_{sims} was set at 5,000 as this was seen to be safely above the threshold of 1,000 to ensure the simulation results were reliable without unnecessarily increasing the simulation time. For the sample size of each simulated sample n, observations of 500 were arbitrarily chosen. This was informed by the fact that $n = 500$ was a large enough figure to ensure that the restricted sample size, n_r, remained large even after the selection ratio of 0.2 (i.e.$\frac{20}{100} * 500 = 100$) was applied during the simulations.

8. **Software, packages and statistical errors**

For the simulation study, the R software (R Core Team, 2014) was used. Within the software, the *Amelia II* package was used for MI based on the EM algorithm (Honaker, King, Blackwell, et al., 2011) while the *mice* package was used for MI based on the MCMC algorithm (Buuren and Groothuis-Oudshoorn, 2011). The Thorndike Case II, Thorndike Case III, Pearson Lawley and FIML corrected estimates for the correlation were obtained using the *selection* package (Dustin Fife, 2016). R software was also used to create graphical outputs of the results. Unless otherwise stated, the type I and type II error rate were (arbitrarily) pre-specified to be 5% and 20% respectively in line with most scientific studies (Banerjee et al., 2009). All conceptual diagrams that do not present results of statistical analyses were constructed in Lucidchart (Lucid Software Inc, 2018).

9. **Simulation time**

   The total simulation time for the different selection ratios under the different selection designs considted of 59 simulation settings conducted over a period of 1,248 hours. The breakdown of the different simulation settings conducted and their respective durations are presented in Table 5.2.

	Predictive validity selection design			
Setting	# Runs	Time per run (hours)	Validation conducted	Total time taken (hours)
Section 5.1.1 (DRR)	4	12	No	48
Section 5.1.2.1 (IRR)	16	18	No	288
Section 5.1.2.2 (IRR)	16	18	No	288
Section 5.2.1 (DRR)	4	12	Yes (Section 5.3.1)	96
Section 5.2.2 (IRR)	4	18	Yes (Section 5.3.2.1)	144
Section 5.2.2 (IRR)	4	18	Yes (Section 5.3.2.2)	144
	Concurrent validity selection design			
Setting	# Runs	Time / run (hours)	Validation conducted	Total time taken (hours)
Section 5.2.4	1	24	Yes (Section 5.3.4)	48
	Two hurdle validity design			
Setting	# Runs	Time / run (hours)	Validation conducted	Total time taken (hours)
Section 5.2.3.1	1	36	Yes (Section 5.3.3.1)	36
Section 5.2.3.2	1	36	Yes (Section5.3.3.2)	36
	Discrepancy of results between testing and validation phase			
Setting	# Runs	Time / run (hours)	Validation conducted	Total time taken (hours)
Section 5.4(DRR)	4	12	No	48
Section 5.4(IRR)	4	18	No	72
Total	59			1,248

Table 5.2.: *Simulation times for the different investigations conducted. Note that a run consists of simulation for all selection ratios (0.2, 0.4, 0.6 and 0.8) evaluated for a particular setting under consideration. In places where validation was conducted using real world data, the total simulation time doubles as both the testing and validation phases involve separate simulations*

5.1. Impact of the magnitude of correlation coefficient on the correction for range restriction bias

5.1.1. Direct range restriction

Assume a predictor (selection test) x and outcome (criterion) y. In this section, the performance of the MI corrections for range restriction bias will be evaluated to determine whether their performance is affected by the magnitude of unrestricted (Pearson) correlation coefficient r^u_{xy}. To determine this, values of r^u_{xy} were set at $\{0.2, 0.4, 0.6, 0.8\}$. Exactly 5,000 Monte Carlo samples of 500 observations were generated as per the multivariate normal distribution in equation 3.1.8. Note that since the mean and variance of the multivariate normal distribution is standardised, all the associations computed are at the level of correlation coefficients. In each of the samples obtained, selection was based on the predictor x for the selection ratios of 0.2, 0.4, 0.6 and 0.8. The resulting Pearson correlation coefficient r^r_{xy} was biased downwards. For the bias correction methods based on MI, a dry run of the simulation indicated that 5 to ten MI yielded similar results. These results remained unchanged even with further increases in the number of imputations, P, for both EM and MCMC algorithm based imputations. The same strategy was employed to determine the number of iterations, n_{iter} for each imputation based on MCMC algorithm. The optimum n_{iter} was found to be 10. However, in order to be defensive, P and n_{iter} were set at 25 and 50 respectively. A choice of higher numbers for P and n_{iter} would considerably increase the simulation time without a justifiable gain in reliability of the results.

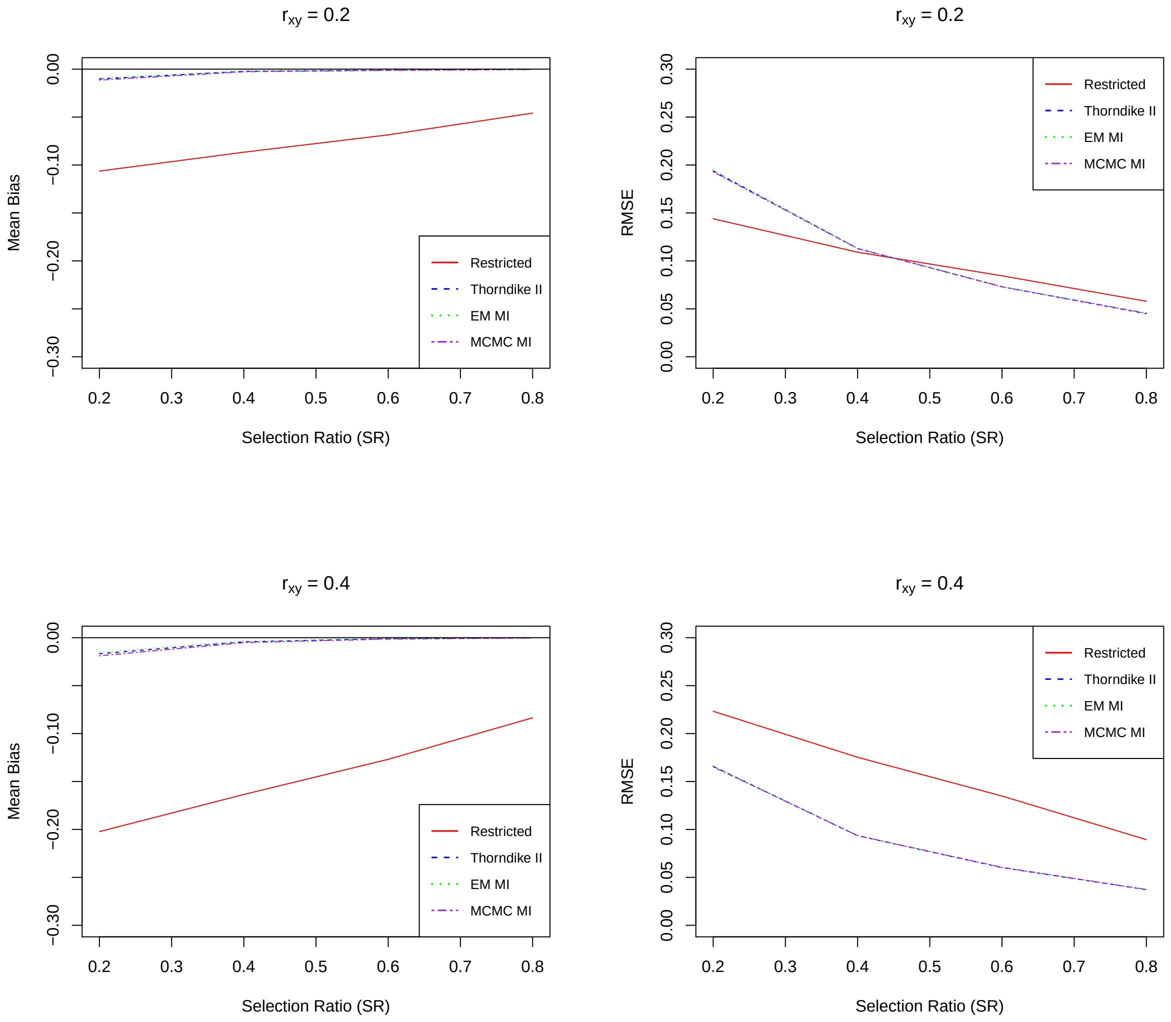
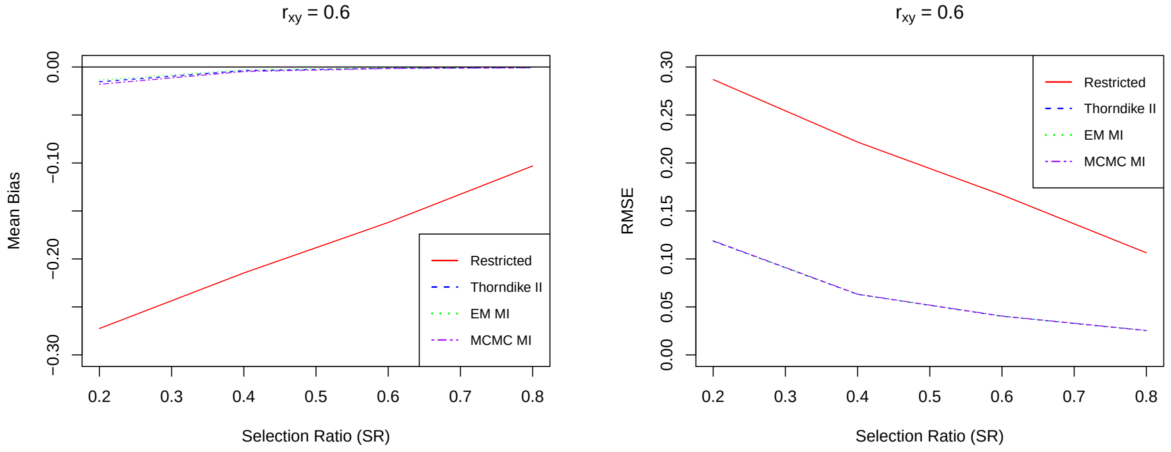


Figure 5.1.: *Results from the simulations exploring the impact of quantity of correlations, r^u_{xy}=0.2 and 0.4, on performance measures mean bias and RMSE. Note that r^u_{xy} is denoted by r_{xy}.*



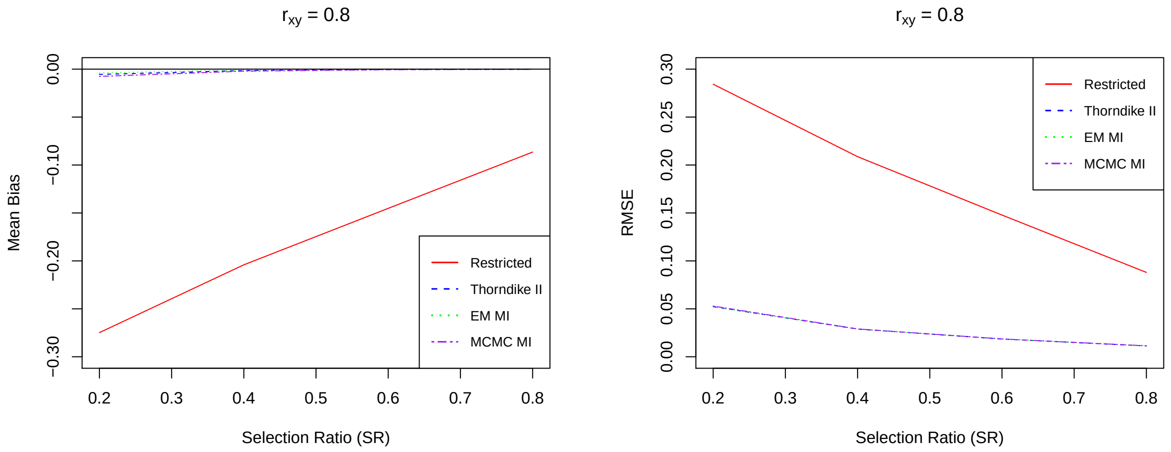


Figure 5.2.: *Results from the simulations exploring the impact of quantity of correlations, r^u_{xy}=0.6 and 0.8, on performance measures mean bias and RMSE. Note that r^u_{xy} is denoted by r_{xy}.*

Figures 5.1 and 5.2 show the results of the simulations with respect to bias and precision. It was noted that the bias and imprecision increased with the magnitude of unrestricted Pearson correlation coefficient r^u_{xy} and strictness of the selection process (lower selection ratios). Compared to the corrected correlations, the uncorrected correlations had higher bias values for all the selection ratios under consideration. Lower unrestricted correlation coefficients ($r^u_{xy} \leq 0.2$) and lower selection ratios (SR of ≤ 0.4) were marked with highly imprecise bias corrections. Lastly, the three corrective methods *Thorndike Case II, EM* and *MCMC MI* had equivalent performance for all the selection ratios and unrestricted correlation values investigated. The bias values observed for these methods was negligible for the selection ratios of ≥ 0.4. Table 5.3 summarises the impact of the strictness of the selection process and magnitude of the unrestricted correlation coefficient. Their effects on the bias and precision are shown before and after corrections for range restriction are applied.

		Increasing	
	Imputation model	Strictness of selection	r^u_{xy}
Mean bias (negative)	No correction	⇑	⇑
Precision	No correction	⇓	⇓
Mean bias (negative)	x (selection test)	⇑	⇓
Precision	x (selection test)	⇓	⇑

Table 5.3.: *Results of the impact of the strictness of the selection process and magnitude of unrestricted correlation r^u_{xy} under a DRR with selection based on x. Note that the bias and loss of precision are mitigated by the MI correction. This mitigation increases with magnitude of r^u_{xy}. The stricter the selection process, the more acute the bias and loss of precision. MI correction mitigates these to a great extent but does not eliminate them entirely.*

5.1.2. Indirect range restriction

5.1.2.1. Imputation with z (selection test) only

Consider a scenario of IRR where selection is based on predictor z but interest is the Pearson correlation coefficient between another predictor x and outcome y. As in section 5.1.1, P and n_{iter} were set at 25 and 50 respectively. A total of 5,000 Monte Carlo samples of 500 observations were generated from the multivariate normal distribution shown in equation 5.1.1. Note that since the multivariate normal distribution is standardised, all the associations computed are at the level of correlation coefficients. The correlation coefficients r^u_{zy}, r^u_{xy} were allowed to vary across the grid shown in Table 5.4. Notice that there were sixteen different combinations of the correlation coeffcients considered. These combinations were evaluated for each of the four selection ratios of $\{0,2,0.4,0.6,0.8\}$. For each sample, the correlation coefficient between x and z was randomly drawn from a uniform distribution $r^u_{xz} \sim U[0.1,0.3]$. This was to ensure a modest level of correlation between the two predictors. The implication of this was that the complications arising from *multicollinearity (intercorrelation)* were averted and that each of the predictor introduced independent information into the samples. The samples were used to determine the impact of the strictness of the selection process and magnitude of r^u_{xy} and r^u_{zy} on the different bias correction methods considered.

$$\begin{pmatrix} z \\ x \\ y \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & r^u_{zx} & r^u_{zy} \\ r^u_{zx} & 1 & r^u_{xy} \\ r^u_{zy} & r^u_{xy} & 1 \end{pmatrix}\right] \tag{5.1.1}$$

		r^u_{zy}			
		=0.2	=0.4	=0.6	=0.8
r^u_{xy}	=0.2	✓	✓	✓	✓
	=0.4	✓	✓	✓	✓
	=0.6	✓	✓	✓	✓
	=0.8	✓	✓	✓	✓

Table 5.4.: *The grid of values for the pairs of correlation r^u_{xy} and r^u_{zy} sampled in the simulation.*

Figures 5.3 to 5.6 show the results of the simulations. Comparisons were made between lack of correction for range restriction on the restricted correlation r^r_{xy} and corrections for range restriction made using *Thorndike Case III* and MI based on *EM* and *MCMC* algorithms. For MI correction, the imputation model consists of the predictor (selection test) z only. Generally, the bias observed was proportional to the magnitude of r^u_{xy} and r^u_{y} considered. The bias observed for the MI correction based on *EM* and *MCMC* algorithms were worse off compared to the restricted correlation (r^r_{xy}) and Thorndike Case III corrections. For the MI correction method, the bias increased with the magnitudes of r^u_{zy} considered, with bias peaking at $r^u_{zy} = 0.6$ and declining for values of $r^u_{zy} > 0.6$. This may be explained by the fact that the highly predictive power of selection variable z ($r^u_{zy} > 0.6$) compensates for the bias introduced by excluding predictor x from the imputation model. With respect to precision, Figures 5.3 to 5.6 show that MI correction based on *EM* and *MCMC* algorithms had the worst performance compared to the Thorndike Case III and restricted correlation. The loss of precision was proportional to the magnitudes of r^u_{xy} and r^u_{zy} considered. As expected, the strictness of the selection process (decreasing selection ratio) had the effect of increasing the bias and reducing the precision for all values of correlations considered for r^u_{xy} and r^u_{zy}.

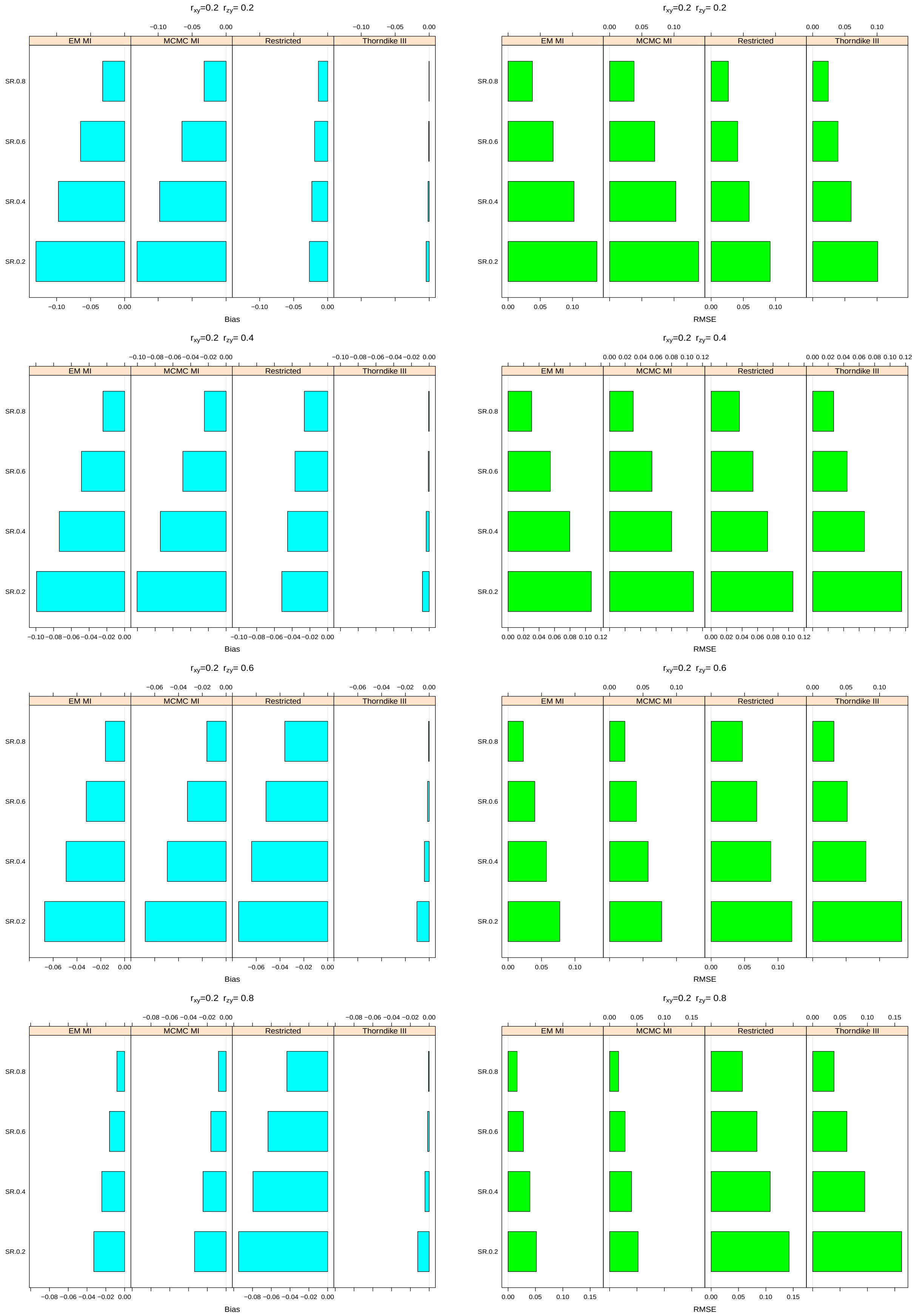


Figure 5.3.: *Results from the simulations exploring the impact of quantity of correlations of interest, r^{u}_{xy}=0.2 and varying magnitude of r^{u}_{zy} on performance measures mean bias and RMSE following selection on variable z and MI based on z only. Note that r^{u}_{xy} and r^{u}_{zy} are denoted by r_{xy} and r_{zy}.*

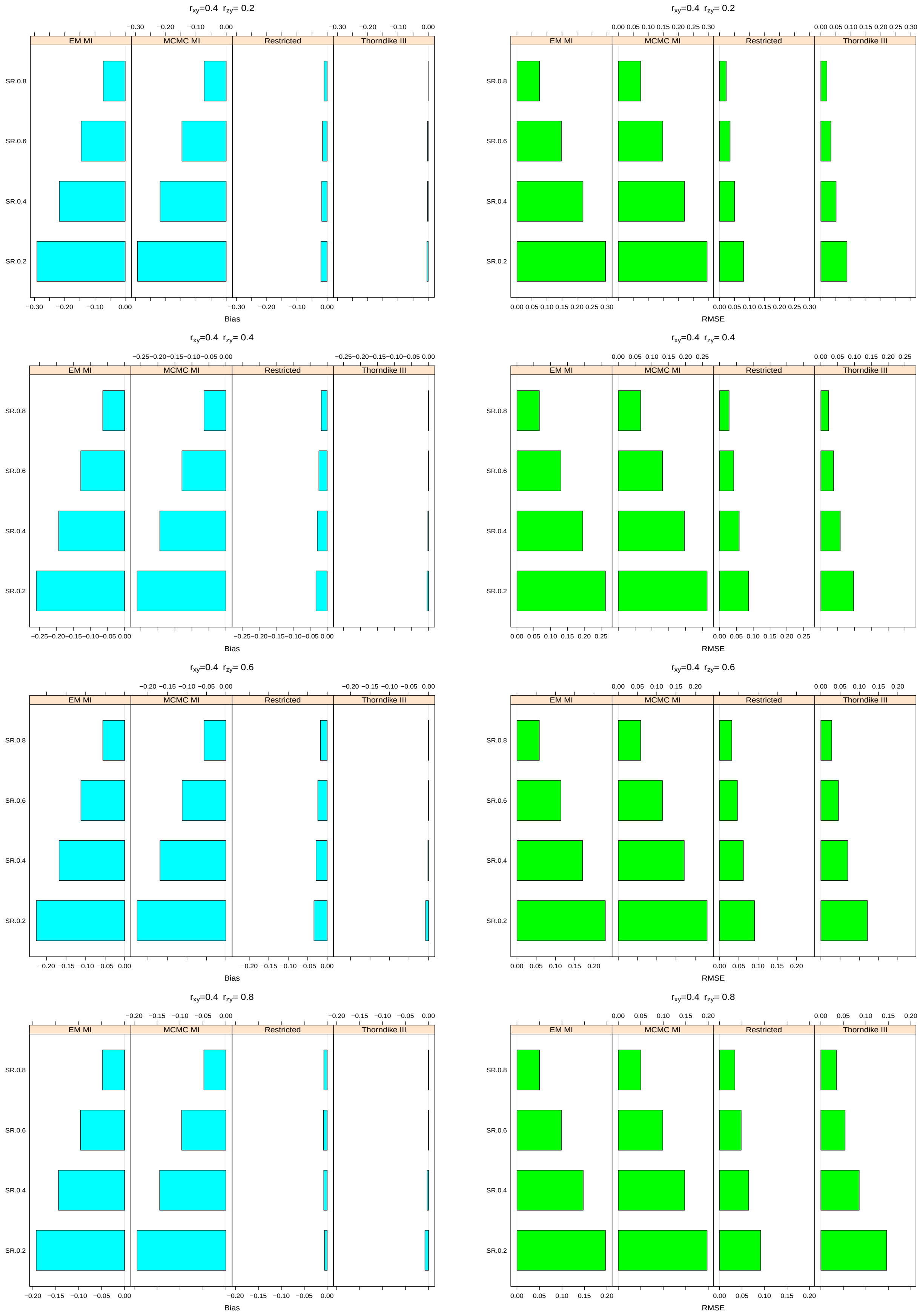


Figure 5.4.: *Results from the simulations exploring the impact of quantity of correlations of interest, r^u_{xy}=0.4 and varying magnitude of r^u_{zy} on performance measures mean bias and RMSE following selection on variable z and MI based on z only. Note that r^u_{xy} and r^u_{zy} are denoted by r_{xy} and r_{zy}.*

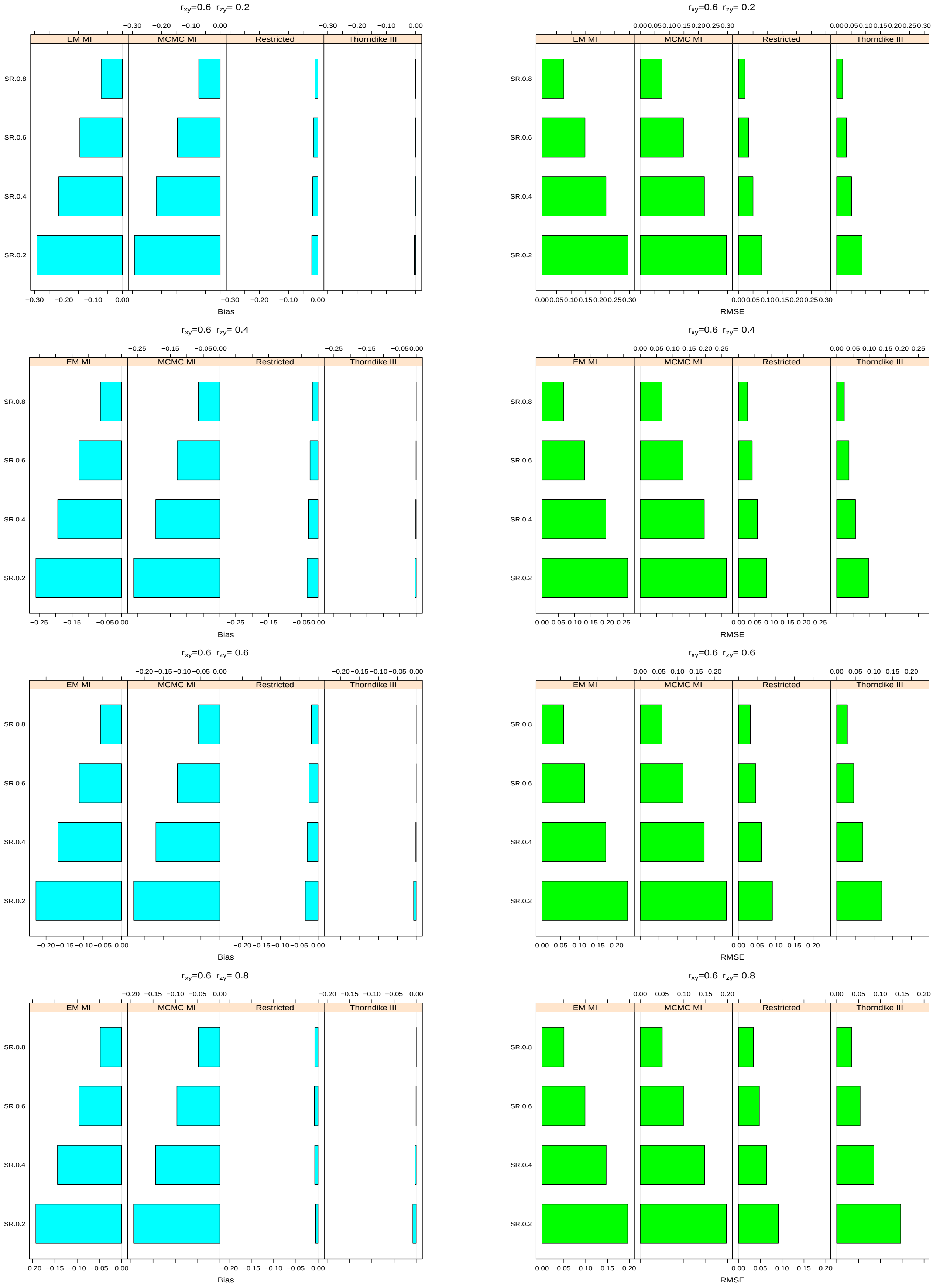


Figure 5.5.: *Results from the simulations exploring the impact of quantity of correlations of interest, r^u_{xy}=0.6 and varying magnitude of r^u_{zy} on performance measures mean bias and RMSE following selection on variable z and MI based on z only. Note that r^u_{xy} and r^u_{zy} are denoted by r_{xy} and r_{zy}.*

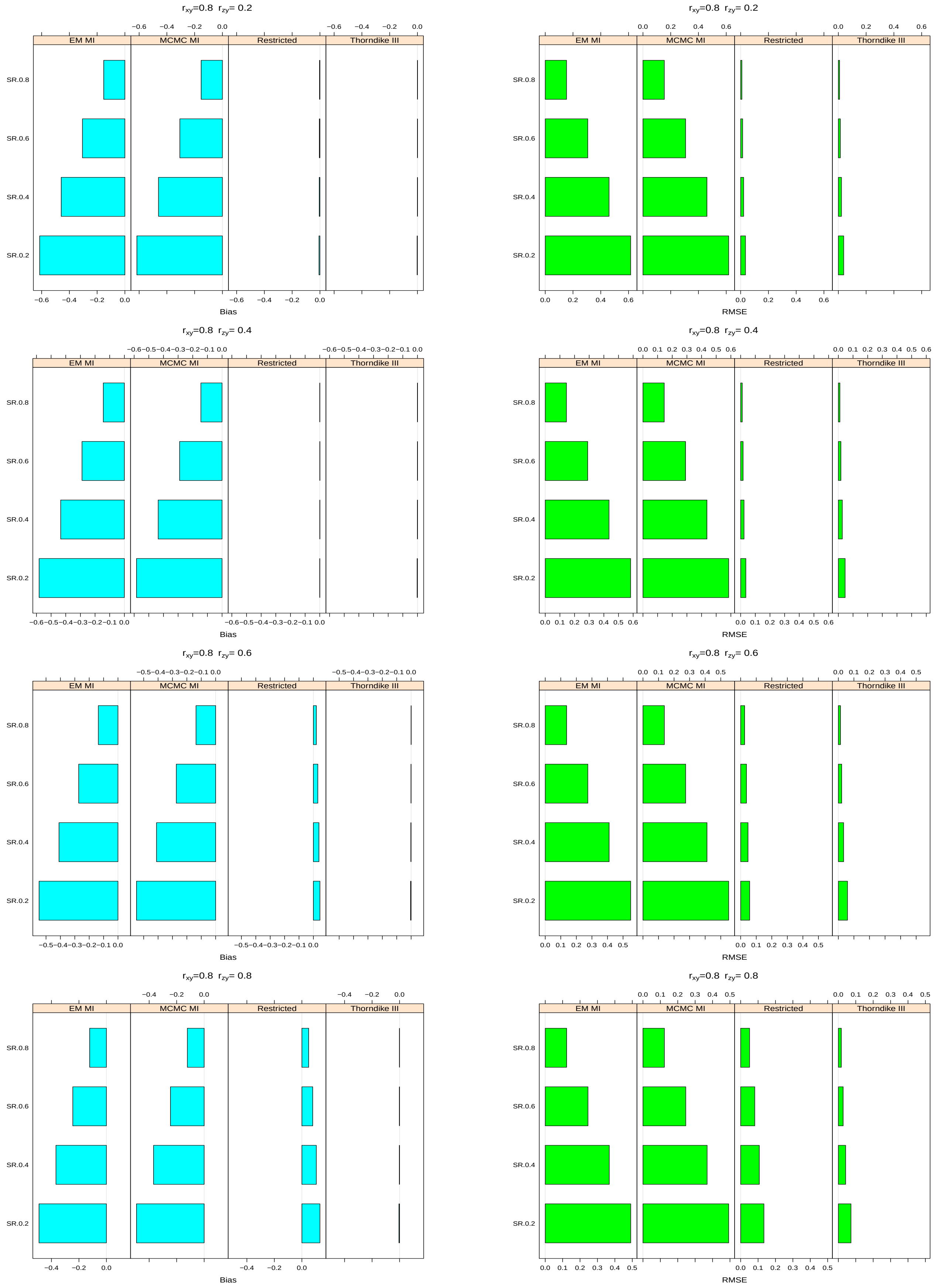


Figure 5.6.: *Results from the simulations exploring the impact of quantity of correlations of interest, r^u_{xy}=0.8 and varying magnitude of r^u_{zy} on performance measures mean bias and RMSE following selection on variable z and MI based on z only. Note that r^u_{xy} and r^u_{zy} are denoted by r_{xy} and r_{zy}.*

5.1.2.2. Imputation with z (selection test) and predictor x

Simulations were conducted as in section 5.1.2.1 but both the selection test z and predictor x were included in the imputation models for the MI correction. Figures 5.7 to 5.10 show the results of the simulations. Unlike in section 5.1.2.1, imputations were based on selection test z and predictor x. With respect to bias, it was observed that the MI corrections outperformed the restricted correlation (r^r_{xy}) and the Thorndike Case III correction as the magnitude of r^u_{xy} and r^u_{zy} were increased. The EM performed better than the MCMC imputation algorithm. With respect to precision, the MI based corrections had the best performance of all the methods considered, followed by the Thorndike Case III correction. The precision of the MI based corrections increased with the magnitude of r^u_{xy} and r^u_{zy}. As expected, there was a loss in precision for all the methods under considerations as the strictness of the selection was increased (decrease in selection ratio).

The results from Figures 5.3 and 5.10 are summarised in Table 5.5. Generally, for the case of IRR, the strictness of the selection process, larger magnitudes of the correlation coefficients between the selection test, predictor and outcome tend to worsen the bias and loss of precision. The correction of range restriction is best handled using a MI strategy that makes use of the selection test and predictor of interest under consideration.

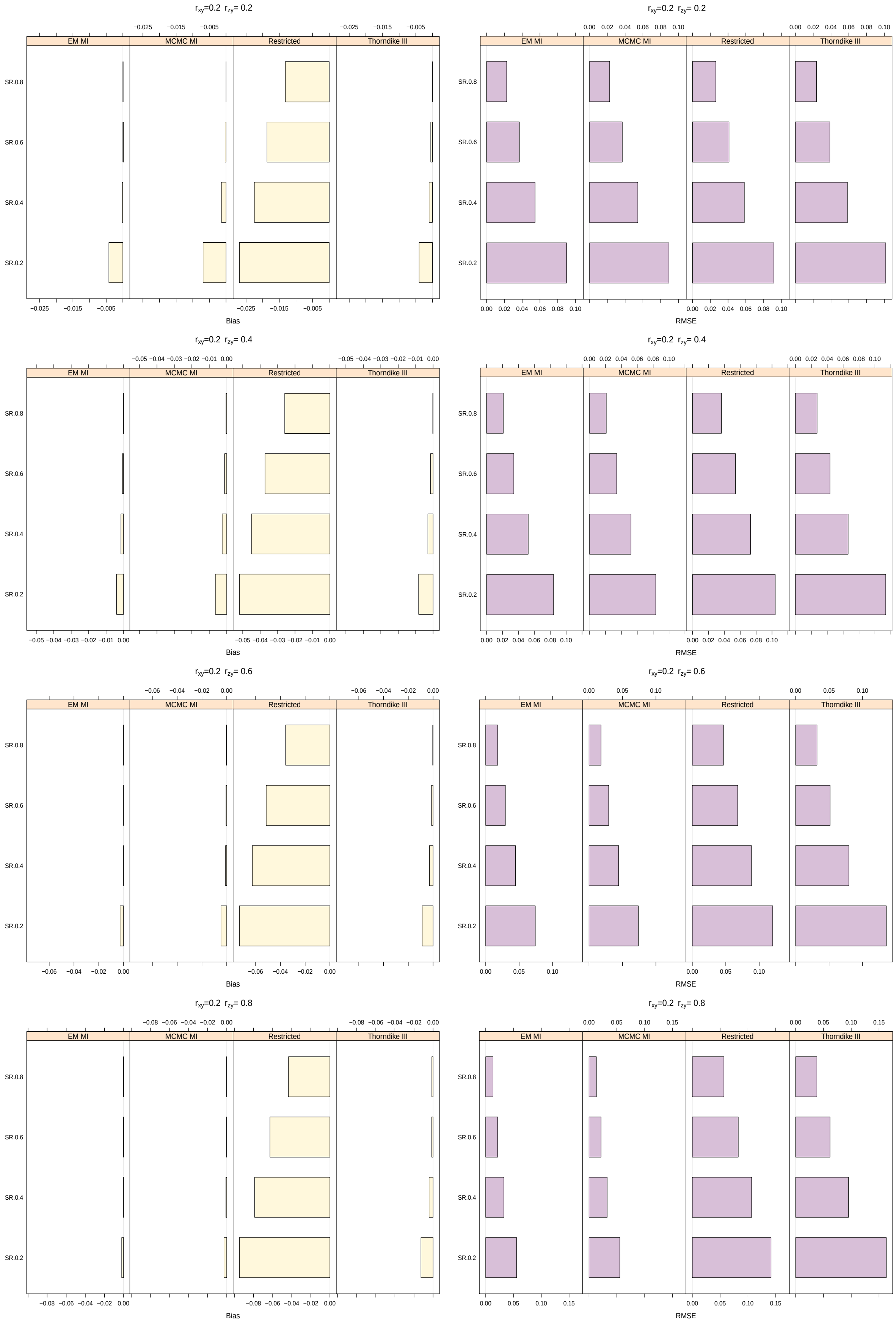


Figure 5.7.: *Results from the simulations exploring the impact of quantity of correlations of interest, r^u_{xy}=0.2 and varying magnitude of r^u_{zy} on performance measures mean bias and RMSE following selection on variable z and MI based on z and x. Note that r^u_{xy} and r^u_{zy} are denoted by r_{xy} and r_{zy}.*

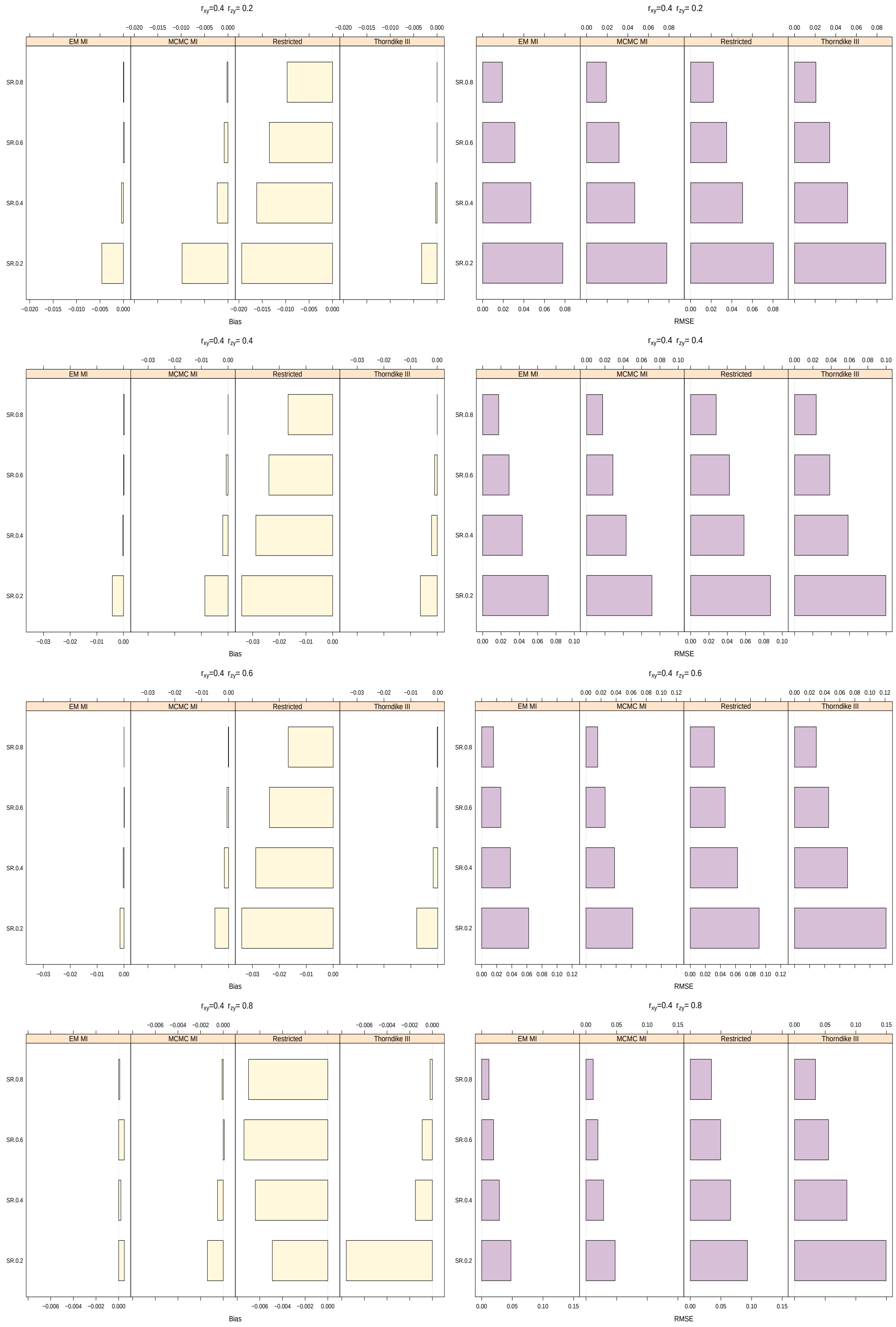


Figure 5.8.: *Results from the simulations exploring the impact of quantity of correlations of interest, r^u_{xy}=0.4 and varying magnitude of r^u_{zy} on performance measures mean bias and RMSE following selection on variable z and MI based on z and x. Note that r^u_{xy} and r^u_{zy} are denoted by r_{xy} and r_{zy}.*

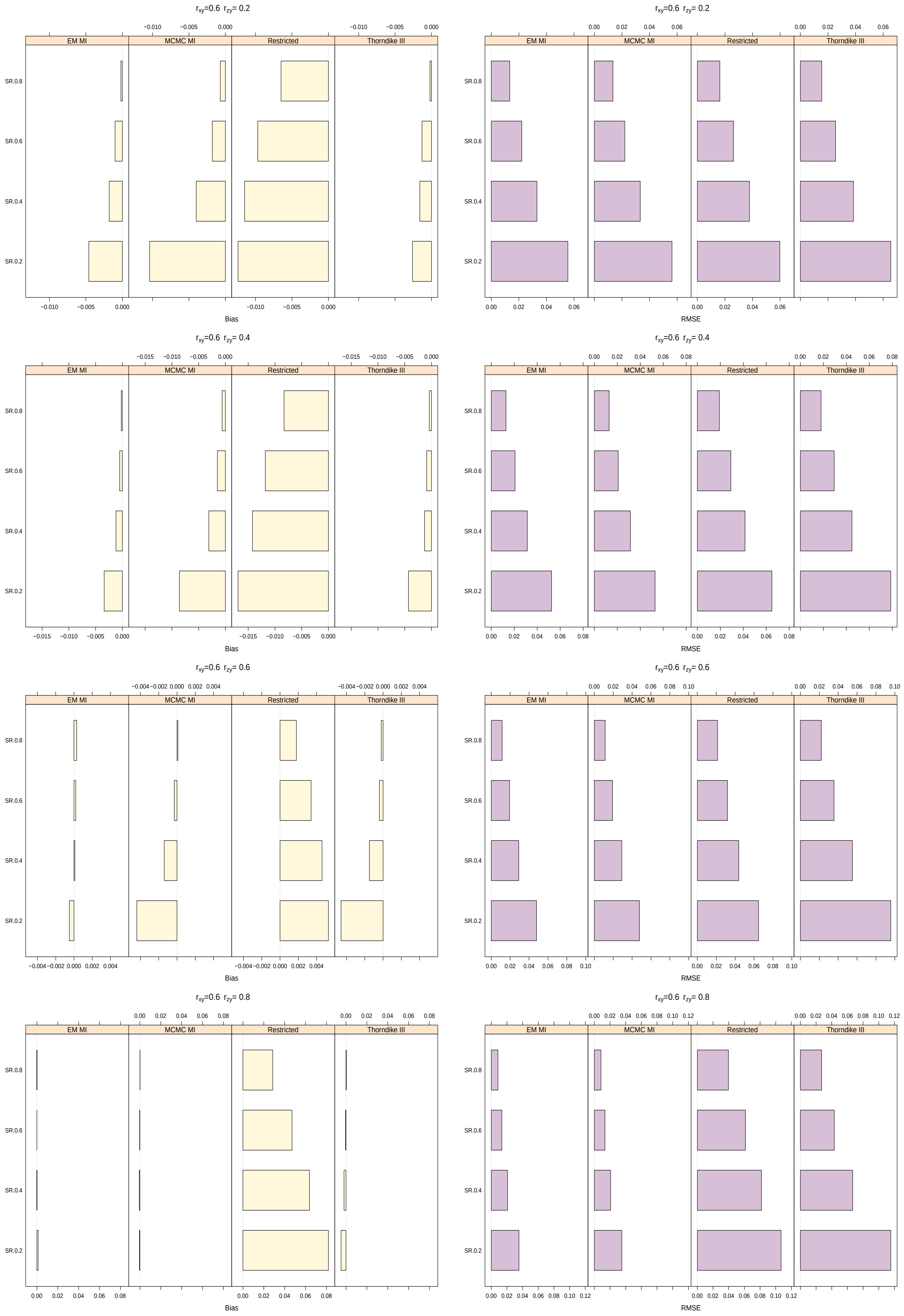


Figure 5.9.: *Results from the simulations exploring the impact of quantity of correlations of interest, r^{u}_{xy}=0.6 and varying magnitude of r^{u}_{zy} on performance measures mean bias and RMSE following selection on variable z and MI based on z and x. Note that r^{u}_{xy} and r^{u}_{zy} are denoted by r_{xy} and r_{zy}.*

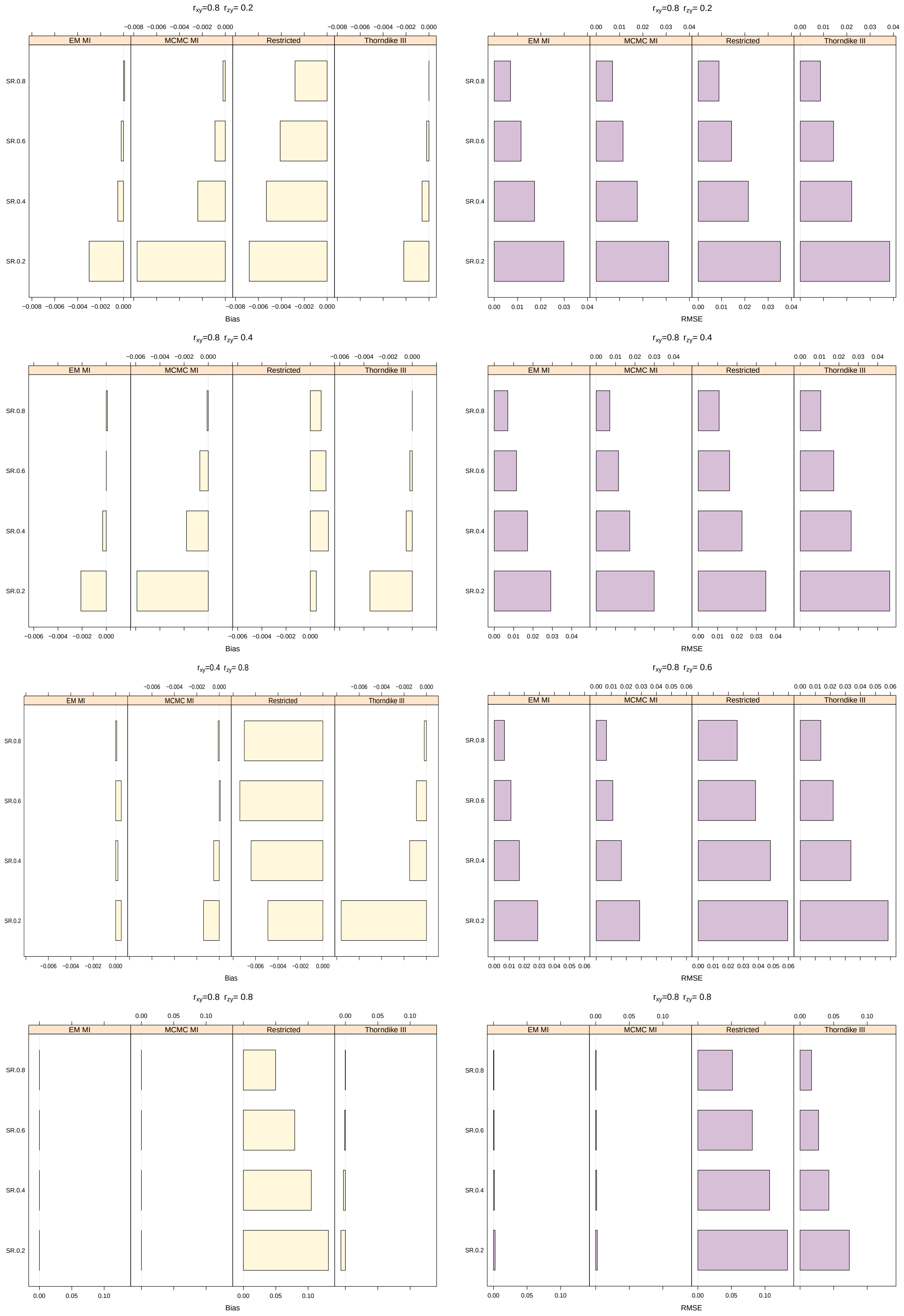


Figure 5.10.: *Results from the simulations exploring the impact of quantity of correlations of interest, r^u_{xy}=0.8 and varying magnitude of r^u_{zy} on performance measures mean bias and RMSE following selection on variable z and MI based on z and x. Note that r^u_{xy} and r^u_{zy} are denoted by r_{xy} and r_{zy}.*

		Increasing		
	Imputation model	Strictness of selection	r^u_{zy}	r^u_{xy}
Mean bias (negative)	No correction	⇑	⇑	⇑
Precision	No correction	⇓	⇓	⇓
Mean bias (negative)	z (selection test)	⇑	⇑	⇑
	z (selection test) and x	⇑	⇓	⇓
Precision	z (selection test)	⇓	⇓	⇓
	z (selection test) and x	⇓	⇑	⇑

Table 5.5.: *Results of the impact of the strictness of the selection process and magnitude of unrestricted correlations r^u_{zy} and r^u_{xy} under an IRR with selection based on z. Note that the bias and loss of precision are mitigated by the MI correction only when both z and x are in the imputation model. This mitigation increases with magnitude of r^u_{zy} and r^u_{xy}. The stricter the selection process, the more acute the bias and loss of precision. MI correction mitigates these to a great extent but does not eliminate them entirely.*

5.2. Testing phase for the expected performance of the different methods for correcting range restriction

In sections 5.1.1 and 5.1.2, the impact of the magnitudes of the unrestricted correlations were evaluated. The magnitude of the unrestricted correlations, under a DRR and IRR, were found to be directly proportional to the bias and loss of precision under all the selection ratios as was summarised in Tables 5.3 and 5.5. In the following sections, the expected (average) performance of the different methods of correcting for range restriction will be evaluated. This will be accomplished by assuming the unrestricted correlations are drawn from a population of plausible values randomly rather than fixing them to $\{0.2, 0.4, 0.6, 0.8\}$. In addition, for the MI correction method, the impact of including *auxiliary variables* (described in section 4.5.7) in the imputation model will be evaluated.

5.2.1. Simulation results for correlation coefficients for the predictive validity selection design: Thorndike Case II vs MI

To empirically explore the performance of the MI correction method, data were simulated consisting of a predictor (selection test) x and outcome (criterion) y. In addition, *auxiliary variables t, u, v* and w were also simulated from the trivariate normal distributions shown in equations (5.2.1), (5.2.2), (5.2.3) and (5.2.4).

$$\begin{pmatrix} t \\ x \\ y \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & r^u_{tx} & 0.2 \\ r^u_{tx} & 1 & r^u_{xy} \\ 0.2 & r^u_{xy} & 1 \end{pmatrix} \right] \tag{5.2.1}$$

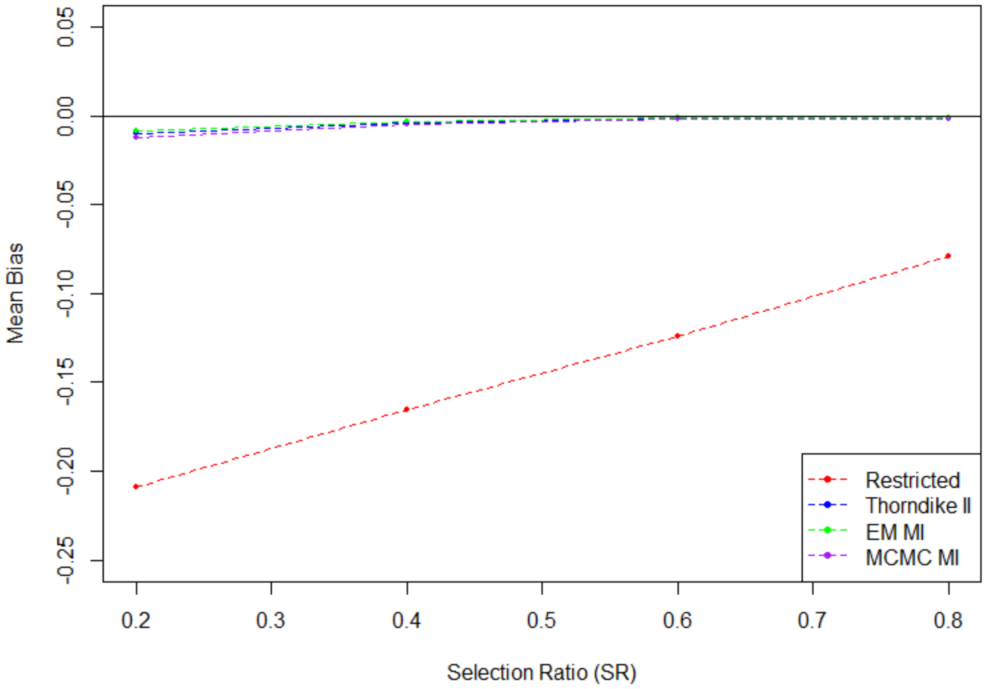
$$\begin{pmatrix} u \\ x \\ y \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & r^u_{ux} & 0.4 \\ r^u_{ux} & 1 & r^u_{xy} \\ 0.4 & r^u_{xy} & 1 \end{pmatrix} \right] \tag{5.2.2}$$

$$\begin{pmatrix} v \\ x \\ y \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & r^u_{vx} & 0.6 \\ r^u_{vx} & 1 & r^u_{xy} \\ 0.6 & r^u_{xy} & 1 \end{pmatrix} \right] \tag{5.2.3}$$

$$\begin{pmatrix} w \\ x \\ y \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & r^u_{wx} & 0.8 \\ r^u_{wx} & 1 & r^u_{xy} \\ 0.8 & r^u_{xy} & 1 \end{pmatrix} \right] \tag{5.2.4}$$

Note that because each of the trivariate normal distribution is standardised, the covariance matrix specified is equal to the correlation matrix. It was assumed that $r^u_{xy} \sim U(0.1, 0.9)$. The correlations $r^u_{ty}, r^u_{uy}, r^u_{vy}, r^u_{wy}$ were fixed at $\{0.2, 0.4, 0.6, 0.8\}$ respectively. The variables (t,u,v,w) were allowed to covary randomly with x uniformly between 0.1 and 0.3. The choice of modest correlations of (t,u,v,w) with x ensured that complications arising from *multicollinearity (intercorrelation)* were averted and that each of the variables introduced independent information into the samples. A total of 5,000 samples of 500 observations were simulated. For the MI correction, P was set at 25. The convergence stopping rule was tolerance of $1 * 10^{-6}$ and $n_{iter} = 50$ for the EM and MCMC algorithms respectively.

Figure 5.11 shows the results from the simulations. The MI correction based on EM and MCMC algorithm had only the selection test x in the imputation model. It was observed that the mean bias was inversely proportional to the selection ratio. This implied that the degree of bias increased with strictness of selection process. Note that the mean bias is negative which is testament to the attenuated correlation, with the degree of attenuation being greatest when no corrective method was used. Precision of the corrective methods was lowest when the selection ratio was the smallest. The lowest precision across the different selection ratios was for the case where no correction for the effect of range restriction was applied. The mean bias and RMSE seemed somewhat equivalent for the three corrective considered.



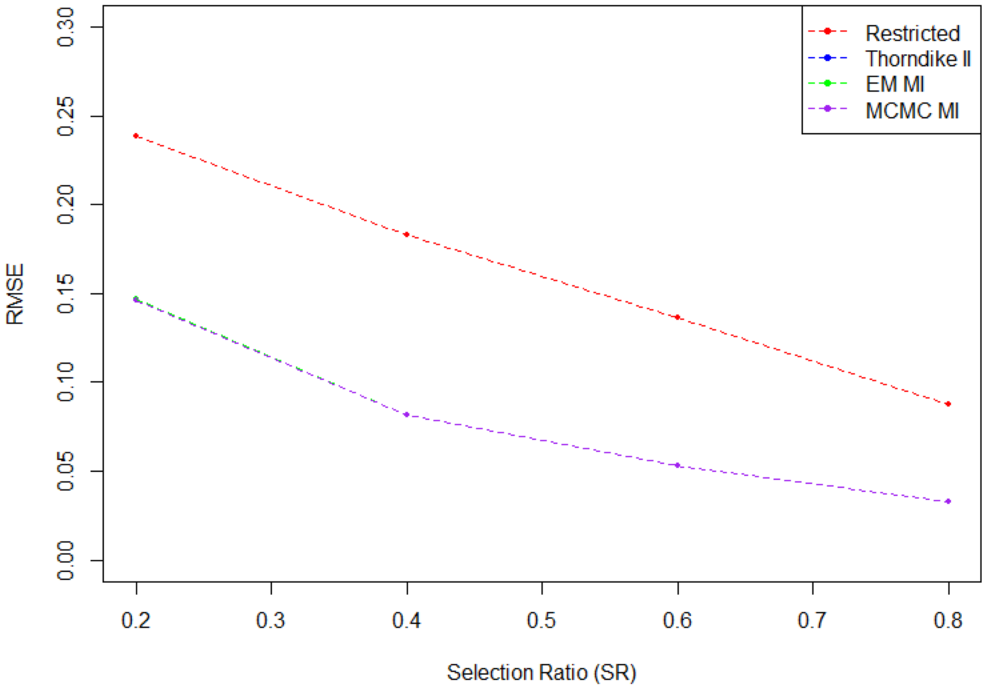


Figure 5.11.: *Mean bias and RMSE for the methods under consideration under direct range restriction for the selection ratios of 0.2, 0.4, 0.6 and 0.8. The imputation model for EM and MCMC MI algorithm contains only variable x*

	T values			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case II	-75.7052	-100.8353	-111.9771	-113.3640
Restricted vs EM MI	-76.0852	-101.0720	-111.9989	-113.1165
Restricted vs MCMC MI	-75.0992	-100.3068	-111.6402	-112.7271
Thorndike Case II vs EM MI	-0.4929	-0.4275	-0.1970	-0.0969
Thorndike Case II vs MCMC MI	0.6726	0.4492	0.2334	0.2938
EM MI vs MCMC MI	1.1650	0.8755	0.4298	0.3894
	F values			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case II	2.6645	5.0319	6.6160	7.2330
Restricted vs EM MI	2.6489	4.9978	6.5745	7.1372
Restricted vs MCMC MI	2.6751	5.0135	6.5865	7.1270
Thorndike Case II vs EM MI	0.9941	0.9932	0.9937	0.9868
Thorndike Case II vs MCMC MI	1.0040	0.9963	0.9955	0.9853
EM MI vs MCMC MI	1.0099	1.0031	1.0018	0.9986

Table 5.6.: *T and F-test comparison of the methods under evaluation under direct range restriction with imputation based only on variable x for bias and MSE respectively. The T and F values highlighted in green were significant with p-values of less than 0.0001*

Table 5.6 shows results of formal statistical comparison of the methods. As already observed in Figure 5.11, the case where no correction for range restriction was applied (restricted correlation) yielded inferior performance compared to the other three methods. The three methods, *Thorndike Case II*, MI correction based on EM and MCMC algorithm were equivalent in performance. Table 8.8 in the Technical Appendices shows the bias of MI correction (based on the EM and MCMC algorithms) compared to the *Thorndike Case II* method and for the uncorrected (restricted correlation). The *auxiliary variables t,u,v* and *w* were had correlations with the criterion (outcome) of $\{0.2, 0.4, 0.6, 0.8\}$ respectively. It was observed from the results of T-tests that, with respect to bias, the *Thorndike Case II* performed equally as well as the MI correction. In addition, the imputation of the criterion with the pair of selection test x and any of the *auxiliary variables* did not improve the performance of the MI correction except for *x, w* combination ($r^u_{wy} = 0.8$). Inclusion of (*x, w*) in the imputation model, resulted in the MI correction having less bias than the *Thorndike Case II* correction for the selection ratios of 0.2 and 0.4.

Table 8.9 in the Technical Appendices shows the performance of correction methods with regards to MSE. It was observed from the results of the F tests that inclusion of (*t, x*) in the imputation model, resulted in MSE that was equivalent to when only x was included in the imputation model. This is because *auxiliary variable*, t, had a low correlation with the outcome ($r^u_{ty} = 0.2$). Therefore, its inclusion in the imputation model did not improve prediction. It was noted that the performance of MI correction was superior to *Thorndike Case II* correction when the imputation was done with each of the remaining *auxiliary variables* alongside the selection test x. This was because $r^u_{uy} = 0.4, r^u_{vy} = 0.6$ and $r^u_{wy} = 0.8$ significantly increased the predictive capability of the imputation model. There was no evidence of difference in performance of EM and MCMC algorithms.

5.2.2. Simulation results for correlation coefficients for the predictive validity selection design: Thorndike Case III vs MI

As before, a total of 5,000 Monte Carlo samples consisting of 500 observations were generated from a multivariate standard normal distribution expressed by equation (5.2.5). Under the IRR, interest was the estimation of the correlation between x and y following selection based on variable z. The *auxiliary variables t, u, v* and w were also simulated to assess the impact of including variables in the imputation model that had varying predictive ability for the outcome (criterion) y. Data for this purpose were generated from the multivariate standard normal distributions expressed in equations (5.2.6) to (5.2.9).

$$\begin{pmatrix} z \\ x \\ y \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & r^u_{zx} & r^u_{zy} \\ r^u_{zx} & 1 & r^u_{xy} \\ r^u_{zy} & r^u_{xy} & 1 \end{pmatrix}\right] \tag{5.2.5}$$

$$\begin{pmatrix} t \\ z \\ x \\ y \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & r^u_{tz} & r^u_{tx} & 0.2 \\ r^u_{tz} & 1 & r^u_{zx} & r^u_{zy} \\ r^u_{tx} & r^u_{zx} & 1 & r^u_{xy} \\ 0.2 & r^u_{zy} & r^u_{xy} & 1 \end{pmatrix}\right] \tag{5.2.6}$$

$$\begin{pmatrix} u \\ z \\ x \\ y \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & r^u_{uz} & r^u_{ux} & 0.4 \\ r^u_{uz} & 1 & r^u_{zx} & r^u_{zy} \\ r^u_{ux} & r^u_{zx} & 1 & r^u_{xy} \\ 0.4 & r^u_{zy} & r^u_{xy} & 1 \end{pmatrix}\right] \tag{5.2.7}$$

$$\begin{pmatrix} v \\ z \\ x \\ y \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & r^u_{vz} & r^u_{vx} & 0.6 \\ r^u_{vz} & 1 & r^u_{zx} & r^u_{zy} \\ r^u_{vx} & r^u_{zx} & 1 & r^u_{xy} \\ 0.6 & r^u_{zy} & r^u_{xy} & 1 \end{pmatrix}\right] \tag{5.2.8}$$

$$\begin{pmatrix} w \\ z \\ x \\ y \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & r^u_{wz} & r^u_{wx} & 0.8 \\ r^u_{wz} & 1 & r^u_{zx} & r^u_{zy} \\ r^u_{wx} & r^u_{zx} & 1 & r^u_{xy} \\ 0.8 & r^u_{zy} & r^u_{xy} & 1 \end{pmatrix} \right] \tag{5.2.9}$$

It was assumed that $r^u_{xy} \sim U(0.1, 0.9)$, $r^u_{zy} \sim U(0.1, 0.9)$ and $r^u_{ty}, r^u_{uy}, r^u_{vy}, r^u_{wy}$ fixed at $\{0.2, 0.4, 0.6, 0.8\}$ respectively. The *auxiliary variables* were drawn from a uniform distribution *t,u,v,w* $\sim U[0.1, 0.3]$. This ensured that multicollinearity problems did not occur and that each of the variables introduced independent information into the data. Selection ratios were set at 0.2, 0.4, 0.6 and 0.8. For the MI correction, P set at 25. The convergence stopping rule was tolerance of $1 * 10^{-6}$ and $n_{iter} = 50$ for the EM and MCMC algorithms respectively. Since the missingness mechanism was assumed to be MAR, all imputation models included z which induced missingness in y through selection. Therefore performance of the imputations were evaluated for the variables *z, xz, xzt, xzu, xzv* and *xzw*.

Figure 5.12 and 5.13 show the performance of the methods when no correction for range restriction was applied (restricted correlation), for *Thorndike Case III* and MI correction. In Figure 5.12, the MI method included only the selection variable z in the imputation model while in Figure 5.13, the MI method included the selection variable z and x in the imputation model. It was observed that including only variable z in the imputation model results in extremely poor performance for the MI method. In fact, imputation with only variable z results in performance poorer than not applying any correction for the effects IRR whatsoever. Imputation with z and x result in superior performance for the MI with respect to RMSE although there seems to be little difference in terms of bias between MI and the *Thorndike Case III* correction. In this instance, MI method performs far better than not correcting for the effects of IRR both in terms of bias and RMSE. Table 5.7 confirms by means of formal statistical T and F-tests that when compared to the *Thorndike Case III* correction, the MI fairs poorly with respect to bias and MSE when imputation is based only on the selection variable z. Under this scenario, the *Thorndike Case III* correction has the least amount of bias followed by ignoring the effect of IRR. For selection ratio of 0.2, the *Thorndike Case III* correction had similar bias levels to that of the restricted correlation (effect of IRR ignored). With respect to MSE, the *Thorndike Case III* correction outperformed all the other corrections apart from when the effect of IRR was ignored. Lastly, as expected, MI based on EM and MCMC algorithms were equivalent in performance.

Table 5.8 shows the results of the simulation when the imputation model contains the selection variable z and x. In line with Figure 5.13, based on the T and F-tests, it was observed that in terms of bias the three methods for correcting for the effects of range restriction were better than ignoring the effects of IRR altogether. The *Thorndike Case III* correction was equivalent to MI based on EM and MCMC algorithms in terms of bias although unexpectedly the EM outperformed the MCMC algorithm for strict selection ratio of 0.2. This may be a result of the different convergence rules adopted for the two algorithms. That is tolerance of $1 * 10^{-6}$ and $n_{iter} = 50$ for the EM and MCMC algorithm respectively. The implication was that the MCMC algorithm needed more than the specified 50 iterations in order to achieve parity with the EM algorithm. Just like in the case of imputation with only the selection test z, imputation based on z and x did not result in significant differences in MSE when the *Thorndike Case III* correction was used compared to when IRR was ignored altogether under strict selection ratios of 0.2. Overall, it was observed that in terms MSE, the *Thorndike Case III* correction was better than ignoring the effects of indirect IRR.

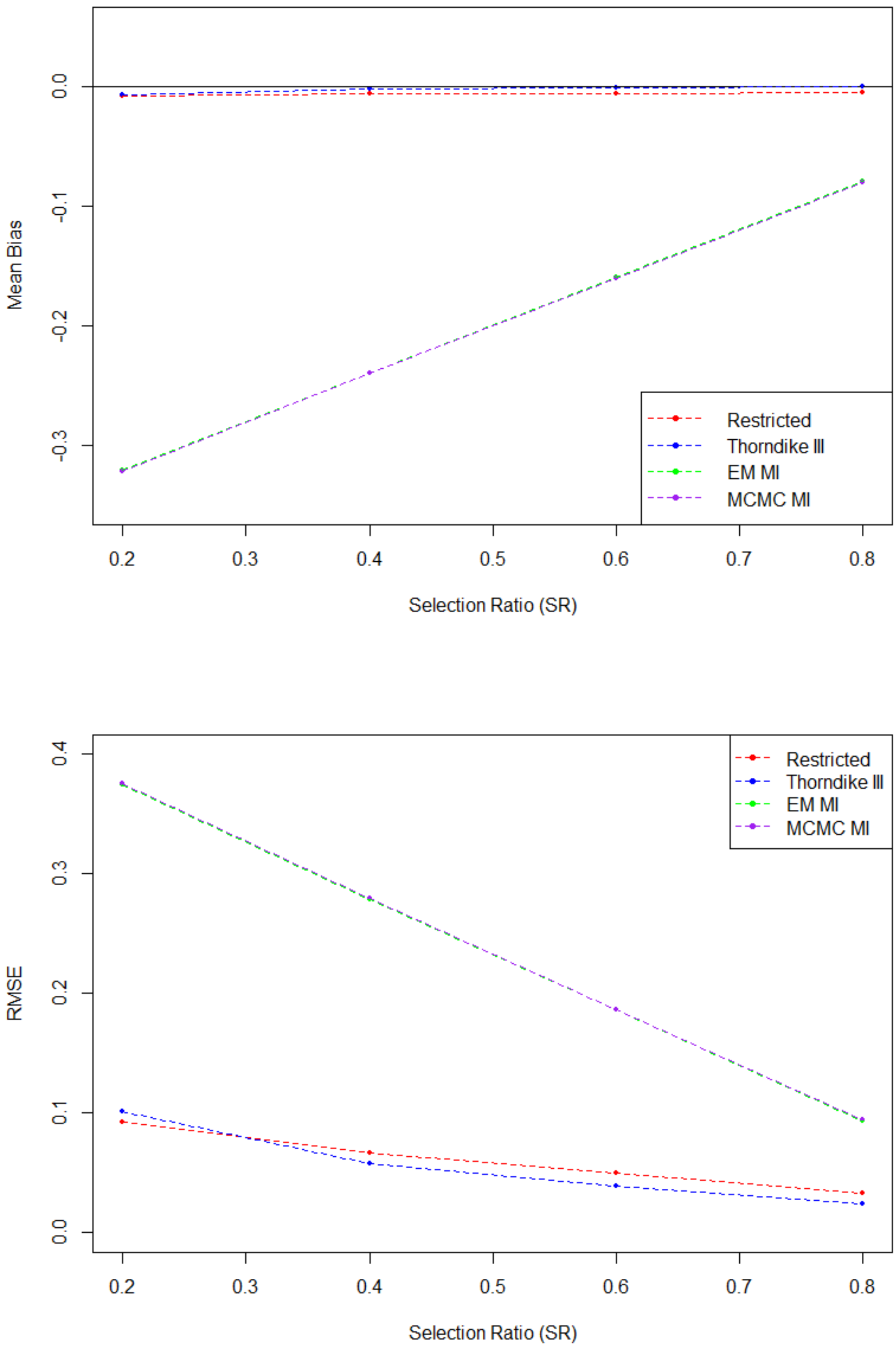


Figure 5.12.: *Mean bias and RMSE for the methods under consideration under indirect range restriction for the selection ratios of 0.2, 0.4, 0.6 and 0.8. The imputation model for EM and MCMC MI algorithm contains only variable z*

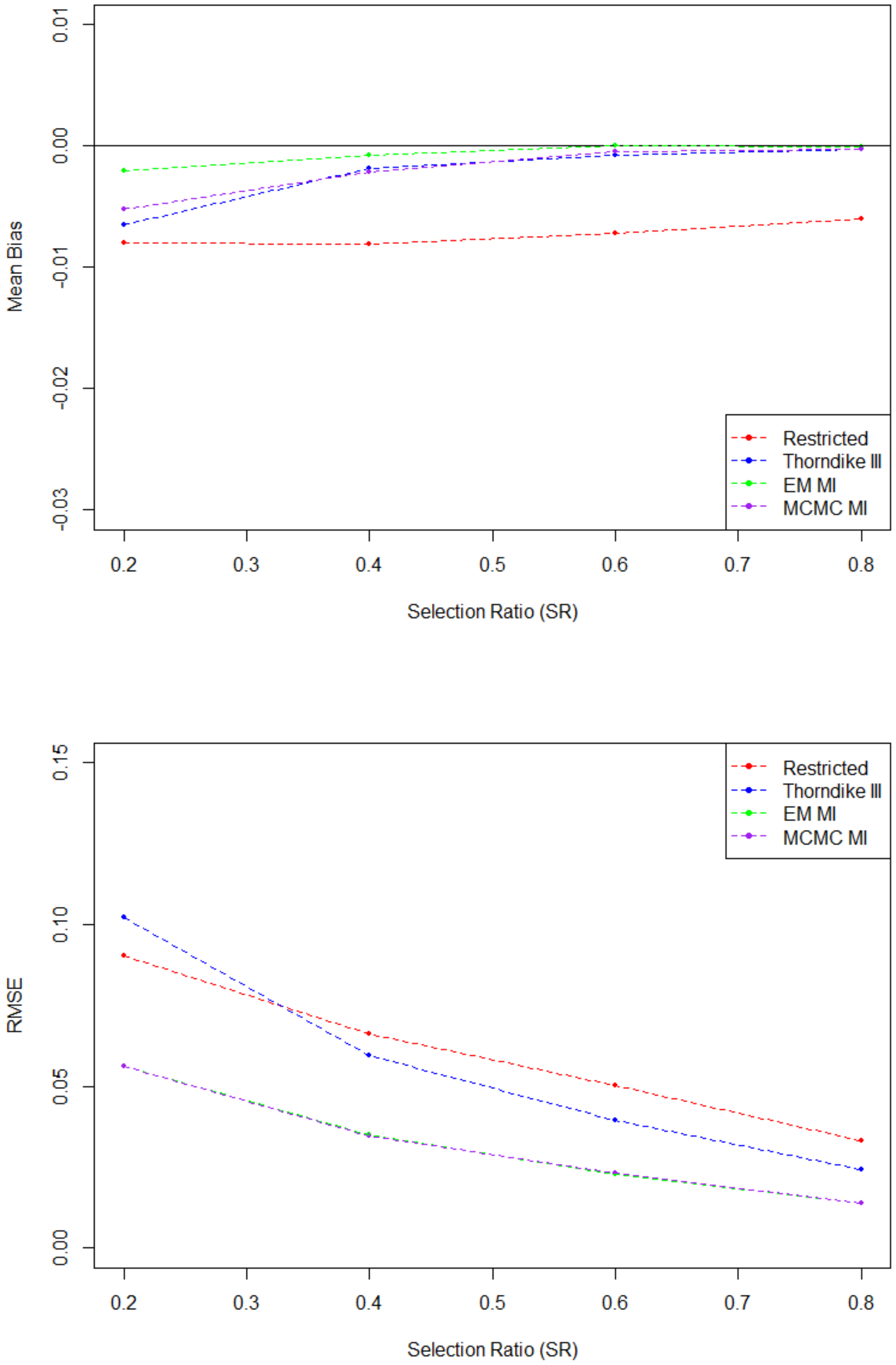


Figure 5.13.: *Mean bias and RMSE for the methods under consideration under indirect range restriction for the selection ratios of 0.2, 0.4, 0.6 and 0.8. The imputation model for EM and MCMC MI algorithm contains variable z and x*

	T values			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case III	-0.7062	-3.8270	-6.4187	-8.7500
Restricted vs EM MI	103.5435	104.1378	99.6600	88.4733
Restricted vs MCMC MI	103.8271	104.1724	99.7073	88.7336
Thorndike Case III vs EM MI	102.2099	108.6502	107.9307	101.9115
Thorndike Case III vs MCMC MI	102.4933	108.6698	107.9463	102.1800
EM MI vs MCMC MI	0.3448	0.2085	0.2355	0.2605

	F values			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case III	0.8446	1.3350	1.6679	1.9085
EM MI vs Restricted	16.2265	17.4203	13.7652	7.9716
MCMC MI vs Restricted	16.3414	17.5096	13.8461	8.0124
EM MI vs Thorndike Case III	13.7042	23.2557	22.9596	15.2138
MCMC MI vs Thorndike Case III	13.8014	23.3750	23.0946	15.2917
EM MI vs MCMC MI	0.9930	0.9949	0.9942	0.9949

Table 5.7.: *T and F-test comparison of the methods under evaluation under indirect range restriction with imputation based only on variable z for bias and MSE respectively. The T and F values highlighted in green were significant with p-values of less than 0.0001*

	T values			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case III	-0.7789	-4.9785	-7.173	-9.9244
Restricted vs EM MI	-4.0094	-6.8764	-9.3873	-11.7522
Restricted vs MCMC MI	-1.8461	-5.7305	-8.7192	-11.3029
Thorndike Case III vs EM MI	-2.7419	-1.0102	-1.2816	-0.4979
Thorndike Case III vs MCMC MI	-0.7700	0.2365	-0.4863	0.0784
EM MI vs MCMC MI	2.8997	1.7527	1.1170	0.8166
	F values			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case III	0.7807	1.2303	1.6161	1.8788
Restricted vs EM MI	2.5800	3.5932	4.7859	5.6493
Restricted vs MCMC MI	2.5905	3.6505	4.7547	5.6746
Thorndike Case III vs EM MI	3.3047	2.9205	2.9614	3.0068
Thorndike Case III vs MCMC MI	3.3182	2.9671	2.9421	3.0202
EM MI vs MCMC MI	1.0041	1.0160	0.9935	1.0045

Table 5.8.: *T and F-test comparison of the methods under evaluation under indirect range restriction with imputation based on variable z and x for bias and MSE respectively. The T and F values highlighted in green were significant with p-values of less than 0.0001*

An interesting question is whether the performance of MI correction based on an imputation model containing z and x may be further improved by including in the imputation model a third variable that is more predictive of the outcome (criterion) y? Therefore comparisons were made for when the imputation model included variables *xz, xzt, xzu, xzv* and *xzw*. Recall that the $r^u_{ty}, r^u_{uy}, r^u_{vy}$ and r^u_{wy} were fixed at $\{0.2, 0.4, 0.6, 0.8\}$ respectively. Tables 8.10 and 8.11 in the Technical Appendices show the results of these comparisons with respect to bias and MSE respectively. For the most part, the successive addition of a third variable more predictive of the outcome (criterion) y in the imputation model had little effect in reducing bias. There were differences in bias between MI based on EM and MCMC algorithms for the selection ratio of

0.2 highlighting the fact that for lower selection ratios, more than 50 iterations were required for the MCMC algorithm to match the performance of the EM algorithm. With respect to MSE, the successive addition of a third variable more predictive of the outcome (criterion) y in the imputation model increased the precision of the estimation only for variables v and w. It was noted that the observed F values for variable w were larger in magnitude than for variable v which is indicative of the precision gain resulting from use of the highly predictive variable w.

5.2.3. Simulation results for correlation coefficients for the two hurdle validity selection design: Pearson Lawley vs FIML vs MI

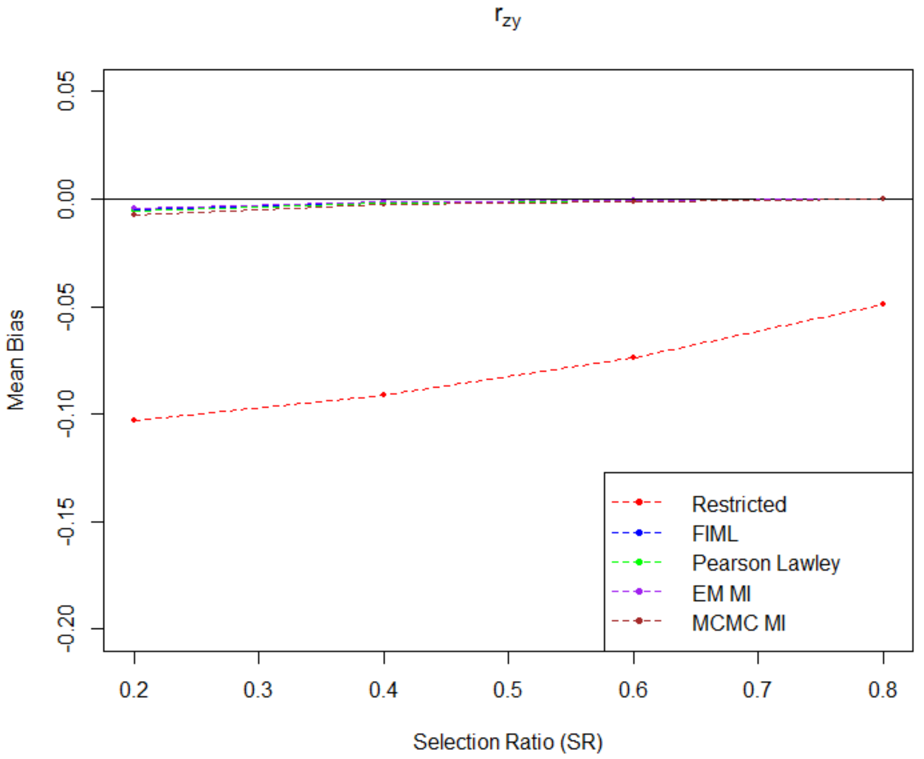
5.2.3.1. Full information on selection tests *z* and *x*

Simulations were conducted as per the conceptual representation of the *two hurdle validity selection design* in Figure 3.7. In the simulations, data were generated from the multivariate standard normal distribution in equation 5.2.10. A total of 5,000 Monte Carlo samples consisting of 500 observations were generated. It was assumed that $r^{u}_{zy} \sim U(0.1, 0.9)$, $r_{xy} \sim U(0.1, 0.9)$ and $r_{zx} \sim U(0.1, 0.3)$. The selection ratios of $\{0.2, 0.4, 0.6, 0.8\}$ were considered. For the MI correction, P set at 25. The convergence stopping rule was tolerance of $1 * 10^{-6}$ and $n_{iter} = 50$ for the EM and MCMC algorithms respectively. It was assumed that from the onset, full information was available for the two tests z and x as a result of all applicants sitting for these tests before the selection process was commenced. Note that the selection of the n_1 entrants was conducted in two stages (hurdles). In the first stage (hurdle), $n_1 + n_2$, applicants were selected based on their scores on test z. In the second stage (hurdle) n_2 applicants were rejected based on their scores on test x thus ending up with n_1 entrants. For example, to achieve a selection ratio of 0.8, ten percent were rejected by first selecting those in the top 90% of the score distribution for test z. Thereafter, of those in the top 90% of the score distribution for test z, the appropriate number were rejected based on their scores on test x to achieve the targeted selection ratio of $\frac{n_1}{n_1+n_2+n_3}$=0.8. For the simulations, interest was the correlation coefficients between the selection tests and the outcome, that is, correlation coefficients containing variables z, y and x, y

respectively. For the MI correction, both z and x were included in the imputation models.

$$\begin{pmatrix} z \\ x \\ y \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & r^u_{xy} & r^u_{zy} \\ r^u_{xy} & 1 & r^u_{xy} \\ r^u_{zy} & r^u_{xy} & 1 \end{pmatrix}\right] \tag{5.2.10}$$

Figure 5.14 and 5.15 depict the bias and the RMSE of the correction methods under consideration. It was observed that the restricted correlation (correlation not corrected for the effect of range restriction) faired the worst in terms of bias and RMSE as expected. All the other methods of correcting for range restriction, that is, *Pearson Lawley*, FIML and MI based on EM and MCMC algorithms were equivalent with respect to the bias and RMSE. Confirmation of this may be viewed in Tables 8.12 and 8.13 in the Technical Appendices which show that there were no statistical significant differences between the correction methods. All of these methods have equivalent performance and are by far better than ignoring the effects of range restriction altogether. The equivalence in performance between the aforementioned methods stem from the fact that they all utilised the full information available on the selection tests z and x.



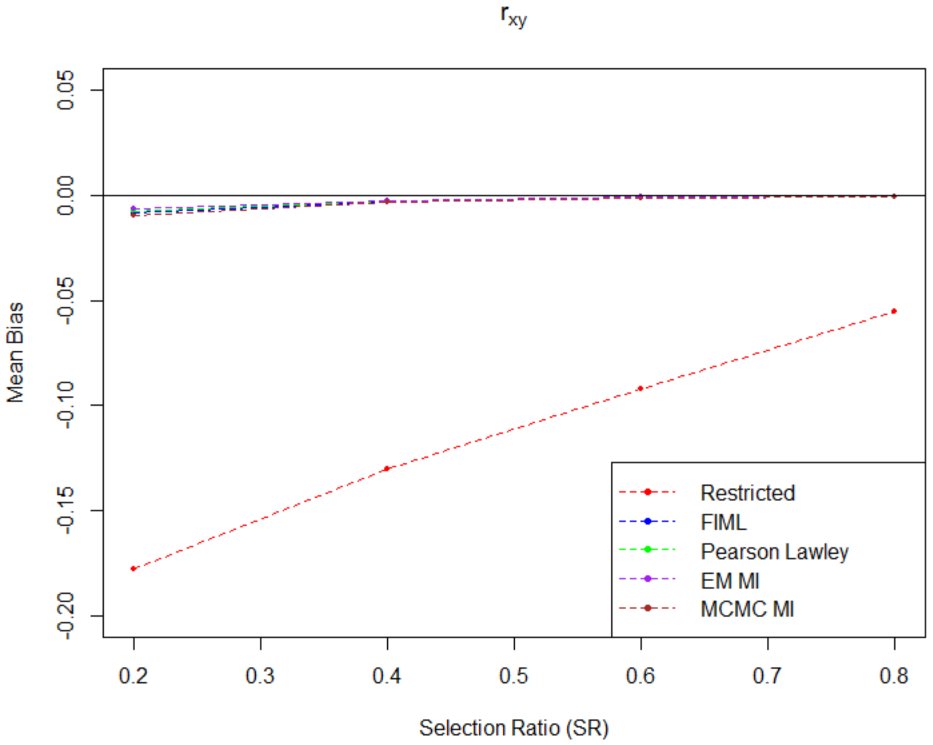
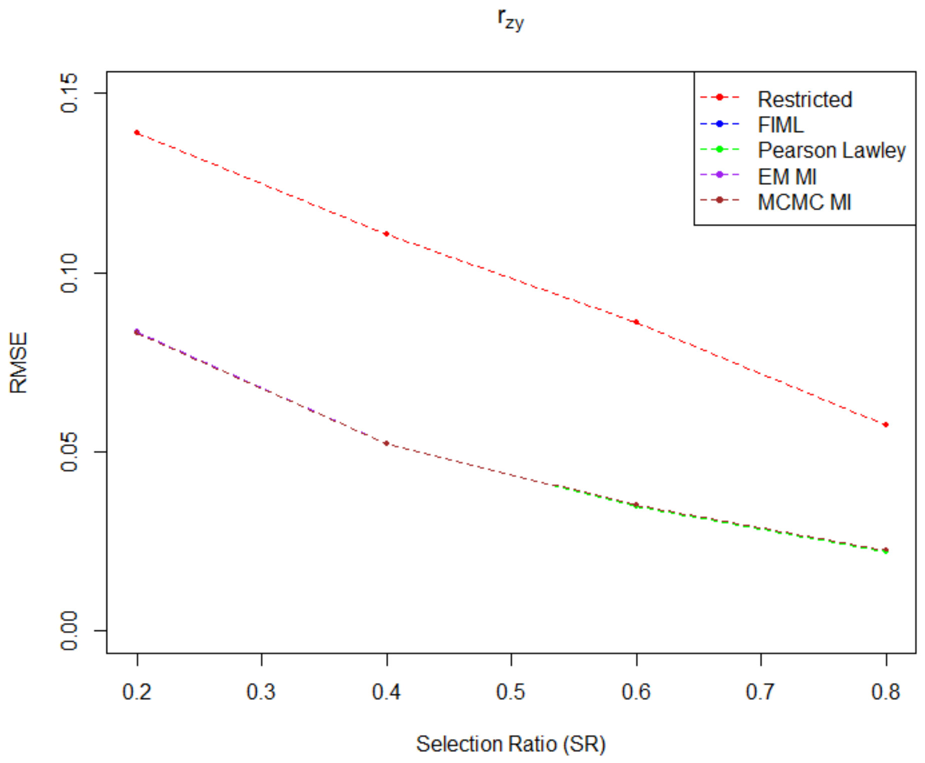


Figure 5.14.: *Mean bias for Pearson Lawley, FIML and MI methods for two hurdle validity design with full information available for variables z and x for the selection ratios of 0.2, 0.4, 0.6 and 0.8. Note that r^u_{zy} and r^u_{xy} are denoted by r_{zy} and r_{xy} respectively.*



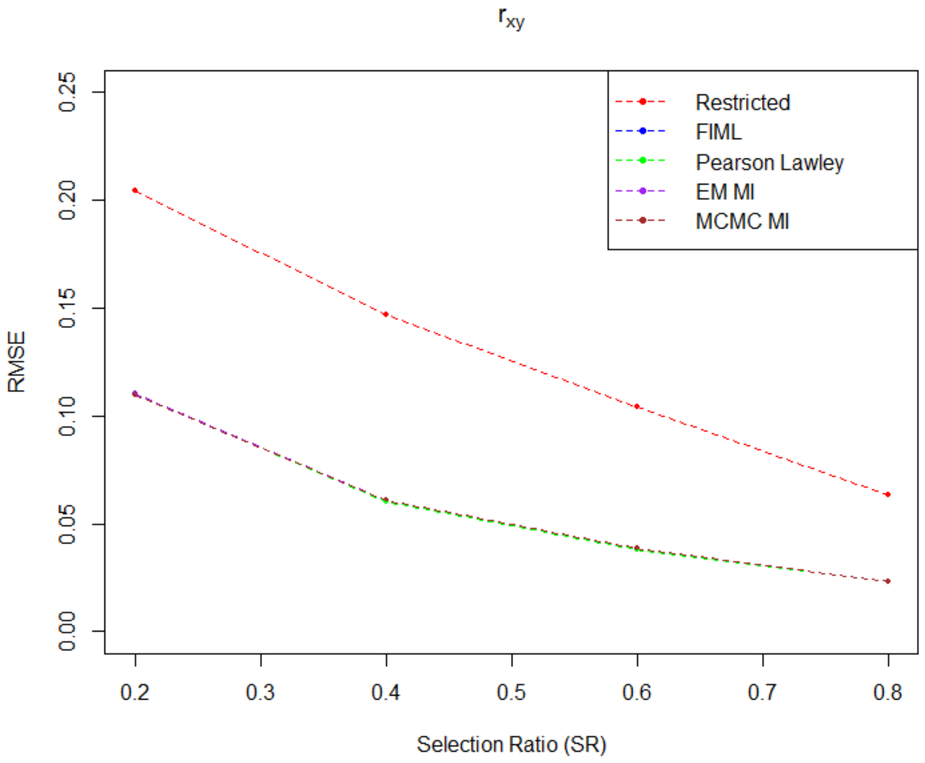


Figure 5.15.: *RMSE for Pearson Lawley, FIML and MI methods for two hurdle validity design with full information available for variables z and x for the selection ratios of 0.2, 0.4, 0.6 and 0.8. Note that r^u_{zy} and r^u_{xy} are denoted by r_{zy} and r_{xy} respectively.*

5.2.3.2. Full information on only selection test *z*

Simulations were conducted as in section 5.2.3.1 with the exception that full information was available only for selection test z. It was assumed that applicants sat for tests z and x in sequence. Selection was then conducted in two stages (hurdles). In the first stage (hurdle), $n_1 + n_2$ (out of the $n_1 + n_2 + n_3$) applicants were selected based on their scores on test z. In the second stage (hurdle), the $n_1 + n_2$ applicants from the first stage were required to sit for test x, after which only n_1 applicants were selected based on their performance. Note that at this point the full score distribution for all the $n_1 + n_2 + n_3$ applicants was only available for selection test z. The MI was conducted to reflect the ordering of the selection tests. That is, z was used to impute missing values for x. Afterwards, both z and x were then used to impute missing values for y.

Figures 5.16 and 5.17 summarise the results of the simulations with respect to bias and RMSE. Bias and RMSE associated with the restricted correlations r^r_{xy} were bigger in magnitude than that of r^r_{zy}. This is because the computation of r^r_{xy} involved using variables, x and y, both of which were subject to range restriction. *Pearson Lawley* was the only (correction) method that made use of the partial information for selection test x. Therefore it had the worst performance when correcting for the bias in r^r_{zy}. The performance of the FIML and MI based on the MCMC algorithm seemed to be at par. Results of formal statistical tests may be viewed in Tables 8.14 and 8.15 in the Technical Appendices. In Table 8.14, the magnitude of T values obtained confirmed that the bias values associated with r^r_{xy} were indeed worse compared to that for r^r_{zy}. Further, for r^r_{zy}, *Pearson Lawley* correction had the worst performance, in fact, it did not result in any statistically significant reduction in bias compared to the restricted case for the selection ratio of 0.8. Overall FIML and MI based on MCMC algorithm were equally effective at correcting for bias induced by range restriction although their performance was as good as that of *Pearson Lawley* correction when correcting r^r_{xy} for selection ratio of 0.2.

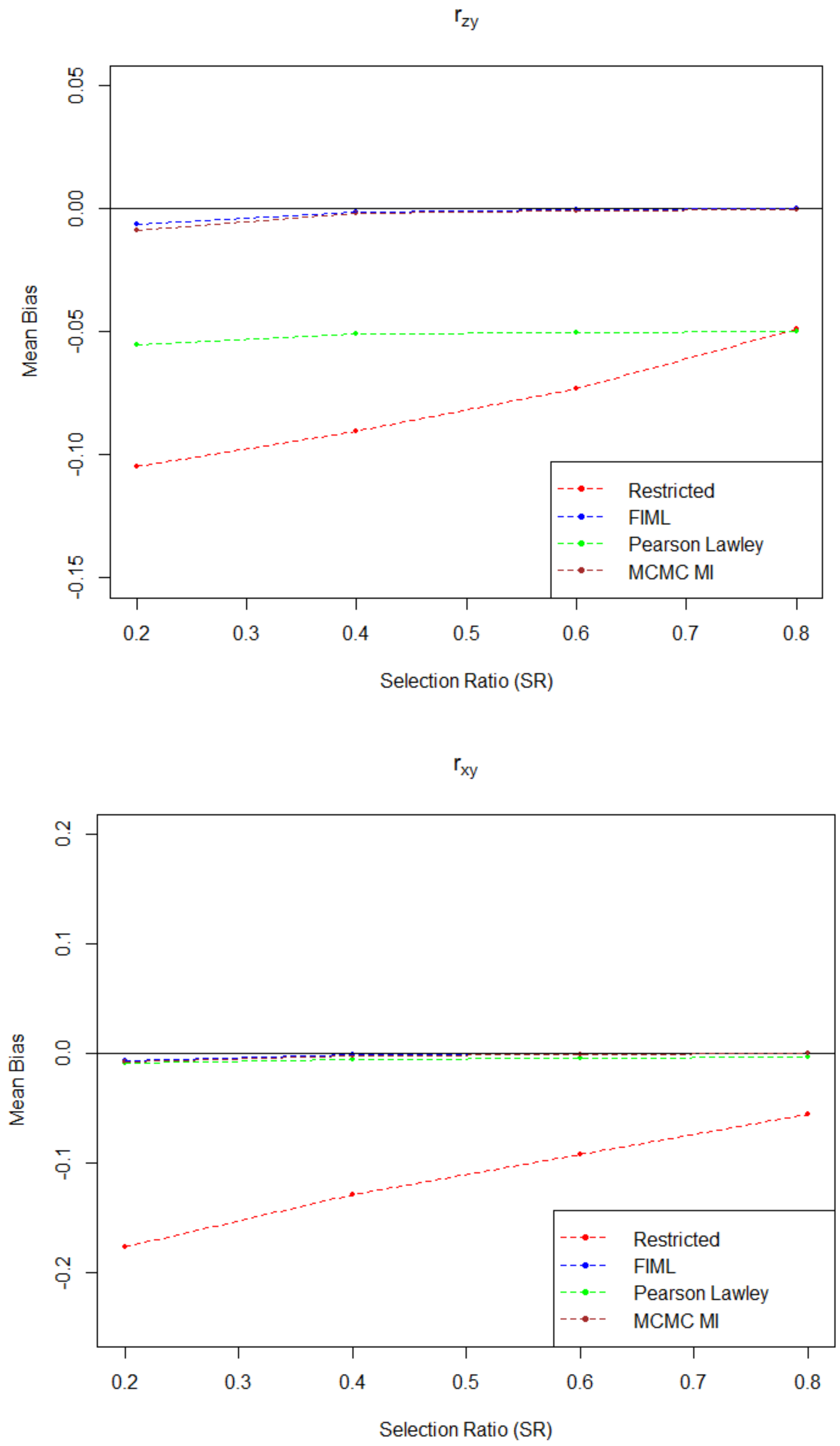


Figure 5.16.: *Mean bias for Pearson Lawley, FIML and MI methods for two hurdle validity design with full information based only on selection test z for the selection ratios of 0.2, 0.4, 0.6 and 0.8. Note that r^u_{zy} and r^u_{xy} are denoted by r_{zy} and r_{xy} respectively.*

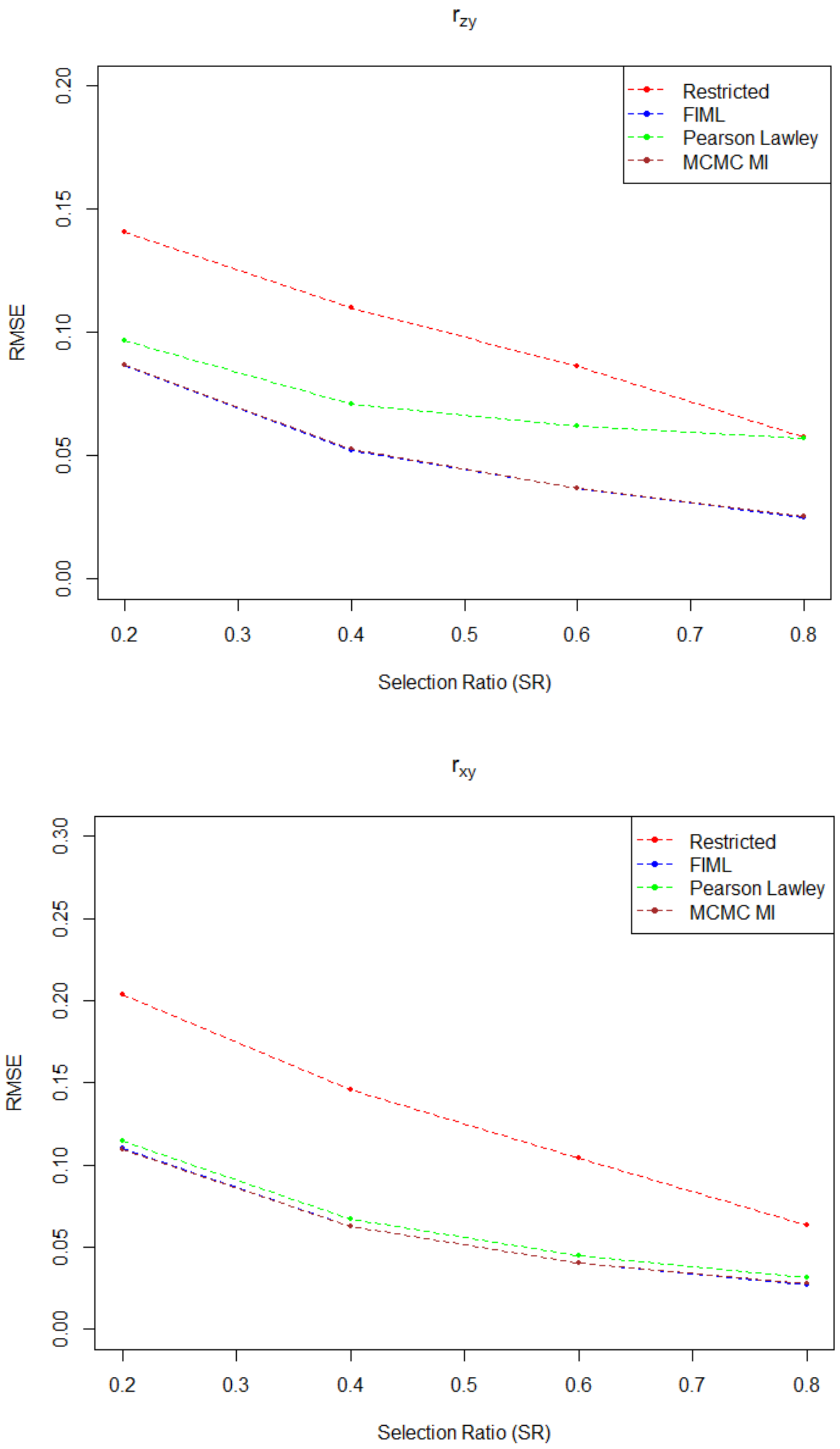


Figure 5.17.: *RMSE for Pearson Lawley, FIML and MI methods for two hurdle validity design with full information based only on selection test z for the selection ratios of 0.2, 0.4, 0.6 and 0.8. Note that r^u_{zy} and r^u_{xy} are denoted by r_{zy} and r_{xy} respectively.*

In Table 8.15, the results of the F-tests revealed a statistically significant gain in precision for all corrective methods for both r^r_{zy} than r^r_{xy}. However there was no statistically significant gain in precision when *Pearson Lawley* correction was applied at the selection ratio of 0.8 compared to the restricted case for r^r_{zy}.

5.2.4. Simulation results for correlation coefficients for the single hurdle concurrent validity selection design: Pearson Lawley vs FIML vs MI

Simulations were conducted as per the conceptual representation of the *single hurdle validity selection design* in Figure 3.5. The selection test was taken as z. The variable x was taken to be another predictor and the outcome (criterion) taken as y. For the simulations, interest was the correlation coefficients containing variables z, y and x, y respectively. For the simulations, 5,000 Monte Carlo samples consisting of 500 observations were obtained from the trivariate standard normal distribution in equation 5.1.1. Selection was based on test z with the selection ratios set at $\{0.2, 0.4, 0.6, 0.8\}$. For the MI correction, the missingness in x was imputed using z while the missingness in y was imputed with both z and x. For the MI based on MCMC algorithm, P and n_{iter} were set at 25 and 50 respectively. Figures 5.18 and 5.19 summarise the results of the simulation. Bias and RMSE were bigger in magnitude for r^r_{zy} than r^r_{xy} due to the fact that computation of former disregarded information from predictor x which had an influence on y. It was also observed that the effect of range restriction on bias and RMSE was not as severe for r^r_{xy} as it was for r^r_{zy}. The different methods for correcting for the effects of range restriction had somewhat similar levels of bias and RMSE.

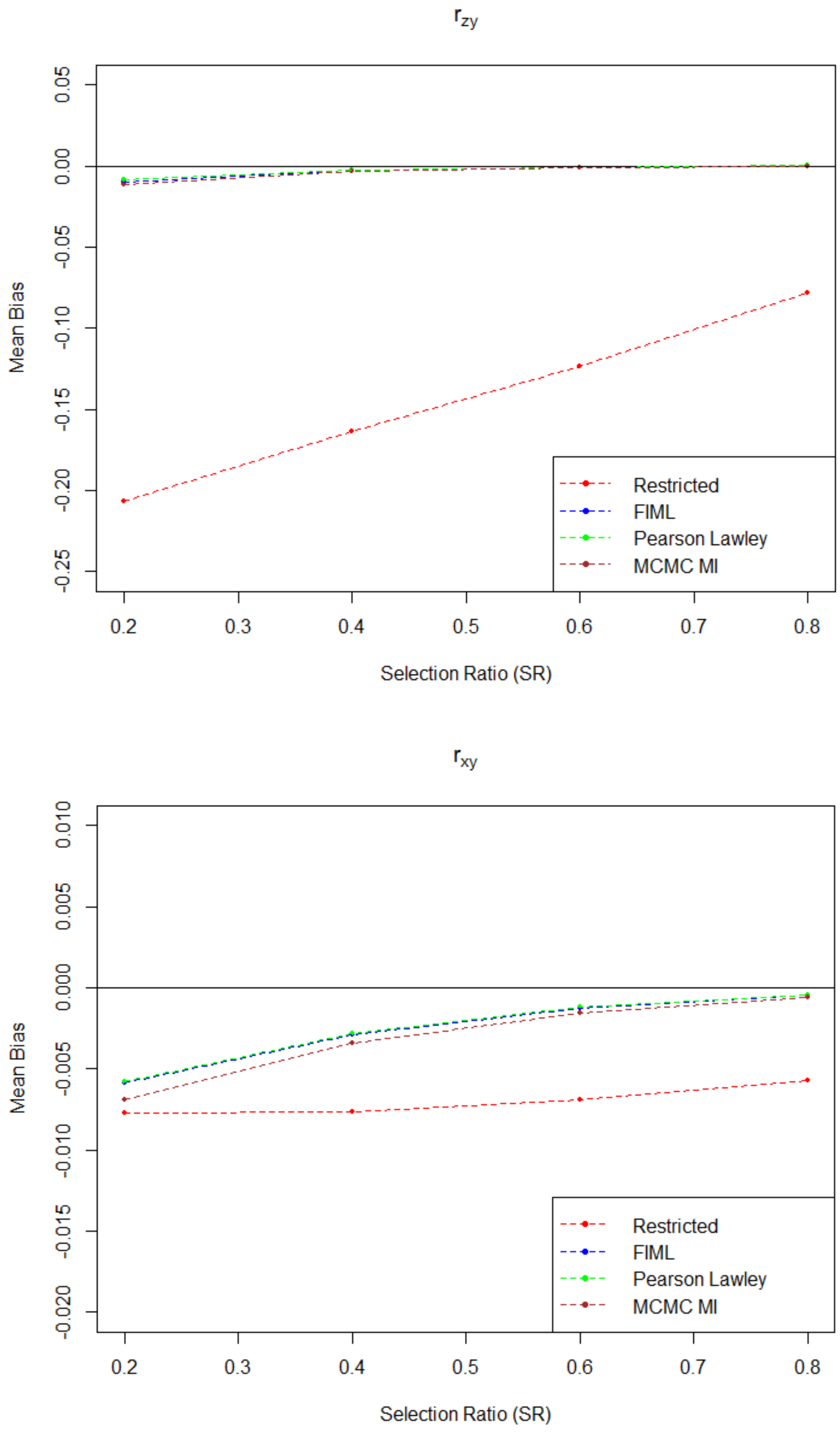
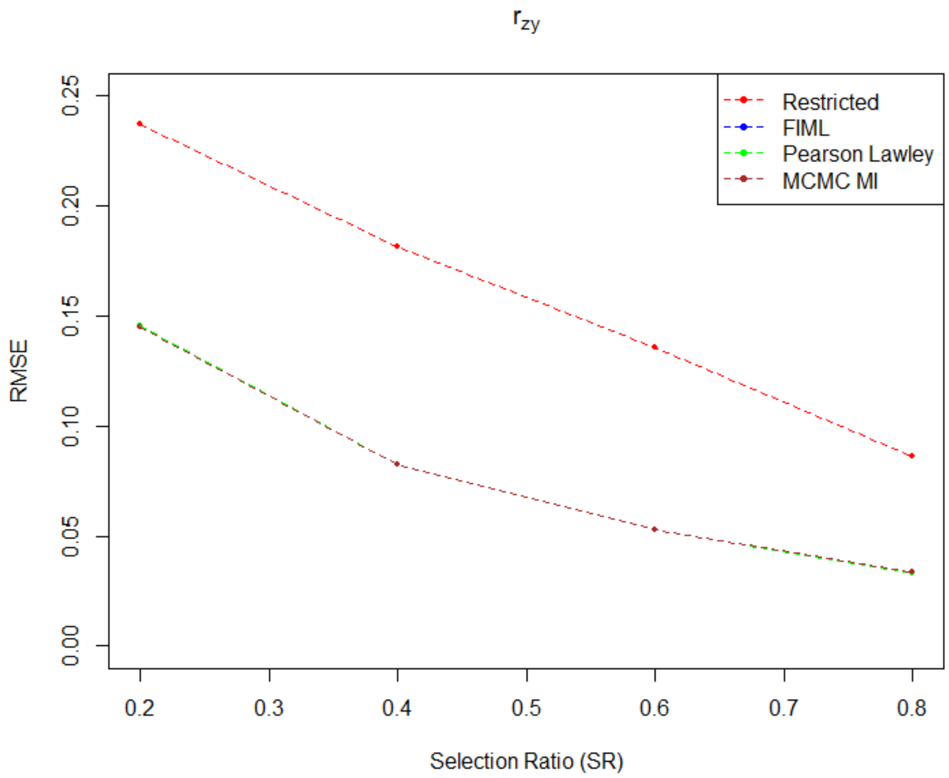


Figure 5.18.: *Mean bias for Pearson Lawley, FIML and MI methods for single hurdle concurrent validity design for the selection ratios of 0.2, 0.4, 0.6 and 0.8. Note that r^u_{zy} and r^u_{xy} are denoted by r_{zy} and r_{xy} respectively.*



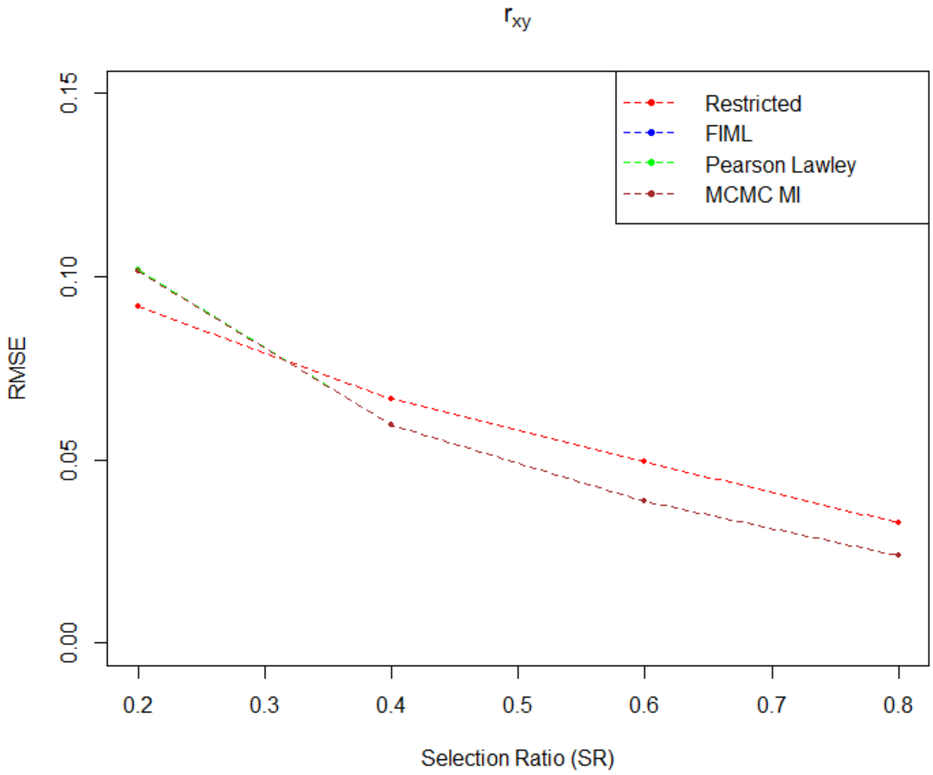


Figure 5.19.: *RMSE for Pearson Lawley, FIML and MI methods for single hurdle concurrent validity design for the selection ratios of 0.2, 0.4, 0.6 and 0.8. Note that r^u_{zy} and r^u_{xy} are denoted by r_{zy} and r_{xy} respectively.*

Table 8.16 in the Technical Appendices summarises the results of the T-tests comparing the bias for the different methods. For r^r_{zy}, all the methods led to significant reduction in bias. All the three correction methods had equal bias levels suggesting that neither of three methods had an edge over the others. For r^r_{xy}, the performance of the corrective methods was similar to the case of r^r_{zy} in all the selection ratios except for the selection ratio of 0.2 where the application of the methods did not result in a statistically significant reduction in bias. Table 8.17 in the Technical Appendices shows the result of the F-tests comparing RMSE of the methods for correcting range restriction under consideration. For r^r_{zy}, the use of the methods resulted in a statistically significant gain in precision compared to the restricted correlation. The three methods, *Pearson Lawley correction*, FIML and MI based on MCMC algorithm achieved the same level of precision. For r^r_{xy}, the performance of the three methods with respect to precision was similar to that of r^r_{zy}. For the selection ratio of 0.2, none of the methods offered any statistically significant gain in performance over the restricted correlation.

5.3. Validation phase for the expected performance of the different methods for correcting range restriction using Professional and Linguistic Assessments Board (PLAB) data

In the *testing phase* covered in section 5.2, the average performance of statistical methods for handling range restriction (*Thorndike Case II, Thorndike Case III* and *Pearson Lawley*) were compared to those based on missing data handling methods (*FIML and MI*). This was done in accordance to objective 1 (a) outlined at the start of this chapter. This however, was done using pseudo-random data. In this section, the focus will turn to the validation of these methods under the same scenarios with the aid of a contrived example using the Professional and Linguistic Assessments Board (PLAB) data. The end goal will be to inform discussion about the expected performance of the missing data methods for handling range restriction bias with real data in the field. This will thus meet objective 1 (b) specified at the start of the chapter.

In order to meet the objective 1 (b) as described, the best approach would be to validate the usefulness of the methods using real world selection data devoid of range restriction. However, within the context of selection, real world data, devoid of range restriction are not easily available. This is because, selection, no matter how liberal always induces range restriction. Therefore to proceed to validate the methods, the next best alternative of a contrived example using real-world data was explored. This involved use of real world data that consisted of variables that would act as "proxies" for a selection test and an outcome (criterion). Having full information for these "proxies" would then enable selection to be simulated thus artificially introducing range restriction. This would subsequently allow validation of the methods to be conducted. This contrived example using real-world data is thus the closest one can get to real-world selection data that has not been subjected to range restriction.

For the contrived example, the Professional and Linguistic Assessments Board (PLAB) test was used. The PLAB test is the main route by which *International Medical Graduates (IMGs)* demonstrate they have the required skills and knowledge to work in the UK (General Medical Council, 2015). The PLAB tests consists of two parts, namely I and II, the first of which covers domains such as *applying knowledge and experience to practice, clinical care, assessment* and *clinical management.* The second part consists of a *practical application of clinical skills* (Tiffin, Paton, et al., 2017). All those IMGs interested in practicing medicine in the UK have to sit for both parts of the test. In a contrived example, PLAB I may be considered to be a selection test and PLAB II the outcome (criterion). The PLAB data used was made available by the General Medical Council (GMC). The data consisted of 30,049 IMGs who sat for the PLAB tests between 2000 and 2011. In the data, only 8,828 IMGs had complete data available for first attempt scores on PLAB I, PLAB II and the *International English Language Testing System (IELTS)* overall band score which measured the English language proficiency of the IMGs. Equation (5.3.1) details the distribution of the three variables for the 8,828 IMGs . The *IELTS* overall score had the least variance while PLAB I had the highest observed mean and variance. The highest covariance was between PLAB I and PLAB II. Tables 5.9 and 5.10 show the correlation structure of the data and frequency distribution of the *IELTS* overall score. The highest observed correlation was between *IELTS* overall score and PLAB II test. Most of the IMGs had a good command of the English language.

$$\begin{pmatrix} Overall\ IELTS\ score \\ PLAB_I \\ PLAB_{II} \end{pmatrix} \sim N \left[\begin{pmatrix} 7.463 \\ 10.01 \\ 6,723 \end{pmatrix}, \begin{pmatrix} 0.2216 & 1.6014 & 0.6051 \\ 1.6014 & 337.5768 & 20.5663 \\ 0.6051 & 20.5663 & 19.7066 \end{pmatrix} \right] \tag{5.3.1}$$

Variable	Overall IELTS score	PLABI	PLABII	Minimum	Maximum
Overall IELTS score	1.0000	0.1852	0.2896	6.5	9
PLABI	0.1852	1.0000	0.2522	-90	59
PLABII	0.2896	0.2522	1.0000	-12.9500	23.4500

Table 5.9.: *Correlation matrix of the three variables for the 8,828 IMGs with their corresponding minimum and maximum values obtained from the PLAB data*

Overall band score	Frequency
6.5	1
7	3,564
7.5	2,994
8	1,697
8.5	508
9	64

Table 5.10.: *Frequency distribution of the overall IELTS score for the 8,828 IMGs from the PLAB data . Note that band 6, 7, 8 and 9 represent "competent","good","very good" and "expert" user of the English language respectively as per the IELTS classification (International English Language Testing System, 2017)*

In order to draw valid conclusions from the validation process using the PLAB data. The (Monte Carlo) simulated results from the *testing phase* were compared to the results obtained from the *validation phase* (using PLAB data) under similar settings (as much as was possible). Therefore this involved validating the methods using the PLAB data for the same number of samples (5,000 simulations), same sample size (500 observations), the selection ratios of $\{0.2, 0.4, 0.6, 0.8\}$ and selection validity designs (*predictive, two hurdle* and *single hurdle concurrent validity* selection design). To achieve a sample size of 500 observations, the 8,828 IMGs were viewed as a representative population of IMGs. This enabled a sample of 500 IMGs to be randomly drawn from the population with replacement. Note that this is a *"modified form of case resampling bootstrap"*. This is because unlike in the usual case *resampling bootstrap*, the resultant samples were made up of 500 observations and were not equal in size to the original population (of 8,828 observations) from which they are drawn. Further details of the *case resampling bootstrap* can be found in section 8.4 of the Technical Appendices. For the MI correction, P was set at 25. The tolerance and n_{iter} were set at $1 * 10 - 6$ and 50 for the EM and MCMC imputation algorithms respectively. The performance of the methods was done by assessing the bias and RMSE summarised over the entire 5,000 simulations.

5.3.1. Validation of results for correlation coefficients for the predictive validity selection design: Thorndike Case II vs MI

Under the DRR, selection was based on PLAB I with the criterion (outcome) taken as PLAB II. Figure 5.20 shows the results of the validation phase. As expected, the restricted correlation (correlation uncorrected for the effect of range restriction) is biased downwards. The magnitude of this bias worsened with increased strictness of selection process. The three corrections considered seemed to be equal in bias with each of the correction resulting in a positive bias which was highest for the selection ratio of 0.6. With regards to RMSE, the restricted correlation had the least RMSE of all the selection ratios considered. All the corrections had inferior performance compared to the restricted correlation. Table 8.18 in the Technical Appendices shows the results of the formal statistical tests which confirm the results depicted in Figure 5.20. The three correction methods were equivalent in bias and precision with the methods exhibiting positive bias and less precision compared to the restricted correlation.

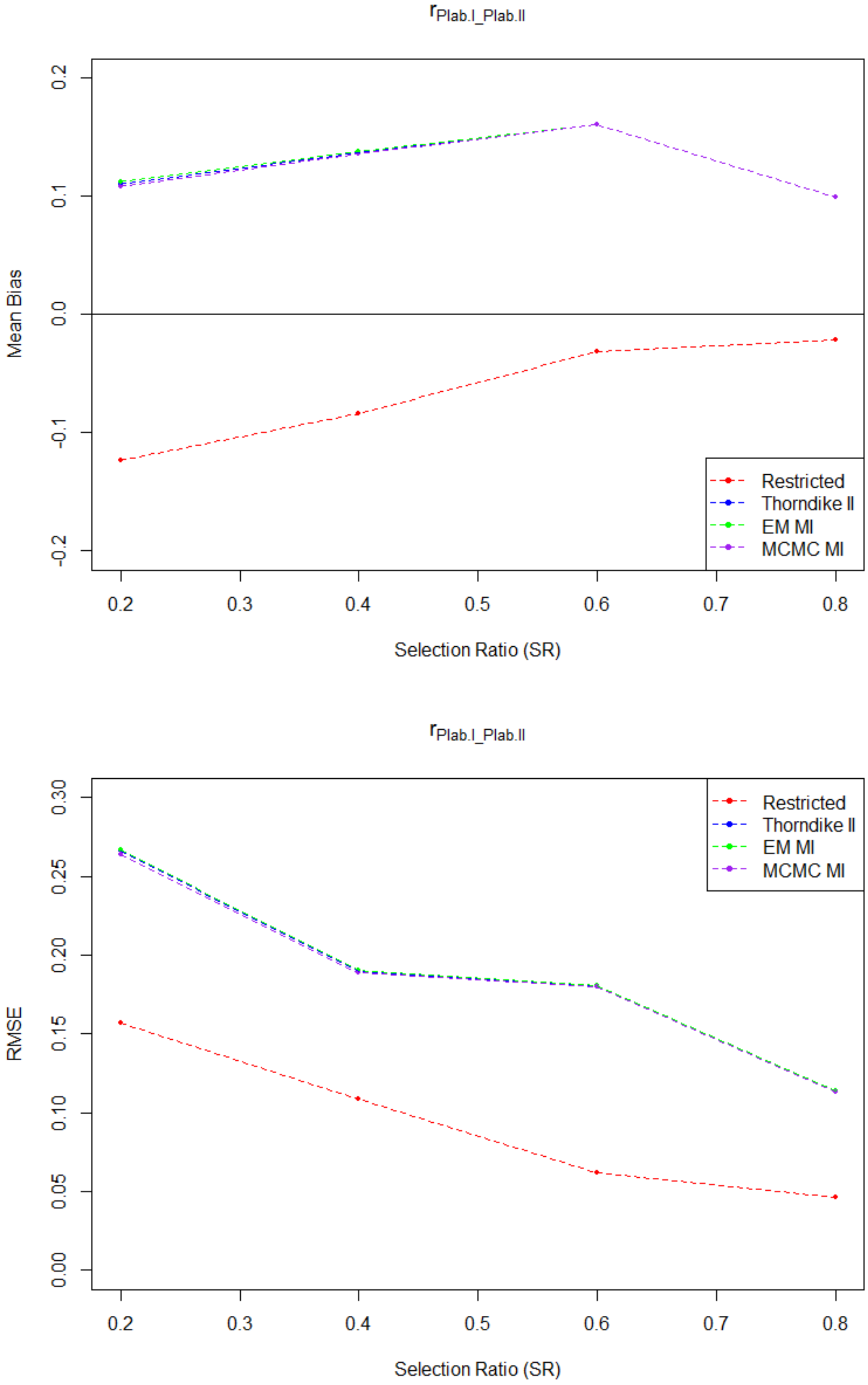


Figure 5.20.: *Performance measures for Thorndike Case II, MI for PLAB II using the selection test, PLAB I, based on EM and MCMC for the predictive validity selection design for the selection ratios of 0.2, 0.4, 0.6 and 0.8. Note that $r_{Plab.I.Plab.II}$ denotes $r^{u}_{Plab.I.Plab.II}$.*

5.3.2. Validation of results for correlation coefficients for the predictive validity selection design: Thorndike Case III vs MI

Under IRR, selection was based on PLAB I with interest being the correlation coefficient between the predictor *IELTS* and the outcome of interest PLAB II. Two scenarios were evaluated. The first involved correcting for indirect range restriction by making use of only the selection variable PLAB I in the imputation model. The second involved making use of both the selection test, PLAB I, and the predictor *IELTS* in the imputation model.

5.3.2.1. MI based only on selection test PLAB I

Figure 5.21 shows the result of the simulation. In terms of bias, the restricted correlation (correlation uncorrected for the effects of range restriction) performed best compared to all the correction under consideration. The MI correction had a downward bias while the *Thorndike Case III* had a positive bias. The stricter the selection, the worse the bias became for the methods as expected. In terms of RMSE, the restricted correlation had the best performance. The MI based correction based on EM and MCMC algorithms were equivalent in performance. The MI correction faired better than the *Thorndike Case III* for less strict selection scenarios (SR of 0.5 and above). For stricter selection scenarios, the *Thorndike Case III* faired better than MI correction. Table 8.20 in the Technical Appendices summarises the results of the statistical tests which confirmed the results in Figure 5.21. The restricted correlation had the least bias and highest precision in all the selection ratios. The MI correction had the worst bias. Both the EM and MCMC imputation algorithms were equivalent in performance. The MI correction was more precise than *Thorndike Case III* only for the higher selection ratios.

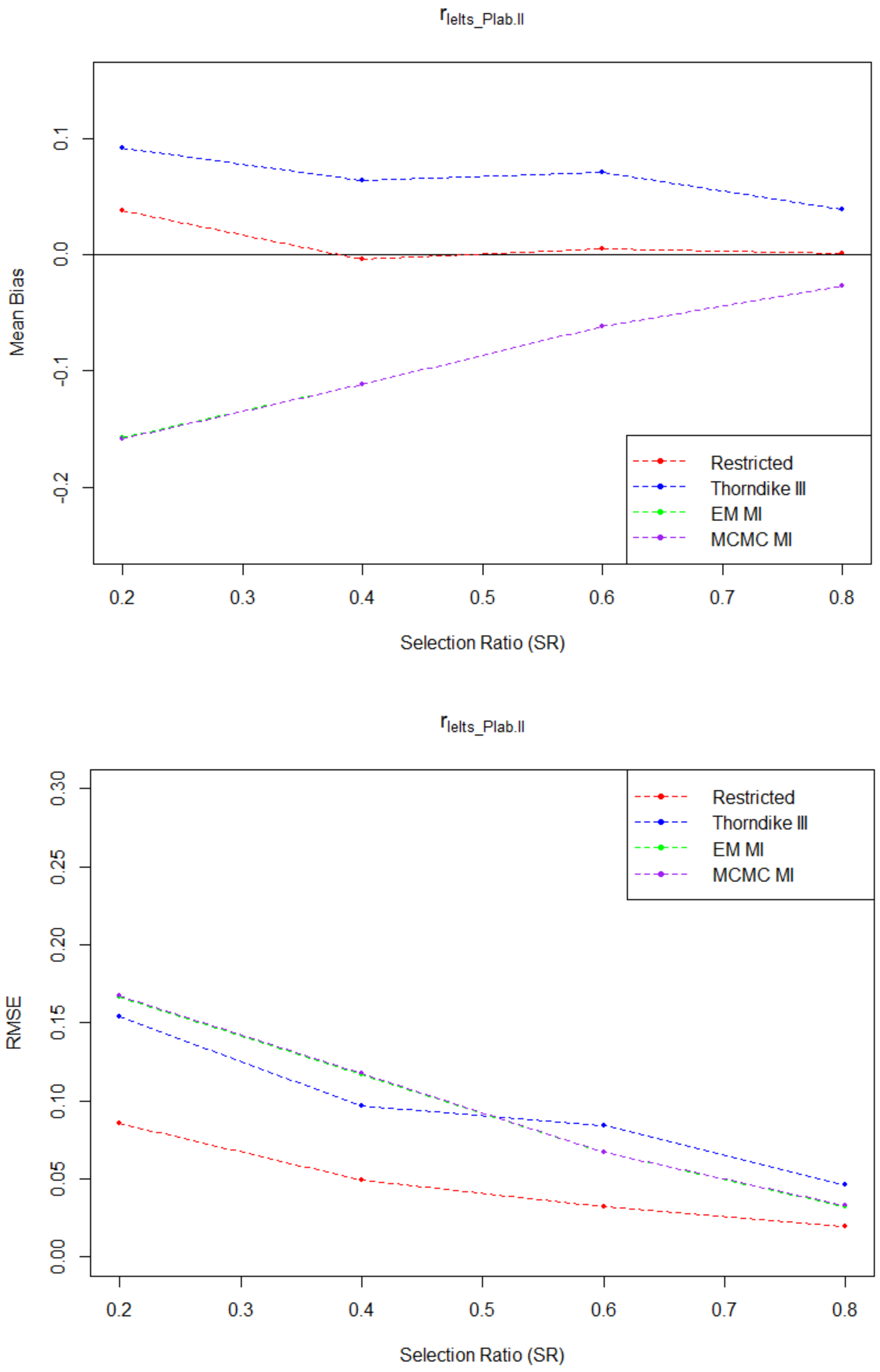


Figure 5.21.: *Performance measures for Thorndike Case III, MI for PLAB II using the selection test, PLAB I, only based on EM and MCMC for the predictive validity selection design for the selection ratios of 0.2, 0.4, 0.6 and 0.8. Note that $r_{Ielts.Plab.II}$ denotes $r^{u}_{Ielts.Plab.II}$.*

5.3.2.2. MI based on selection test PLAB I and predictor *IELTS*

Figures 5.22 and 5.23 show the results of the simulations when MI correction for the effect of range restriction when both the selection test PLAB I and the predictor *IELTS* were included in the imputation model. With respect to bias, compared to Figure 5.21 (which shows results when only the selection test PLAB I was included in the imputation model), the performance of the MI correction shown in Figure 5.22 improves dramatically. The MI correction was somewhat at par with the restricted correlation but it outperformed the *Thorndike Case III* correction. The same trend was observed for RMSE. Table 8.19 in the Technical Appendices confirms through formal statistical testing that the restricted correlation and the MI correction outperformed the *Thorndike Case III* both in terms of bias and precision. Generally the restricted correlation outperformed the MI correction with respect to both bias and precision except for the selection ratio of 0.2 where MI correction had bias at par with the restricted correlation. With regards to precision, the MI correction was superior to the restricted correlation for the selection ratios of 0.4 and below.

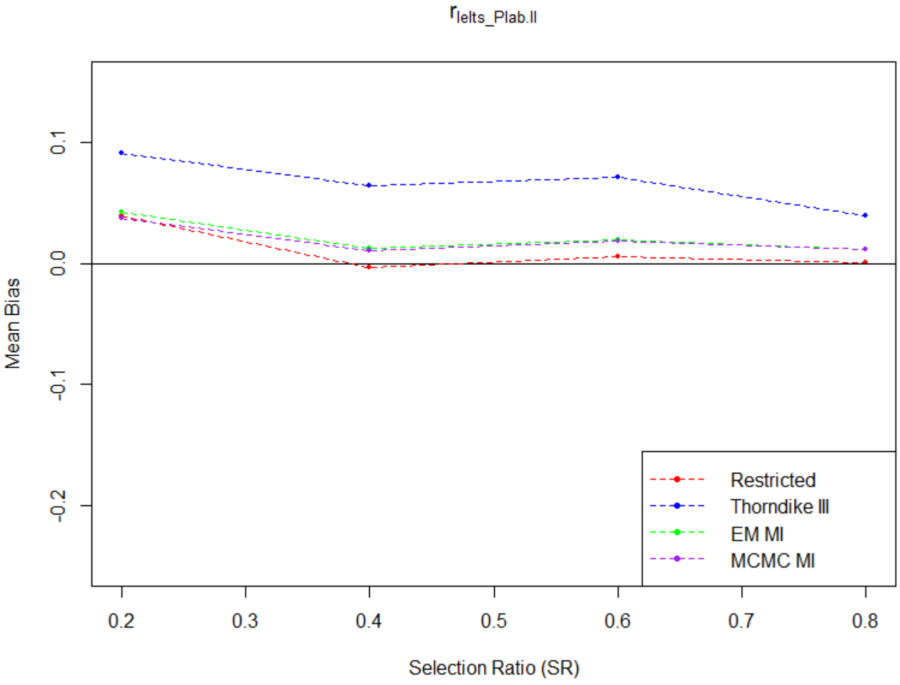


Figure 5.22.: *Bias for Thorndike Case III, MI for PLAB II using the selection test, PLAB I, and predictor, IELTS, based on EM and MCMC for the predictive validity selection design for the selection ratios of 0.2, 0.4, 0.6 and 0.8. Note that* $r_{Ielts.Plab.II}$ *denotes* $r^{u}_{Ielts.Plab.II}$.

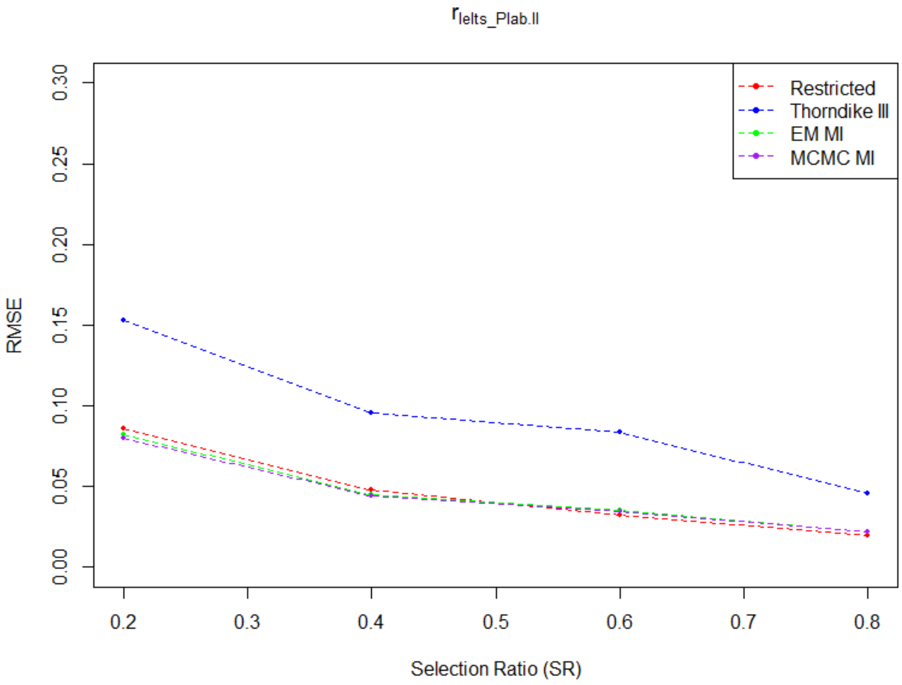


Figure 5.23.: *RMSE for Thorndike Case III, MI for PLAB II using the selection variable, PLAB I, and incidentally selected variable, IELTS, based on EM and MCMC for the predictive validity selection design for the selection ratios of 0.2, 0.4, 0.6 and 0.8. Note that $r_{Ielts.Plab.II}$ denotes $r^u_{Ielts.Plab.II}$.*

5.3.3. Validation of results for correlation coefficients for the two hurdle validity selection design: Pearson Lawley vs FIML vs MI

5.3.3.1. Full information on selection tests PLAB I and *IELTS*

Figure 5.24 shows the results of the bias for the methods from the simulation. Note that the selection was sequentially conducted using the two selection tests PLAB I and *IELTS*. This was done as described in section 5.2.3.1 with z and x being analogous to PLAB I and *IELTS* respectively. It was observed that the bias from the uncorrected correlation was worse when estimating $r^u_{Ielts.Plab.II}$ compared to $r^u_{Plab.I.Plab.II}$. All the correction methods were positively biased with the bias increasing with degree of strictness in selection. For $r^u_{Ielts.Plab.II}$, the correction methods faired better when the performance of the restricted correlation was worse. Results from formal statistical testing may be viewed in Table 8.23 in the Technical Appendices. The results confirm that indeed there was a statistically significant difference in bias between the restricted correlation and the correction methods considered. There were no statistically significant differences in bias between the correction methods.

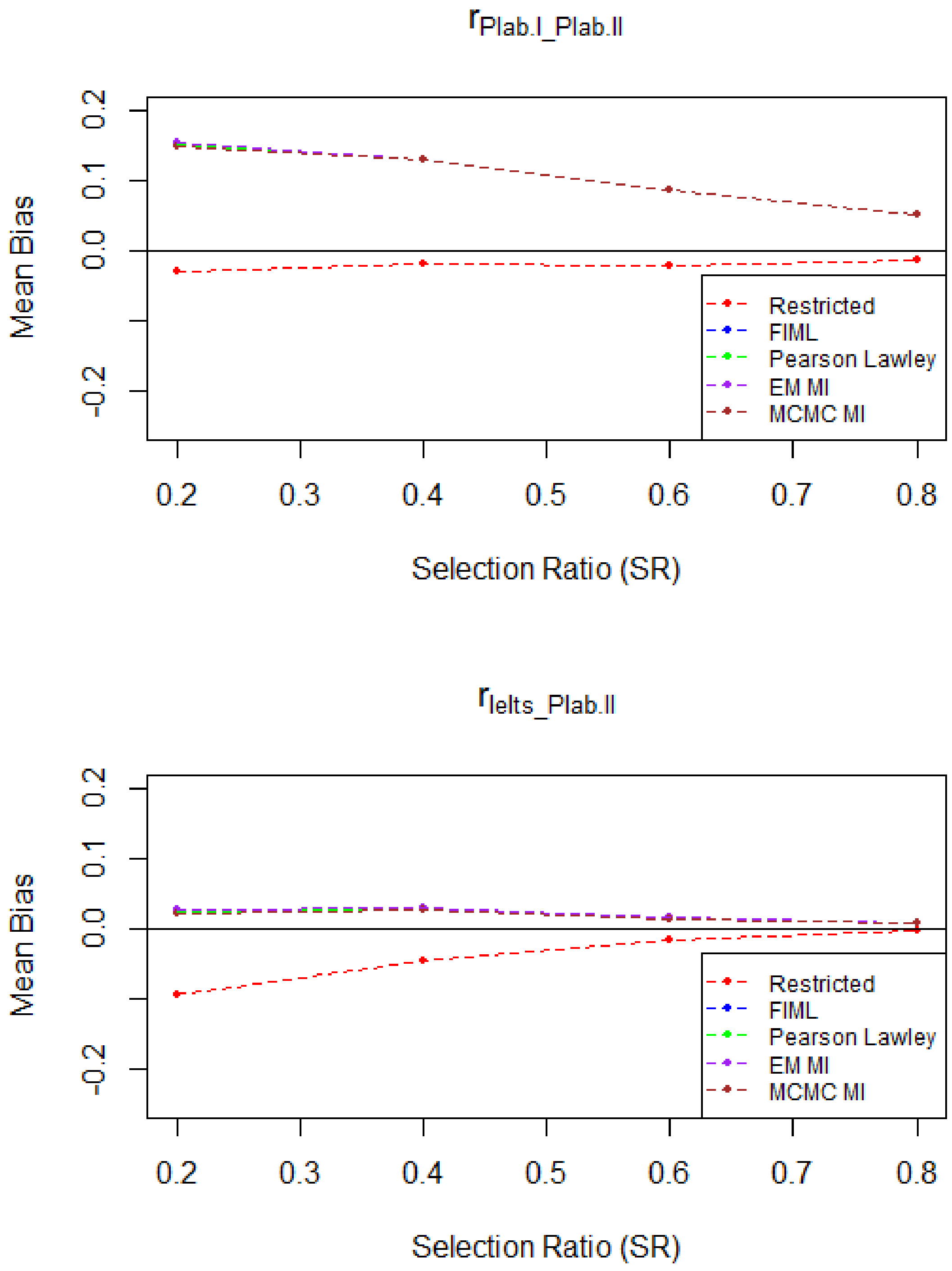


Figure 5.24.: *Bias for FIML, Pearson Lawley, MI for PLAB II using the selection tests, PLAB I and IELTS based on EM and MCMC for the two hurdle selection design for the selection ratios of 0.2, 0.4, 0.6 and 0.8. Note that $r_{Plab.I.Plab.II}$ and $r_{Ielts.Plab.II}$ denote $r^{u}_{Plab.I.Plab.II}$ and $r^{u}_{Ielts.Plab.II}$ respectively.*

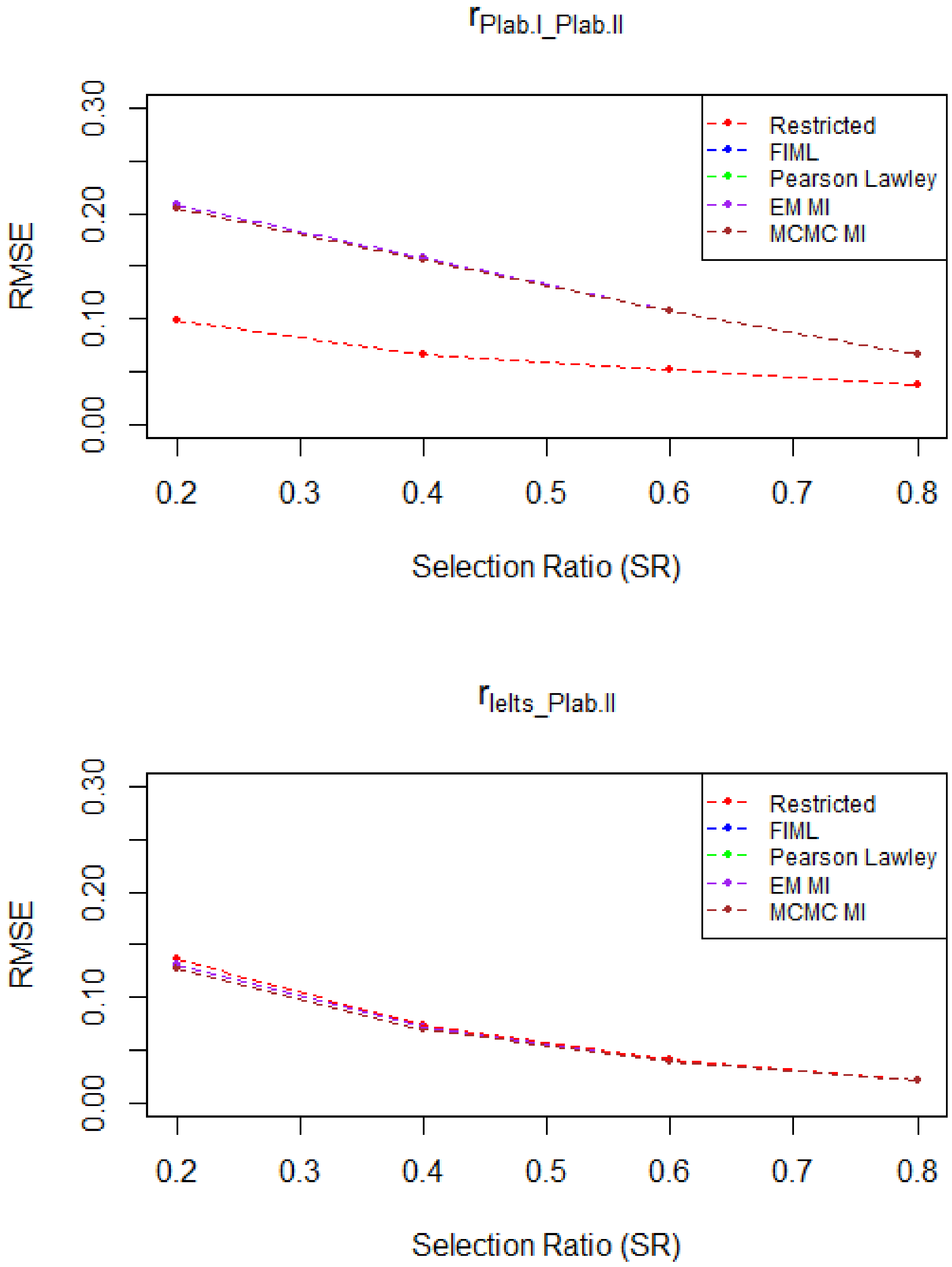


Figure 5.25.: *RMSE for FIML, Pearson Lawley, MI for PLAB II using the selection tests, PLAB I and IELTS based on EM and MCMC for the two hurdle selection design for the selection ratios of 0.2, 0.4, 0.6 and 0.8. Note that* $r_{Plab.I.Plab.II}$ *and* $r_{Ielts.Plab.II}$ *denote* $r^{u}_{Plab.I.Plab.II}$ *and* $r^{u}_{Ielts.Plab.II}$ *respectively.*

Figure 5.25 shows the results for precision from the simulations, similar to what was observed for the bias, the greater the strictness of selection, the greater the loss of precision in all of the methods considered. Overall, there seemed to be little difference in precision between the methods when estimating $r^{u}_{Ielts.Plab.II}$ with a slight difference observed for selection ratio ≤ 0.4. This was confirmed through formal statistical testing whose results are available in Table 8.24 in the Technical Appendices. For $r^{u}_{Plab.I.Plab.II}$, all the correction methods faired worse than the restricted correlation. There seemed to be no statistically significant differences between the correction methods.

5.3.3.2. Full information only on the selection test PLAB I

Figures 5.26 and 5.27 show the results of the bias and RMSE from the simulation when full information is present only for selection test PLAB I. Selection was conducted as described in section 5.2.3.2 with z and x being analogous to PLAB I and *IELTS* respectively. Note that PLAB I was first used to impute missing values for *IELTS*. Subsequently, both PLAB I and *IELTS* were used to impute missing values for PLAB II. Compared to Figures 5.26 and 5.27, the partial loss of information for the selection test, *IELTS*, leads to little (if any) change in performance of the MI correction and FIML methods with respect to bias and precision. The Pearson Lawley correction exhibited an improvement in performance for both correlations considered. Tables 8.21 and 8.22 in the Technical Appendices show the results of the formal statistical testing which confirms that for $r^{u}_{Plab.I.Plab.II}$, all the correction methods faired badly compared to the uncorrected correlation. Pearson Lawley was the second best performer in terms of bias and precision after the restricted correlation. These two were equivalent for selection ratio of 0.8 for bias and selection ratio ≥ 0.6 for precision respectively. For $r^{u}_{Ielts.Plab.II}$, Pearson Lawley correction faired better in terms of bias than all the correction methods including the restricted correlation. With respect to precision, all the methods were equivalent for the selection ratio of 0.2. Pearson Lawley was equivalent to the restricted correlation for all the selection ratios and was better than the FIML and MI correction for the selection ratio of ≥ 0.4.

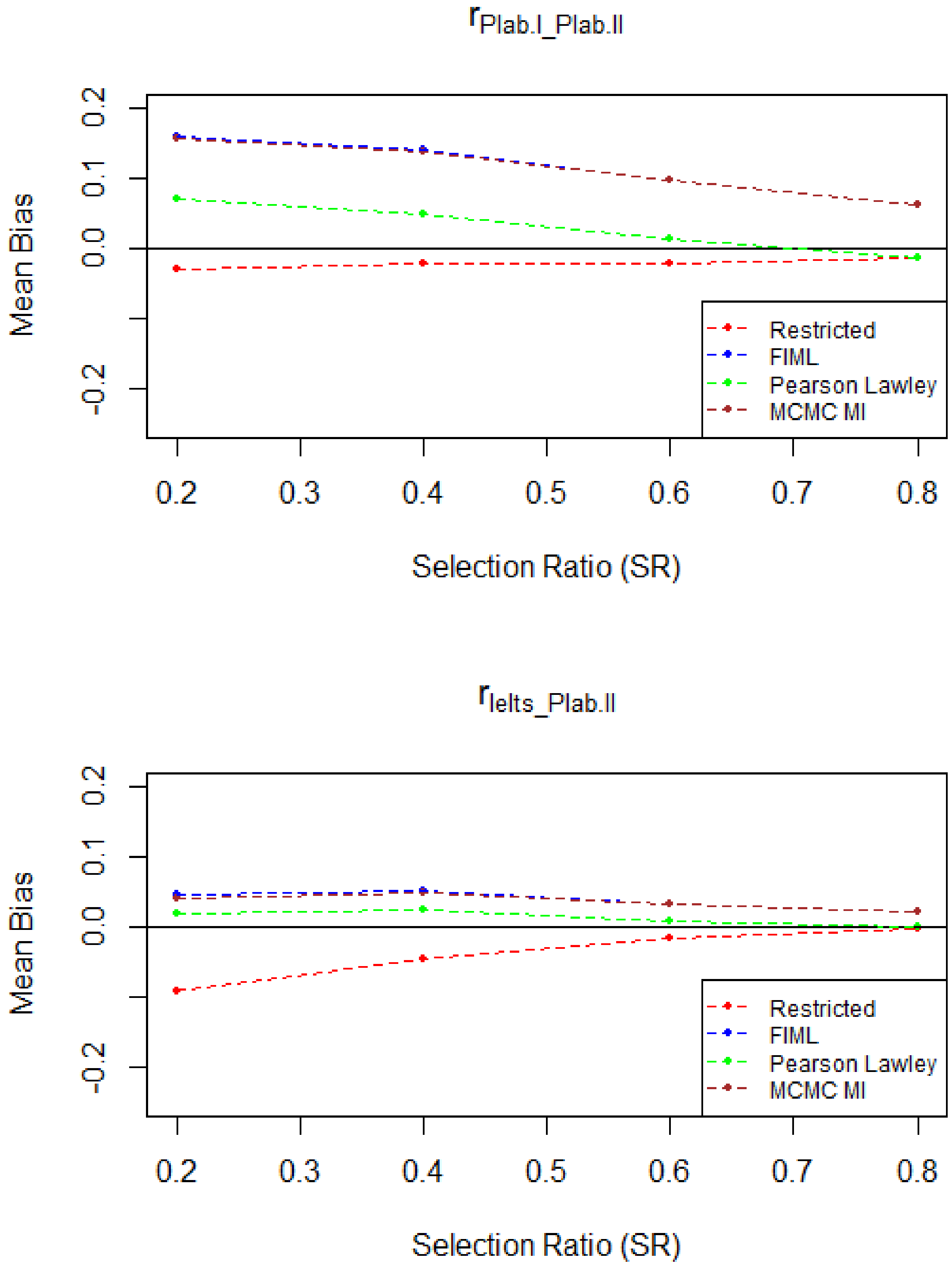


Figure 5.26.: *Bias for FIML, Pearson Lawley, MI for PLAB II using the selection tests, PLAB I and IELTS based on MCMC for the two hurdle selection design for the selection ratios of 0.2, 0.4, 0.6 and 0.8 with full information for PLAB I only. Note that* $r_{Plab.I.Plab.II}$ *and* $r_{Ielts.Plab.II}$ *denote* $r^u_{Plab.I.Plab.II}$ *and* $r^u_{Ielts.Plab.II}$ *respectively.*

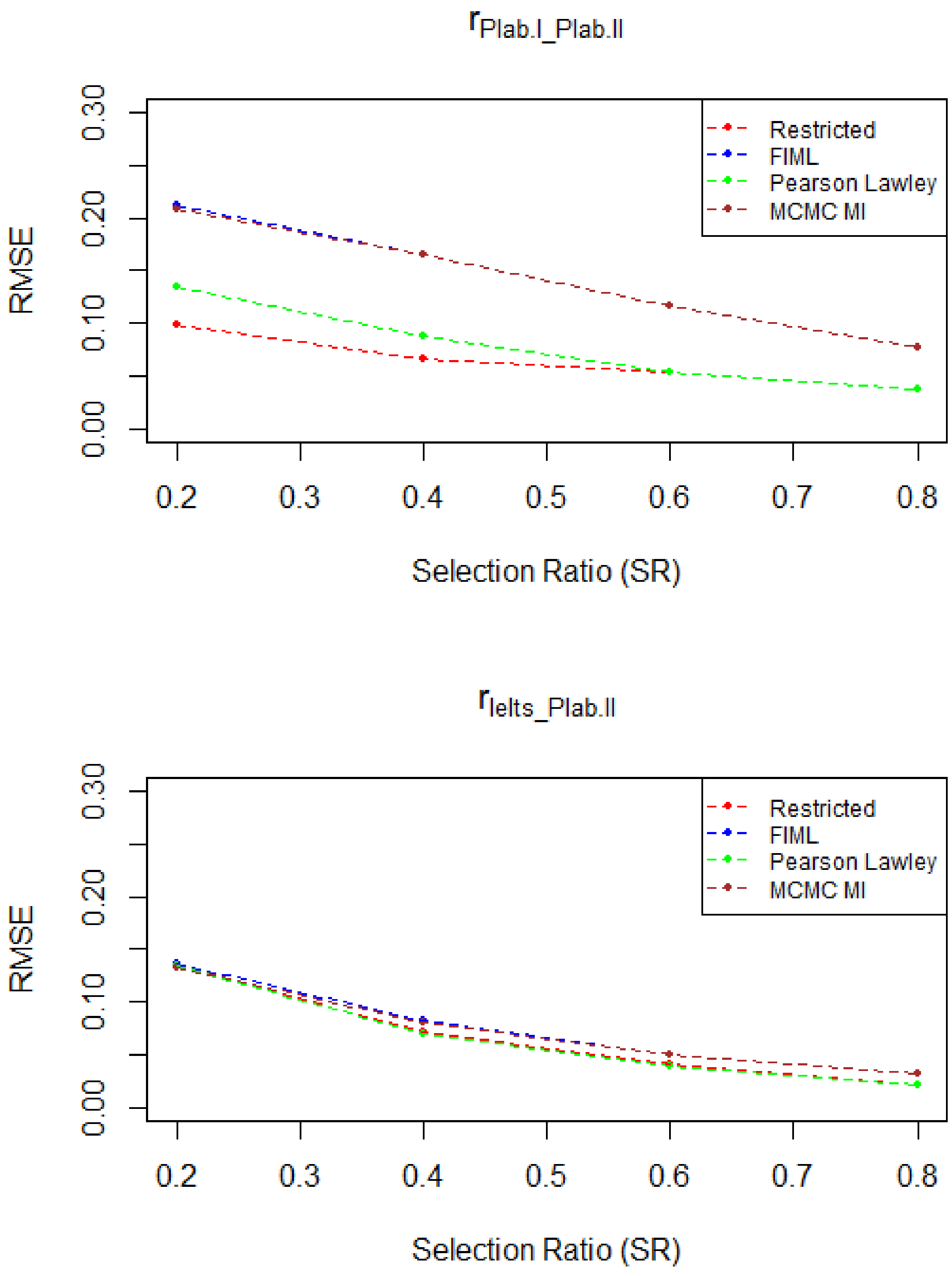


Figure 5.27.: *RMSE for FIML, Pearson Lawley, MI for PLAB II using the selection tests, PLAB I and IELTS based on MCMC for the two hurdle selection design for the selection ratios of 0.2, 0.4, 0.6 and 0.8 with full information for PLAB I only. Note that* $r_{Plab.I.Plab.II}$ *and* $r_{Ielts.Plab.II}$ *denote* $r^{u}_{Plab.I.Plab.II}$ *and* $r^{u}_{Ielts.Plab.II}$ *respectively.*

5.3.4. Validation of results for correlation coefficients for the single hurdle concurrent validity selection design: Pearson Lawley vs FIML vs MI

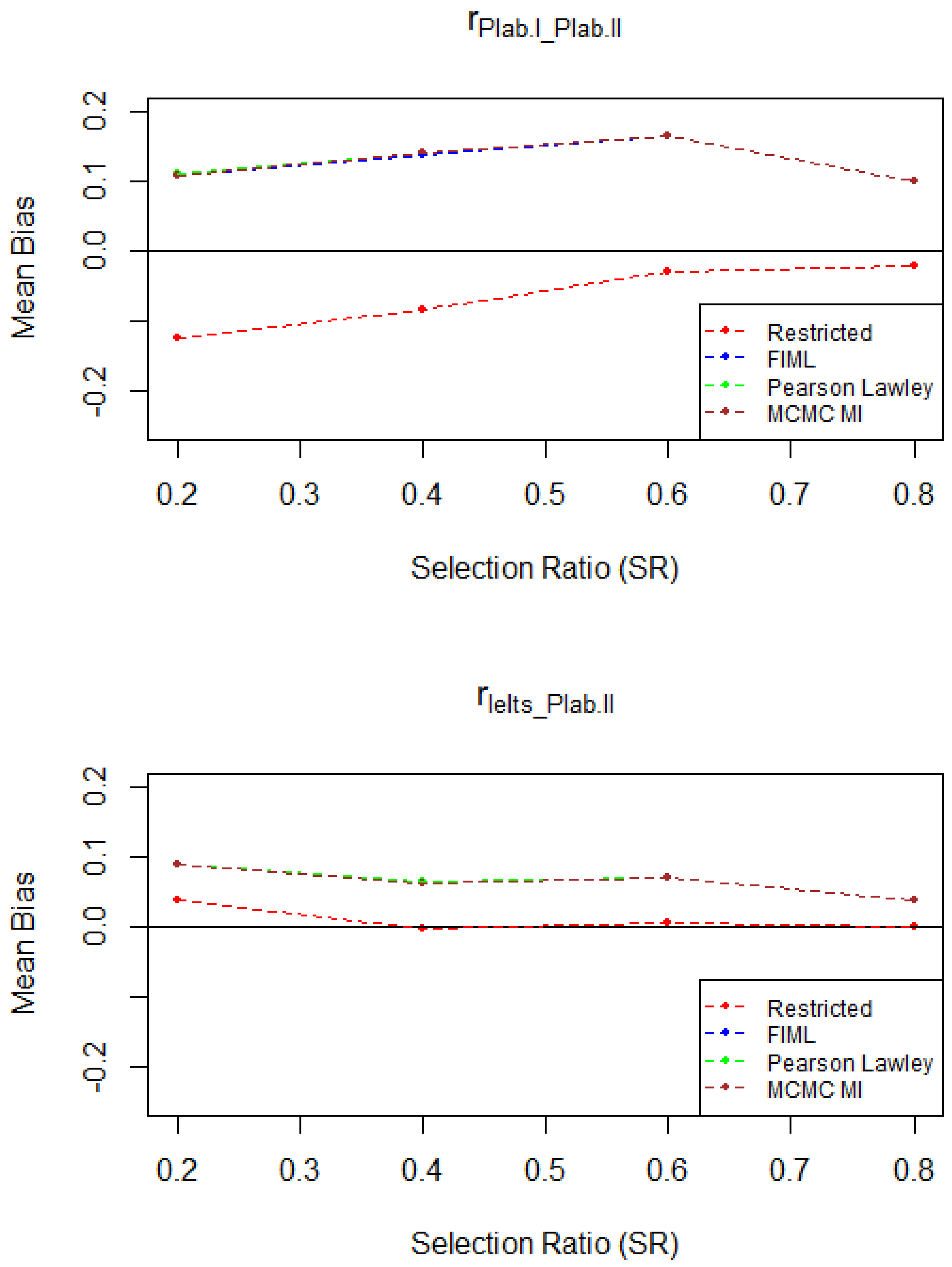


Figure 5.28.: *Bias for FIML, Pearson Lawley, MI for PLAB II using the selection test PLAB I and outcome IELTS based on MCMC for the single hurdle concurrent selection design for the selection ratios of 0.2, 0.4, 0.6 and 0.8. Note that* $r_{Plab.I.Plab.II}$ *and* $r_{Ielts.Plab.II}$ *denote* $r^{u}_{Plab.I.Plab.II}$ *and* $r^{u}_{Ielts.Plab.II}$ *respectively.*

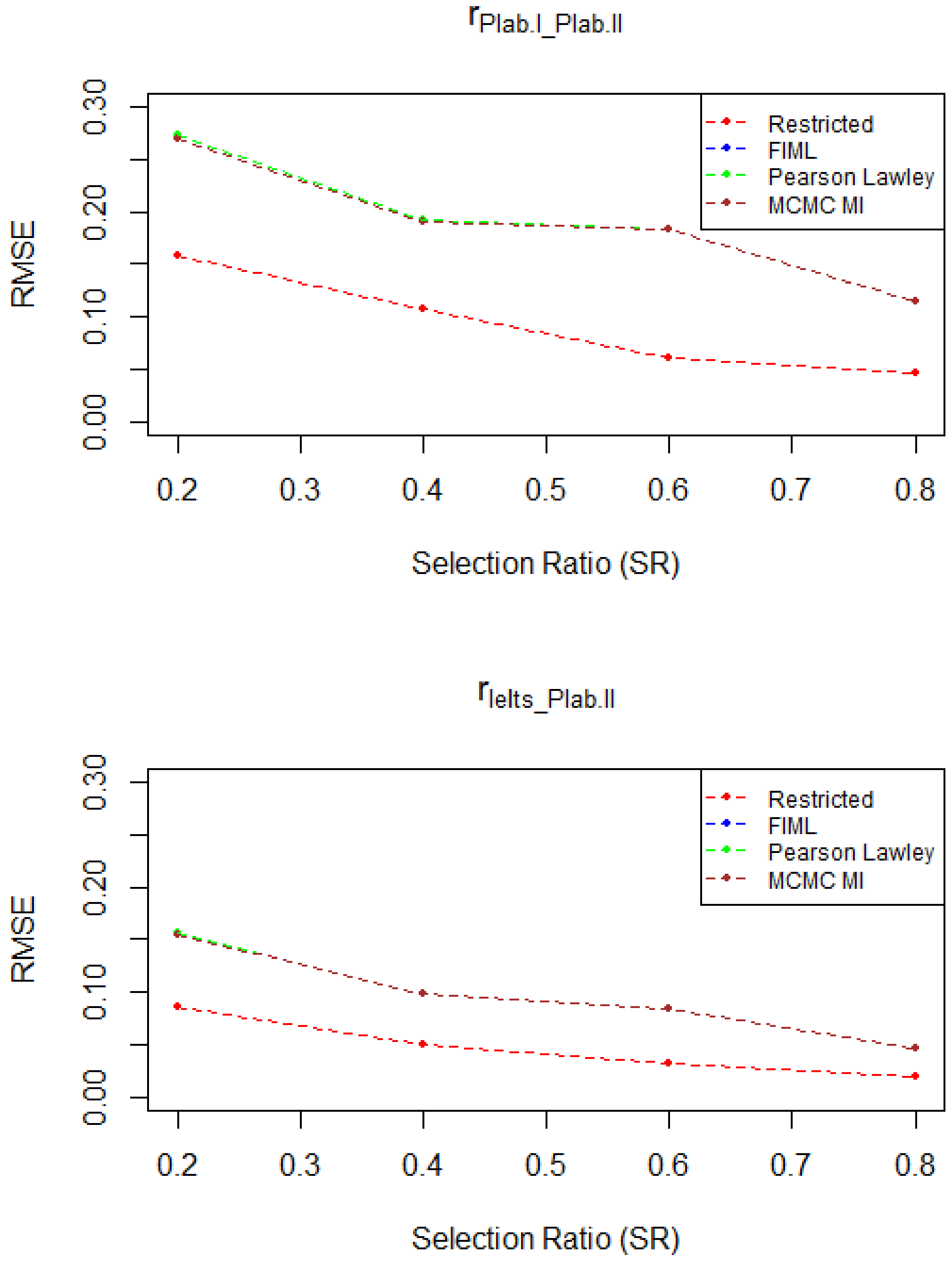


Figure 5.29.: *RMSE for FIML, Pearson Lawley, MI for PLAB II using the selection test PLAB I and outcome IELTS based on MCMC for the single hurdle concurrent selection design for the selection ratios of 0.2, 0.4, 0.6 and 0.8. Note that* $r_{Plab.I.Plab.II}$ *and* $r_{Ielts.Plab.II}$ *denote* $r^{u}_{Plab.I.Plab.II}$ *and* $r^{u}_{Ielts.Plab.II}$ *respectively.*

Figure 5.28 and 5.29 show the results of the simulations. Note that PLAB I was used to impute missing values for *IELTS*. Subsequently, both PLAB I and *IELTS* were used to impute missing values for PLAB II. The bias from the restricted correlation was more pronounced for

$r^r_{Plab.I.Plab.II}$ than for $r^r_{Ielts.Plab.II}$. The bias was negligible for selection ratio ≥ 0.4 and positive for selection ratio ≤ 0.4 for $r^r_{Ielts.Plab.II}$. With respect to precision, $r^r_{Ielts.Plab.II}$ was associated with higher precision levels compared to $r^r_{Plab.I.Plab.II}$. All correction methods were equivalent in performance in terms of bias and precision. This was also confirmed through formal statistical testing, the results of which may be viewed in Tables 8.25 and 8.26 in the Technical Appendices.

5.4. Discrepancy in performance between the testing and validation phases

So far, the the different methods for correcting range restriction have been evaluated conducted under the *testing phase* using Monte Carlo simulations (section 5.2). The results from the Monte Carlo simulations suggested that the missing data handling methods fair comparatively well in performance when juxtaposed with the formula based methods for correcting range restriction. As an extension of the study, in order to be confident that these missing data handling methods would perform well in practice, the performance of the missing data handling methods was evaluated with real world data (PLAB provided by the GMC). This constituted the *validation* phase which entailed making use of a modified bootstrapping (section 5.3). As may be seen from the results, the missing data methods (together with the formula based methods) clearly underperform with respect to bias reduction and precision. In fact, from the validation phase, in most instances, the restricted correlations had the least bias and greatest precision implying that it would have been better not to apply any type of correction for range restriction at all. This conclusion from the *validation* phase is completely at odds with what was observed from the *testing* phase. This off course begs the question-why is there a discrepancy between the performance of the methods when applied to pseudo-random data vis-á-vis real world contrived data?

The answer to the question related to the data generating mechanisms adopted for the *testing* and *validation* phase. In the *testing* phase, data were generated from Multivariate Normal

(MVN) distributions whilst in the the *validation phase* data were used with no regard whatsoever as to whether they were from were MVN distributed or not. This necessitates the investigation of the performance of the methods for correcting for range restriction in instances where data deviate from multivariate normality (that is not MVN distributed). To achieve this, data will be generated from Multivariate Skew Normal (MSN) (Azzalini and Capitanio, 1999; Wang and Genton, 2006) using the *sn* R package (Azzalini, 2018). The MSN distribution has three parameters, $\boldsymbol{\mu}$, Σ and $\boldsymbol{\eta}$ as shown in equation (5.4.1). Notice that the first two parameters $\boldsymbol{\mu}$ and Σ are the mean vector and symmetric covariance matrix of the data from a MVN distribution. For the purpose of demonstration, consider the dimensions of μ and Σ to be p (number of variables in the distribution, if p=1, then data $\boldsymbol{D}$ are of univariate nature) and $p{*}p$ respectively. The parameter $\boldsymbol{\eta}$ has a dimension of p and contains perturbation values $(\eta_1, \eta_2, ..., \eta_p)$ which skew the variables of the MVN distribution to the left or right when $\eta_i < 0$ and $\eta_i > 0$ respectively. Notice also that when $\eta_i = 0$, the data are MVN distributed (no-skewness).

$$\boldsymbol{D} \sim MSN(\boldsymbol{\mu}, \Sigma, \boldsymbol{\eta}) \tag{5.4.1}$$

In selection settings, predictor and outcome variable are typically left-skewed as may be seen from histograms (Figures 5.30 and 5.31) of actual predictor and outcome variables from data in medical school selection setting in the UK. Therefore for the purposes of the simulations, $\boldsymbol{\eta}$ may be considered to contain values are negative which act to induce left-skewness on Multivariate Normal (MVN) distributed data, $\boldsymbol{D}$, generated under Direct Range Restriction (DRR) and Indirect Range Restriction (IRR) selection designs based on equations 3.1.8 and 5.1.1 respectively. Four scenarios of deviations from MVN distributed data (alternatively MSN distributed data) were simulated. These consisted of $\boldsymbol{\eta}$ containing values -0.5, -0.75, -1 and -2 for each set of the 2 (*x, y*) and 3 (*x, y, z*) variables under the DRR and IRR selection designs respectively. The pairs of values contained in $\boldsymbol{\eta}$ may be thought of as perturbations of deviation from MVN distributed data with a category of *"Minor","Mild","Moderate"* and *"Severe"*. In the simulations, the estimand of interest is the correlation coefficient between x and y for both DRR and IRR selection designs.

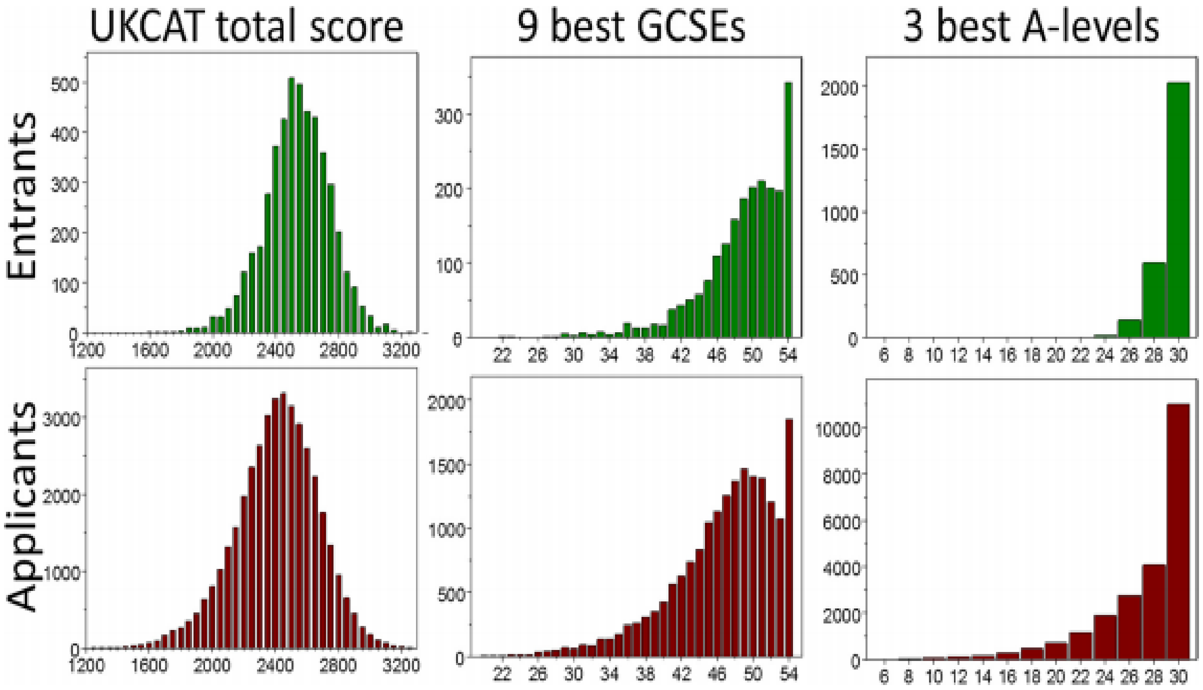


Figure 5.30.: *Distributions of the total United Kingdom Clinical Aptitude Test (UKCAT) scores, the nine best General Certificate of Secondary Education (GCSE) and the three General Certificate of Education Advanced Level (A-level) in entrants (top) and applicants (bottom) for undergraduate medical school entry cohorts of 2007, 2008 and 2009 in the UK. Figure adapted from McManus, Dewberry, Nicholson, and Dowell, 2013*

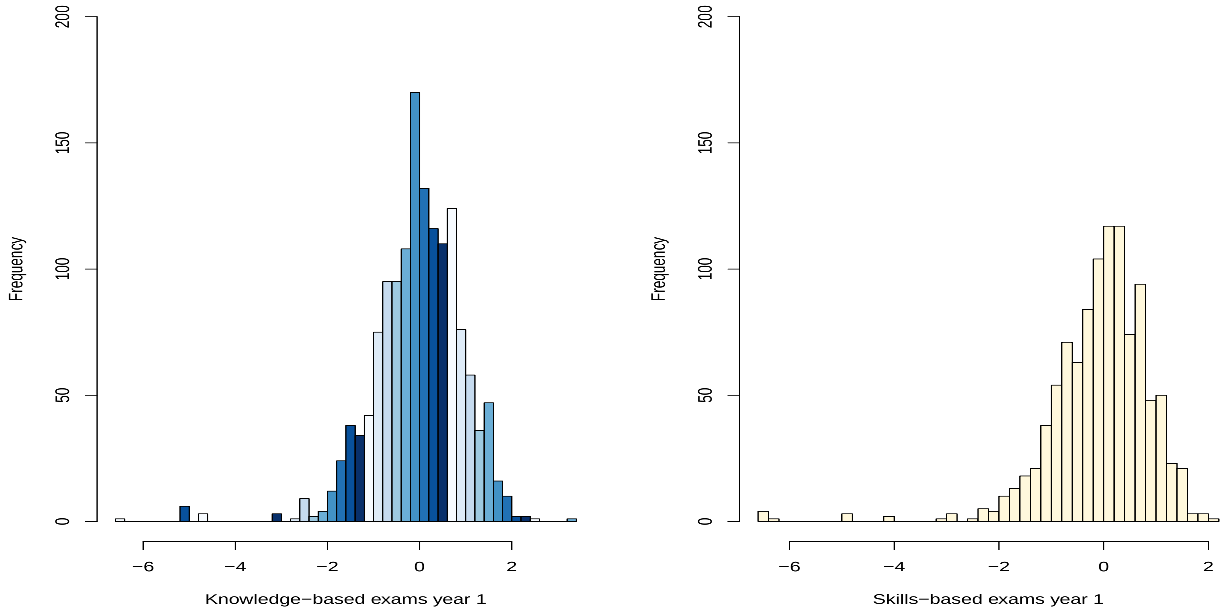


Figure 5.31.: *Distributions of the standardised undergraduate knowledge and skills-based exam outcomes across 18 medical schools in the UK for the 2007 and 2008 undergraduate medical school entry cohorts*

As in sections 5.2 and 5.3, unrestricted correlation coefficients generating data from MVN were sampled from uniform distribution, r^u_{zy} (under IRR) and r^u_{xy} (under DRR and IRR) $\sim U[0.1, 0.9]$. The number of imputations P and n_{iter} were set at 25 and 50 respectively for the EM and MCMC based imputation algorithms respectively. A total of 5,000 Monte Carlo samples of 500 observations were generated for evaluations at each of the four selection ratios of $\{0, 2, 0.4, 0.6, 0.8\}$. For each sample under IRR selection design, the correlation coefficient between x and z was randomly drawn from a uniform distribution $r^u_{xz} \sim U[0.1, 0.3]$. This was to ensure a modest level of correlation between the two predictors. The implication of this was that the complications arising from *multicollinearity (intercorrelation)* were averted and that each of the predictor introduced independent information into the samples. To keep track of the effect of skewness introduced by through the η parameter of the MSN distributions, test of multivariate normality was conducted using the *Mardia* test of multivariate normality (Kankainen, Taskinen, and Oja, 2004; Kres, 1983) implemented in the R package *MVN* (Korkmaz, Goksuluk, and Zararsiz, 2014) for each of the 5,000 MSN samples. This was done in order to determine the proportion of the 5,000 samples which deviated from multivariate normality at particular parameter values contained in η. In addition, for each of the 5,000 samples, the effect of the skewness was also tracked by evaluating (prior to introducing range restriction at different selection ratios), the proportion of samples in which skewness induced attenuation of the unrestricted correlation coefficients. The results of these tests are presented in Figure 5.32. From the top panel, it was observed that the proportion of simulated samples that deviated from multivariate normality increased as degree of (left) skewness was increased with all simulated sample deviating from multivariate normality when the degree of (left) skewness was severe. This trend was more pronounced for the IRR selection design. From the bottom panel, it was observed that skewness had an adverse effect of attenuating the unrestricted correlation coefficient r^u_{xy} even before selection was induced (before range restriction was introduced). This effect was observed for both DRR and IRR regardless of the degree of (left) skewness introduced into the simulation.

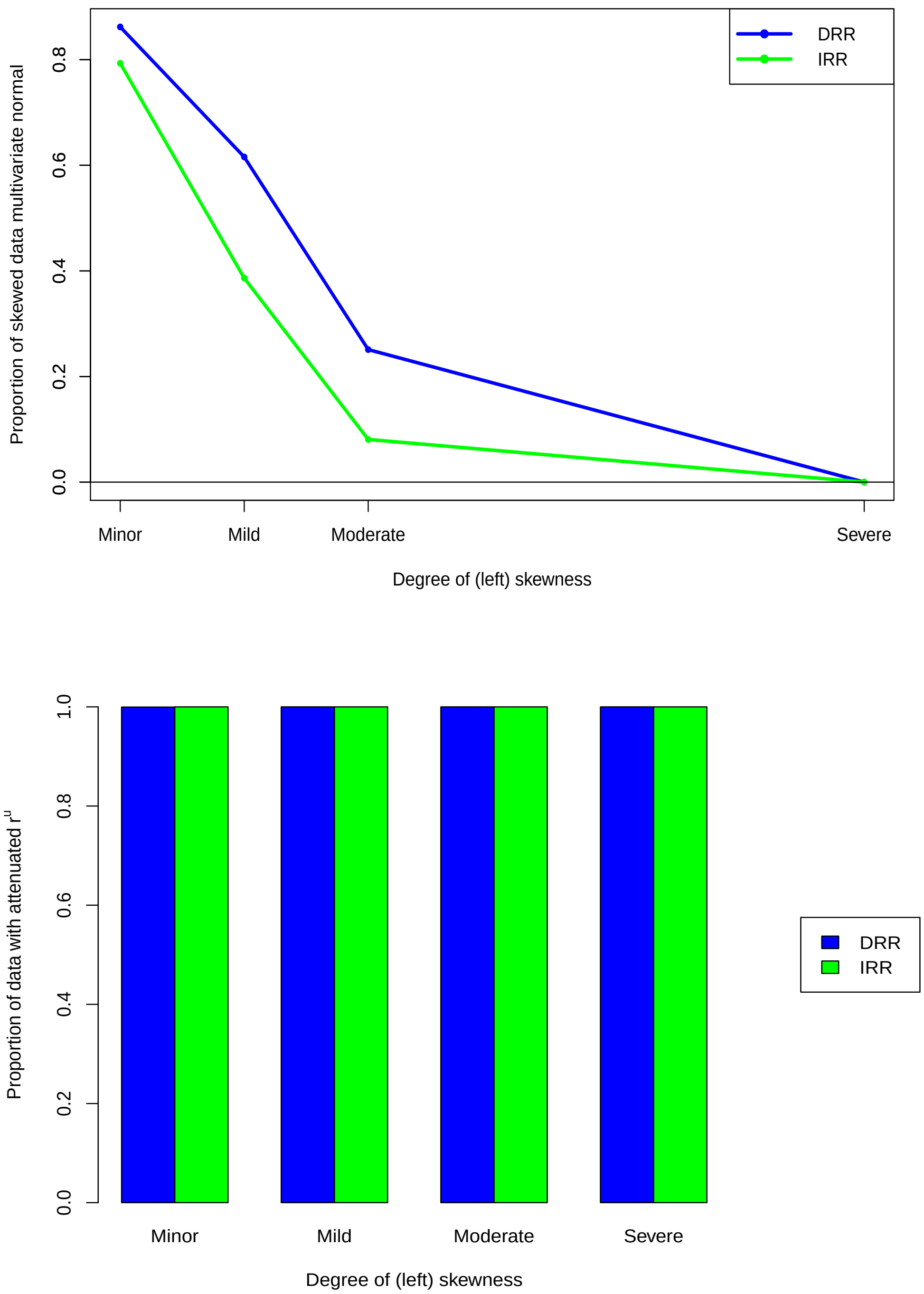


Figure 5.32.: *Proportions of the 5,000 simulated data from the MSN distributions with varying amounts of left skewesss* $(\eta_x, \eta_y) = \{(-0.5,-0.5),(-0.75,-0.75),(-1,-1),(-2,-2)\}$ *under DRR selection design where selection is based on x and* $(\eta_x, \eta_z, \eta_y) = \{(-0.5,-0.5,-0.5),(-0.75,-0.75,-0.75),(-1,-1,-1),(-2,-2,-2)\}$ *for the IRR selection design where selection is based on z (top panel). The bottom panel shows the proportion of the 5,000 simulated data from the MSN distributions with varying amounts of the specified left skewness that resulted in attenuated the restricted correlation* (r^u_{xy}) *under both under the DRR and IRR selection designs. Note that the pair of values in* η *of* $-0.5, -0.75, -1$ *and* -2 *denote "Minor","Mild","Moderate" and "Severe" departure from multivariate normality respectively*

With respect to the PLAB data used in the *validation* phase (section 5.3), all of the simulated samples were found to deviate from multivariate normality (based on results from *Mardia* test of multivariate normality). Therefore, it is reasonable to conclude based on the these findings (see also Figure 5.32) that the PLAB data were not MVN distributed due to severe left skewness.

Figure 5.33 shows the effects of the degree of (left) on bias and precision under the DRR selection design. It was observed that the increase in the degree of (left) skewness on both the selection and criterion variables, x and y, result in an increase of bias and loss of precision. Generally, the methods for correcting the effect of range restriction had less bias than the restricted case (effect for range restriction not corrected). However, the methods did not fair as well as when (left) skewness was absent (that is when data were MVN distributed, see section 5.2). For severe departures from multivariate normality ($\eta_x = -2$, $\eta_y = -2$), all the methods were worse off in bias than the restricted case except for selection ratio 0.8. With regard to precision, all the methods had relatively more precision compared to the restricted case. However for severe departures from multivariate normality, the restricted case faired better. These observations were formalised by statistical T and F-tests whose results are presented in Tables 5.11 and 5.12.

Figure 5.34 shows the effects of (left) skewness on bias and precision under IRR selection design. Compared to the DRR selection design, it was observed that the effect of skewness under IRR was worse. For all selection ratios and degrees of (left) skewness considered, the restricted case (effect of range restriction ignored) had the least bias and greatest precision followed by the Thorndike Case III correction formula. The MI had the worst bias and precision with both the EM and MCMC algorithms the same performance. An increase in the degree of (left) skewness resulted in a decrease in performance for all the methods for correcting range restriction under evaluation. These observations were confirmed by statistical T and F-tests whose results are presented in Tables 5.13 and 5.14.

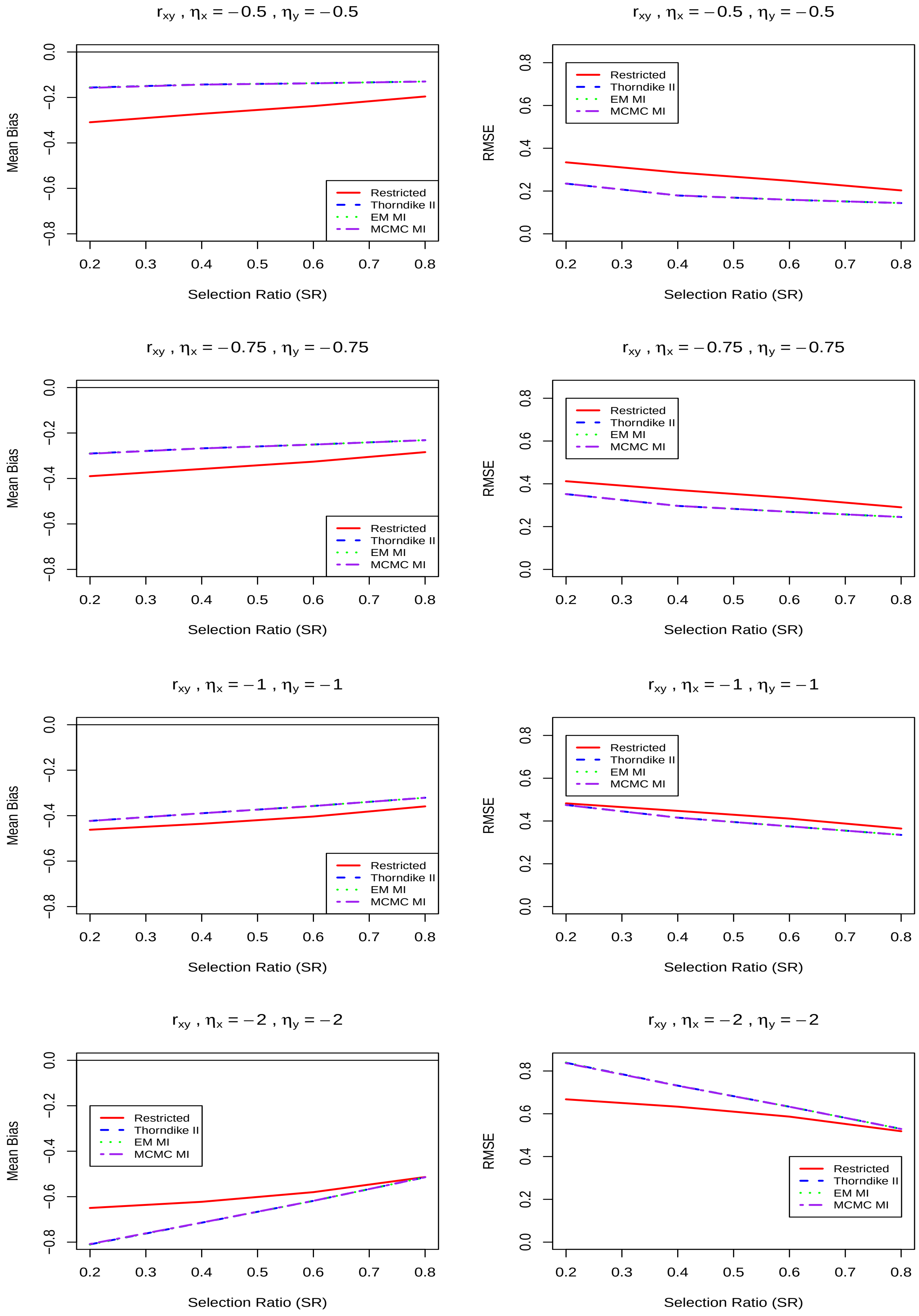


Figure 5.33.: *Bias and RMSE for the Thorndike Case II, MI based on EM and MCMC correction for DRR on data drawn from a Multivariate Skew Normal (MSN) distribution with varying amounts of skewness* η.

	T values (Bias)			
Comparison ($\eta_x = -0.5,\ \eta_y = -0.5$)	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case II	✓	✓	✓	✓
Restricted vs EM MI	✓	✓	✓	✓
Restricted vs MCMC MI	✓	✓	✓	✓
Thorndike Case II vs EM MI	✗	✗	✗	✗
Thorndike Case II vs MCMC MI	✗	✗	✗	✗
EM MI vs MCMC MI	✗	✗	✗	✗
	T values (Bias)			
Comparison ($\eta_x = -0.75,\ \eta_y = -0.75$)	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case II	✓	✓	✓	✓
Restricted vs EM MI	✓	✓	✓	✓
Restricted vs MCMC MI	✓	✓	✓	✓
Thorndike Case II vs EM MI	✗	✗	✗	✗
Thorndike Case II vs MCMC MI	✗	✗	✗	✗
EM MI vs MCMC MI	✗	✗	✗	✗
	T values (Bias)			
Comparison ($\eta_x = -1,\ \eta_y = -1$)	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case II	✓	✓	✓	✓
Restricted vs EM MI	✓	✓	✓	✓
Restricted vs MCMC MI	✓	✓	✓	✓
Thorndike Case II vs EM MI	✗	✗	✗	✗
Thorndike Case II vs MCMC MI	✗	✗	✗	✗
EM MI vs MCMC MI	✗	✗	✗	✗
	T values (Bias)			
Comparison ($\eta_x = -2,\ \eta_y = -2$)	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case II	✓	✓	✓	✗
Restricted vs EM MI	✓	✓	✓	✗
Restricted vs MCMC MI	✓	✓	✓	✗
Thorndike Case II vs EM MI	✗	✗	✗	✗
Thorndike Case II vs MCMC MI	✗	✗	✗	✗
EM MI vs MCMC MI	✗	✗	✗	✗

Table 5.11.: *T-tests comparing bias for the different methods of correcting range restriction under DRR selection design for varying degrees of left skewness. Note that ✓and ✗ denote statistically significant and non-significant difference respectively between the two methods under comparison*

	F values (Precision)			
Comparison ($\eta_x = -0.5,\ \eta_y = -0.5$)	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case II	✓	✓	✓	✓
Restricted vs EM MI	✓	✓	✓	✓
Restricted vs MCMC MI	✓	✓	✓	✓
Thorndike Case II vs EM MI	✗	✗	✗	✗
Thorndike Case II vs MCMC MI	✗	✗	✗	✗
EM MI vs MCMC MI	✗	✗	✗	✗

	F values (Precision)			
Comparison ($\eta_x = -0.75,\ \eta_y = -0.75$)	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case II	✓	✓	✓	✓
Restricted vs EM MI	✓	✓	✓	✓
Restricted vs MCMC MI	✓	✓	✓	✓
Thorndike Case II vs EM MI	✗	✗	✗	✗
Thorndike Case II vs MCMC MI	✗	✗	✗	✗
EM MI vs MCMC MI	✗	✗	✗	✗

	F values (Precision)			
Comparison ($\eta_x = -1,\ \eta_y = -1$)	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case II	✗	✓	✓	✓
Restricted vs EM MI	✗	✓	✓	✓
Restricted vs MCMC MI	✗	✓	✓	✓
Thorndike Case II vs EM MI	✗	✗	✗	✗
Thorndike Case II vs MCMC MI	✗	✗	✗	✗
EM MI vs MCMC MI	✗	✗	✗	✗

	F values (Precision)			
Comparison ($\eta_x = -2,\ \eta_y = -2$)	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case II	✓	✓	✓	✗
Restricted vs EM MI	✓	✓	✓	✗
Restricted vs MCMC MI	✓	✓	✓	✗
Thorndike Case II vs EM MI	✗	✗	✗	✗
Thorndike Case II vs MCMC MI	✗	✗	✗	✗
EM MI vs MCMC MI	✗	✗	✗	✗

Table 5.12.: *F-tests comparing precision for the different methods of correcting range restriction under DRR selection design for varying degrees of left skewness. Note that ✓and ✗ denote statistically significant and non-significant difference respectively between the two methods under comparison*

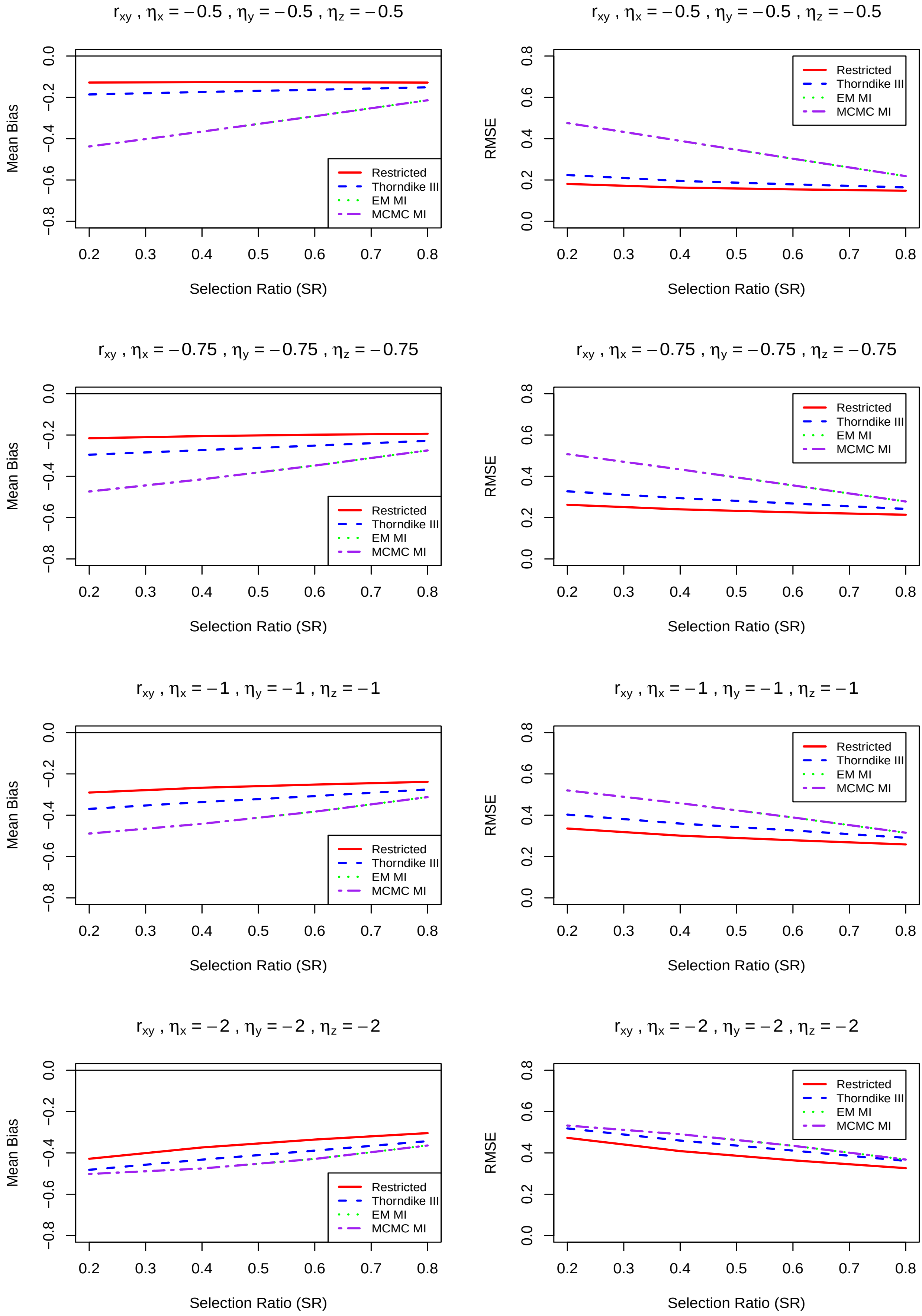


Figure 5.34.: *Bias and RMSE for the Thorndike Case III, MI based on EM and MCMC correction for IRR on data drawn from a Multivariate Skew Normal (MSN) distribution with varying amounts of skewness η.*

	T values (Bias)			
Comparison ($\eta_x = -0.5,\ \eta_y = -0.5,\ \eta_z = -0.5$)	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case III	✓	✓	✓	✓
Restricted vs EM MI	✓	✓	✓	✓
Restricted vs MCMC MI	✓	✓	✓	✓
Thorndike Case III vs EM MI	✓	✓	✓	✓
Thorndike Case III vs MCMC MI	✓	✓	✓	✓
EM MI vs MCMC MI	✗	✗	✗	✗
	T values (Bias)			
Comparison ($\eta_x = -0.75,\ \eta_y = -0.75,\ \eta_z = -0.75$)	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case III	✓	✓	✓	✓
Restricted vs EM MI	✓	✓	✓	✓
Restricted vs MCMC MI	✓	✓	✓	✓
Thorndike Case III vs EM MI	✓	✓	✓	✓
Thorndike Case III vs MCMC MI	✓	✓	✓	✓
EM MI vs MCMC MI	✗	✗	✗	✗
	T values (Bias)			
Comparison ($\eta_x = -1,\ \eta_y = -1,\ \eta_z = -1$)	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case III	✓	✓	✓	✓
Restricted vs EM MI	✓	✓	✓	✓
Restricted vs MCMC MI	✓	✓	✓	✓
Thorndike Case III vs EM MI	✓	✓	✓	✓
Thorndike Case III vs MCMC MI	✓	✓	✓	✓
EM MI vs MCMC MI	✗	✗	✗	✗
	T values (Bias)			
Comparison ($\eta_x = -2,\ \eta_y = -2,\ \eta_z = -2$)	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case III	✓	✓	✓	✓
Restricted vs EM MI	✓	✓	✓	✓
Restricted vs MCMC MI	✓	✓	✓	✓
Thorndike Case III vs EM MI	✓	✓	✓	✓
Thorndike Case III vs MCMC MI	✓	✓	✓	✓
EM MI vs MCMC MI	✗	✗	✗	✗

Table 5.13.: *T-tests comparing bias for the different methods of correcting range restriction under IRR selection design for varying degrees of left skewness. Note that ✓and ✗ denote statistically significant and non-significant difference respectively between the two methods under comparison*

	F values (Precision)			
Comparison ($\eta_x = -0.5,\ \eta_y = -0.5,\ \eta_z = -0.5$)	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case III	✓	✓	✓	✓
Restricted vs EM MI	✓	✓	✓	✓
Restricted vs MCMC MI	✓	✓	✓	✓
EM MI vs Thorndike Case III	✓	✓	✓	✓
MCMC MI vs Thorndike Case III	✓	✓	✓	✓
EM MI vs MCMC MI	✗	✗	✗	✗
	F values (Precision)			
Comparison ($\eta_x = -0.75,\ \eta_y = -0.75,\ \eta_z = -0.75$)	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case III	✓	✓	✓	✓
Restricted vs EM MI	✓	✓	✓	✓
Restricted vs MCMC MI	✓	✓	✓	✓
EM MI vs Thorndike Case III	✓	✓	✓	✓
MCMC MI vs Thorndike Case III	✓	✓	✓	✓
EM MI vs MCMC MI	✗	✗	✗	✗
	F values (Precision)			
Comparison ($\eta_x = -1,\ \eta_y = -1,\ \eta_z = -1$)	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case III	✓	✓	✓	✓
Restricted vs EM MI	✓	✓	✓	✓
Restricted vs MCMC MI	✓	✓	✓	✓
EM MI vs Thorndike Case III	✓	✓	✓	✓
MCMC MI vs Thorndike Case III	✓	✓	✓	✓
EM MI vs MCMC MI	✗	✗	✗	✗
	F values (Precision)			
Comparison ($\eta_x = -2,\ \eta_y = -2,\ \eta_z = -2$)	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case III	✓	✓	✓	✓
Restricted vs EM MI	✓	✓	✓	✓
Restricted vs MCMC MI	✓	✓	✓	✓
EM MI vs Thorndike Case III	✓	✓	✓	✗
MCMC MI vs Thorndike Case III	✓	✓	✓	✗
EM MI vs MCMC MI	✗	✗	✗	✗

Table 5.14.: *F-tests comparing precision for the different methods of correcting range restriction under IRR selection design for varying degrees of left skewness. Note that ✓and ✗ denote statistically significant and non-significant difference respectively between the two methods under comparison*

5.5. Chapter summary

In this chapter, the potential of missing data handling methods for dealing with the downward bias in the Pearson correlation coefficients due to range restriction was explored. This was done with the aid of simulated data drawn from multivariate standard normal distributions under a variety of settings. This addressed the objective 2 outlined in section 1.4. Further, the performance of the missing data handling methods was evaluated under a variety of settings with the aid of a contrived example that made use of PLAB data. This was done in accordance with objective 3 outlined in section 1.4.

The results from the *testing phase* (section 5.2) relating to objective 2 of the thesis suggest that missing data handling methods are potentially useful in correcting for bias due to range restriction following selection. However, their potential utility is restricted to data that are multivariate normal. This may be seen from the poor performance of the missing data handling methods when applied to a contrived selection example that made use of PLAB data that was not multivariate normal (section 5.3). This was further confirmed by the results obtained in section 5.4 where the missing data handling methods were used to correct for range restriction in data that deviated from multivariate normality to varying degrees. In the next chapter, missing data handling methods will be used to address some aspects of Number Needed to Reject (NNR) and Peer Competition Rescaling (PCR) in accordance to objective 4 outlined in section 1.4.

6. Estimating uncertainty about the estimate of Number Needed to Reject (NNR) and proof of concept for “Peer Competition Rescaling (PCR)”

6.1. Estimating uncertainty about the estimate of Number Needed to Reject (NNR) using resampling methods

6.1.1. Introduction

In biomedical health research, the *Number Needed to Treat (NNT)* (or *Number Needed to Harm (NNH)* when NNT is negative) is often used to convey the potential benefits or risks that a drug or intervention would have for an individual patient. It is thus used in to inform decisions about the effectiveness of a drug. NNT may estimated by *Odds Ratios (OR), Relative Risks (RR)* or as the inverse of the *Absolute Risk Reduction (ARR)*. However, the computation of NNT is typically computed from ARR. This is because the computation of NNT from OR and RR is arduous. A nomogram developed by *Chatellier et al. (1996)* may be used to compute NNT from OR and RR with relative ease. Due to the simplicity of computing NNT from ARR, the focus of this thesis will now shift to the ARR. By definition, ARR is the difference between the proportion of events for those in the treatment group and proportion of events for those in the control group (Les Irwig and Sweet, 2008; Szumilas, 2010). To demonstrate its use in computing NNT, consider a hypothetical trial conducted to compare the effects of *Cognitive Behavioural Therapy (CBT)* versus the *Interpersonal Psychotherapy (IP)* for patients with *bulimia*. If 5% of those in the *CBT* group relapsed after treatment compared to 25% of those in the *IP* group. Then the ARR would be $25\% - 5\% = 20\%$, the NNT would thus be $\frac{1}{\frac{20}{100}} = 5$. This would mean that *five patients would have to be treated with CBT rather than IP to prevent one relapse*. The ideal NNT of 1 would mean that each patient treated with the CBT would see a beneficial effect. In practice though, a NNT of 2 to 5 is typically deemed to be reasonable evidence for the effectiveness of a drug or therapy (Centre for Evidence-Based Medicine, 2017; Chatellier et al., 1996; Cook and Sackett, 1995; Public Health Action Support Team (PHAST), 2017).

The uncertainty associated with NNT is commonly expressed in confidence intervals computed from the inverse of the limits for the ARR. If the treatment effect is statistically significant at say $\alpha = 5\%$, the 95% confidence intervals for ARR will not include zero. In turn, the 95% confidence intervals for NNT will not include infinity (∞). For example, if the ARR is statistically significant at 10% then the NNT would be $\frac{1}{\frac{10}{100}} = 10$. If the 95% confidence interval for ARR is 5% to 15%, then the 95% confidence interval for the NNT would be $\frac{1}{\frac{15}{100}} = 6.67$ to $\frac{1}{\frac{5}{100}} = 20$. However, if the ARR is statistically non-significant at 10% then the NNT would still be $\frac{1}{\frac{10}{100}} = 10$. If the 95% confidence interval for ARR was -5% to 25%, then the 95% confidence interval for the NNT would be $\frac{1}{\frac{-5}{100}} = -20$ to $\frac{1}{\frac{25}{100}} = 4$. Note that the negative value for NNT would imply that the treatment would do harm to the patient. This however does not automatically follow from the statistically non-significant result obtained. In addition, the confidence interval is problematic since its range does not actually include the actual point estimate for the NNT. Therefore the computation of the uncertainty about NNT may be problematic when the treatment effect is statistically non-significant (Altman, 1998; Muthu, 2003; Sedgwick, 2013).

Within the context of selection, *Tiffin, Mwandigha, et al. (2016)* developed a concept for estimating the of effectiveness of selection tests that is analogous to NNT. This concept, referred to Number Needed to Reject (NNR), estimates the *number of good candidates that would have to be rejected in order to get rid of one poor candidate during the selection process*. The term "good candidates" in this context refers to *those applicants who are at a very low risk of a specified adverse outcome (i.e. failing at least one year at undergraduate medical school).* Good selection tests would thus be expected to have a low value for NNR. An ideal selection test would be expected to have a NNR value of 0. Recall that a NNT value of 2 to 5 is considered to be reasonable evidence for the effectiveness of a drug or therapy (Public Health Action Support Team (PHAST), 2017). Since in practice, selection occurs and is evaluated in non-ideal settings, it will be reasonable to adopt a similar classification rule for a good selection test; a NNR value of less than 5. The estimation of the NNR is demonstrated next by use of a motivating example adopted from Tiffin, Mwandigha, et al., 2016 as shown in Table 6.1. The UKCAT is used to screen entrants (at a standardised threshold of $z \leq 0$) who are likely to fail to pass at least one end year exams at medical school at first sitting. The outcome (criterion) of interest is the determination of entrants who *fail in any of their end year medical exams at first sitting during the five years of undergraduate training at medical school.* It is possible to derive well known quantities associated with screening tests, for example *Sensitivity* is $\frac{4{,}903}{9{,}273} = 0.5287$, *Specificity* is $\frac{9{,}560}{15{,}360} = 0.6223$, *Positive Predictive Value (PPV)* is $\frac{4{,}903}{10{,}693} = 0.4585$ and *Negative Predictive Value (NPV)* is $\frac{9{,}560}{13{,}930} = 0.7140$. However, the quantity of interest in this thesis is the NNR expressed as $\frac{5{,}790}{4{,}903} = 1.1809$. When rounded down, assuming the application of the screening test at selection, the value of NNR would translate to one good candidate being wrongly rejected for every poor candidate rightly rejected. Since the NNR value is less than 5, it may be concluded that the UKCAT forms a good selection test at the chosen threshold.

	Fail at least one exam at first sitting		
Screened to Fail by UKCAT ($z \leq 0$)	No	Yes	Total
No	9,560	4,370	13,930
Yes	5,790	4,903	10,693
Total	15,350	9,273	24, 623

Table 6.1.: *A scenario for the distribution of numbers following the use of total UKCAT score to screen out potential failing candidates in undergraduate medical school*

Just like the NNT, the NNR does not automatically come with a measure of its uncertainty. In fact, *Tiffin, Mwandigha, et al. (2016)* did not provide any standard errors or confidence intervals associated with their estimates of NNR. In this thesis chapter, the measure of uncertainty about NNR will be estimated by employing *resampling methods* whilst accounting for the complicated data structures that may be encountered in selection. These include the presence of incomplete data and clustering of observations dealt with in section 6.1.3.

6.1.2. Data

For the purpose of determining the uncertainty about the NNR, data were made available for the undergraduate medical school entry cohort of 2007 and 2008 who sat for the UKCAT in 2006 and 2007 respectively. The data included total UKCAT scores, socio-demographic variables such as age (dichotomised as mature when aged 21 years and above), sex (male or female), professional background, type of school attended (selective for independent and grammar schools and non-selective for state schools and sixth form colleges), ethnicity (dichotomised into white and non-white) and registration status for the UKCAT (dichotomised to identify entrants who qualified to sit for the United Kingdom Clinical Aptitude Test for Special Educational Needs (UKCATSEN) and were thus permitted additional time to complete the UKCAT test). The progression data for the entry cohorts were also available and were reported for each academic year as *graduated first sitting, left course (for academic, health, personal or other reasons), passed after first sitting, proceeded after resit* and *repeated academic year.* The progression outcomes were dichotomised into a variable called *pass* which was assigned a value of "1" for the favourable academic outcomes *graduated first sitting* and *passed after first sitting* and assigned a value of "0" for all other progression outcomes. The data consisted of 6,812 undergraduate medical school entrants in 18 universities across the UK.

6.1.3. Methods

Given the data, the total UKCAT scores were cross classified (tabulated) based on a specified threshold (thus utilising it as a screening tool) against the outcome of interest. This cross classification generated a table similar to Table 6.1. Thereafter, NNR was computed. To determine the required UKCAT threshold, the total UKCAT scores were standardised as shown on equation 6.1.1. Note that, the *"i"* denotes an entrant, *"j"* the year of sitting for UKCAT, 2007 or 2008. As implied by the equation, the standardisation was done by year of sitting of the UKCAT test. The UKCAT means, ($\overline{UKCAT_{ij}}$), and respective standard deviations (denominator in equation 6.1.1) utilised were computed from all medical school applicants (rather than entrants) who sat for the UKCAT in 2007 and 2008. The distribution of the UKCAT scores for all applicants is shown in Table 6.3. Subsequently, different thresholds (cutoffs) for screening out potential poor candidates were set at -1, -0.5, 0, 0.5 and 1. Entrants below those standardised thresholds were screened out and essentially considered to have been predicted to fail at least one exam in (undergraduate) medical school at first sitting. For the actual outcome, entrants were deemed to have failed at least one exam at first sitting if the sum of the *pass* variable (described in section 6.1.2) during their five years of undergraduate medical school training was less than five.

$$Z_i = \frac{UKCAT_{ij} - \overline{UKCAT_{ij}}}{\sqrt{(Variance(UKCAT_{ij}))}} \tag{6.1.1}$$

Determination of the uncertainty about the NNR entailed the estimation of the empirical distribution of the NNR. The resampling methods described in detail in section 8.3 of the Technical Appendices were considered. Since the data was already available to begin with, the resampling methods that lent themselves to being useful were the *bootstrap* and *jacknife*. The bootstrap is much more versatile and can potentially yield many more samples compared to the jackknife. For this reason, the bootstrap was chosen. The idea behind the bootstrap was to generate B (say) samples of same size as the original data. Thereafter, compute NNR for each of the samples. This would then enable the derivation of the empirical distribution for the NNR at each screening threshold of choice. To obtain a measure of uncertainty, the *percentile bootstrap confidence interval* described in section 8.5 of the Technical Appendices was used. The percentile bootstrap confidence interval was an appealing choice since it was easy to implement. In addition, the resulting confidence limits obtained were within the range of allowable values for the empirical distribution of NNR.

This bootstrap approach in the form described, given the data available, would understate the uncertainty about NNR. This is because of three aspects of the data. Firstly, the presence of missing data would lead to biased results for NNR and the associated percentile confidence intervals. Secondly, the undergraduate medical schools outcomes (criterion) in participating universities were local rather than national measures. Thirdly, entrants in each undergraduate medical school within a participating university were more likely to be similar due to a shared academic experience. The second and third aspect described induce a clustering effect which has been demonstrated to artificially narrow confidence intervals. This may (potentially) lead to a false estimate of uncertainty about NNR (Verbeke and Molenberghs, 2009). Therefore the issue of missing data and clustering would have to be dealt with so as to obtain good estimates for the uncertainty about the NNR.

To address the missing data, MI based on MCMC algorithm was used as described in section 4.7.2. Since the missingness affected the predictors and outcomes of interest, both of these would have to be imputed. The imputation involved the derivation of the *imputation model* for each of variables affected by missingness. For the outcome *pass*, the predictors to be included in the imputation model were determined using a *(General) Linear Mixed Model (LMM)* with a random intercept. The random intercept was modelled at the university-level with variables having a statistically significant relationship with the *pass* outcome being selected for inclusion into the *imputation model*. For the *nominal variables*, (non-white ethnicity, non-selective school attended, registered as special education needs candidate for the UKCAT) the variables to be included in their respective *imputation models* were determined by use of the χ^2 *test of* *independence*. Based on the χ^2 test, the nominal variables which had a statistically significant relationship with the nominal variable under consideration were included as predictors in its *imputation model*.

The *ordinal variables* (age $>=$ 21 years and non-professional educational background) had their respective imputation models decided based on *tetrachoric correlation* and *linear trend test*. The latter is a statistical test akin to χ^2 test of independence modified for use when one or both of the categorical variables under consideration are ordinal in nature. *Tetrachoric correlation* is the correlation coefficient computed for two ordinal variables. Variables that had a statistically signficant relationship with variable of interest were included as predictors in its *imputation model*. For the *continuous variable*, advanced qualification, the predictors for its *imputation model* were determined using a *linear regression model* as no clustering effects in the data were detected. Lastly, the inclusion of *continuous variable*, total UKCAT score (which was not affected by missingness) and advanced qualification into the imputation model of the ordinal and nominal variables were determined by *biserial* and *point biserial correlations* respectively. The variables that had a statistically significant relationship were included as predictors in the *imputation model* for the variable under consideration (Smith Hall, 2005). A summary of correlation types based on their respective variable types may be viewed in Table 8.1 in the Technical Appendices. Table 6.2 shows the summary of *imputation models* derived for each of the variables with missing values in the data.

		Missing								
		Non-white ethnicity	Non-selective school attended	Non-professional educational background	Advanced qualification	Pass at year 1	Pass at year 2	Pass at year 3	Pass at year 4	Pass at year 5
Impute with	Non-white ethnicity		✓	✓	✗	✓	✓	✓	✓	✓
	Non-selective school attended	✓		✓	✗	✗	✓	✗	✗	✗
	Male sex	✓	✗	✗	✗	✗	✓	✓	✓	✓
	Registered as special education needs for UKCAT	✓	✗	✗	✗	✗	✗	✗	✗	✗
	Age >= 21 at entry	✗	✓	✓	✗	✗	✗	✗	✗	✗
	Non professional educational background	✓	✓		✗	✗	✗	✗	✗	✗
	Advanced qualification	✗	✓	✓		✓	✓	✓	✓	✓
	Total UKCAT score	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Pass at year 1	✗	✗	✗	✗		✓	✓	✓	✓
	Pass at year 2	✗	✗	✗	✗	✗		✓	✓	✓
	Pass at year 3	✗	✗	✗	✗	✗	✗		✓	✓
	Pass at year 4	✗	✗	✗	✗	✗	✗	✗		✓
	Pass at year 5	✗	✗	✗	✗	✗	✗	✗	✗	

Table 6.2.: *Imputation matrix for the predictors and outcomes of interest*

Based on the imputation models, MI based on MCMC algorithm was conducted with both P and n_{iter} set at 50. Subsequently, NNR was computed for the different chosen threshold values of the total UKCAT scores for each sample. This resulted in 50 NNR values which were taken to represent the empirical distribution of the NNR. The point estimate of NNR was computed as the arithmetic mean across all the 50 values and the 95% confidence interval derived from the 2.5^{th} and 97.5^{th} percentiles of the empirical distribution. Note that if data is affected by missingness, then MI presents an alternative method to the bootstrap method. The MI approach in this case would however underestimate the uncertainty about the NNR for the present data since the clustering nature of the data is ignored. To correct for this, one approach would be to conduct MI within each cluster and then proceed to compute NNR and 95 % confidence interval derived from the 2.5^{th} and 97.5^{th} percentiles of the empirical distribution as before.

The bootstrap method is useful in instances where it may not be desirable (due to choice) or even possible (as is the case where data is complete) to use MI. For this thesis, the *case resampling bootstrap* was implemented as described in sections 8.3 and 8.4 of the Technical Appendices. The case resampling bootstrap was modified to account for the clustering (medical school within universities) in the data by sampling at each level of the data. Firstly, the cluster (university) which was the level 2 unit was sampled with replacement from the pool of available universities in the data. Secondly, *case resampling bootstrap* was implemented from the sampled university thus sampling level 1 units (students). In order to have a complete data prior to implementing the case resampling bootstrap modified for clustered the data, single imputation was conducted based on the derived *imputation models* in Table 6.2. Thereafter, the NNR and its 95% confidence intervals were derived from the 2.5^{th} and 97.5^{th} percentiles of the empirical distribution obtained.

6.1.4. Results

6.1.4.1. Descriptive statistics

Table 6.3 shows the summary statistics for the total UKCAT scores. In 2006, the mean score was lower with higher standard deviation compared to 2007. It was observed in Table 6.4 that the cohort that sat for the UKCAT in 2006 and enrolled into undergraduate medical school in 2007 was better represented in the data with respect to higher numbers and lower data attrition (missing values) rates. Both cohorts suffered the highest data attrition (missing values) rates in year four and five of medical school which can also be seen from the low number of universities that participated in the study in those years. This was mainly attributed to medical schools deciding not to return entrants' performance data as the study progressed (later years four and five) rather than entrants leaving or dropping out of medical school as described by Mwandigha et al., 2018.

	Year of sitting=2006		Year of sitting=2007	
Scale	Mean	SD	Mean	SD
Total UKCAT score	2,480.52	216.35	2,521.30	198.78

Table 6.3.: *Means and standard deviation of the UKCAT scores for medical school applicants who sat for the UKCAT in 2006 to 2007*

Year of Program	Year of entry=2007			Year of entry=2008		
	No. of universities	No. of students	% Attrition	No. of universities	No. of students	% Attrition
1	17	3,505	-	18	2,866	-
2	18	3,426	2.25	18	2,810	1.95
3	18	3,158	9.90	17	2,366	17.45
4	16	2,636	24.79	15	1,977	31.02
5	13	1,944	44.54	9	950	66.85

Table 6.4.: *Data attrition (missing values) rates for the two cohorts (2007 and 2008 entrants)*

.

Table 6.5 summarises the distribution of the socio-demograhic variables of the two entry cohorts. A majority of entrants were of white ethnicity, female, from non-selective schools, younger than 21 years of age with a professional socio-economic background. Three variables, namely, sex, age and UKCAT registration were complete with no missingness observed.

Variable	Proportion (%)	Missing (%)
Non-white ethnicity	2,053 / 6,714 (30.58)	98 / 6812 (1.17)
Male sex	2,874 / 6,812 (42.19)	0 / 6812 (0.00)
Non-selective school attended	3,097 / 5,725 (54.10)	1,087 / 6812 (15.96)
Non-professional socio-economic background	125 / 5,653 (2.21)	1,159 / 6812 (17.01)
Age>=21 years at entry	1,147 / 6,812 (16.84)	0 / 6812 (0.00)
Registered as special educational needs for UKCAT	65 / 6,812 (0.95)	0 / 6,812 (0.00)

Table 6.5.: *Socio-demographic characteristics of the two entry cohorts*

Table 6.6 provides further information regarding the pattern of missingness for the socio-demographic variables. It was observed that about 61% of the observations in the data had complete information. The most frequently occurring missing profile was for observations with data on all variables but the non-professional socio-economic background variable. The least frequently occurring missing profile was for observations with data on all the variables but advanced qualification and non-white ethnicity. Table 6.7 shows the pattern of missingness for the *pass* outcome. It was observed that only 39.97% of the outcome data was complete. There were 48.75% observations with monotone pattern of missingness with the most frequently occurring profile having data missing only for year 5 of medical school. The most frequently occurring arbitrary pattern of missingness had missingness in year 3 and 5 of medical school.

Variables								Count	%
Group	Non-white ethnicity	Male sex	Non selective school attended	Non-professional socio-economic background	Registered as special educational needs for UKCAT	Age>=21 years at entry	Advanced qualification		
Complete									
1	O	O	O	O	O	O	O	4,135	60.70
Missing									
2	O	O	O	O	O	O	M	608	8.93
3	O	O	O	M	O	O	O	762	11.19
4	O	O	O	M	O	O	M	141	2.07
5	O	M	O	O	O	O	O	287	4.21
6	O	O	M	O	O	O	M	601	8.82
7	O	O	M	M	O	O	O	53	0.78
8	O	O	M	M	O	O	M	127	1.86
9	M	O	O	O	O	O	O	14	0.21
10	M	O	O	O	O	O	M	1	0.01
11	M	O	O	M	O	O	O	57	0.84
12	M	O	O	M	O	O	M	7	0.10
13	M	O	M	O	O	O	O	3	0.04
14	M	O	M	O	O	O	M	4	0.06
15	M	O	M	M	O	O	O	3	0.04
16	M	O	M	M	O	O	M	9	0.13
Total								6,812	100

Table 6.6.: *Patterns of missingness in the variables. Each "O" and "M" represents each instance where data are present and absent respectively (Note that the first row represents the proportion of cases with no missing data).*

Outcome						Count	%
Group	Year 1	Year 2	Year 3	Year 4	Year 5		
Complete							
1	O	O	O	O	O	2,723	39.97
Monotone pattern of missingness							
2	O	O	O	O	M	1,280	18.79
3	O	O	O	M	M	1,089	15.94
4	O	O	M	M	M	748	10.98
5	O	M	M	M	M	153	2.25
6	M	M	M	M	M	54	0.79
Arbitrary pattern of missingness							
7	O	O	O	M	O	113	1.66
8	O	O	M	O	O	3	0.04
9	O	O	M	O	M	231	3.39
10	O	M	O	O	O	6	0.09
11	O	M	O	O	O	9	0.13
12	O	M	O	M	M	14	0.21
13	O	M	M	O	O	4	0.06
14	O	M	M	O	M	1	0.01
15	M	O	O	O	O	17	0.25
16	M	O	O	O	M	14	1.21
17	M	O	O	M	M	5	0.07
18	M	O	M	O	M	1	1.01
19	M	O	M	M	M	15	0.22
20	M	M	O	O	O	28	0.41
21	M	M	O	O	M	218	3.20
22	M	M	O	M	M	11	0.16
23	M	M	M	O	M	78	1.15
Total						6,812	100

Table 6.7.: *Patterns of missingness for pass each year. Each “O” and “M” represents each instance where data are present and absent respectively (Note that the first row represents the proportion of cases with no missing data). Patterns are categorised as either monotone (i.e. where data relating to all subsequent years are missing after the initial missing data year) or arbitrary (i.e. non-monotone)*

6.1.4.2. Inferential statistics

Figure 6.1 shows the *sensitivity* (probability of a true fail in at least one exam at first sitting in undergraduate medical school) and *specificity* (probability of a true pass in all exams at first sitting in undergraduate medical school). These were obtained from the MI approach conducted whilst disregarding the clustered (hierarchical) nature of the data. It was observed that, for lower thresholds of total UKCAT scores, the *sensitivity* was lower. However, this *sensitivity* increased with the raising of the total UKCAT scores threshold. Conversely, at lower thresholds of total UKCAT scores, the *specificity* was higher. However, this *specificity* decreased with the raising of the UKCAT threshold.

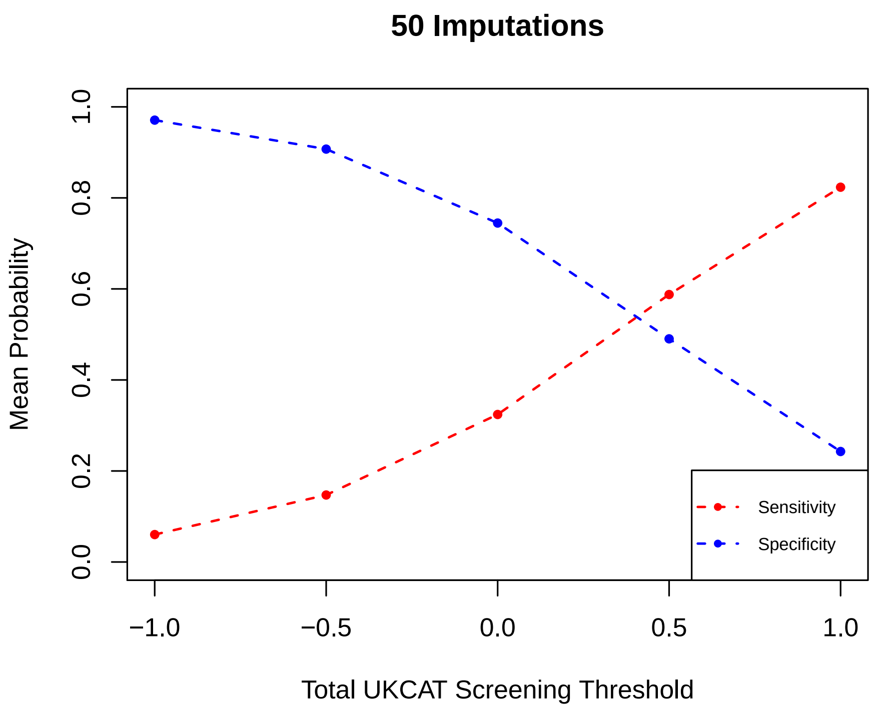


Figure 6.1.: *Sensitivity and specificity based on multiply imputed data for different thresholds of total UKCAT scores. MI was conducted without regard to university effect.*

Figure 6.2 shows the PPV and NNR from the MI approach conducted whilst disregarding the clustered (hierarchical) nature of the data. It was observed that PPV was much higher for lower threshold values of total UKCAT scores. However, the PPV declined as the total UKCAT scores threshold was raised. The opposite trend was observed for NNR with lower values observed for lower threshold values of total UKCAT scores. The NNR increased as the total UKCAT scores threshold was raised. Lower thresholds of total UKCAT scores were associated with wider confidence intervals for both the PPV and NNR.

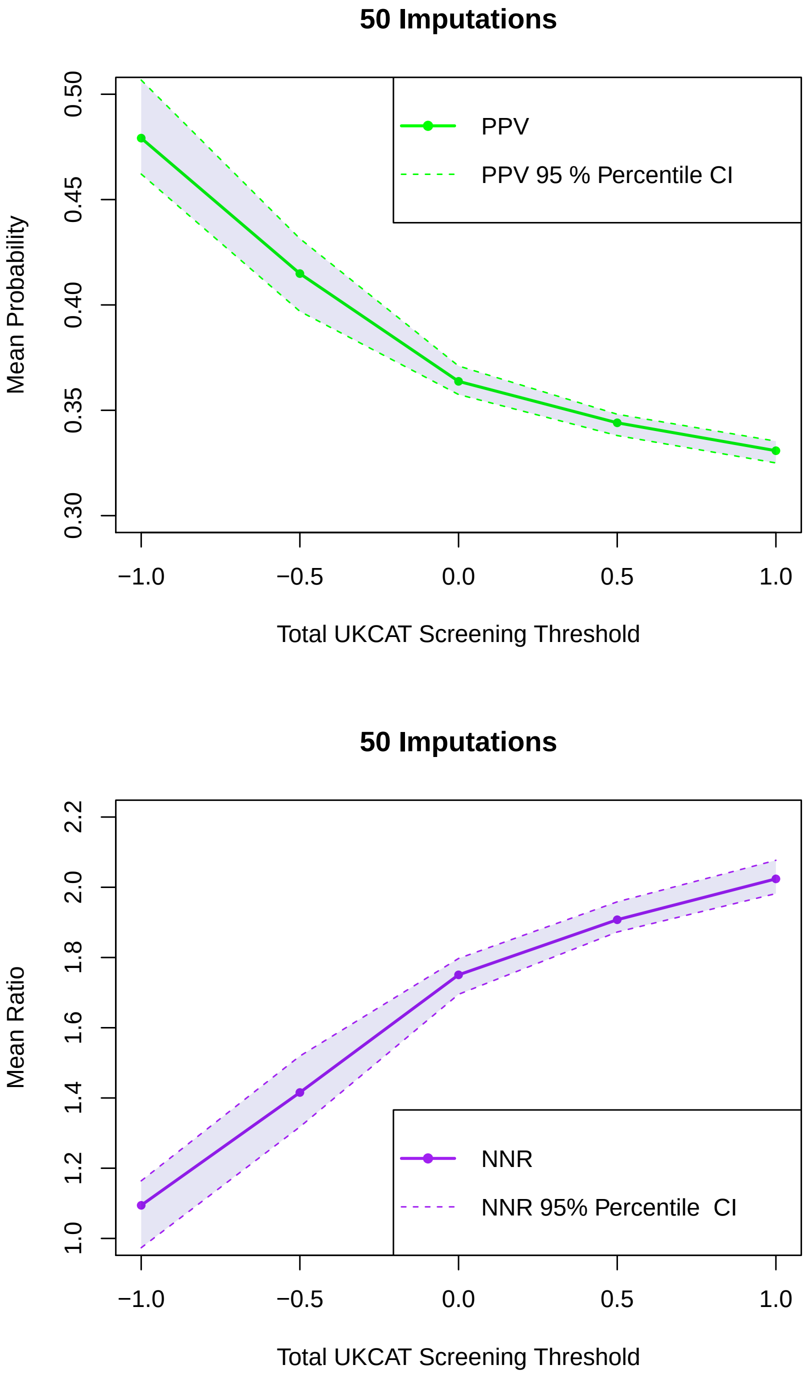


Figure 6.2.: *Positive Predictive Value (PPV) and Number Needed to Reject (NNR) based on multiply imputed data for different thresholds of total UKCAT scores. MI was conducted without regard to university (cluster) effect.*

An attempt to conduct MI within each cluster (university) so as to account for the hierarchical nature of the data was abortive. Further investigations revealed that this was caused by what may be referred to as *“complete separation of data points”*. Table 6.8 shows the clusters (universities) associated with this problem. As may be seen in the Table, values for the *pass* outcomes for these universities were either completely non existent for all observations or had all observed values available for only one level of the variable (i.e. Pass or Fail). Therefore, for these (clusters) universities, it was not possible to derive the predictive distribution of the pass outcome given the variables in the *imputation model* so as to conduct MI.

	Outcome of year 1 of medical school			Outcome of year 5 of medical school		
University identifier	Pass	Fail	Missing	Pass	Fail	Missing
136	1	-	322	-	-	323

	Outcome of year 3 of medical school		
University identifier	Pass	Fail	Missing
168	247	-	18

	Outcome of year 4 of medical school			Outcome of year 5 of medical school		
University identifier	Pass	Fail	Missing	Pass	Fail	Missing
108				48	-	220
148	-	-	389	-	-	389
156	-	-	265	-	-	265
160	18	-	295	13	-	300
180				-	-	298
188				-	-	416

Table 6.8.: *Complete separation of data points within the universities in the data.*

In order to proceed with MI within each cluster (university), the 8 universities affected with complete separation of data points were excluded from the data. This left the data consisting of 10 universities unaffected with complete separation of data points. The data had a total of 4,275 observations. Note that this represented a loss of 2,537 observations or 37.24% of the data. Subsequent results from the MI conducted using MCMC algorithm based on *imputation models* in Table 6.2 are available in Figures 6.3 and 6.4. It was observed that the trends associated with *sensitivity, specificity*, PPV and NNR were similar to those observed with MI conducted across all clusters while disregarding the hierarchical nature of the data. Unexpectedly, the percentile confidence intervals resulting from the MI conducted within each cluster were narrower.

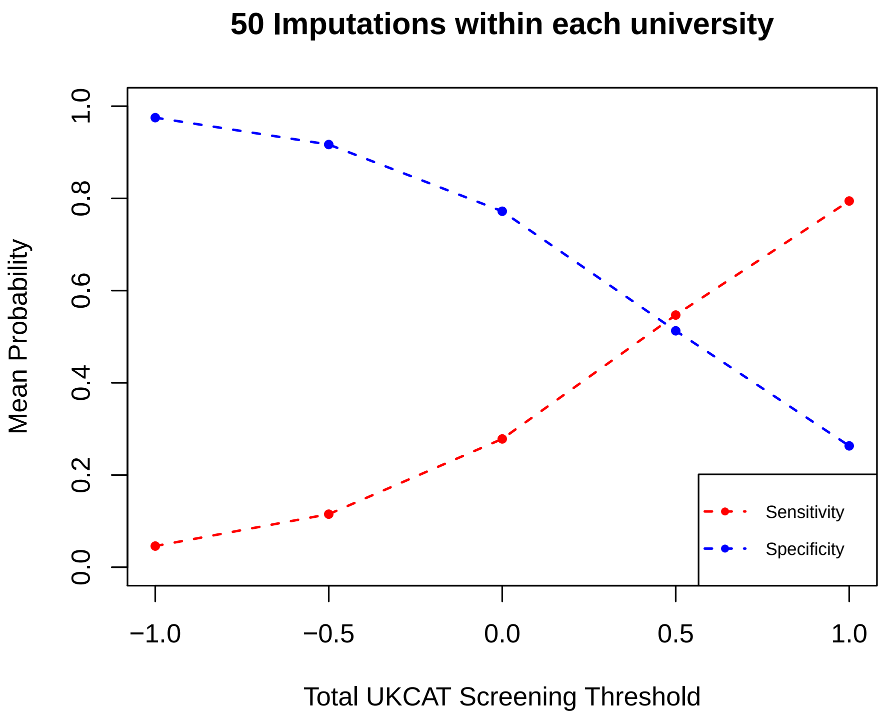


Figure 6.3.: *Sensitivity and specificity based on MI implemented within each cluster (university) for different thresholds of the total UKCAT scores.*

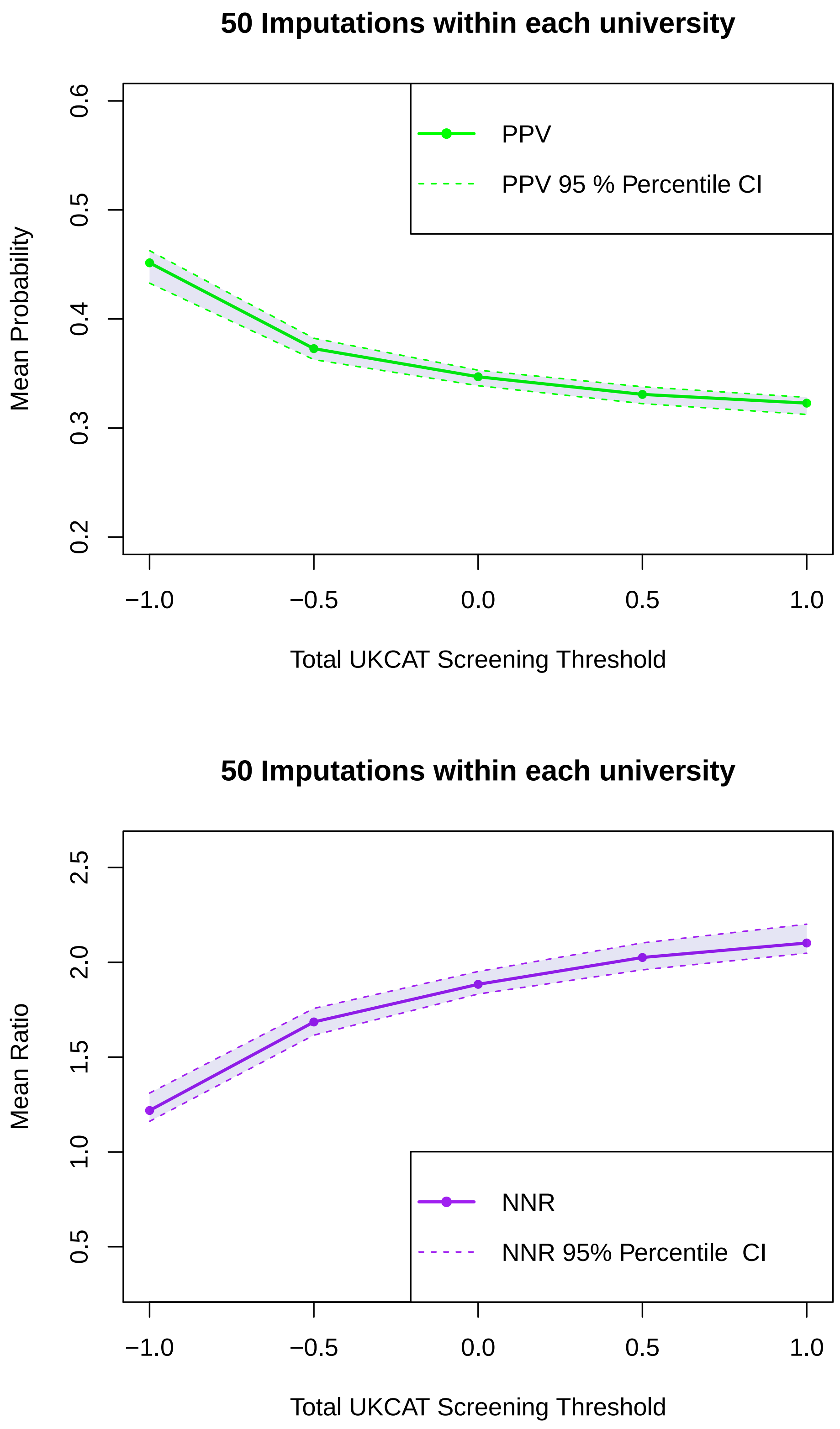


Figure 6.4.: *Positive Predictive Value (PPV) and Number Needed to Reject (NNR) based on MI implemented within each cluster (university) for different thresholds of the total UKCAT scores.*

A different approach that took into account the clustering in the data consisting of all 6,812 observations in the 18 universities was considered. This lead to the use of *case resampling bootstrap* which was implemented at levels 1 and 2 of the data. The results of *sensitivity, specificity*, PPV and NNR may be viewed in Figures 6.5 and 6.6. With the exception of confidence intervals for PPV and NNR, the trends for *sensitivity, specificity*, PPV and NNR were similar to those observed for MI conducted while disregarding the hierarchical nature of the data and MI conducted within each cluster. There was a difference in the confidence intervals for PPV and NNR obtained from the three approaches. The *case resampling bootstrap* implemented at levels 1 and 2 of the data had the widest confidence intervals of the three approaches across all thresholds of the total UKCAT scores. In addition, it had confidence intervals that were approximately the same width across all thresholds of the total UKCAT scores. For the MI approaches conducted, the confidence intervals were much wider for lower threshold values of the total UKCAT scores. The MI approach conducted within each cluster resulted in the narrowest confidence intervals of the three approaches.

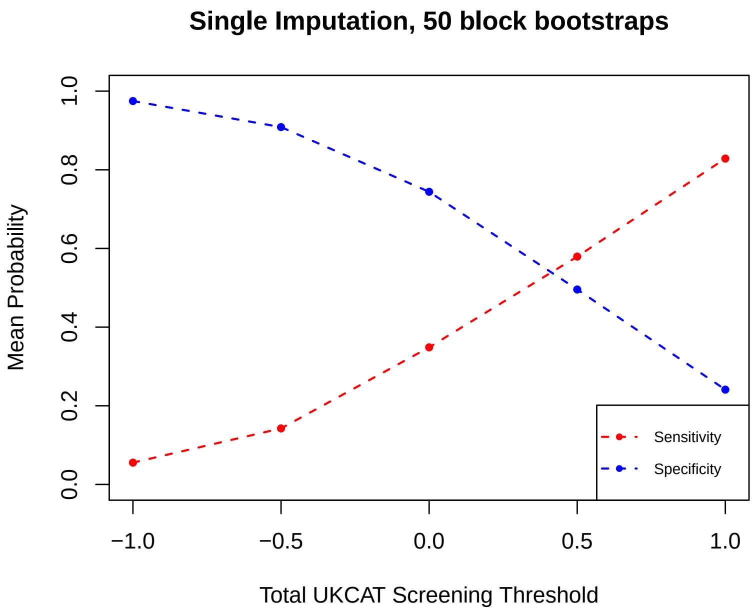


Figure 6.5.: *Sensitivity and specificity based on single imputation followed by case resampling bootstrap implemented at student and university levels of the hierarchical data for different thresholds of the total UKCAT scores.*

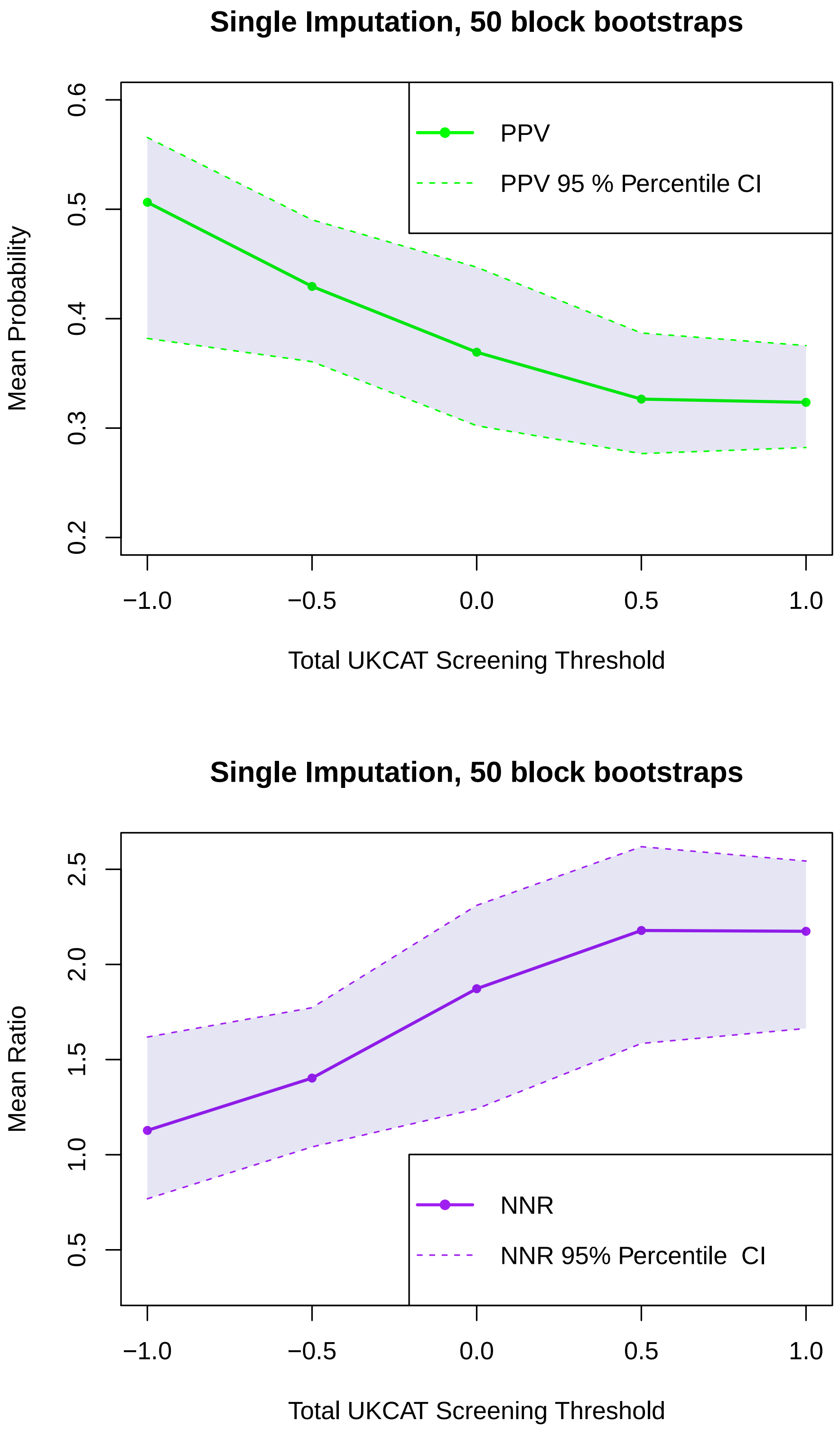


Figure 6.6.: *Positive Predictive Value (PPV) and Number Needed to Reject (NNR) based on single imputation followed by case resampling bootstrap implemented at student and university-levels of the hierarchical data for different thresholds of the total UKCAT scores.*

To conclude this section, it has been demonstrated that MI and *case resampling bootstrap* are useful methods for estimating the uncertainty about NNR. Although these methods may be computationally intensive, the pay-offs are great. This is because these methods can handle data complexities such as incompleteness and clustering. In addition, unlike in the case of the NNT computed from the inverse of the ARR, the confidence intervals obtained for the NNT always contain plausible point estimates of the Number Needed to Reject (NNR).

6.2. Proof of concept for Peer Competition Rescaling: “Nationalising” local outcome measures

6.2.1. Introduction

In medical education in the UK, national decisions are partly based on local outcomes. For example, the Educational Performance Measure (EPM) is a medical school measure based on a ranking of a graduate within their medical school (Medical School Council, 2017a). However, this measure is partly used to select candidates for the national *Foundation Programme* (Foundation Programme, 2017). Likewise, attempts to establish the predictive validity of predictors make use of local measures such as *knowledge* and *skills*-based medical school exams. The use of local measures of performance as proxies for national ones may lead to bias and incorrect inference. To correct for this, it is worthwhile to explore methods for adjusting local outcomes in order to “nationalise” them. This may be considered something of a case of a “missing data” problem as the national measure is non-existent but is to be estimated from local measures. Local performance is typically determined by the overall level of ability of a medical school entrant within the entrant’s medical school cohort. This is particularly true where percentile rankings are used for an entrant.

It may be possible to use national, observable, predictors such as the UKCAT in order to assess the amount of “competition” a entrant may face when being judged on a local performance.

This introduced information on "peer competition" could be used to weigh or rescale local measures to get a better estimate of how such a entrant may have performed on a national-level measure if one existed. This estimate of performance would be fairer than the current Educational Performance Measure (EPM) as medical school graduates would be compared on the same scale across all medical schools in the UK. For this reason, this concept introduced by *Tiffin and Paton (2017)* is aptly referred to as *"Peer Competition Rescaling (PCR)"*. Conceptually, the idea demonstrated in Figure 6.7, will entail using path *a* to develop path *b* based on some guidelines. Subsequently attempts may be made to empirically deduce whether statistically, path *b* and *c* are equivalent.

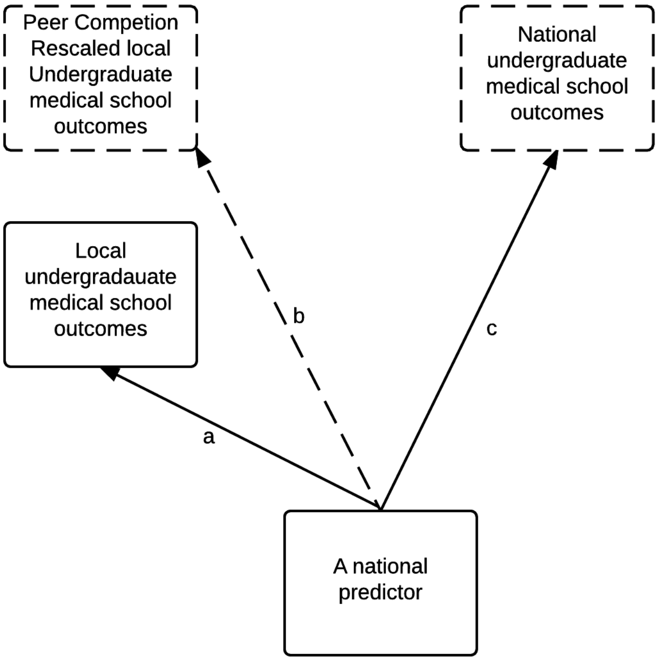

Peer Competion Rescaled local Undergraduate medical school outcomes
National undergraduate medical school outcomes
Local undergradauate medical school outcomes
b
c
a
A national predictor

Figure 6.7.: *Conceptual diagram for the Peer Competition Rescaling*

6.2.2. Data

In order to examine PCR, the available data described in sections 6.1.2 and 6.1.4.1 were used. For the medical school progression data, *knowledge* and *skills*-based exams for all the five years of undergraduate medical school training were used.

6.2.3. Methods

In order to provide proof of concept and validation for PCR, it was important to make use of a national observable predictor and local outcomes to predict performance on a national outcome. The national predictor considered was the total UKCAT scores. The local outcomes considered were *knowledge* and *skills*-based exam outcomes in (undergraduate) medical school. The national observable predictor, the UKCAT, was used to adjust the readily available *knowledge* and *skills* based local outcomes in order to achieve PCR. The modelling was conducted using the mixed model approach where both fixed and random effects were modelled using the (General) Linear Mixed Model (LMM) (Verbeke and Molenberghs, 2009). This was necessitated by the hierarchical nature of the data consisting of *knowledge* and *skills*-based local outcomes nested within the 18 universities present in the data.

Using the LMM, univariable and multivariable models were fitted for each of the local outcomes, *knowledge* and *skills*-based exams, for each of the five years of medical school training. The outcomes were allowed to vary by university through incorporating a random intercept at university level thus accounting for the hierarchical nature of the data. The univariable model was used to assess the association between total UKCAT scores and the outcomes of interest. The multivariable model went further by assessing the association between UKCAT scores and the outcomes of interest while controlling for various predictors. The predictors considered were advanced qualifications, non-white ethnicity, male sex, non-selective school attended, non-professional socio-economic background and age $>=$ 21 years. The univariable and multivariable models were fitted in two steps to enable comparisons before and after PCR was conducted. The first step made use of the reported local outcomes while the second step made use of the reported local outcomes adjusted for PCR as previously described by Tiffin and Paton (2017). This was accomplished by first standardising the total UKCAT applicants’ scores for each cohort as shown in equation (6.1.1). The undergraduate medical school outcomes (*knowledge* and *skills*-based exams) were then rescaled by adding to them a mean of the total UKCAT score for each medical school cohort. Subsequently, the outcomes were then divided by the standard deviation of the total UKCAT score specific to that medical school year. Thus, the standardisation which occurred within each medical school year was adjusted for by how well a student's peers performed on the UKCAT. A more detailed explanation for computing PCR is provided in section 8.7 in the Technical Appendices for the interested reader. The explanation is split into three sections, the first section 8.7.1 enumerates the four key guidelines that should be followed when computing PCR. The application of these guidelines are shown through a simulated example in R software that involves generation of selection data for 100 applicants applying to two hypothetical medical schools in section 8.7.2. Thereafter in section 8.7.3, using the generated selection data with one predictor and outcome, PCR is conducted thus generating a re-scaled outcome. The R code provided in sections 8.7.2 and 8.7.3 may be copy pasted or typed and run in R as a reproducible example or modified to cater for more selection complex scenarios (e.g. more than the two medical schools considered). Key assumptions like the mean and covariance structure of the selection data, selection ratio, number of applicants to medical schools and variability between medical schools may be changed in the R code as desired.

6.2.4. Results

6.2.4.1. Descriptive statistics

The descriptive statistics for the predictors in the data may be viewed in section 6.1.4.1. For the progression data from undergraduate medical school, the descriptive statistics are as follows. In Table 6.9, it was observed that only 23.18% of the observations had complete data for the *knowledge*-based exam outcomes, with 44.91% having monotone pattern of missingness. The most frequently occurring monotone pattern of missingness profile had drop out at the fifth year while the least frequently occurring monotone pattern of missingness profile had data only for the first year of undergraduate medical school training. There were 19 different arbitrary pattern of missingness profiles with the most frequently occurring profile having data available for year 4 and 5 only.

In Table 6.10, it was noted that complete information for the *skills*-based exam outcome score were available for only about 20% of the data. About 38% of the missingness was monotone pattern of missingness with the most frequently occurring monotone pattern of missingness profile having no data across the 5 years of undergraduate medical school training. There were twenty one different arbitrary pattern of missingness profiles for the *skills*-based exam outcome score with the most frequently occurring profile having data only for year 5 of medical school.

Outcome						Count	%
Group	Year 1	Year 2	Year 3	Year 4	Year 5		
Complete							
1	O	O	O	O	O	1,579	23.18
Monotone pattern of missingness							
2	O	O	O	O	M	1,035	15.19
3	O	O	O	M	M	870	12.77
4	O	O	M	M	M	549	8.06
5	O	M	M	M	M	133	1.95
6	M	M	M	M	M	473	6.94
Arbitrary pattern of missingness							
7	O	O	O	M	O	111	1.63
8	O	O	M	O	O	351	5.15
9	O	O	M	O	M	265	3.89
10	O	O	M	M	O	2	0.03
11	O	M	O	O	O	6	0.09
12	O	M	O	O	M	3	0.04
13	O	M	O	M	M	17	0.25
14	O	M	M	M	O	1	0.01
15	M	O	O	O	O	4	0.06
16	M	O	O	O	M	14	0.21
17	M	O	O	M	M	9	0.13
18	M	O	M	O	O	13	0.19
19	M	O	M	M	M	13	0.19
20	M	M	O	O	O	1	0.01
21	M	M	O	O	M	458	6.72
22	M	M	O	M	M	28	0.41
23	M	M	M	O	O	476	6.99
24	M	M	M	O	M	168	2.47
25	M	M	M	M	O	233	3.42
Total						6,812	100

Table 6.9.: *Patterns of missingness for knowledge-based exam scores. Each “O” and “M” represents each instance where data are present and absent respectively (Note that the first row represents the proportion of cases with no missing data). Patterns are categorised as either monotone (i.e. where data relating to all subsequent years are missing after the initial missing data year) or arbitrary (i.e. non-monotone).*

Outcome						Count	%
Group	Year 1	Year 2	Year 3	Year 4	Year 5		
Complete							
1	O	O	O	O	O	1,338	19.64
Monotone pattern of missingness							
2	O	O	O	O	M	672	9.86
3	O	O	O	M	M	413	6.06
4	O	O	M	M	M	671	9.85
5	O	M	M	M	M	99	1.45
6	M	M	M	M	M	699	10.26
Arbitrary pattern of missingness							
7	O	O	O	M	O	110	1.61
8	O	O	M	O	O	2	0.03
9	O	O	M	O	M	205	3.01
10	O	O	M	M	O	3	0.04
11	O	M	O	O	M	3	0.04
12	O	M	O	M	M	15	0.22
13	O	M	M	O	O	5	0.07
14	O	M	M	O	M	1	0.01
15	M	M	M	M	O	1	0.01
16	M	O	O	O	O	106	1.56
17	M	O	O	O	M	193	2.83
18	M	O	O	M	M	107	1.57
19	M	O	M	O	M	77	1.13
20	M	O	M	M	M	228	3.35
21	M	M	O	O	O	407	5.97
22	M	M	O	O	M	357	5.24
23	M	M	O	M	O	2	0.03
24	M	M	O	M	M	351	5.15
25	M	M	M	O	O	147	2.16
26	M	M	M	O	M	81	1.19
27	M	M	M	M	O	519	7.62
Total						6,812	100

Table 6.10.: *Patterns of missingness for skills-based exam scores. Each "O" and "M" represents each instance where data are present and absent respectively (note that the first row represents the proportion of cases with no missing data). Patterns are categorised as either monotone (i.e. where data relating to all subsequent years are missing after the initial missing data year) or arbitrary (i.e. non-monotone).*

6.2.4.2. Inferential statistics

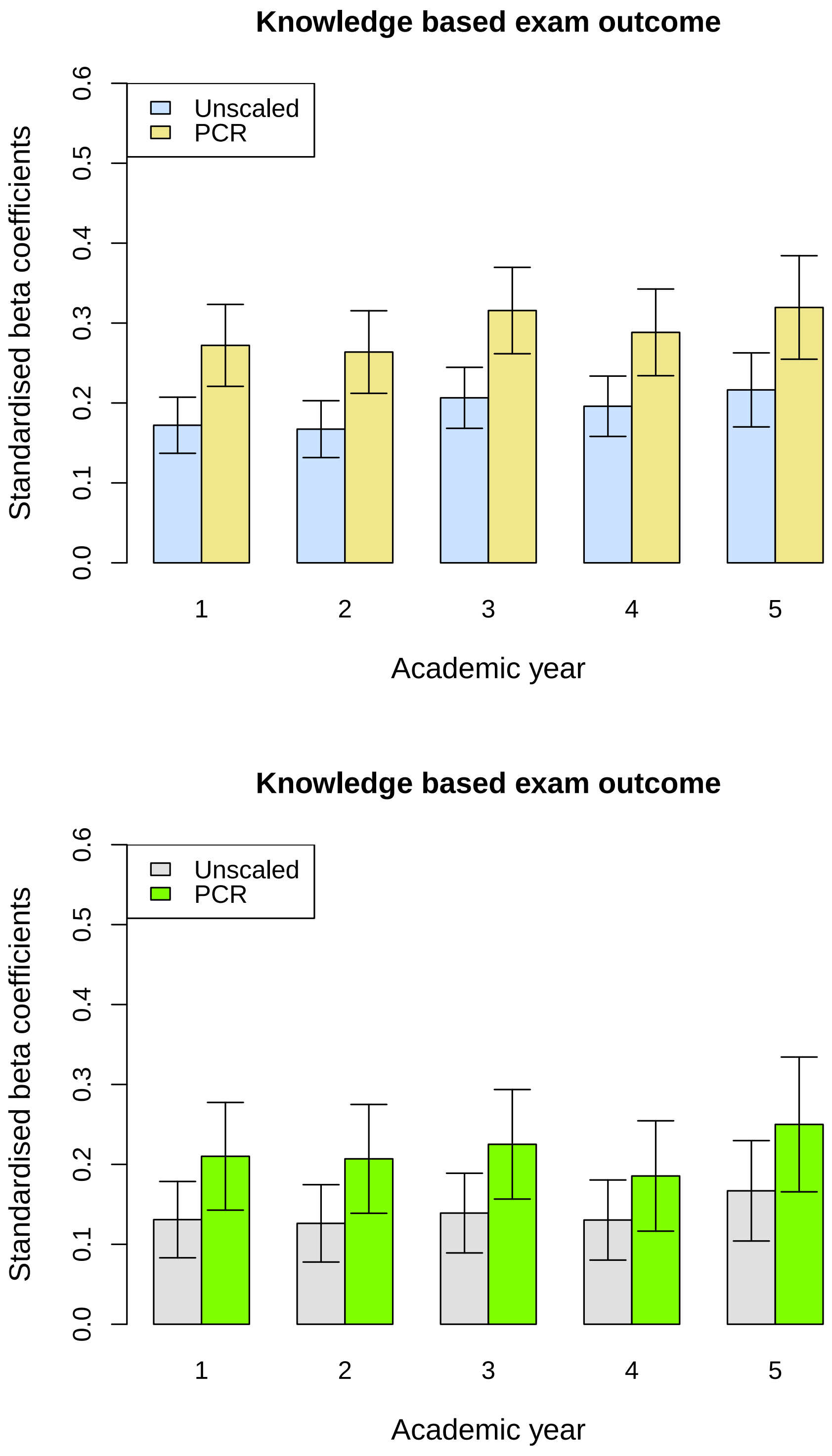


Figure 6.8.: *Results on Unscaled and and Peer Competition Rescaling (PCR) from the univariable (top panel) and multivariable (bottom panel) linear mixed models for knowledge-based exam outcomes. The size of the bars represent the magnitude of the (un)scaled coefficients while the line segments on the bars represent the estimated 95% confidence intervals.*

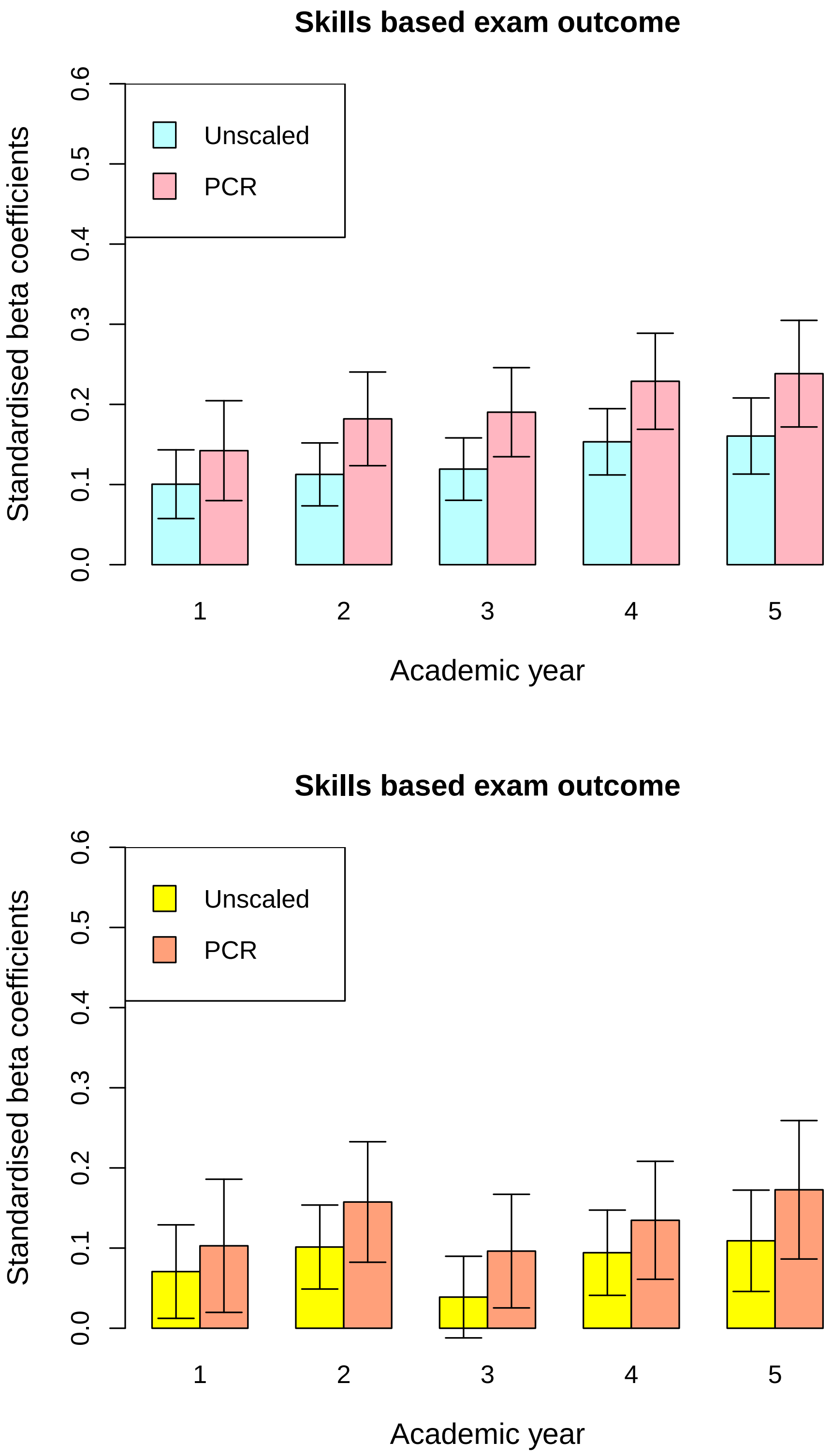


Figure 6.9.: *Results on Unscaled and and Peer Competition Rescaling (PCR) from the univariable (top panel) and multivariable (bottom panel) linear mixed models for skills-based exam outcomes. The size of the bars represent the magnitude of the (un)scaled coefficients while the line segments on the bars represent the estimated 95% confidence intervals.*

Figures 6.8 and 6.9 show the results of the models fitted before and after peer competition rescaling for *knowledge* and *skills*-based exams outcomes uncontrolled and controlled for predictors of interest. It was observed that, all the magnitudes of the β coefficients for the *knowledge*-based exam outcomes were bigger than those for *skills*-based outcomes. For both outcomes considered, the PCR coefficients were bigger in magnitude than those without the rescaling. In some instances, like the third year of medical school training, the PCR converted the statistically non-significant multivariable association between the standardised total UKCAT scores and *skills*-based outcome into a statistically significant association.

6.3. Chapter summary

In this thesis chapter, the uncertainty about the NNR was addressed using MI and resampling methods. The estimation of the uncertainty about NNR was conducted for clustered and non-clustered data. The confidence intervals for NNR obtained not only contained a plausible point estimate of the NNR but the confidence limits were themselves plausible values of NNR. The problem of "nationalising" local outcomes was also dealt with by means of "Peer Competition Rescaling (PCR)". The validity estimates obtained after PCR were larger in magnitude compared to the unscaled validity estimates. This suggests that the use of local outcome measures to make national decisions in medical education may need further scrutiny. This work presented in thesis chapter was done in accordance to objective 4 outlined in section 1.4. In the next chapter, this thesis will be concluded by summarising the key aspects of the thesis. Implications for selection practice and policy will also be discussed.

7. Discussion

In this chapter, this thesis will be concluded. This will entail revisiting the contents of the past chapters and showing how they line up with the objectives that were outlined in section 1.4. Further, findings from the thesis will also be discussed with (potential) future practice and policy implications examined in section 7.5.

7.1. Main findings

7.1.1. Construct-level predictive validity

In chapter 1, the focus and scope of the thesis was introduced. This was broken down into a series of four objectives to be dealt with. These are outlined in section 1.4. The main thrust of the thesis dealt with the use of selection tests (aptitude tests) for undergraduate medical selection. In accordance to objective 1(a), the predictors of medical undergraduate medical school performance used in selection were reviewed. These predictors were discussed in chapter 2. Based on a review of existing literature, these predictors were classified into two, the *short-listing* and *final* stage selection predictors. As a foundation for further work based on objective 1(b) of the thesis outlined in section 1.4, special consideration was given to the short-listing predictor-selection tests (aptitude tests). Selection tests are widely used across the world for undergraduate medical school selection. Several studies have researched the usefulness of selection tests. These studies are known as predictive validity studies. Specifically, predictive validity is the estimate of the association between the selection test and outcome (criterion) of interest. Thus, it is a type of *criterion related validity* (NOVA South Eastern University, 2017; Statistics How To, 2017).

In predictive validity studies, the outcome (criterion) is only observed for the entrants rather than the whole range of applicants. This is referred to as range restriction, which leads to a type of sample (selection) bias. This is because inferences based on predictive validity estimates are made from a sample that is not selected randomly from a population of applicants. Thus, the predictive validity estimates lead to inferences about the applicants made from the entrants. This violates a key requirement for *external validity* (generalisability) since the sample of selected entrants is not representative of the population of applicants for which inference is to be made (Bracht and Glass, 1968; Godwin et al., 2003; Lynch Jr, 1982; Rothwell, 2005). In chapter 3, it was demonstrated from a review of literature and Monte Carlo simulations that in predictive validity studies, only correlation coefficients (rather than regression coefficients and odds ratios) are biased downwards (Bengt O. Muthén and Asparouhov, 2016, pp 443-445; Fife, Mendoza, and Terry, 2013). This is because one of the key assumptions underpinning the estimation of correlation coefficients is bivariate normality (Albers and Kallenberg, 1994; Eric W Weisstein, 2017; Fosdick and Raftery, 2012). Therefore range restriction on the selection test or outcome distorts bivariate normality which in turn affects the accuracy of the estimates for correlation coefficients. Note that range restriction may lead to attenuated regression coefficients only when the selection test and outcome (criterion) are standardised prior to regression modelling. This is because, the standardised estimates from the regression model would be (partial) correlation coefficients (Bengt O. Muthén and Asparouhov, 2016, pp 443-445; Burt, 1943; General Medical Council, 1973; McManus, Dewberry, Nicholson, Dowell, et al., 2013). Further, it is worth mentioning that the standardisation prior to regression modelling would only be problematic if it was done using information from the restricted sample (entrants) rather than from all applicants.

A review of literature conducted in accordance to objective 1(b) whose results are summarised in Table 2.3 revealed that a vast majority (about 80%) of the predictive validity estimates from the studies were correlation coefficients. Only about 15% of these studies were corrected for the downward bias due to range restriction. This means that most of the correlation coefficients (predictive validity estimates) reported in literature are, in effect, underestimated. Therefore, a review of the literature was conducted in accordance to objective 1(c) to determine the statistical methods used to correct for this downward bias due to range restriction. The results of this review of the literature were presented in chapter 3. The statistical methods presented include *Thorndike Case I, Thorndike Case II, Thorndike Case III* and *Pearson Lawley* corrections (Allen and Dunbar, 1989; Alliger, 1987; Duan and Dunlap, 1997; Dunbar and Linn, 1991; Fife, Hunter, and Mendoza, 2016; Held and Foley, 1994; Li, Chan, and Cui, 2011; Saupe and Eimers, 2010; Schmidt, Oh, and Le, 2006; Sjöberg et al., 2012). In addition, it was found that missing data handling methods, Full Information Maximum Likelihood (FIML) and Multiple Imputation (MI), are now increasingly used to correct for bias due to range restriction (Mendoza, Bard, et al., 2004; Pfaffel, Kollmayer, et al., 2016; Pfaffel, Schober, and Spiel, 2016; Pfaffel and Spiel, 2016; Wiberg and Sundström, 2009). These missing data handling methods were introduced in chapter 3 but covered more extensively in chapter 4. The performance of all the statistical methods for correcting for the downward bias due to range restriction was evaluated in accordance to objective (2) and (3) by means of simulated data (*testing* phase) and contrived example using real-world data (*validation* phase) respectively. The effectiveness of these statistical methods was evaluated under a variety of conditions, the results of which are included in chapter 5.

Figure 7.1 gives a summary of the results from the *testing* phase under the Direct Range Restriction (DRR) and the *predictive validity* selection design. Selection was based on selection test *x*. It was observed that in general, the use of the MI correction based on EM or MCMC algorithm performed better than the restricted correlation coefficient. The MI correction was equivalent to the *Thorndike Case II*, in terms of bias even after the addition of *auxiliary variables* in the *imputation model*. Formal statistical testing in Table 8.8 and 8.9 (in the Technical Appendices) confirmed the trend but showed that the MI correction outperformed *Thorndike Case II* in some instances where the selection ratio was low (SR <= 0.4). For SR > 0.4, the

MI correction was equivalent in bias and precision compared to the *Thorndike Case II*. The inclusion of *auxiliary variables* in the *imputation model* led to better performance in terms of precision only for *auxiliary variables* that have correlation coefficient with outcome of >=0.4.

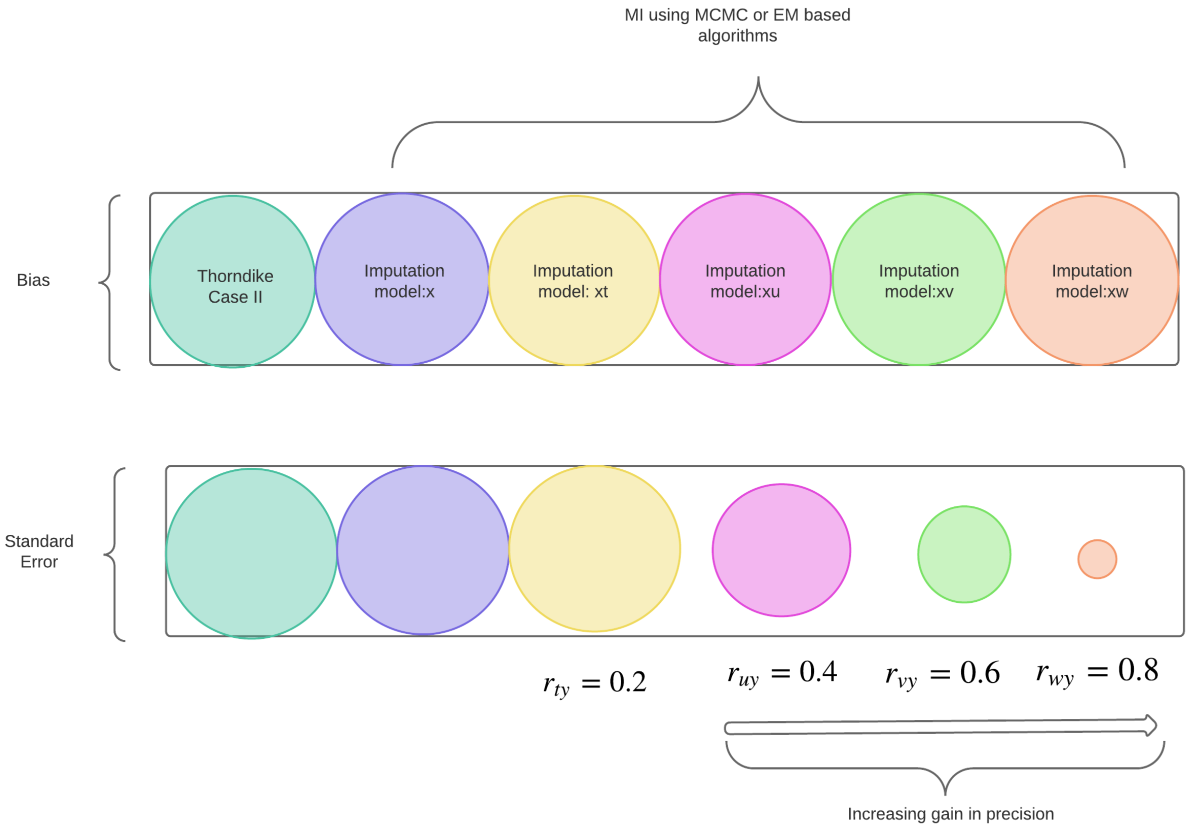


Figure 7.1.: *Conceptual representation of the performance measures of the Multiple Imputation (MI) correction for the effects under the Direct Range Restriction (DRR) over the Thorndike Case II correction for the predictive validity selection design for SR > 0.4. For SR <= 0.4, the trend is the same save for the bias which is significantly lower for MI correction than for Thorndike Case II. This trend becomes more pronounced with inclusion of highly predictive auxiliary variables in the imputation model. The variables x and y are the selection test and outcome (criterion) respectively while t,u,v and w are auxiliary variables which have increasing magnitude of correlation with the outcome variable.*

Figure 7.2 summarises the results for the *testing* phase under Indirect Range Restriction (IRR) for the *predictive validity* selection design. It was observed that excluding the indirectly selected predictor *x* in the *imputation model* results in high levels of bias and loss of precision. Inclusion of both the selection test *z* and indirectly selected predictor *x* in the *imputation model* dramatically improves performance with respect to bias. All the correction methods perform better than the restricted correlation coefficient. The performance of the MI correction was equivalent to *Thorndike Case III*. This trend was also confirmed through formal statistical testing whose results are available in Tables 8.10 and 8.11 (in the Technical Appendices).

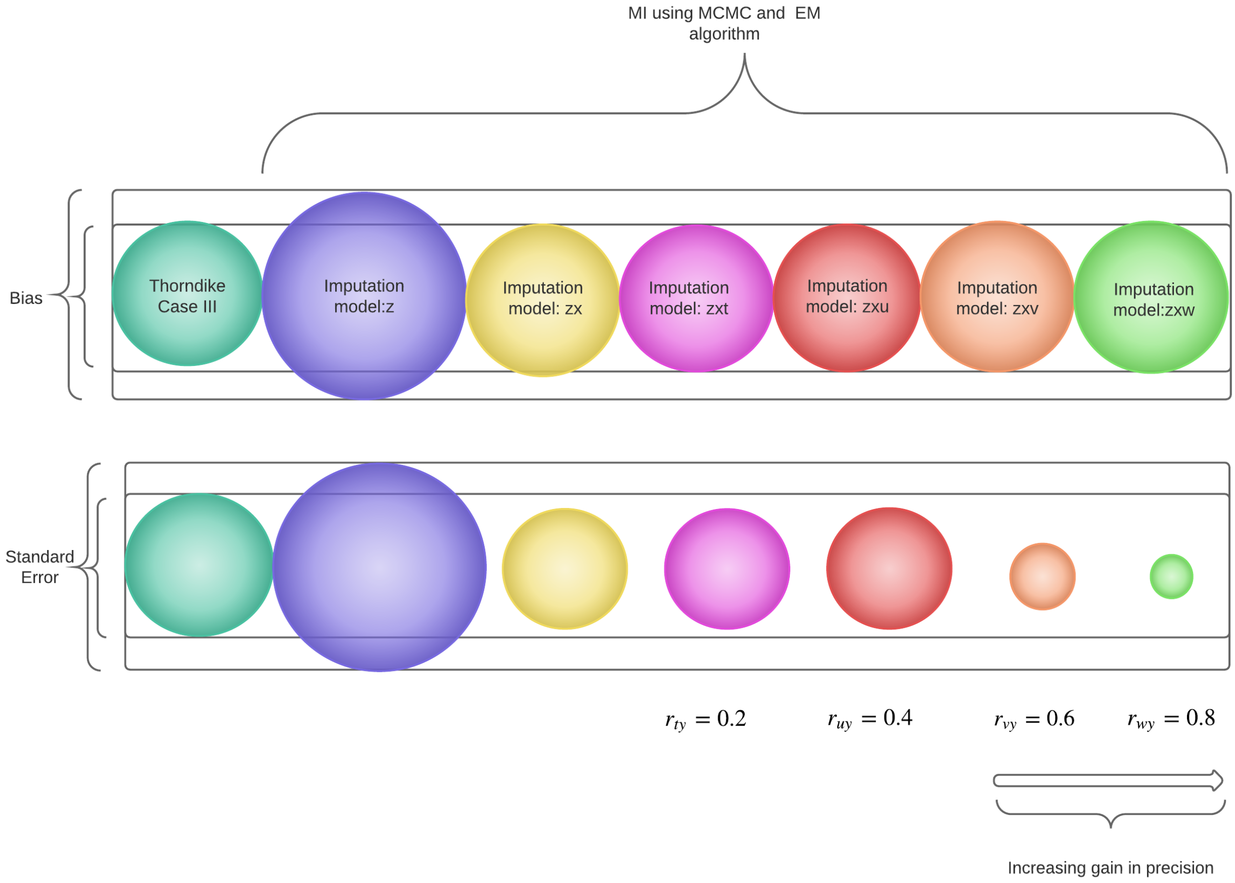


Figure 7.2.: *Conceptual representation of the performance measures of the MI correction under Indirect Range Restriction (IRR) over the Thorndike Case III for the predictive validity selection design. The selection test, indirectly selected predictor and outcome are denoted by z, x and y respectively. The variables t, u, v and w are auxiliary variables which have increasing magnitude of correlation with the outcome variable.*

For instances involving lower selection ratios (SR $<= 0.4$), there were discrepancies between the performance of MCMC and EM algorithm with the former prone to convergence difficulties. With respect to precision, inclusion of both the selection test z and indirectly selected predictor x in the *imputation model* resulted in more precise estimates compared to *Thorndike Case III*. The precision was further improved for the MI correction through the use of highly predictive *auxiliary variables* having correlation coefficient with the outcome variable of $>=$

0.6. The *validation* with the aid of a contrived example using PLAB data confirmed that correcting for range restriction using the selection test z whilst excluding the indirectly selected predictor x in the *imputation model* undermines performance of the MI correction. In fact, the resulting performance is worse than when the effect of range restriction is ignored altogether. The inclusion of both the selection test z and indirectly selected predictor x in the *imputation model* lead to performance that is superior to that of *Thorndike Case III*. However there were no discernible differences between the MI correction and the restricted correlation coefficient.

Figure 7.3 summarises the results of the *testing* phase under the *two hurdle validity* selection design. Selection was assumed to be a two stage process in which two selection tests z and x were used in sequence. Two scenarios were investigated. In the first scenario, full information was available for both selection tests. In the second scenario, full information was only available for the first selection test z (and partial information available only for the second selection test x). In both scenarios, correlation between the selection test and the outcome were considered. It was observed that for the scenario of full information for both selection variables that the three correction methods, FIML, Pearson Lawley and the MI correction were equivalent in performance with respect to both bias and precision. The three correction methods were all better than the restricted correlation coefficients considered. For the scenario in which full information was only available for the first selection test z, bias and precision results are summarised in Figure 7.4. For the correlation between the first selection test z and the outcome y, the performance of the MI correction based on MCMC algorithm was equivalent to that of FIML but better than that of the restricted correlation and *Pearson Lawley* corrections with respect to both bias and precision. Figure 7.3 also shows results for the *testing* phase under the *single hurdle concurrent validity* selection design. The correlations considered were (i) between the selection test z and the predictor x, (ii) between the selection test z and outcome y. The MI correction, FIML and *Pearson Lawley* were generally at par in terms of bias and precision for the two correlation coefficients considered.

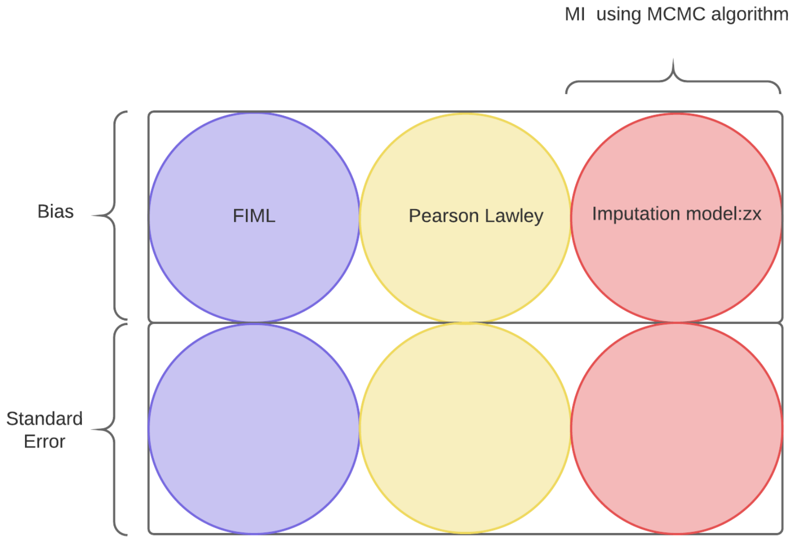


Figure 7.3.: *Conceptual representation of the performance measures of the MI correction for the effects of range restriction over the FIML and Pearson Lawley correction under the two hurdle validity selection design (with full information on the two selection tests z and y) and single hurdle concurrent validity selection design (with MI based only on MCMC algorithm).*

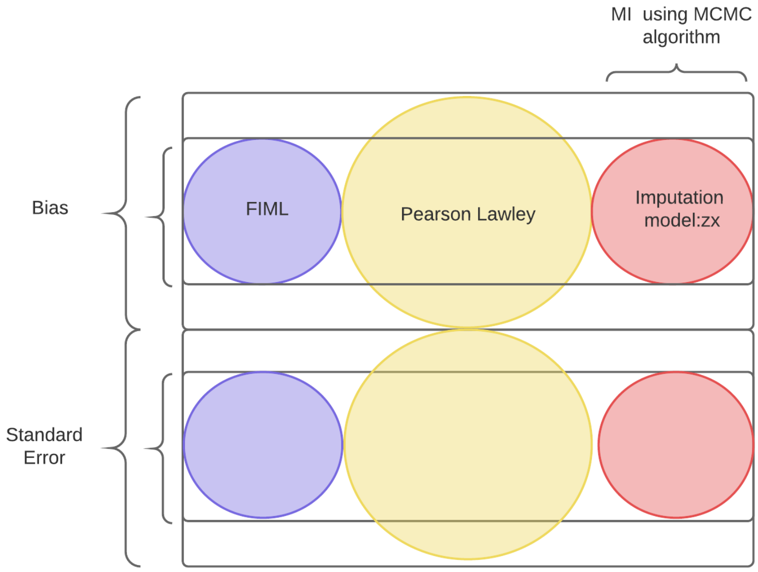


Figure 7.4.: *Conceptual representation of the performance measures of the MI correction for the effects of direct range restriction for the FIML and Pearson Lawley corrections under the two hurdle validity design with full information on only selection test z with MI based only on MCMC algorithm.*

The performance of the methods for correcting range restriction in the *testing* phase under the *predictive validity, single hurdle concurrent* and *two hurdle validity* selection design was promising. However, these methods underperformed in the *validation* phase when evaluated (under the same selection design as in the *testing* phase) in the context of a contrived example using Professional and Linguistic Assessments Board (PLAB) data. The results found in section 5.3 show that the methods were unreliable and inconsistent. Further examination of the PLAB data revealed that the data were not multivariate normal. The impact of violation of multivariate normality on the performance of the methods was investigated in section 5.4 using a simulation study under the *predictive validity* selection design assuming Multivariate Skew Normal (MSN) distributed data with varying amounts of skewness both for the DRR and IRR. The results of the simulation study are also presented in section 5.4 reveal that when the selection data are not multivariate normal, the performance of the methods for correcting range restriction is unrealisable and inconsistent. The performance of the methods worsen in terms of increase in bias and loss of precision with increase in the degree of departure (skewness) from multivariate normality.

7.1.2. Estimating uncertainty about the Number Needed to Reject (NNR)

Work related to objective 4 outlined in section 1.4 dealing with estimation of the uncertainty about Number Needed to Reject (NNR) was presented in chapter 6. The NNR, concept introduced by *Tiffin, Mwandigha, et al. (2016)*, estimates the *number of good candidates that would be rejected in order to get rid of one poor candidate during the selection process*. The term “good candidates” in this context refers to *those applicants who are at a very low risk of a specified adverse outcome (i.e. failing at least one year at undergraduate medical school)*. The concept of NNR is analogous to the *Number Needed to Treat (NNT)* (or *Number Needed to Harm (NNH)* when NNT is negative) in biomedical health research. The NNT conveys the usefulness of a drug or therapy in quantifiable terms. Specifically, it estimates *the number of patients that would have to be treated with a particular therapy to prevent one adverse event*. Low values for NNR and NNT are preferred (Chatellier et al., 1996; Cook and Sackett, 1995;

Public Health Action Support Team (PHAST), 2017). The uncertainty about NNT is normally expressed in terms of confidence intervals. As was shown in section 6.1.1, these confidence intervals may exclude the estimate of the NNT when they are estimated from the inverse of the ARR and the treatment effect is statistically non-significant (Altman, 1998; Muthu, 2003; Sedgwick, 2013). This same challenge is faced when estimating confidence intervals for the NNR. This challenge was addressed in chapter 6 by use of *resampling methods* that took into account the completeness and nature of the data structure. This approach guarantees that the percentile confidence intervals obtained always contain the estimate of the NNR computed.

The selection data made available for the computation of uncertainty about NNR was incomplete and hierarchical (clustered). The incompleteness was due to missing data and the hierarchy was in two levels, student (entrant)-level and university-level data. The computation of the uncertainty about NNR was conducted under three settings. (i) The hierarchical (clustered) nature of the data was ignored by employing MI as though the data were of single-level. (ii) Putting into account the clustered nature of the data by conducting MI within each participating university. (iii) Putting into account the clustered nature of the data by conducting single imputation within each participating university. This was then followed by *case resampling bootstrap* implemented at each of the two levels of the data. All the three setting yielded several samples which were used to compute quantities associated with screening tests such as *sensitivity, specificity*, PPV and NNR. All the three approaches in the settings described demonstrated the same trend for *sensitivity, specificity*, PPV and NNR for the different thresholds of total UKCAT scores considered. As expected, compared to the *case resampling bootstrap*, the MI that ignored the hierarchical structure of the data yielded percentile confidence intervals for NNR that were artificially narrower. An attempt to correct this drawback of the MI approach by conducting MI within each participating university led to the percentile confidence interval for NNR being narrowed even further. This may be explained by the fact that over a third of the data had to be excluded from use as it was not possible to conduct MI in over 40% of the participating universities.

7.1.3. Proof of concept for "Peer Competition Rescaling"

Work related to objective 4 outlined in section 1.4 dealing with the controversial idea of using local measures of undergraduate medical school performance to make national decisions was presented in section 6.2. In the UK selection for undergraduate medical school is at least partly based on a national predictor like the UKCAT (UKCAT, 2017). Therefore, at the point of entry into medical school, potential undergraduate medical school applicants and indeed entrants are compared on a common scale. However, the progression outcomes in undergraduate medical school (such as *knowledge* and *skills*-based exams) are localised. This is because each undergraduate medical school has the autonomy to set its own exams (criterion measures) thus introducing differentiated scales. Therefore, a comparison of medical school entrants from different undergraduate medical schools is akin to a comparison of oranges to apples. For this reason, the present ranking of graduates for national opportunities based on their performance in the local medical school they attended, the Educational Performance Measure (EPM) (Medical School Council, 2017a) is problematic at best. The EPM is used to compare graduates from different medical schools and is partly used to select candidates for the national *Foundation Programme* (Foundation Programme, 2017).

The possibility of using national, observable, predictors such as the UKCAT in order to assess the amount of "competition" a candidate may face when being judged on a local *knowledge* and *skills*-based exam was explored. This introduced information on "peer competition" was used to weight (rescale) local measures to get a better estimate of how such an entrant (graduate) may have performed on a national level measure if one existed. This obtained estimate of performance would be fairer than the current EPM as medical school entrants (and graduates) would be compared on the same scale across all medical schools in the UK. For this reason, this concept introduced by *Tiffin and Paton (2017)* is referred to as *"Peer Competition Rescaling (PCR)"*. The results from the *"Peer Competition Rescaling (PCR)"* show associations between the national predictor, UKCAT and rescaled outcomes, *knowledge* and *skills*-based exams, that were bigger in magnitude than those prior to rescaling.

7.2. Comparison with findings from other studies

With respect to construct-level predictive validity, this thesis builds on previous research and seeks to add to previously published findings in three important ways. Firstly, previous studies have investigated the correction of the effects of range restriction for correlation coefficients primarily under the *predictive validity selection design*. This present study extends the investigation to selection under the *single hurdle concurrent* and *two hurdle validity designs*. Secondly, for theoretical and practical considerations, the impact of including *auxiliary variables* in the *imputation model* on the overall performance of the MI correction was evaluated. In this case, the *auxiliary variables* were defined as variables which have varying magnitude of correlation with the outcome (criterion) of interest. Thirdly, this thesis evaluates the performance of the methods for correcting range restriction in two phases, the *testing* and *validation* phase which made use of Monte Carlo simulated data and contrived example using real-world data respectively. The *testing* and *validation* phase encompassed the three selection designs described, selection ratios $\{0.2, 0.4, 0.6, 0.8\}$ and the different statistical methods for correcting bias due to range restriction in the correlation coefficient. The findings from the Monte Carlo simulation (*testing* phase) under the *predictive validity* design are in line with what other simulation studies have found. *Wiberg and Sundström (2009)* conducted a selection validation study in Sweden. In their study, data were available for both *theory* and *practical* components of a driving test. Testees were allowed to sit for the *theory* test followed by a *practical* test regardless of whether they failed or passed the *theory* test. Therefore, the obtained sample containing the *theory* and *practical* tests was unrestricted. Subsequently, based on the performance of the *theory* test (pass or fail), the data were selected into a "restricted" sample. It was found that a single imputation based on EM algorithm was more accurate at estimating construct-level predictive validity from the "restricted sample" than *Thorndike Case II* correction. A major limitation of the study was that a single imputation (rather than MI) was used to impute missing values of the *practical* test in the "restricted sample". This means that the uncertainty about the correction for range restriction was underestimated. *Pfaffel, Schober, and Spiel (2016)* conducted a simulation study to determine the performance of methods for estimating construct-level predictive validity. Under Direct Range Restriction (DRR), the *Thorndike Case II*, FIML and the MI correction based on MCMC algorithm were found to have equivalent performance in terms of bias and precision. Under Indirect Range Restriction (IRR), FIML and the MI correction based on MCMC algorithm were found to be more precise than *Thorndike Case III* correction. This precision waned with the lowering of the selection ratio. *Mendoza, Bard, et al. (2004)* conducted a simulation study in which methods for estimating construct-level predictive validity were evaluated for a *two hurdle selection* design. In the study, FIML and Bayesian MI were found to be equal in performance in terms of bias. Both of the methods were superior to *listwise deletion* although the Bayesian MI was more advantageous because standard errors for the estimated construct-level predictive validity were easily obtainable.

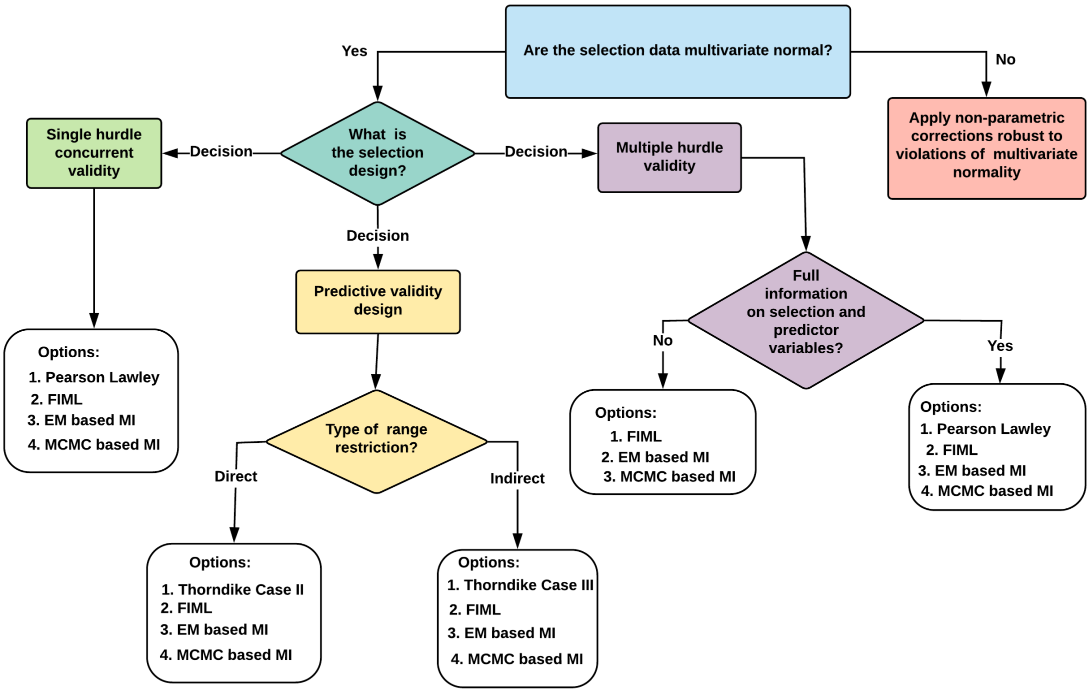
The results obtained during the study of the Number Needed to Reject (NNR) line up with a study by *Tiffin, Mwandigha, et al. (2016).* In that study, the trends for the *sensitivity, specificity*, PPV and NNR were similar to those obtained in this thesis. However, in the study, the uncertainty about the NNR was not estimated. In addition, the clustered nature of the data was not accounted for in the study although the missing values in the *pass* outcome was imputed. Therefore, this thesis extends that study in two important aspects. (i) The uncertainty about the NNR was estimated. (ii) Data structures that may be encountered in practice, incompleteness and clusters, were accounted for.

In this thesis, the *"Peer Competition Rescaling (PCR)"* resulted in associations between the national predictor, UKCAT and rescaled undergraduate medical school outcomes, *knowledge* and *skills*-based exams, that were bigger in magnitude than those prior to rescaling. These results are similar to those found by *Tiffin and Paton (2017)* who demonstrated that the use of local medical school outcomes in predictive validity studies resulted in attenuated associations. In addition, *"Peer Competition Rescaling (PCR)"* was shown to be capable of disattenuating these associations thereby "nationalising" them.

7.3. Inferences and interpretation of findings

From the results for construct-level predictive validity obtained through Monte Carlo simulation in the *testing* phase. It may be concluded that the MI correction is a viable alternative to the *Thorndike* formulas under the *predictive validity* design. Multiple Imputation (MI) performs best in instances of low selection ratios where the magnitude of correlation between the selection test/predictor and outcome is high (see Tables 5.3 and 5.5). The bias and precision for the MI correction was equal to that of *Thorndike Case II*. The MI correction was equal to *Thorndike Case III* with respect to bias but the MI correction had greater precision. The bias and loss of precision in the estimate of predictive validity due to range restriction is mitigated to a great extent by the MI correction but not eliminated entirely. From a theoretical standpoint, the use of highly predictive *auxiliary variables* in the *imputation model* for the MI correction significantly improved the precision in all selection ratios but reduced bias only for strict selection scenarios. From a practical standpoint the availability of highly predictive *auxiliary variables* may be difficult to find in practice.

For the *single hurdle concurrent validity* and *two hurdle validity design*, the MI correction performed as well as FIML and *Pearson Lawley*. However, this was only when (i) full information was available on the two selection tests for the *two hurdle selection design* (ii) correlation considered was between second selection test and the outcome of interest. For the *single hurdle selection* design, the MI correction performed as well as FIML and *Pearson Lawley* only for the correlation between the predictor (not selection test) and outcome considered. For the correction using MI, the performance of EM algorithm may differ from that of MCMC algorithm for very strict selection (low selection ratios). In such instances, MCMC algorithm may require exceedingly many more iterations to converge to the same solution as the EM algorithm. For this reason, it is recommended that all convergence diagnostics for the MCMC imputation be evaluated before the correction is adopted. However, as was observed from the *validation* phase (see section 5.3) and the evaluation of the discrepancy between the performance of the methods under the *testing* and *validation* phases (section 5.4), even slight departures from multivariate normality were observed to have an adverse effect on the performance of the MI correction especially under IRR. This implies that the methods considered (including MI) may hold very little utility in correcting for range restriction bias in practice since selection data are likely not to be multivariate normal (a key assumption underlying their use). To aid in decision making as to when the methods for correcting range restriction bias may be useful, a decision flow chart is presented in Figure 7.5.



Note that for Direct range restriction

1. MI produces the most precise estimates of all (options) methods when auxilliary variable(s) highly predictive of the outcome (correlation of 0.2 and above) are available.
2. When selection is strict (selection ratio of 0.4 and below), MI produces estimates that have the least bias and highest precision.

Note that for Indirect range restriction

1. MI produces the most precise estimates of all (options) methods. This precision improves when auxilliary variable(s) highly predictive of outcome (correlation of 0.6 and above) are available.

Figure 7.5.: *Decision tree for achieving construct-level predictive validity in a variety of selection scenarios*

With regards to the results of the estimate for uncertainty about Number Needed to Reject (NNR) from the three scenarios investigated. It was revealed that in instances where the data is not hierarchical, MI (when data is incomplete) and case resampling bootstrap (when data is complete) are reliable methods of estimating the uncertainty about NNR. A combination of imputation and *case resampling bootstrap* may also be used for instances of incomplete data by first conducting a single imputation to form a complete data followed by *case resampling bootstrap*. For hierarchical data, the *case resampling bootstrap* applied at all levels of the data may be used when data is complete, when the data is incomplete single imputation may be employed first followed by *case resampling bootstrap* implemented at all levels of the data. Therefore, the method to be used in estimating the uncertainty for NNR should factor in the complexity of the data structure in terms of hierarchy and missing values in the selection data as summarised in Figure 7.6.

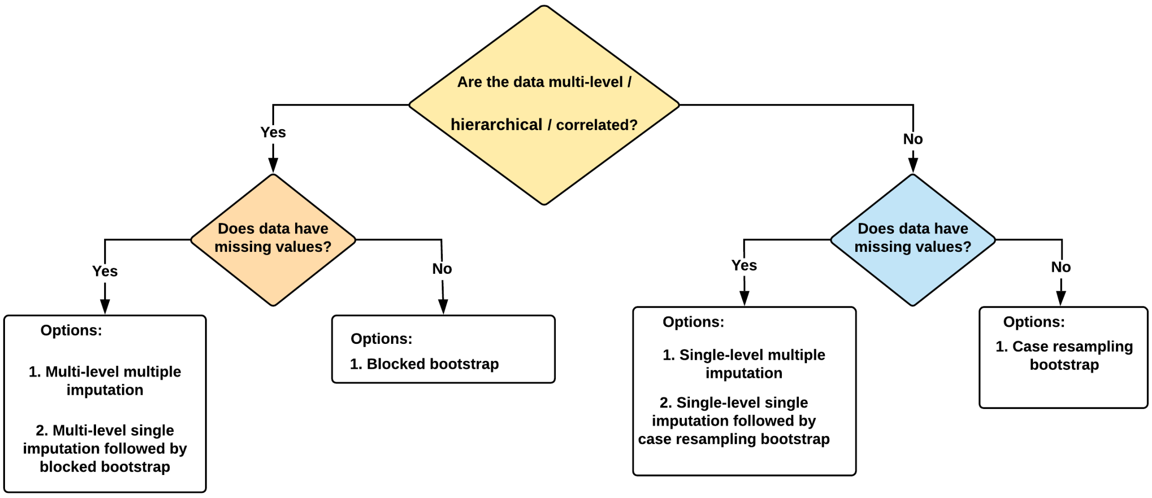

Are the data multi-level / hierarchical / correlated?
Yes
No
Does data have missing values?
Yes
No
Options:
1. Multi-level multiple imputation
2. Multi-level single imputation followed by blocked bootstrap
Options:
1. Blocked bootstrap
Does data have missing values?
Yes
No
Options:
1. Single-level multiple imputation
2. Single-level single imputation followed by case resampling bootstrap
Options:
1. Case resampling bootstrap

Figure 7.6.: *Decision tree for estimating uncertainty for NNR for different data structures*

With regards to the operational usefulness of the UKCAT as a screening test for potentially poor candidates during selection, a few issues need to be considered. Based on the results obtained, regardless of approach applied, the *sensitivity* of the UKCAT increased in tandem with the threshold UKCAT used. In other words, as selection criterion became strict, the ability of the UKCAT to detect a genuinely poor candidate who had failed at least one exam at first sitting during their undergraduate medical school training increased. In fact overall, the UKCAT was demonstrated to be a useful screening test since all the NNR values obtained were less than 5, the classification rule adopted in this thesis for determining a good selection test(for further details, see section 6.1.1) (Public Health Action Support Team (PHAST), 2017). Although the UKCAT's *sensitivity* increased in tandem with the threshold of the UKCAT used, the PPV declined as *sensitivity* increased. *Sensitivity* by definition, in this selection context, answers the question-given that a candidate is poor based on available results from his undergraduate medical school performance, what is the probability that the UKCAT will detect him as a poor candidate? However, it should be noted that *sensitivity* would not be possible to evaluate during the selection process as the future undergraduate medical school performance for the applicants would need to be known apriori (that is good versus poor candidate status-the very thing that is to be screened for). What is important however is what may be inferred regarding the ability of the UKCAT to detect poor candidates at point of selection who would end up failing at least one exam at first sitting in undergraduate medical school. That is, given that a candidate has been detected to be poor, what is the probability that they will end up failing at least one exam at first sitting at undergraduate medical school? This is called the PPV, a quantity that is seen to diminish when selection criteria is made more strict by raising the UKCAT threshold. The value of the PPV depends on *sensitivity, specificity* and *prevalence*, the proportion of candidates classified as poor, (Altman and Bland, 1994) as shown in equation (7.3.1). Notice that the observed declining *specificity* and low prevalence value (a result of effective selection) works to diminish the PPV. This explains why the UKCAT has poor PPV values but is still a good screeening test based on the NNR values obtained.

$$PPV = \frac{sensitivity * prevalence}{sensitivity * prevalence + (1 - specificity) * (1 - prevalence)} \tag{7.3.1}$$

7.4. Strengths and limitations

In this thesis, the evaluation of the methods for achieving construct-level predictive validity was subject to the following limitations. (i) The different simulations considered only four selection ratios of $\{0.2, 0.4, 0.6, 0.8\}$. This was done to limit the number of simulations that were to be conducted with respect to five predictive *auxiliary variables*, three selection designs and four statistical methods under investigation for both the *testing* and *validation* phase of the simulations. *Pfaffel, Schober, and Spiel (2016)*, in their simulation study to determine the performance of methods for estimating construct-level predictive validity under Direct Range Restriction (DRR) and Indirect Range Restriction (IRR), used selection ratios of $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6,$

$0.7, 0.8, 0.9\}$. The conclusions from their study were in line with those made in this thesis. Therefore, it is reasonable to conclude that the consideration of four selection ratios in this thesis was not detrimental. (ii) The selection test, predictor and outcome (criterion) featured in the study were assumed to be fully observed and measurable. This may not reflect how selection is done in practice. For example, the selection test may be a latent variable as is the case when *Thorndike Case IV* correction for range restriction is used. A simulation study conducted by *Pfaffel, Kollmayer, et al. (2016)* to determine the performance of methods for estimating construct-level predictive validity under Direct Range Restriction (DRR) and Indirect Range Restriction (IRR) for a dichotomous outcome (criterion) had positive results. Their findings suggested that the MI correction performed better in terms of bias and precision than the *Thorndike* corrections. This is similar to the findings in this thesis which used a continuous criterion. This implies that if the criterion in this thesis had been dichotomous, the missing data handling methods would have performed just as well. (iii) Simulations assumed also that selection test, predictor and criterion of interest were continuous. This necessitated the computation of the Pearson correlation coefficient. Based on limitation (ii), *Pfaffel, Kollmayer, et al. (2016)* demonstrated that the construct-level predictive validity based on a *polyserial correlation* (see Table 8.1 in the Technical Appendices for correlation types) may be achieved by the MI correction. This implies that the MI correction for range restriction is expected to perform just as well for a non-continuous criterion. Presently, there is a dearth of literature on construct-level predictive validity in instances where both the selection test and outcome (criterion) are categorical in nature. (iv) The imputations were conducted assuming a fixed sample size of 500 observations. This was chosen so as to enable a large number of observations to be available for use even after selection ratio of 0.2 had been applied ($\frac{20}{100} * 500 = 100$). *Pfaffel and Spiel (2016)* found that for the *predictive validity* selection design, low selection ratios of ≤ 0.2 and small to moderate sample sizes (50 to 100 observations before selection) led to biased estimates of construct-level predictive validity when the outcome (criterion) was dichotomous. When the outcome (criterion) was continuous, the estimate of construct-level predictive validity was found to more accurate for moderate sample sizes (observations of about 100 before selection). This accuracy increased with increase in selection ratio and magnitude of unrestricted correlation coefficient. Therefore, since in this thesis, the criterion was continuous, use of sample sizes smaller than one used is not expected to lead to a change in the conclusions. (v) Only four correlation values between the predictive auxiliary variables and outcome (criterion) of 0,2, 0.4, 0.6 and 0.8 were considered. Based on the limitations (i) to (iv), there is need for future research to assess the impact of different values of predictive auxiliary variables, latent selection tests, polytomous predictors and outcomes (criterion) on the missing data handling method for estimating construct-level predictive validity. These need to be examined for differing sample sizes for the different selection designs.

A number of limitations encountered when estimating uncertainty about NNR bear mentioning. (i) The PPV associated with the UKCAT as a screening test depends on the *prevalence value* (proportion of applicants who actually fail in at least one exam at first sitting in undergraduate medical school) which was unknown due to range restriction (only outcome data for the entrants was available). The "restricted sample prevalence" was $\frac{674}{2{,}723}$ (no fail or pass information was available for at least one of the five years of undergraduate medical school training for 4,089 candidates). This may not be reflective of the actual "unrestricted sample prevalence". (ii) The data consisted of high attrition rates especially for later years of undergraduate medical school training. This severely compromised the performance of the MI approach conducted within the universities as over 40% of the participating universities were excluded from analysis. Therefore no generalisations can be made for the affected MI approach. (iii) The criterion (outcome) was observed in only those who were selected which relates to the limitation introduced by range restriction. Note that this contrasts with *Number Needed to Treat (NNT)* in biomedical health research where the outcomes are observed in many, if not most, cases. Further, since the criterion (outcome) for the NNR was non-continuous (pass or fail), the issue of estimating construct-level predictive validity for this type of criterion (outcome) was beyond the scope of this thesis. Based on work by *Pfaffel, Kollmayer, et al. (2016)* (see limitation (ii) under construct-level predictive validity), it is reasonable to state that it is possible for this limitation to be addressed by MI and FIML.

The results from the *"Peer Competition Rescaling (PCR)"* seem promising. However, they need to be interpreted with a caveat since the validation of the approach in the UK would require the use of national measure of performance that is actually being approximated by the *"Peer Competition Rescaling (PCR)"*. Unfortunately, such an undergraduate medical school national measure in the UK does not exist. A close alternative that may be used for validation of the *"Peer Competition Rescaling (PCR)"* is the *Royal College of Physicians* membership exams currently sat for by a subset of postgraduate medical students. Outside the UK, data from the USA may be used. This may be obtained from all medical graduates expected to sit the United States Medical Licensing Examination (USMLE) (Federation of State Medical Boards, 2017). This, with the existence of local performance measures from different medical schools and a national predictor such as the Medical College Admissions Test (MCAT) may be used for validation of the *"Peer Competition Rescaling (PCR)"*. Presently, data on the United States Medical Licensing Examination (USMLE) are not accessible. However, data on the *Royal College of Physicians* has recently been collated (Steve Thornton, 2017; Thornton, 2017). Access to the data has been applied for via the United Kingdom Medical Education Database (UKMED)(Medical School Council, 2017b; UK Medical Education Database, 2017). Once access to the data is approved, the completion of the proof of concept for the *"Peer Competition Rescaling (PCR)"* will follow as a post-doctoral collaboration with *Tiffin and Paton (2017).*

7.5. Implication for selection practice and policy in the UK

To conclude this thesis without summarising the lessons learnt would be remiss. In chapter 1 the methodological focus of the thesis was introduced. Therein, the background, context and aims for the thesis were specified using motivating examples from the UK medical school selection. It is known that, in the UK, competition for undergraduate medical school places is extremely fierce with approximately 11 applicants per place available to study medicine. In addition, it has been established that 80% of those studying medicine in the UK applied from only 20% of the country's secondary schools, a vast majority of which are selective schools (independent or grammar schools) (Medical School Council, 2014). These selective secondary schools tend to be located in more highly resourced and affluent areas. It has thus been argued that the higher performance and selection rates of students from these secondary schools reflect resource deprivation rather than ability differences (McManus, Dewberry, Nicholson, and Dowell, 2013).

Therefore, there have been several attempts to address this as two-pronged objective. (i) By selecting the most able applicants for medicine using a host of predictors covered in chapter 2 which include the UKCAT. (ii) Diversifying the intake by selecting many more applicants from non-selective schools background and other under-represented groups without adversely impacting objective (i) (King's College London, 2017; University of Birmingham, 2017a,b). Recently, *De Corte, Sackett, and Lievens (2011)* introduced a novel approach in selection borrowed from mathematics and economics that approached the problem as a *Multiple Objective Optimisation Problem (MOOP)*. This approach is a culmination (of refinements following their previous works in the area) that enables selection decisions to be made subject to trade-offs in levels of each objective that would best suit a medical school's priorities and defined constraints (time, money or staff etc). This approach is thus more pragmatic in achieving selection optimality in a given local selection context (De Corte, Lievens, and Sackett, 2007, 2008; De Corte, Sackett, and Lievens, 2011; De Corte, Sackett, and Lievens, 2010). Their proposed approach would need a battery of information regarding the *effect size* (defined as standardised difference in performance between the majority and under-represented (minority) groups for each predictor under consideration), the *correlation coefficients* between the predictors, *medical school constraints* and *predictive validity* estimates of each of the predictor. Notice that if inaccurate predictive validity estimates were used in the MOOP suggested by *De Corte, Sackett, and Lievens (2011)*, the selection decisions made would be less than optimal. Therefore the estimation of construct-level predictive validity (one of the methodological aim of the thesis, see section 1.3) would be crucial in meeting the selection objectives (i) and (ii) of the MOOP. Perhaps the most challenging aspect of selection not captured by the MOOP is the "human factor". For instance, in the UK, emerging evidence suggests that all undergraduate medical schools want to select high UKCAT and Situation Judgment Tests (SJTs) scoring candidates. Thus applicants tend to get either three (to four) offers or none at all. This means that those with offers are much more likely to turn them down thus having medical schools scrambling to fill places through “clearing”. In fact, medicine places were filled by clearing for the first time in 2016 (David Millett, 2016; Matthew Pinchard, 2015; Sofia Lind, 2016; St George’s University of London, 2016). This was replicated in 2017 and seems to be a trend that is set to continue in future if the current selection environment prevails (Lucinda Borrell, 2017; Richard Adams, 2017).

The operational usefulness of aptitude (selection) tests is invariably determined from their predictive validity estimates-that is association between the aptitude (selection) tests and future outcome (criterion) measures-with high predictive validity estimates deemed to suggest a much more useful aptitude (selection) test. In chapter 3, it was demonstrated from literature and Monte Carlo simulations that selection induces range restriction which has an effect of attenuating predictive validity estimates derived from correlation rather than regression coefficients or odds ratios. The review of literature conducted in chapter 2 whose results are summarised in Table 2.3 shows a vast majority (about 80%) of the predictive validity estimates from the studies were correlation coefficients. Only about 15% of these studies were corrected for the attenuating effects of range restriction. This means that most of the correlation coefficients (predictive validity estimates) for aptitude tests (including the UKCAT) reported in literature are, in effect, underestimated. It is thus recommended that future UKCAT predictive validity estimated from correlation coefficients be corrected for range restriction bias. This would be best practice as it would allow medical school applicants, medical schools and policy makes to make informed decisons regarding the operational usefulness of the UKCAT. In the thesis it was found that generally, Multiple Imputation (MI) and Full Information Maximum Likelihood (FIML) were found to be equivalent in performance but superior to other methods for selection ratios $\leq 20\%$. Thus these methods are recommended for use in estimating construct-level predictive validity of the UKCAT in practice only when the data are determined to be multivariate normal for the case of continuous variables. As was seen in sections 5.3 and 5.4, even slight departures in the data from from multivariate normality (which is widespread in practice) renders the application of methods for correcting range restriction ineffective.

In addition, even instances where distribution assumptions are satisfied, the following issues would have to be acknowledged and possibly addressed within the medical selection context. (i) The use of the UKCAT in selection across undergraduate medical schools is not uniform. For example, as shown by *Adam, Dowell, and Greatrix (2011)*, UKCAT may be used in selection as a “borderline method” (to discriminate amongst a small number of applicants lying at a decision borderline, who are otherwise indistinguishable on the medical school’s other selection criteria), “factor method” (add an applicant’s UKCAT score or a proxy for that score to the score the applicant obtains in the medical school’s usual method of selection, to provide a total score), “threshold method” (minimum or threshold UKCAT score adopted to create a hurdle that an applicant must cross to reach the next stage in the selection process) and “rescue” (to compensate for an applicants who would otherwise be rejected on account of their score on other selection criteria). Information on how a specific medical school used the UKCAT in selection is collected when medical school UKCAT selection data is aggregated (Tiffin, Dowell, and McLachlan, 2012). However, the precise impact of the varied use of the UKCAT on its predictive validity and construct-level predictive validity estimates is unknown. This is because, unlike for the case of the Medical College Admissions Test (MCAT) in the USA, the varied use of the UKCAT in the UK has not yet been modelled or controlled for in UKCAT selection validity studies (Albanese, Farrell, and Dottl, 2005a; Albanese, Farrell, and Dottl, 2005b; Zhao et al., 2010). (ii) Provisional offers to medical school applicants in the UK are made on the basis of their General Certificate of Secondary Education (GCSE) results, predicted General Certificate of Education Advanced Level (A-level) grades and actual UKCAT scores. Applicants must then achieve the required A-level grades, specified by the medical school at the time of provisional offer, in order to matriculate. There have been reports of some secondary schools “gaming the system” by overestimating the predicted A-level grades to increase the odds of provisional offers made to their students (Kim Catcheside, 2012; Nuala Burgess, 2017). In addition, it has been argued that this system favours more ambitious, better-informed applicants, from wealthier families and high-achieving secondary schools. For example, in 2017, it was reported that 73% of predicted A-level grades for 18-year-old applicants turned out to be higher than their actual results with only about 16% of the applicants achieving the A-level grades that they were predicted to achieve (Katherine Sellgren, 2016; Telegraph Reporters, 2016; Wyness, 2016). Reportedly, this means that by the time actual A-level grade results are received many decisions will already have been made by universities and students about their offers and choices. Better-off applicants, even if they miss out on grades, still tend to get on to more sought-after courses (Sean Coughlan, 2017). Thus, it has been suggested that basing offers on actual rather than predicted A-level grades would level the playing field. This system has however been defended by universities which state that over 70% of applicants are placed at their first choice and those who want to apply post A-level results can easily do so. To end the controversy, it is worthwhile for further research to be conducted to determine the precise impact this system of selection has on fairness. Nevertheless, given the predicted A-level grades, GCSE and UKCAT results used in selection, the predictive validity and construct-level predictive validity estimates for the UKCAT would need the effect of the other predictors partialled out to accurately determine its operational usefulness. In other words, the use of partial correlations would be needed. The estimation of construct-level predictive validity by partial correlation coefficients was outside the scope of this thesis. However, as was shown in chapter 3, the use of regression modelling would tackle this problem since the effects of the other predictors would be controlled for and range restriction would not attenuate the predictive validity estimated. (iii) The predictive validity and construct-level predictive validity of the UKCAT's Situation Judgment Tests (SJTs) component would be problematic to estimate if the criterion of interest was a binary outcome (pass or fail) at undergraduate medical school. This is because both the criterion (pass or fail) and SJTs band scores 1=highest to 4=lowest are categorical variables (see section 1.1 for details on the UKCAT's SJTs). Therefore, the estimated construct-level predictive validity would require computation of Spearman rho / Polychoric correlation coefficients (see Table 8.1 for correlation types). However, presently, there is a dearth of literature relating to the use of FIML, MI and other statistical methods in estimating construct-level predictive validity in instances where both the selection test and criterion (outcome) are dichotomous, polytomous or in instances where the data deviate from multivariate normality.

As was discussed under 7.1.2, the operational usefulness of the UKCAT as a screening test for potentially poor candidates during selection is impressive. Although it was seen that the UKCAT had a high *sensitivity* for identifying entrants who were known to have failed (apriori) for higher threshold of the UKCAT, at the point of selection this would not be meaningful. This is because the computation of *sensitivity* would require the candidate status (good versus bad candidate) to be known before hand, the very thing that is to be screened for by UKCAT. In addition, the PPV depends on the *prevalence value* (proportion of applicants who actually fail in at least one exam at first sitting in undergraduate medical school). The "restricted sample prevalence" was $\frac{674}{2{,}723} = 0.2475$ (no fail or pass information was available for at least one of the five years of undergraduate medical school training for 4,089 candidates). Whether this "restricted sample prevalence" is reflective of the prevalence value that is representative of the population of interest (all undergraduate medical school applicants-not just entrants) is a subject of construct-level predictive validity beyond the scope of the thesis (more on that later). What can be said for certain is that the unknown "unrestricted sample prevalence value" even with the high *sensitivity* was not adequate to translate into higher values of the PPV for higher threshold of the total UKCAT scores considered. Fortunately, this has no adverse effect on the capability of the UKCAT to be a good screening test.

It is worth mentioning that the results obtained in this thesis are a function of the criterion (outcome) as was defined by each medical school. In addition, the criterion (outcome) were observed in only those who were selected which relates to the limitation introduced by range restriction. Since, the criterion (outcome) was non-continuous (pass or fail), the issue of estimating construct-level predictive validity for this type of criterion (outcome) was beyond the scope of this thesis. Therefore, this is acknowledged as a limitation that would need to be addressed in other follow up studies so as to offer firm guidance on this issue. With respect to the definition of criterion (outcome) measures, as was discussed in section 7.1.3. These were locally defined by each medical school. This is an important issue that is yet to be addressed (as far as is known) in all UKCAT predictive validity studies that make use of outcomes collated from multiple undergraduate medical schools. It is worth bearing in mind that at point of selection for medical school, applicants are usually evaluated based on national predictors like the GCSE results, A-level grades and actual UKCAT scores. However, after graduation, selection decisions for national opportunities like the *Foundation Programme* (Foundation Programme, 2017) are partly based on a ranking of achievement in *knowledge* and *skills*-based exams at the local medical school attended. This contributes to what is referred to as the Educational Performance Measure (EPM), a local measure of undergraduate performance for a graduate within their medical school up to the point of application to the *Foundation Programme* as shown in Table 7.1. As may be observed, the local medical measures of performance contributes 86% of the maximum possible EPM scores. Table 7.2 shows the assignment of scores for the "medical school performance" component of the EPM. Graduates are assigned scores based on their decile rank of performance within the undergraduate medical school they attended. This introduces a non-common scale of comparison among the medical school graduates from different undergraduate medical schools.

EPM component	Number of points
Medical school performance (calculated in deciles)	34-43
Additional degrees	0-5
Other educational achievements	0-2
Maximum points available	50

Table 7.1.: *The components that contribute towards the Educational Performance Measure (EPM) scores used to selection for the Foundation Programme.* [1]

[1] Table adapted from "Educational Performance Measure (EPM) Framework"(Imperial Centre for Endocrinology, 2013)

Decile rank	Number of points
1	43
2	42
3	41
4	40
5	39
6	38
7	37
8	36
9	35
10	34

Table 7.2.: *The assignment of points based on decile of performance in local medical school exams.* [2]

The Peer Competition Rescaling (PCR), may be a viable alternative which may be used to "nationalise" such local measures to enable fairer comparisons of graduates by placing them on a common scale. The results from section 6.2.4.2 revealed that validity estimates from scaled local outcomes were of larger magnitude compared to the validity estimates from unscaled local outcomes. This implies that the use of local outcomes has the detrimental effect of attenuating validity estimates. Further, it may be argued that the PCR is capable of disattenuating the validity estimates from unscaled local outcomes. However, caution needs to be exercised since the PCR is still a "work in progress".

The validation of the operational usefulness of PCR in levelling the playing field for postgraduate school national opportunities requires data on a national outcome measure which does not presently exist. That said, General Medical Council (GMC), is considering proposals (based on a consultation concluded in April 2017) to introduce a single, common objective measure that those applying for registration with a licence to practise medicine in the UK must meet to ensure safe practice. This national measure would be known as the *Medical Licensing Assessment (MLA)* (General Medical Council, 2017b). Fortunately, in the meantime, alternative data

[2] Table adapted from "Educational Performance Measure (EPM) Framework"(Imperial Centre for Endocrinology, 2013)

that may be used relate to the *Royal College of Physicians* membership exams currently sat for by a subset of postgraduate medical students. These data may be made available in the future to facilitate the validation of PCR.

The implementation of the concepts covered in the thesis, like missing data handling methods for establishing construct-level predictive validity, the resampling methods for the estimation of uncertainty about NNR and PCR require statistical erudition. This may need to be contextualised with the understanding of the "criterion problem" as highlighted by *Cleland et al. (2012)*. For example, it is common place to use *knowledge* and *skills*-based undergraduate medical outcomes to determine the predictive validity of the UKCAT. However, this is a small piece of the jigsaw puzzle of medical practice. There are other non-academic outcomes like *professional* and *behavioural* attributes that are predictive of successful job performance that may be of interest to different stakeholders in medical education. Therefore, the use of the statistical methods evaluated and developed in this thesis will need to be contextualised for the outcome of interest. Given the complexity of the issue, it does seem like the development of an *R package* to tackle this problem would be a worthwhile endeavour. This would make it easier for selection teams (universities or employers) based on available data to run simulations that would present solutions like the ones developed based on MOOP by *De Corte, Sackett, and Lievens (2011)*. This would also help disseminate the methods widely and make them of routine use in medical selection. For these reasons, there may be need for further training for key staff in selection teams at medical school whose contribution would be mastery of the intricacies of the concepts so as to offer practical advice and guidance whenever needed to improve selection.

7.6. Directions for future research

Given the issues highlighted in the preceding sections, a few future research opportunities are reiterated briefly. This thesis was focused on the use of missing data handling methods in achieving construct-level predictive validity given a continuous selection test and criterion.

Owing to the death of literature on construct-level predictive validity involving continuous data that deviate from multivariate normality and those that contain a non-continuous selection test and criterion, there are research opportunities related to construct-level predictive validity in instances where both the selection tests (like the UKCAT's SJTs) and criterion (like pass or fail in undergraduate medical school) are categorical. Further, this may be extended to the estimation of construct-level predictive validity in instances where the continuous variables are not multivariate normal or when both the selection test and criterion are latent variables, these may encompass the Five Factor predictors (*conscientiousness, extraversion, agreeableness, neuroticism* and *openness to experience*(Costa et al., 2014)) described in section 2.1.4. Their construct-level predictive validity for future *professional* and *behavioural* attributes highlighted as being predictive of successful medical practice may be assessed (Cleland et al., 2012). Certainly, these may need to be examined for differing sample sizes for the different selection designs. With regard to NNR, given the complex statistical methods (imputations and resampling methods) used, for the different data structures (incompleteness and clustering), it may be practical to develop software solutions like *R packages* that to some degree automate the process. Although this approach may be prone to misuse, it is the best avenue for disseminating the methods widely so as to make them common place in medical selection. Lastly, pending the creation of the national measure *Medical Licensing Assessment (MLA)* (General Medical Council, 2017b), validation of the "Peer Competition Rescaling (PCR)" may be accomplished using data about the *Royal College of Physicians* membership exams currently sat for by a subset of postgraduate medical students.

8. Technical Appendices

8.1. Correcting for bias in the Pearson correlation coefficient due to measurement error

Assume that there are two observable variables of interest, *x* and *y*, representing the predictor (selection test) and outcome (criterion) respectively and that further, correlation between the two variables is of particular importance. If at least one of the variables is non-continuous the *Pearson product moment correlation* computed will be attenuated even if measurement error is non-existent. In such instances, the consideration of the scale of the variable should lead to the appropriate correlation coefficient computed based in Table 8.1. The non-continuous variables are generally assumed to represent underlying continuous dimensions (Olsson, Drasgow, and Dorans, 1982; Poon and Lee, 1987; Theresa Gillian and Nelson, 2010).

	Scale of *y*		
Scale of *x*	Ratio / Interval	Ordinal	Nominal
Ratio / Interval	Pearson product	(Bi)Polyserial	Point (Bi)Polyserial
Ordinal	(Bi)Polyserial	Spearman rho / Poly(Tetra)choric	Rank (Bi)Polyserial
Nominal	Point (Bi)Polyserial	Rank (Bi)Polyserial	Phi, L, C, Lambda

Table 8.1.: *Types of correlation to be computed depending on the scales of the variables of interest. Note that the term "Bi" or "Tetra" implies that at least one of the variables is dichotomous*[1].

In the thesis, *x* and *y*, were assumed to be bivariate normal, linearly related and homoscedastic (variables have common error variance). These underlying assumptions support the computa-

[1] Table adapted from *"More Correlation Coefficients"*(Smith Hall, 2005)

tion of the Pearson correlation coefficient (Held and Foley, 1994; InfluentialPoints, 2017; Laerd Statistics, 2013a,b). Therefore the methods of tackling measurement error are briefly presented for the Pearson correlation coefficient.

8.1.1. Traditional approach

As already documented in chapter 3 from the review of literature and the Monte Carlo simulation (see Figure 3.1), in the presence of measurement error, r_{xy}, is prone to attenuation (downward bias). Formal treatment of this attenuation (downward bias) due to the effect of measurement error is traced back to *Spearman, Charles (1904)*, who proposed the formula in equation 8.1.1 as a correction for the dis-attenuation of the Pearson correlation coefficient thus resulting in what was termed as the true or corrected correlation $r_{t_x t_y}$ (Lovie and Lovie, 2010; Spiegelman, 2010):

$$r_{t_x t_y} = \frac{r_{xy}}{\sqrt{r_{xx} r_{yy}}} \tag{8.1.1}$$

The terms r_{xx}, r_{yy} are the reliability estimates of x and y respectively, these are classically defined as proportions of variance devoid of measurement error in x and y respectively. They may be theoretically computed by using the formula 8.1.2 and 8.1.3 respectively. Note that $0 \leq r_{xx}, r_{yy} \leq 1$.

$$r_{xx} = \frac{\sigma_{t_x}^2}{\sigma_{t_x}^2 + \sigma_{\varepsilon_x}^2} \tag{8.1.2}$$

$$r_{yy} = \frac{\sigma_{t_y}^2}{\sigma_{t_y}^2 + \sigma_{\varepsilon_y}^2} \tag{8.1.3}$$

There is a major hindrance in the application of the correction in equation 8.1.1, this is because the numerator terms in equation 8.1.2 and 8.1.3 are always unknown. Thus the reliability estimates r_{xx} and r_{yy} are not directly obtainable as expressed by the formulas. Consequently alternatives have been proposed, they include *Spearman-Brown corrected split-half coefficients, Test-Retest reliability, Parallel Forms reliability* and *Inter-Rater reliability*. For the interested reader, detailed explanations of these methods can be found in the following references (Johnson, 1944; Webb, Shavelson, and Haertel, 2006; William M.K. Trochim, 2008; Zimmerman and Williams, 1997).

8.1.2. Structural Equation Modeling (SEM) approach

The crucial issue in the approaches discussed in 8.1.1 for dealing with attenuated correlation is accurate estimation of reliability estimates. The SEM approach when used with continuous data for the predictor (selection test), x and y, is more accurate in dealing with attenuated correlation due to measurement error. This is because less than perfect measurements of x and y may be modelled thus resulting in what are considered to be accurate values of true correlation $r_{t_x t_y}$ (Charles, 2005; Jason W Osborne, 2015). In SEM, variables are either *measured (observed, manifest)* or *latent (unobserved).* The SEM framework consists of two types of models, *path analysis* which contain manifest variables only and a *measurement model* which describes the relationship between a set of observed and unobserved variables. In SEM notation, observed (measured) variables are represented by rectangle or square box, and latent (unmeasured) variables by circle or ellipse. Single headed arrow is used to define a causal relationship in the model, with the variable at the tail of the arrow causing the variable at the point. Double headed curved arrows indicate covariances or correlations, without a causal interpretation. The terms *endogenous* and *exogenous* variables are frequently used to describe variables in the model that are dependent on other variables and those that are not dependent on (independent of) other variables in the model respectively. Endogenous variables will have a directed arrow entering into them (i.e. prediction) both from the predictors and a residual term that represents the variance not explained by the predictors. On the other hand, an independent (exogenous) variable is a one that has causes assumed to be external to the model and not influenced by any other variable in the model. Exogenous variables can only have double headed arrows (i.e. covariance or correlation) going into them (Hox and Bechger, 2007; Mwandigha, 2014; Statsoft, 2017).

In dealing with attenuated correlation, SEM utilises measurement model whereby x and y are observed scores for the predictor (selection test) and outcome (criterion) respectively. The terms t_x and t_y are the unobserved (latent) true scores just like in the traditional approach of section 8.1.1. Further, structurally, it is assumed that the observed variables x and y have error terms ε_x and ε_y respectively which represent the measurement error between the measured (x, y) variables and latent variables (t_x, t_y) (see equation 3.1.1 and 3.1.2). The latent variables t_x and t_y also have their respective error terms ε_{t_x} and ε_{t_y} which represent the variance unaccounted for in the latent variable by the measured variable. It is also assumed that there is a relationship between the observed and latent variables (x and t_x , y and t_y). This is diagrammatically represented in Figure 8.1 which includes an extra predictor, z, so as to make the model identifiable. Note that λ_x and λ_y represent the factor loadings accounted for by t_x and t_y. The parameter of interest $r_{t_x t_y}$ (true correlation devoid of the effect of attenuation) is thus essentially estimated free from residual variances (Bedeian, Day, and Kelloway, 1997; Cote and Greenberg, 1990; Lomax, 1986).

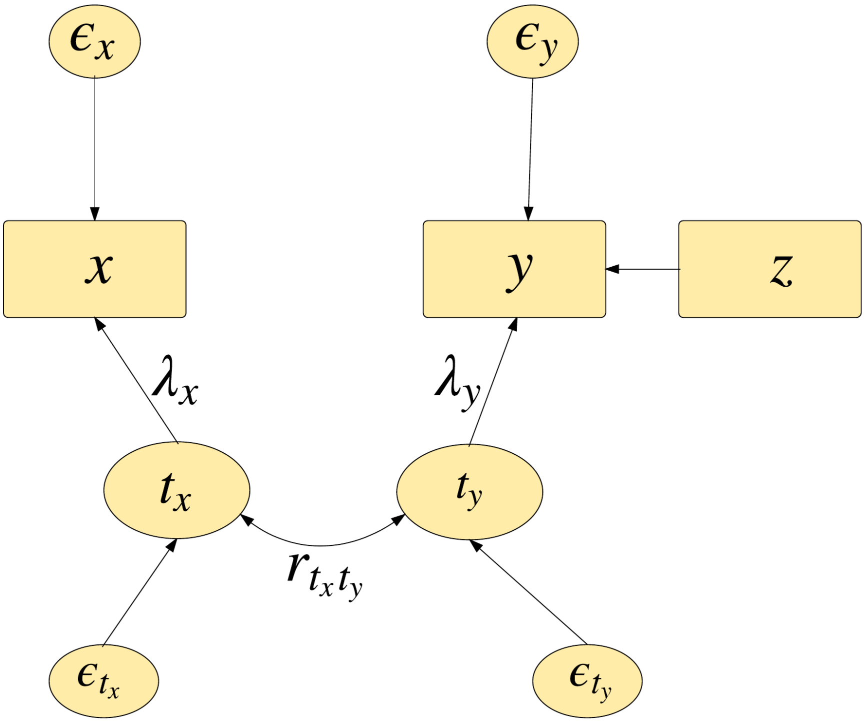


Figure 8.1.: *SEM approach to dealing with attenuation, Note that the predictor z is included in the model so as to make it estimable.*

8.1.3. Bayesian approach

Another alternative to the traditional and SEM approach to dealing with attenuated correlation due to measurement error is the Bayesian approach proposed by *Behseta et al. (2009).* The advantage of the Bayesian approach is that it can handle complex modelling frameworks while incorporating researchers' prior knowledge into the statistical analysis (Lesaffre and Lawson, 2012). The problem of attenuation is dealt with by use of a Bayesian Hierarchical Model (BHM) since the data is assumed to constitute a *hierarchical (random effects)* structure and all parameters are given a *prior distribution*. Take equations 3.1.1 and 3.1.2, where x and y represent the observed predictor (selection test) and outcome (criterion) respectively. The terms t_x, t_y, ε_x and ε_y are unobserved (latent) variables true score for x, true score for y, measurement error for x and y respectively. The hierarchical structure follows from the fact that the observed variables are taken as level 1 and unobserved variables as level 2. To estimate the true correlation, $r_{t_x t_y}$ which is free from attenuation, the multivariate normal distribution of true scores is defined in equation 8.1.4 with the parameters of the distribution modelled as shown in equation 8.1.5 and 8.1.6.

$$\begin{bmatrix} t_x \\ t_y \end{bmatrix} \sim N(\mu, \Sigma) \tag{8.1.4}$$

where

$$\mu = \begin{bmatrix} \mu_{t_x} \\ \mu_{t_y} \end{bmatrix} \tag{8.1.5}$$

and symmetric matrix

$$\Sigma = \begin{bmatrix} \sigma_{t_x}^2 & \sigma_{t_x t_y} \\ & \sigma_{t_y}^2 \end{bmatrix} \tag{8.1.6}$$

Note that, $\sigma_{t_x t_y}$, the covariance of the true score of x and y may be expressed as $r_{t_x t_y}\sigma_{t_x}\sigma_{t_y}$. Therefore the symmetric matrix may be expressed as shown in equation 8.1.7. The terms μ_{t_x}, μ_{t_y} are the means while $\sigma_{t_x}^2$ and $\sigma_{t_y}^2$ are the variances of true scores t_x and t_y respectively. Next, priors are assigned to all the parameters as shown in equations 8.1.8 and 8.1.9

$$\Sigma = \begin{bmatrix} \sigma_{t_x}^2 & r_{t_x t_y}\sigma_{t_x}\sigma_{t_y} \\ & \sigma_{t_y}^2 \end{bmatrix} \tag{8.1.7}$$

$$\varepsilon_x \sim N(0, \sigma_{\varepsilon_x}^2) \tag{8.1.8}$$

$$\varepsilon_y \sim N(0, \sigma_{\varepsilon_y}^2) \tag{8.1.9}$$

Assuming that the priors chosen are to express ignorance regarding prior knowledge of the value of the parameter $\sigma_{\varepsilon_x}^2$ and $\sigma_{\varepsilon_y}^2$, the popular choice of *Jeffreys priors*, $p(\sigma_{\varepsilon_x}^2) \propto \sigma_{\varepsilon_x}^{-2}$ and $p(\sigma_{\varepsilon_y}^2) \propto \sigma_{\varepsilon_y}^{-2}$ have been demonstrated to be improper priors since their Area Under the Curve (AUC) is $\neq 1$. Thus, their use may lead to an improper posterior distributions. To prevent this, locally uniform prior, (Inverse Gaussian) IG $(\varepsilon, \varepsilon)$ would have to be used which would approximate Jeffreys priors $p(\sigma_{\varepsilon_x}^2) \propto \sigma_{\varepsilon_x}^{-2}$ and $p(\sigma_{\varepsilon_y}^2) \propto \sigma_{\varepsilon_y}^{-2}$ when $\varepsilon \to 0$. Similarly, to express ignorance regarding prior knowledge of the parameter values for μ_{t_x} and μ_{t_y}, the priors $N(0, \sigma_0^2)$ may be used with σ_0 being a very large number (e.g.10, 000). Finally, the prior for Σ may be chosen as the Invert Wishart distribution. Alternatively, the elements of the matrix Σ may have priors attached to each of them separately. In that case, use of IG $(\varepsilon, \varepsilon)$ with $\varepsilon \to 0$ for each component would be a poor choice since posterior distribution of the parameters would depend heavily on the choice of ε. Therefore uniform prior $U(0, c)$ with c being a very large number (e.g.10, 000) is recommended. The quantity of interest $r_{t_x t_y}$ in $r_{t_x t_y}\sigma_{t_x}\sigma_{t_y}$ is the correlation devoid of attenuation. It may be obtained by sampling from its posterior distribution together with Bayesian credibility intervals for the purpose of hypothesis testing (Lesaffre and Lawson, 2012).

8.2. Correcting for bias in the (Pearson) correlation coefficient due to range restriction and measurement error

Assume that there is direct selection on, x the predictor (selection test), and interest lies in estimating the correlation coefficient between x and outcome (criterion) variable y. Further, take that x and y are measured with error. Following a selection process, the sample of entrants will be subject to range restriction and measurement error. Thus the resulting Pearson correlation coefficient, r^r_{xy}, will be attenuated (biased downwards) as shown in Figures 3.1 and 3.2. To correct for the effects of both range restriction and measurement error, the correction in equation 8.2.1 may be used. The corrected Pearson correlation coefficient is denoted by $r^c_{t_x t_y}$. Note that $U_x = \frac{1}{u_x}$ where $u_x = sd_x/SD_x$ and r_{xx} and r_{yy} are the respective reliabilities of x and y (Mendoza and Mumford, 1987b). Using the Delta method, it may be shown that the sampling variability of $r^c_{t_x t_y}$ may be obtained as shown on equation 8.2.2 where $\hat{W} = r_{xx}r_{yy} - (r^r_{xy})^2 + U^2_x r^r_{xy}$ and n_r is the restricted sample size (Hakstian, Schroeder, and Rogers, 1988, 1989; Raju and Brand, 2003).

$$r^c_{t_x t_y} = \frac{U_x r^r_{xy}}{\sqrt{r_{xx}r_{yy} - (r^r_{xy})^2 + U^2_x (r^r_{xy})^2)}} \tag{8.2.1}$$

$$\widehat{V(r^c_{t_x t_y})} = \frac{U^2_x r_{xx} r_{yy} (r_{xx} - (r^r_{xy})^2)(r_{yy} - r^r_{xy})^2)}{(n_r - 1)\hat{W}^3} \tag{8.2.2}$$

Publication	Year	Methodology	Participants	Conclusion
Does the UKCAT predict Year 1 performance in medical school (*Lynch et al. (2009)*)	2009	Complete Case analysis (missing data handling technique) ,Pearson and Spearman rank correlation correlation, Multiple Logistic Regression	Year 1 medical students in academic year 2007-2008 at University of Dundee and University of Aberdeen	Negative • *Overall UKCAT scores do NOT predict performance in year 1 of medical school* • *Students who resat examinations had high scores on UKCAT's decision analysis*
Comparison of A-level and UKCAT performance in students applying to UK medical and dental schools in 2006: cohort study (*James, Yates, and Nicholson (2010)*)	2010	Complete Case analysis (missing data handling technique), Spearman rank correlation, Hierarchical logistic regression	Students with at at least 3 passes at A-level who applied to 23 UK medical schools in 2006, data obtained from UKCAT and UCAS	Positive • *Significant correlation between A-level and UKCAT thus UKCAT may be used as proxy for A-level during selection* Negative • *Inherent favourable bias to applicants who are males, from independent schools and from a higher social class*
Has the UK Clinical Aptitude Test improved medical student selection? (*Wright and Bradley (2010)*)	2010	Complete Case analysis (missing data handling technique), Multiple Regression	UKCAT entering University of Newcastle medical school in academic year 2007-2008 (cohort 1) and 2008-2009 (cohort 2)	Positive • *UKCAT significant predictor of performance in all but one knowledge based examination* Negative • *UKCAT's predictive validity declines with subsequent years of study in medical school*

Publication	Year	Methodology	Participants	Conclusion
The UKCAT-12 study: educational attainment, aptitude test performance, demographic and socio-economic contextual factors as predictors of first year outcome in a cross-sectional collaborative study of 12 UK medical schools (*McManus, Dewberry, Nicholson, and Dowell (2013)*)	2013	EM for handling missing data, hierarchical modelling	12 UK medical schools taking the UKCAT as part of the application process from 2006 to 2008, for whom Year 1 medical results were available in 2008 to 2010	Positive • *UKCAT scores have predictive validity for Year 1 medical school performance* • *UKCAT has incremental validity after considering A-level*
Construct-level predictive validity of educational attainment and intellectual aptitude tests in medical student selection: meta-regression of six UK longitudinal studies (*McManus, Dewberry, Nicholson, Dowell, et al. (2013)*)	2013	EM for handling missing data, Pearson, Tetrachoric and Polychoric correlations for grouped data, MCMC for construct-level predictive validity, Meta Regression	12 UK medical schools (4 in Scotland) that use UKCAT as part of the selection process in 2007, 2008 and 2009	Negative • *A-level have higher construct-level predictive validity for postgraduate performance compared to UKCAT*
Validity of the UKCAT in Applicant Selection and Predicting Exam Performance inUK Dental Students (*Lala, Wood, and Baker (2013)*)	2013	Pearson and Spearman correlation, Multiple Regression, Logistic Regression	2008-2009 and 2009-2010 UKCAT entrants into Sheffield dental school	Negative • *UKCAT unable to predict Year 1 dental examination performance* • *Concerns over equity with females less likely to do well in UKCAT*
Predictive power of UKCAT and other pre-admission measures for performance in a medical school in Glasgow: a cohort study (*Sartania et al. (2014)*)	2014	Pearson correlation, Anova, Multiple Regression	Retrospective study of 2007-2008 UKCAT entrants into Glasgow Medical School	Positive • *UKCAT has predictive validity for Year 1 and final course outcomes*
Predictive validity of the UK clinical aptitude test in the final years of medical school: a prospective cohort study (*Husbands, Mathieson, et al. (2014)*)	2014	Pearson correlation, Multiple Regression	Year 4-5 who commenced studies in 2007 at Aberdeen and Dundee University. NOTE : UKCAT only used at Dundee to rank applicants near cut-point for offers	Positive • *UKCAT has predictive validity for performance in Year 4 and 5 of medical school*

Table 8.2.: *Some publications on the UKCAT's predictive validity*

8.3. Simulation techniques: resampling methods

Simulation studies are computer intensive approaches in which different performance measures are used to evaluate statistical methods under a variety of conditions. Simulation studies have become increasingly useful in testing of statistical hypotheses, estimation of standard errors, and/or creation of confidence intervals (Rodgers, Joseph Lee, 1999). Simulation studies involve the use of *resampling methods* which include *Monte Carlo methods, permutation (or randomisation) tests, jacknife* and *bootstrap*. These methods are termed as resampling methods as they involve repeatedly drawing samples from the already available data (such as in randomisation, jackknife and bootstrap). Monte Carlo method involves resampling from a stochastic process (probability distribution) known a priori to generate the data required (Crowley, 1992). The distinction between the resampling methods with respect to sampling approach is shown in Table 8.3. As may be observed, resampling methods that utilise available data may further be categorised on the basis of (i) whether sampling is done *with or without replacement* and (ii) *sample size* categorised on the basis of whether the whole available data or a subset of it is utilised for resampling.

		Sample size	
		Sub-data	Full-data
Sampling approach	Sample without replacement	Jackknife	Permutation
	Sample with replacement		*Bootstrap*

Table 8.3.: *Classification of resampling approaches based on already available data*[2]

Resampling methods are advantageous over traditional statistical methods because they do not require that the distribution of the data be normal or that the sample size be large. In addition, they are potentially highly accurate, widely generalisable and enable validation of theoretical concepts (Hesterberg et al., 2005). Table 8.4 summarises the basic assumptions normally made under each of the resampling methods.

[2] Table adapted from *"The Bootstrap, the Jackknife, and the Randomization Test: A Sampling Taxonomy" by Rodgers, Joseph Lee (1999)*

Assumptions	Resampling methods			
	Permutation	Monte Carlo	Bootstrap	Jackknife
Statistical independent data	✓	✓	✓	✓
Particular underlying distribution(s)	✗	✓	✗	✗
Empirical samples must be random	✗	✓	✓	✓
Relative sensitive to outliers	✗	✓	✓	✓
Data ranked (reduced to rank)	✗	✗	✗	✗
No values in data cause problems	✗	✗	✗	✗
Ties in data cause problems	✗	✗	✗	✗

Table 8.4.: *Basic assumptions of resampling methods data*[3]

8.3.1. Monte Carlo method

Monte Carlo is the only resampling method that does not involve randomly drawing the required samples from already available data. It involves randomly drawing repeatedly samples from a known stochastic process (probability distribution). This typically involves specification of the parameters, sample size and number of samples to be drawn from the distribution (Mahadevan, 1997; Mooney, 1997). The resulting Monte Carlo samples may be used to evaluate bias of estimators, variance of estimator, construct confidence intervals and hypothesis testing. Since the Monte Carlo method draws samples from a probability distribution, the sample size and the number of samples to be drawn are unrestricted. Monte Carlo resampling yields statistically independent samples, to obtain dependent samples, Markov Chain Monte Carlo (MCMC) sampling may be utilised (Lesaffre and Lawson, 2012; Robert, 2004).

8.3.2. Permutation (randomisation)

The randomisation or permutation test is used in the determination of the distribution of a test statistic under the Null Hypothesis, H_0. The test statistic is computed for the original data (taken to be the population) then the data are permuted by shuffling the available data

[3] Table adapted from *"Resampling Methods for Computation-Intensive Data Analysis in Ecology and Evolution" by Crowley (1992)*

numerous times according to the random assignment procedure without replacement. Each time a new sample is created of the same size as the original data. The test statistic is then computed for each permutation (Manly, 2006; Voet, 1994). These data permutations, including the one representing the original results, become the reference set for determining statistical significance. The proportion of data that yield test statistic values greater than or equal to the value of the original test statistic becomes the pvalue. The number of permutations that may be done is dependent on the available data. If all possible permutations for the data are examined then an *exact test* is said to be derived otherwise the result is a *randomisation test* (Good, 2013; Rodgers, Joseph Lee, 1999).

8.3.3. Jackknife

The jackknife was originally developed to assess stability and bias of estimators. The variance of estimators obtained from this method may be used to construct confidence intervals or conduct hypothesis testing. Just like in the permutation or randomisation, the original data (of size n say) is taken to be the population (Efron, 1982; Miller, 1974; Rao and Shao, 1992). The jackknife is implemented by creating samples that are a subset of the original data. This is accomplished by typically drawing from the original data without replacement, samples are generated by leaving out a data point each time sampling is done thus ending up with n samples of size $(n - 1)$. If desired, the jackknife resampling method may be implemented by leaving out two or a group of data points instead of a single data point (Efron, 1992; Rodgers, Joseph Lee, 1999).

8.4. Types of bootstrap

Bootstrap resampling methods may be broadly classified into two groups; *non-parametric* and *parametric bootstrap*.

8.4.1. Non-parametric bootstrap

Non-parametric bootstrap techniques are so referred to because they do not involve making any distributional assumptions relating to data of interest. Instead, the empirical distribution of the data is approximated by resampling without replacement from the already available data which is taken to be the population (Efron, 1982; Efron, Tibshirani, et al., 1986; Van Den Noortgate and Onghena, 2005).

8.4.1.1. Case resampling bootstrap

This bootstrap method is also known as the *ordinary non-parametric (iid) bootstrap*. The term iid means independent and identically distributed. It involves generating samples of the same size as the available data through random draws of cases (whole row of observation) with replacement from the available data (taken to be the population). Because sampling is done with replacement, some of the observations from the available data, of size n say, may appear more than once while others may be omitted altogether in a bootstrap sample. As a consequence, the bootstrap samples are a subset of all possible samples, $\binom{2n-1}{n}$, of the same size, n, as the already available data with n copies of each observation. This means that the bootstrap estimates of bias, standard error, and confidence interval limits are random variables. The variability of the estimates from the non-parametric bootstrap may be reduced by increasing the number of bootstrap samples (n_{sims}) (Rodgers, Joseph Lee, 1999).

The resampling bootstrap may be extended to sample from hierarchical data such as in *clustered* or *longitudinal data*. When the data is hierarchical, from a balanced design with homoscedasticity (common error variance), the simple case resampling done without regard to the hierarchical nature of the data may still perform as well as other (non-)parametric methods for hierarchical data (Thai et al., 2013). For illustration, assume a 2 level hierarchical data structure, where students are clustered within schools, case resampling for hierarchical data may be accomplished in two steps, in the first step, random draws with replacement of the level-2 units, schools, in the data is done followed in the second step by randomly drawing with replacement level-1 units, students, from the schools selected in the first step. In effect, resampling is randomly done with replacement at each level of the hierarchical data in blocks. This has however been shown to lead to biased inference when dealing with hierarchical regression models (Carpenter, Goldstein, and Rasbash, 1999; Van Den Noortgate and Onghena, 2005).

8.4.1.2. Balanced bootstrap

Instead of using the case resampling bootstrap with huge number of samples, n_{sims}, to reduce variability of the estimates from the bootstrap samples, the balanced bootstrap may be used. In the *balanced bootstrap*, each observation is forced to occur in frequency that is equal to the number of bootstrap samples n_{sims} (Dvison, Hinkley, and Schechtman, 1986; Gleason, 1988). This does not however force each bootstrap sample to contain all observations. It may be, for example that the first observation occurs twice in the first bootstrap sample but is excluded in the second bootstrap sample while the second observation may occur in all bootstrap samples. While the balanced bootstrap may decrease the variance of the estimated bias and standard error, it is less useful for estimating confidence intervals (Dixon, 2002; Efron, 1981).

8.4.1.3. Residual (error) resampling bootstrap

The *residual (error) resampling bootstrap* is useful in estimating bias, standard errors and /or confidence intervals for linear regression model problems where the error terms, ε_i, are assumed to be independent and identically distributed (iid) with a common variance σ^2 (homoscedastic) (MacKinnon, 2006). For illustration, suppose, interest lies in fitting the model shown in equation 8.4.1 for the data of size n (note that i=1,2,3,....n) consisting of a predictor x and outcome y. Because the error terms, ε_i, are not supposed to come from the normal or any other distribution, the residual (error) resampling bootstrap is non-parametric (because a model is used however, some refer to it as a type of semi-parametric bootstrap (Carpenter and Bithell, 2000)). The residual (error) resampling bootstrap involves fitting a model (linear regression model in this case) and subsequently obtaining the estimates of error terms ε_i and $\beta_{y|x}$, that is the residuals e_i and $\hat{\beta_{y|x}}$. Thereafter the residuals are sampled with replacement with probability $\frac{1}{n}$ as is the case while using the case resampling bootstrap. To obtain new variable y_i^*, the randomly sampled residuals are added to the original predicted value of y_i from the model as shown in equation 8.4.2 thus creating a bootstrap sample (Dixon, 2002).

$$y_i = \beta_0 + \beta_{y|x} * x_i + \varepsilon_i \tag{8.4.1}$$

$$y_i^* = \hat{\beta_0} + \hat{\beta_{y|x}} * x_i + e_i \tag{8.4.2}$$

The residuals used are unfortunately biased downwards due to a discrepancy between estimated and empirical variance of residuals. To address the bias, it is recommended that the residuals be rescaled so that they have the right variance before the resampling is done. This may be accomplished by multiplying the residuals with a factor equal to $\sqrt{\frac{n-p}{n}}$ where n and p are the number of observations in the data and number of parameters being estimated from the model respectively. Alternatively, the residuals may be multiplied with $\sqrt{\frac{1}{1-h_{ii}}}$ where h_{ii} is the i^{th} diagonal element of the hat matrix which maps the vectors of observed values to the vector of fitted values (Thai et al., 2013). For instances when hierarchical data is in use, the residual (error) resampling bootstrap may still be used. To illustrate, consider the (linear) mixed model for hierarchical data in equation 8.4.3. The model has a random intercept and slope in addition to the fixed parameters being estimated. To implement the residual (error) resampling bootstrap in this context, the model is fitted to obtain the random effects (b_{0i}, b_{1i}), fixed effects (β_0, $\beta_{y|x}$) and residual error ε_i. Subsequently random draws with replacement from the random effects and error terms are done so as obtain b_{0i}^*, b_{1i}^* and e_i^*. These are then used to generate bootstrap samples as shown in equation 8.4.4 (Carpenter, Goldstein, and Rasbash, 2003).

$$y_i = \beta_0 + b_{0i} + \beta_{y|x} * x_i + b_{1i} * x_i + \varepsilon_i \tag{8.4.3}$$

$$y_i^* = \hat{\beta}_0 + b_{0i}^* + \widehat{\beta_{y|x}} * x_i + b_{1i}^* * x_i + e_i^* \tag{8.4.4}$$

The drawback of sampling from $\hat{b_i} = \{\widehat{b_{0i}}, \widehat{b_{1i}}\}$ is that the random effects are shrunk towards zero thus artificially underestimating the variance. To address this shortcoming, a transformation on $\hat{b_i}$ is applied before resampling. A proposed transformation is shown in equation 8.4.5. It involves centering the random effects so as enable resampling from a distribution with mean zero and subsequently multiplying them by the ratio of their corresponding estimated and empirical variance-covariance matrices (Thai et al., 2013). Note that A is an upper triangular matrix of order equal to the number of random effects. Resampling may thus be done from b_i^{**} rather than b_i to obtain bootstrap samples as shown in equation 8.4.6 (Carpenter, Goldstein, and Rasbash, 1999).

$$\hat{b_i}^{**} = \hat{b_i} * A \tag{8.4.5}$$

$$y_i^{**} = \hat{\beta}_0 + b_{0i}^{**} + \widehat{\beta_{y|x}} * x_i + b_{1i}^{**} * x_i + e_i^* \tag{8.4.6}$$

8.4.1.4. Wild and Pairs bootstrap

The *wild bootstrap* and the *pairs bootstrap* are two separate bootstrap methods. They are useful in estimating bias, standard errors and /or confidence intervals for linear regression model problems where the error terms, ε_i, have independence and heteroscedasticity (that is not having a common error variance). Heteroscedasticity in regression models leads to estimates of the covariances of the regression coefficients that are biased and inconsistent. Consequently, the conventional tests are not T and F distributed, even asymptotically (Feng, He, and Hu, 2011; Flachaire, 2005). To illustrate, assume that interest lies in fitting the model in equation 8.4.1. If the error terms, ε_i, have heteroscedasticity then bootstrap samples may be generated the same way as in residual (error) resampling bootstrap as described in section 8.4.1.3 using equation 8.4.7. The term $f(\hat{u}_i)$ is a transformation of the i^{th} residual, $f(\hat{u}_i) = \frac{e_i}{\sqrt{1-h_{ii}}}$, where h_{ii} represents

i^{th} diagonal element of the hat matrix. The term v_i^* is a random variable with possible mean 0 and variance 1. Other better distributions for v_i^* have been proposed such as the one in equation 8.4.8. This has been demonstrated to perform well when the conditional distribution of the error term is approximately symmetric. When the distribution of the error term is asymmetric the distribution for v_i^* shown in equation 8.4.9 is preferred (Davidson and Flachaire, 2008; Mammen, 1993).

$$y_i^* = \hat{\beta}_0 + \widehat{\beta_{y|x}} * x_i + f(\hat{u}_i) v_i^* \tag{8.4.7}$$

$$v_i^* = \begin{cases} 1 \ with\ probability\ \frac{1}{2} \\ -1 \ with\ probability\ \frac{1}{2} \end{cases} \tag{8.4.8}$$

$$v_i^* = \begin{cases} \frac{-(\sqrt{5}-1)}{2} \ with\ probability\ \frac{(\sqrt{5}+1)}{2\sqrt{5}} \\ \frac{(\sqrt{5}+1)}{2} \ with\ probability\ \frac{(\sqrt{5}-1)}{2\sqrt{5}} \end{cases} \tag{8.4.9}$$

The pairs bootstrap, unlike the wild bootstrap, resamples the data rather than residuals. The outcome y_i and predictor x_i are sampled in pairs by selecting a row of observation from the data independently with replacement. Note that the pairs bootstrap is the *case resampling bootstrap* applied to regression problems (Camponovo, 2015; Thai et al., 2013). Consequently, unlike the residual and wild bootstrap, each pairs bootstrap sample drawn has the outcome, y regressed on x as shown in equation 8.4.1. Each draw of the variables in pairs bootstrap may be considered as a draw from a multivariate distribution which does not require the assumption of homoscedasticity. The drawback with the pairs bootstrap is that the predictor x is truly not independent of y. Thus, the choice of the x inadvertently leads to the choice of residual e_i^*. Since the two are related, then $E(e_i^* \neq 0)$, which violates one of the core assumptions of regression. For this reason, the pairs bootstrap which is relatively easy to implement, performs poorly compared to the wild bootstrap when used for linear regression problems (Flachaire, 2005; MacKinnon, 2006).

8.4.1.5. Moving blocks and stationary bootstrap

This bootstrap method works best for *time series data* in which the series of observations are divided into *b* non-overlapping blocks of *l* sequential observations. Observations are assumed to be strongly correlated within a block but weakly correlated between blocks. Each bootstrap sample is constructed by randomly sampling *b* blocks with replacement and combining them into a series of *bl* observations (Paparoditis and Politis, 2001; Politis and White, 2004). Caution is recommended when deciding on the choice of *l*. A small *l* will lead to a big blocks of *b* with observations between the blocks being strongly correlated. If *l* is large, then there is risk that *b* will be small with resulting bootstraps samples not being unique. If the size of *l* is left to be random, then the moving block bootstrap becomes *stationary bootstrap* which is a variant of the moving tiles bootstrap. The major draw back of the moving blocks bootstrap is that it yields data that are less correlated compared to the original data because the blocks are assumed to be independent. Therefore the inference made may be biased even if *l* is appropriately chosen (Dixon, 2002; Lahiri, 1999; Lahiri, 1993).

8.4.1.6. Moving tiles bootstrap

Moving tiles bootstrap methods is used to derive multiple samples for use from spatially correlated data. It is a non-parametric bootstrap method implemented in the same way as moving block bootstrap method. For this reason, it is also referred to as *spatial block bootstrap* method. Just like the moving block bootstrap method, the moving tiles bootstrap is prone to bias resulting to the distortion of the correlation structure in the data (Dixon, 2002; Nordman, Lahiri, and Fridley, 2007).

8.4.2. Parametric bootstrap

The parametric bootstrap generates data by utilising the full distribution of the random component of a *generalised linear model (GLM)* (Pardoe and Weisberg, 2001) thus utilising Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML) estimation. First the model shown in equation 8.4.1 is modelled to obtain residuals. The bootstrap samples are then generated by sampling the residuals from the model based on the specified normal distribution $e_i^* \sim N(0, \sigma^2)$. Thereafter i^{th} sampled residual is added to the i^{th} predicted value of y as shown in equation 8.4.10 (Dixon, 2002; Genest, Rémillard, et al., 2008; MacKinnon, 2006).

$$y_i^* = \hat{\beta}_0 + \widehat{\beta_{y|x}} * x_i + e_i^* \tag{8.4.10}$$

Generally, there is normally little difference between inferences about regression coefficients based on *residual (error) non-parametric bootstrap* or the *parametric bootstrap* with normal errors. This applies regardless of whether or not the error terms are actually normally distributed. The aspect of the model that may lead to discrepancy in results is the assumption (or violation) of homoscedasticity. In order to obtain bootstrap samples, the random effects and the residual errors, are resampled from the normal distributions, $\hat{b_i^*} = \{\hat{b_{0i}^*},\ \hat{b_{1i}^*}\} \sim N(0, D)$ and $\varepsilon_i^* \sim N(0, \Sigma_i)$. The terms D and Σ are the covariance matrices of $\hat{b_i^*}$ and ε_i^* respectively. $\hat{b_i^*}$ and ε_i^*, assumed to be independent, are used to generate new samples as shown in equation 8.4.11 (Carpenter, Goldstein, and Rasbash, 1999; Thai et al., 2013).

$$y_i^* = \hat{\beta}_0 + \hat{b_{0i}^*} + \widehat{\beta_{y|x}} * x_i + \hat{b_{1i}^*} * x_i + e_i^* \tag{8.4.11}$$

8.5. Bootstrap confidence intervals

Confidence intervals combine parameter point estimation and hypothesis testing into a single inferential statement that contains a set of plausible values for the parameter of interest (DiCiccio and Efron, 1996; Wehrens, Putter, and Buydens, 2000). The set of plausible values is called a region which is taken to include the true value of the parameter of interest with a specified probability (Carpenter and Bithell, 2000). Classical asymptotic confidence intervals for a parameter of interest, $\hat{\theta}$ say, may be constructed by a pivotal method in which quantiles of a known distribution are utilised. For example, consider a normal distribution, the confidence interval for $\hat{\theta}$ is shown on equation 8.5.1.

$$\hat{\theta} \pm Z_{(1-\frac{\alpha}{2})} * S(\hat{\theta}) \tag{8.5.1}$$

The terms $\hat{\theta}$ and $S(\hat{\theta})$ are the point estimate and standard error of the point estimate of the parameter respectively. The term $Z_{(1-\frac{\alpha}{2})}$ represents the quantile of the normal distribution, taking $\alpha = 5\%$, the lower and upper quantiles assume the values of ± 1.9604 respectively. Bootstraps confidence intervals may be categorised under three broad family of methods namely *pivotal, non-pivotal* and *test inversion*. Pivotal methods use quantiles estimated from bootstrap methods. Non-pivotal methods are less intuitive while test-inversion methods may be considered a hybrid between confidence intervals and tests useful in (semi-)parametric bootstrap resampling approaches. A summary of the methods are presented in Table 8.5.

Pivotal family	Non-pivotal family	Test-inversion family
Non-studentised pivotal method	Percentile method	Test Inversion Bootstrap (TIB)
Studentised pivotal method	Bias Corrected (BC) percentile method	Studentised Test Inversion Bootstrap (STIB)
	Bias Corrected and accelerated (BCa) method	

Table 8.5.: *Classification of methods for constructing bootstrap confidence interval* [4]

These bootstrap confidence interval methods are briefly explained next. For illustration, the parameter of interest is taken to be $\hat{\theta}$, the true and known value computed from the available data (taken as population). The estimate of this parameter from a bootstrap sample is denoted as $\hat{\theta}_i^*$ with number of bootstrap samples, i, ranging from 1 to n_{sims}. The n_{sims} bootstraps estimates are then sorted in ascending order such that $\hat{\theta}_1^* \leq \hat{\theta}_2^* \leq \hat{\theta}_3^*$.... (Carpenter and Bithell, 2000; DiCiccio and Romano, 1988; Wehrens, Putter, and Buydens, 2000).

[4] Table adapted from *"Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians" by Carpenter and Bithell (2000)*

8.5.1. Non-studentised pivotal method

This is also referred to as a *basic method* or *simple method*. In this method, the difference $d_i^* = (\hat{\theta}_i^* - \hat{\theta})$ is obtained for all the ordered n_{sims} bootstraps estimates. To estimate the upper and lower confidence limits for the confidence interval say for $\alpha = 0.05$ and $n_{sims} = 1000$, the following calculations follow. For the lower confidence limit, $0.025 * 1000 = 25$, and upper confidence limit, $0.975 * 1000 = 975$, resulting in the confidence interval shown in equation 8.5.2. This method is easy to implement but suffers from risk of generating confidence limits that may be outside the permissible range of values of the parameter of interest. For example if $\hat{\theta}$ is known to always exceed zero (like an odds ratio (Bland and Altman, 2000)), the lower confidence limit may be a negative value which would be invalid (Carpenter and Bithell, 2000; DiCiccio and Efron, 1996).

$$(\hat{\theta} - d_{25}^*, \hat{\theta} - d_{975}^*) \tag{8.5.2}$$

8.5.2. Studentised pivotal method

This is also known as *bootstrap t method* or *percentile t method*. It is an improvement over the non-studentised pivotal method. This is because instead of just utilising $d_i^* = (\hat{\theta}_i^* - \hat{\theta})$ to construct the confidence interval, the standard error of $\hat{\theta}_i^*$ is also used, as shown in equation 8.5.3. Thereafter, the t_i^*'s are ordered in ascending order. To obtain confidence intervals at $\alpha = 0.05$, the lower limit and upper limit are computed as in the case for the non-studentised pivotal method but multiplied by the standard error of $\hat{\theta}$, that is $s(\hat{\theta})$, to end up with the desired confidence interval as shown in equation 8.5.4 (Carpenter and Bithell, 2000; DiCiccio and Efron, 1996).

$$t_i^* = \frac{(\hat{\theta}_i^* - \hat{\theta})}{s(\hat{\theta}_i^*)} \tag{8.5.3}$$

$$(\hat{\theta} - s(\hat{\theta}) * t^*_{25}, \hat{\theta} - s(\hat{\theta}) * t^*_{975}) \tag{8.5.4}$$

This method has been shown to perform a kind of bias correction implicitly. When the median of the t^*_i's is positive(negative), the studentised pivotal method tends to shift to the left(right) relative to the classical asymptotic confidence interval based on quantiles from a normal or student's t-distribution. Thus it is more robust to skewness. The studentised pivotal method achieves greater accuracy compared to the asymptotic or simple bootstrap confidence interval. This means that it has higher coverage for comparatively lower sample sizes. The main drawback of this method is that $s(\hat{\theta})$ has to be known (or easily estimable), independent of $\hat{\theta}$ and reliable otherwise the results will be inaccurate. It also suffers from the chief disadvantage of non-studentised confidence interval in that there is a risk of generating confidence limits that may be outside the permissible range of values of the parameter of interest (MacKinnon, 2006).

8.5.3. Percentile method

This method is part of the *non-pivotal* methods which unlike pivotal methods do not require the use of standard error of $\hat{\theta}$, that is, $s(\hat{\theta})$. In addition, unlike pivotal methods, there is no risk of invalid parameter values being included in the confidence interval. The resulting confidence intervals are also invariant to transformation (this means given confidence interval [low,upper] and monotone transformation U, then under the transformation, the confidence interval becomes [U(lower), U(upper)]). The method is also very easy to implement. To illustrate, consider the n_{sims} bootstraps estimates sorted in ascending order such that $\hat{\theta^*_1} \leq \hat{\theta^*_2} \leq \hat{\theta^*_3}$.... and so forth. Assuming that $n_{sims} = 1000$ and $\alpha = 0.05$ then the lower and upper confidence limits are simply the 25^{th} and 975^{th} elements of the ordered sequence. The resulting confidence interval is shown in equation 8.5.5. The major problem of this method is that it yields inaccurate limits of the confidence interval when the distribution of $\hat{\theta}$ is not nearly symmetric (Carpenter and Bithell, 2000; DiCiccio and Romano, 1988; Wehrens, Putter, and Buydens, 2000).

$$(\hat{\theta^*_{25}}, \hat{\theta^*_{975}}) \tag{8.5.5}$$

8.5.4. Bias Corrected (BC) percentile method

The *bias corrected percentile method* is an improvement over the percentile method in that it is a modified percentile method for asymetric distributions of $\hat{\theta}$. As was the case under the percentile method, the n_{sims} bootstrap estimates are sorted in ascending order such that $\hat{\theta^*_1} \leq \hat{\theta^*_2} \leq \hat{\theta^*_3}$.... and so forth. Next, the number of times the bootstrap estimates from the ordered sequence are less than or equal to $\hat{\theta}$, which may be denoted by, $\#(\hat{\theta^*_i} \leq \hat{\theta})$, is obtained (Stata, 2016). This is used to compute the term shown in equation 8.5.6. Note that Φ is the standard cumulative normal. Now taking the desired significance level α, the percentiles of the bootstrap distribution, denoted as p_1 and p_2, may be computed by formulas shown in equation 8.5.7 and 8.5.8. The term $z_{(1-\frac{\alpha}{2})}$ is the $(1-\frac{\alpha}{2})^{th}$ quantile of the normal distribution. To construct the confidence interval, the computed percentiles are then used to construct the confidence intervals in equation 8.5.9 (Carpenter and Bithell, 2000; DiCiccio and Romano, 1988; Wehrens, Putter, and Buydens, 2000).

$$z_0 = \Phi\left(\frac{\#(\hat{\theta^*_i} \leq \hat{\theta})}{n_{nsims}}\right) \tag{8.5.6}$$

$$p_1 = \Phi\left(2(z_0) - z_{(1-\frac{\alpha}{2})}\right) \tag{8.5.7}$$

$$p_2 = \Phi\left(2(z_0) + z_{(1-\frac{\alpha}{2})}\right) \tag{8.5.8}$$

$$(\hat{\theta^*_{p_1}}, \hat{\theta^*_{p_2}}) \tag{8.5.9}$$

8.5.5. Bias Corrected and accelerated (BCa) percentile method

This is an extension of the *bias corrected percentile method* which accounts not only for asymmetry but also kurtosis in the distribution of $\hat{\theta}$. As is in the case of the bias corrected percentile method, the term z_0 is computed as shown in equation 8.5.6. In addition, a variable, which may be denoted by a is estimated from the available data by jackknife method. This is accomplished using the formula in equation 8.5.10. Note that n is taken to be the sample size of the available data, θ_i^{jk} and $\tilde{\theta}$ are the jackknife estimate for θ (from jackknife sample i) and mean of the n jackknife estimates for θ, that is θ_i^{jk}'s, respectively.

$$a = \frac{\sum_i^n (\tilde{\theta} - \theta_i^{jk})^3}{6[\sum_i^n (\tilde{\theta} - \theta_i^{jk})^2]^{\frac{3}{2}}} \tag{8.5.10}$$

The percentiles are then computed by formulas in equation 8.5.11 and 8.5.12. Subsequently, the confidence intervals are constructed using the final procedure of the bias corrected confidence interval shown in equation 8.5.13.

$$p_1 = \Phi\left(z_0 + \frac{z_0 - z_{(1-\frac{\alpha}{2})}}{1 - a[z_0 - z_{(1-\frac{\alpha}{2})}]}\right) \tag{8.5.11}$$

$$p_2 = \Phi\left(z_0 + \frac{z_0 + z_{(1-\frac{\alpha}{2})}}{1 - a[z_0 + z_{(1-\frac{\alpha}{2})}]}\right) \tag{8.5.12}$$

$$(\hat{\theta_{p_1}^*}, \hat{\theta_{p_2}^*}) \tag{8.5.13}$$

Note that the Bias Corrected and accelerated (BCa) percentile method is a special case of Bias Corrected (BC) percentile method when a is zero. When both a and z_0 are zero, the resulting confidence interval is equivalent to the Percentile method (Stata, 2016). The (BCa) percentile method has all the strengths of the Bias Corrected (BC) percentile method and Percentile method in addition to having the narrowest confidence interval of all the the non-pivotal methods. There is a price to pay for this however, the computation of the variable a is ardours and time consuming (Carpenter and Bithell, 2000; DiCiccio and Romano, 1988; Wehrens, Putter, and Buydens, 2000).

8.5.6. Test Inversion Bootstrap (TIB) method

Test inversion methods are a hybrid between confidence intervals methods and statistical tests. They involve the use of (semi-)parametric bootstrap resampling approaches. One of the methods within the test inversion family is the *test inversion bootstrap method.* To illustrate how this works, consider *residual (error) resampling bootstrap* which would normally involve fitting the model in equation 8.4.2. During the construction of the test inversion bootstrap confidence interval, the intercept $\hat{\beta}_0$ is retained but $\widehat{\beta_{y|x}}$ is replaced with an educated guess of the lower and upper confidence limit β_l and β_u which are evaluated separately but in the same way. Therefore, it will suffice to illustrate how only one of the limits, the upper confidence limit, is obtained. The residuals are generated from the model in equation 8.5.14 then resampled to obtain bootstrap samples for $\beta_{y|x}$, that is, $\hat{\beta}^*_{y|x}$. The samples are then used to assess the association in equation 8.5.15. If the term on the left hand side is less than $\frac{\alpha}{2}$, β_u is decreased otherwise it is increased until a solution is found. The appeal of this method is that the standard error of the parameter of interest is not needed. In addition, there is no risk of generating confidence limits that may be outside the permissible range of values of the parameter of interest. The drawback for this method is that it requires twice as many bootstrap samples (compared to the non-pivotal methods) as each of the confidence limits is evaluated separately (Carpenter and Bithell, 2000; DiCiccio and Romano, 1988; Wehrens, Putter, and Buydens, 2000).

$$y_i^* = \hat{\beta}_0 + \hat{\beta}_u * x_i + e_i \tag{8.5.14}$$

$$Pr(\hat{\beta}^*_{y|x} < \widehat{\beta_{y|x}} | \beta_{y|x} = \beta_u) = \frac{\alpha}{2} \tag{8.5.15}$$

8.5.7. Studentised Test Inversion Bootstrap (STIB) method

The rationale of this method is to increase the coverage of Test Inversion Bootstrap (TIB) method by using a studentised statistic such that the inequality in equation 8.5.16 holds. This method is generally more accurate than the Test Inversion Bootstrap (TIB) but it is very computationally intensive. It has the added requirement of utilising $s(\beta_{y|x})$ and $s(\hat{\beta}^*_{y|x})$ which may not be easily estimable (Carpenter and Bithell, 2000; DiCiccio and Romano, 1988; Wehrens, Putter, and Buydens, 2000).

$$\hat{\beta}^*_{y|x} \leq \beta_{y|x} + \frac{(\hat{\beta_{yx}} - \beta_{y|x})s(\hat{\beta}^*_{y|x})}{s(\hat{\beta}_{y|x})} \tag{8.5.16}$$

8.6. Choice of bootstrap and bootstrap confidence interval

As they are several bootstrap methods applicable in different situations, the choice of a particular bootstrap for statistical analysis is crucial. This is because the wrong choice will lead to incorrect inference. To assist in determining which bootstrap method should be used, the properties of the underlying distribution of the data from which the parameters are to be estimated should be considered. Specifically, at least *normality* and *homoscedasticity* should be considered. Both of these properties can be determined by use of statistical tests like *Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors, Anderson-Darling* tests for normality and *White*, *Breusch-Pagan*, *Brown-Forsythe* tests for homoscedasticity (Baser, Crown, and Pollicino, 2006; Carpenter and Bithell, 2000; Neter et al., 1996). Table 8.6 summarises the choice of bootstrap methods under different assumptions for non-hierarchical data structure. The choice of bootstrap confidence interval method to be used will rely on the bootstrap method chosen. Table 8.7 summarises the choice of bootstrap confidence interval for the different bootstrap methods.

Normality	Homoscedasticity	Consistency of estimator
✓	✓	• Parametric bootstrap • Case (pairs) resampling bootstrap • Residual (error) resampling bootstrap
✗	✓	• Case (pairs) resampling bootstrap • Residual (error) resampling bootstrap • Wild bootstrap
✓	✗	• Case (pairs) resampling bootstrap • Wild bootstrap
✗	✗	• Case (pairs) resampling bootstrap • Wild bootstrap

Table 8.6.: *Appropriate bootstrap method based on underlying distributional assumptions (of residuals) of the data*[5]

Method	Properties			
	Transformation respecting	Use with parametric bootstrap	Use with nonparametric bootstrap	$\mathbf{s}(\hat{\theta})$ required
Non studentised pivotal	✗	✓	✓	✗
Studentised pivotal	✗	✓	✓	✓
Percentile	✓	✓	✓	✗
BC percentile	✓	✓	✓	✗
BCa percentile	✓	✓	✓	✗
Test inversion	✓	✓	✗	✗
Studentised test inversion	✗	✓	✗	✓

Table 8.7.: *Properties of bootstrap confidence intervals*[6]

[5] Table adapted from *"Guidelines for selecting among different types of bootstraps" by Baser, Crown, and Pollicino (2006)*

[6] Table adapted from *"Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians" by Carpenter and Bithell (2000)*

	Imputation variables: t and x (NB: $r^u_{ty} = 0.2$)			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case II	-76.5777	-101.8121	-113.8748	-111.8355
Restricted vs EM MI	-77.3078	-102.3017	-113.8748	-112.0387
Restricted vs MCMC MI	-76.0242	-101.5208	-112.9758	-111.7299
Thorndike Case II vs EM MI	-0.5480	-0.3068	-0.3739	-0.1263
Thorndike Case II vs MCMC MI	0.9817	0.7373	0.5286	0.4054
EM MI vs MCMC MI	1.5341	1.0468	0.9027	0.5324
	Imputation variables: u and x (NB: $r^u_{uy} = 0.4$)			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case II	-74.6712	-100.2946	-111.5573	-109.7982
Restricted vs EM MI	-78.0746	-104.2512	-115.0098	-113.5626
Restricted vs MCMC MI	-77.0830	-103.3288	-114.6060	-113.2345
Thorndike Case II vs EM MI	-0.6486	-0.4374	-0.4680	-0.3641
Thorndike Case II vs MCMC	0.6303	0.6012	0.2314	0.0145
EM MI vs MCMC MI	1.3209	1.0765	0.7721	0.3921
	Imputation variables: v and x (NB: $r^u_{vy} = 0.6$)			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case II	-76.0574	-101.6489	-112.4214	-110.9553
Restricted vs EM MI	-87.1524	-111.6869	-122.9365	-119.6702
Restricted vs MCMC MI	-86.3508	-110.9319	-122.5684	-119.8027
Thorndike Case II vs EM MI	-2.3215	-0.6393	-0.6285	0.6196
Thorndike Case II vs MCMC MI	-1.3251	0.1318	-0.0808	-0.6336
EM MI vs MCMC MI	1.1090	0.8543	0.6136	0.147
	Imputation variables: w and x (NB: $r^u_{wy} = 0.8$)			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case II	-78.8676	-102.4855	-115.7457	-115.5726
Restricted vs EM MI	-105.888	-129.2059	-140.6797	-138.2299
Restricted vs MCMC MI	-105.4688	-128.8657	-140.4738	-138.1075
Thorndike Case II vs EM MI	-3.7892	-3.1494	-2.2908	-1.5786
Thorndike Case II vs MCMC MI	-3.3387	-2.8088	-2.0676	-1.4834
EM MI vs MCMC MI	0.6344	0.4817	0.3131	0.1308

Table 8.8.: *T-test comparison of the methods under direct range restriction with imputation based on selection test x together with predictive variables t, u, v and w. The T values highlighted in green and blue were statistically significant with p-values of less than 0.0001 and 0.05 respectively*

	Imputation variables: t and x (NB: $r^{u}_{ty} = 0.2$)			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case II	2.7811	5.1433	6.9106	7.0295
Restricted vs EM MI	2.8007	5.1811	6.9236	7.0544
Restricted vs MCMC MI	2.8343	5.2255	6.9053	7.1040
Thorndike Case II vs EM MI	1.0070	1.0073	1.0019	1.0035
Thorndike Case II vs MCMC MI	1.0191	1.0160	0.9992	1.0106
EM MI vs MCMC MI	1.0120	1.0086	0.9974	1.0070
	Imputation variables: u and x (NB: $r^{u}_{uy} = 0.4$)			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case II	2.5542	4.8127	6.5454	6.8132
Restricted vs EM MI	2.8700	5.5164	7.3619	7.8585
Restricted vs MCMC MI	2.9160	5.5443	7.4388	7.8625
Thorndike Case II vs EM MI	1.1237	1.1462	1.1248	1.1534
Thorndike Case II vs MCMC MI	1.1417	1.1520	1.1365	1.1540
EM MI vs MCMC MI	1.0160	1.0051	1.0040	1.0050
	Imputation variables: v and x (NB: $r^{u}_{vy} = 0.6$)			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case II	2.6713	5.0414	6.6854	6.8702
Restricted vs EM MI	3.8621	7.3323	10.0381	10.3796
Restricted vs MCMC MI	3.9103	7.3339	10.1030	10.4534
Thorndike Case II vs EM MI	1.4457	1.4544	1.5015	1.5108
Thorndike Case II vs MCMC MI	1.4638	1.4547	1.5112	1.5216
EM MI vs MCMC MI	1.0124	1.0002	1.0065	1.0071
	Imputation variables: w and x (NB: $r^{u}_{wy} = 0.8$)			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case II	2.8918	5.2724	7.1976	7.3518
Restricted vs EM MI	8.4514	15.8270	21.1496	20.6853
Restricted vs MCMC MI	8.4993	15.8340	21.1546	20.6391
Thorndike Case II vs EM MI	2.9225	3.0018	2.9385	2.8137
Thorndike Case II vs MCMC MI	2.9391	3.0032	2.9391	2.8074
EM MI vs MCMC MI	1.0057	1.0004	1.0002	0.9978

Table 8.9.: *F-test comparison of the methods under direct range restriction with imputation based on selection test x together with predictive variables t, u, v and w. The F values highlighted in green were significant with p-values of less than 0.0001*

	Imputation variables: t, z and x (NB: $r^{u}_{ty} = 0.2$)			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
EM MI (zx) vs MCMC MI (zx)	2.6885	1.7719	1.1097	0.8087
EM MI (zxt) vs MCMC MI (zxt)	3.4448	2.1801	1.1.216	0.7009
EM MI (zx) vs EM MI (zxt)	0.0185	0.1195	0.0995	0.0999
EM MI (zx) vs MCMC MI (zxt)	3.4608	2.2977	1.2225	0.8011
MCMC MI (zx) vs EM MI (zxt)	-2.6723	-1.6542	-1.0090	-0.7092
MCMC MI (zx) vs MCMC MI (zxt)	0.7666	0.5207	0.1132	-0.0131
	Imputation variables: u, z and x (NB: $r^{u}_{uy} = 0.4$)			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
EM MI (zx) vs MCMC MI (zx)	0.1567	0.1965	0.0044	-0.1063
EM MI (zxu) vs MCMC MI (zxu)	0.3272	0.3242	0.0649	-0.0979
EM MI (zx) vs EM MI (zxu)	0.0948	0.3493	0.2201	-0.0242
EM MI (zx) vs MCMC MI (zxu)	0.4094	0.6568	0.2821	-0.1252
MCMC MI (zx) vs EM MI (zxu)	-0.0667	0.1429	0.2129	0.0791
MCMC MI (zx) vs MCMC MI (zxu)	0.2520	0.4504	0.2742	-0.0205
	Imputation variables: v, z and x (NB: $r^{u}_{vy} = 0.6$)			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
EM MI (zx) vs MCMC MI (zx)	2.7661	1.7703	1.1350	0.8006
EM MI (zxv) vs MCMC MI (zxv)	2.3797	1.2336	0.9445	0.8251
EM MI (zx) vs EM MI (zxv)	-0.0368	0.5749	0.6826	0.3550
EM MI (zx) vs MCMC MI (zxv)	2.120	1.6606	1.5032	1.0708
MCMC MI (zx) vs EM MI (zxv)	-3.0364	-1.3856	0.5855	-0.5398
MCMC MI (zx) vs MCMC MI (zxv)	-0.9528	-0.3049	0.2315	0.1730
	Imputation variables: w, z and x (NB: $r^{u}_{wy} = 0.8$)			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
EM MI (zx) vs MCMC MI (zx)	2.9848	1.9746	1.2821	0.8296
EM MI (zxw) vs MCMC MI (zxw)	1.2033	0.9079	0.7745	0.5694
EM MI (zx) vs EM MI (zxw)	-2.4108	-1.3873	-1.6626	-0.3316
EM MI (zx) vs MCMC MI (zxw)	1.6424	-0.8196	-1.1780	0.0306
MCMC MI (zx) vs EM MI (zxw)	-6.1950	-3.8948	-3.2957	-1.3775
MCMC MI (zx) vs MCMC MI (zxw)	-5.4253	-3.3248	-2.8117	-1.0158

Table 8.10.: *T-test comparison of the methods under indirect range restriction with imputation based on selection test z and predictor x together with predictive t, u, v and w. The T values highlighted in blue and green were significant with p-values of less than 0.05 and 0.0001 respectively*

	Imputation variables: t, z and x (NB: $r^{u}_{ty} = 0.2$)			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
EM MI (zx) vs MCMC MI (zx)	1.0009	0.9953	0.9949	0.9962
EM MI (zxt) vs MCMC MI (zxt)	1.0045	1.0010	0.9978	1.0217
EM MI (zx) vs EM MI (zxt)	1.0035	1.0037	0.9950	1.0018
EM MI (zx) vs MCMC MI (zxt)	1.0080	1.0047	0.9927	1.0237
MCMC MI (zx) vs EM MI (zxt)	1.0026	1.0085	1.0000	1.0057
MCMC MI (zx) vs MCMC MI (zxt)	1.0070	1.0095	0.9978	1.0275
	Imputation variables: u, z and x (NB: $r^{u}_{uy} = 0.4$)			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
EM MI (zx) vs MCMC MI (zx)	1.0450	0.9865	0.9556	0.9523
EM MI (zxu) vs MCMC MI (zxu)	0.9879	0.9662	0.9865	1.0038
EM MI (zx) vs EM MI (zxu)	1.1592	1.1935	1.1447	0.8880
EM MI (zx) vs MCMC MI (zxu)	1.1451	1.1532	1.1293	0.8914
MCMC MI (zx) vs EM MI (zxu)	1.1093	1.2100	1.1979	0.9326
MCMC MI (zx) vs MCMC MI (zxu)	1.0958	1.1690	1.1818	0.9361
	Imputation variables: v, z and x (NB: $r^{u}_{vy} = 0.6$)			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
EM MI (zx) vs MCMC MI (zx)	1.0061	0.9961	0.9907	0.991
EM MI (zxv) vs MCMC MI (zxv)	1.0070	1.0096	1.0047	1.0061
EM MI (zx) vs EM MI (zxv)	1.6061	1.5842	1.6518	1.6577
EM MI (zx) vs MCMC MI (zxv)	1.6174	1.5994	1.6596	1.6679
MCMC MI (zx) vs EM MI (zxv)	1.5964	1.5905	1.6673	1.6728
MCMC MI (zx) vs MCMC MI (zxv)	1.6076	1.6057	1.6751	1.683
	Imputation variables: w, z and x (NB: $r^{u}_{wy} = 0.8$)			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
EM MI (zx) vs MCMC MI (zx)	1.0038	0.9972	1.0093	0.9911
EM MI (zxw) vs MCMC MI (zxw)	1.0051	0.9910	1.0172	0.9926
EM MI (zx) vs EM MI (zxw)	3.8803	4.1498	4.0260	3.9591
EM MI (zx) vs MCMC MI (zxw)	3.8999	4.1126	4.0952	3.9298
MCMC MI (zx) vs EM MI (zxw)	3.8656	4.1615	3.9890	3.9945
MCMC MI (zx) vs MCMC MI (zxw)	3.8851	4.1241	4.0576	3.9650

Table 8.11.: *F-test comparison of the methods under indirect range restriction with imputation based on selection test z and predictor x together with predictive t, u, v and w. The F values highlighted in green were significant with p-values of less than 0.0001*

	T values (r^u_{zy})			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs FIML	-55.5210	-76.8729	-90.5876	-91.6708
Restricted vs Pearson Lawley	-55.4312	-76.828	-90.5527	-91.6407
Restricted vs EM MI	-55.9014	-77.135	-90.4398	-91.6903
Restricted vs MCMC MI	-54.2992	-76.1189	-89.9025	-91.0169
FIML vs Pearson Lawley	0.1748	0.0951	0.0655	0.0448
FIML vs EM MI	-0.4060	-0.3468	-0.0975	-0.2769
FIML vs MCMC MI	1.4639	0.8652	0.4824	0.3909
Pearson Lawley vs EM MI	-0.5814	-0.4421	-0.1629	-0.3215
Pearson Lawley vs MCMC MI	1.2908	0.7706	0.4173	0.3464
EM MI vs MCMC MI	1.8711	1.2114	0.5780	0.6651
	T values (r^u_{xy})			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs FIML	-79.628	-99.7147	-105.9907	-101.0506
Restricted vs Pearson Lawley	-80.2164	-99.9061	-106.0330	-101.0407
Restricted vs EM MI	-80.5609	-100.0748	-106.0519	-100.9078
Restricted vs MCMC MI	-79.4439	-99.3575	-105.4240	-100.2339
FIML vs Pearson Lawley	-0.4461	-0.1244	-0.0170	0.0230
FIML vs EM MI	-0.8626	-0.4610	-0.3284	-0.1699
FIML vs MCMC MI	0.4765	0.4128	0.3948	0.4535
Pearson Lawley vs EM MI	-0.4176	-0.3371	-0.3116	-0.1928
Pearson Lawley vs MCMC MI	0.9249	0.5377	0.4118	0.4308
EM MI vs MCMC MI	1.3424	0.8737	0.7213	0.6206

Table 8.12.: *T-tests for comparing the mean bias of Pearson Lawley, FIML and MI based on EM and MCMC algorithm under the two hurdle selection validity selection design with full information on both selection tests z and x. The T values highlighted in green were significant with p-values of less than 0.0001*

	F values (r^{u}_{zy})			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs FIML	2.7657	4.5138	6.1154	6.7123
Restricted vs Pearson Lawley	2.7813	4.5251	6.1230	6.7153
Restricted vs EM MI	2.7676	4.5012	6.0331	6.6244
Restricted vs MCMC MI	2.7891	4.5143	6.0179	6.5761
FIML vs Pearson Lawley	1.0056	1.0025	1.0012	1.0004
FIML vs EM MI	1.0007	0.9972	0.9865	0.9869
FIML vs MCMC MI	1.0085	1.0001	0.9840	0.9797
Pearson Lawley vs EM MI	0.9951	0.9947	0.9853	0.9865
Pearson Lawley vs MCMC MI	1.0028	0.9976	0.9828	0.9793
EM MI vs MCMC MI	1.0078	1.0029	0.9975	0.9927

	F values (r^{u}_{xy})			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs FIML	3.4270	5.8484	7.4191	7.6913
Restricted vs Pearson Lawley	3.4514	5.8690	7.4290	7.6951
Restricted vs EM MI	3.4411	5.8334	7.3407	7.5719
Restricted vs MCMC MI	3.4712	5.852	7.3382	7.5045
FIML vs Pearson Lawley	1.0071	1.0035	1.0013	1.0005
FIML vs EM MI	1.0041	0.9974	0.9894	0.9845
FIML vs MCMC MI	1.0129	1.0006	0.9891	0.9757
Pearson Lawley vs EM MI	0.9970	0.9939	0.9881	0.9840
Pearson Lawley vs MCMC MI	1.0057	0.9971	0.9878	0.9752
EM MI vs MCMC MI	1.0087	1.0032	0.9997	0.9911

Table 8.13.: *F-tests for comparing the mean bias of Pearson Lawley, FIML and MI based on EM and MCMC algorithm under the two hurdle selection validity design with full information on both selection tests z and x. The F values highlighted in green were significant with p-values of less than 0.0001*

	T values (r^{u}_{zy})			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs FIML	-54.5802	-77.1082	-87.6812	-88.1806
Restricted vs Pearson Lawley	-28.591	-35.0536	-27.8401	1.5932
Restricted vs MCMC MI	-53.3897	-76.3053	-87.0370	-87.5785
Pearson Lawley vs FIML	-29.2863	-48.5311	-67.5211	-94.2514
Pearson Lawley vs MCMC MI	-27.9387	-47.6495	-66.8104	-93.5762
FIML vs MCMC MI	1.3703	0.7875	0.6040	0.2829
	T values (r^{u}_{xy})			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs FIML	-80.7907	-98.7152	-104.6165	-94.7052
Restricted vs Pearson Lawley	-77.7333	-93.1747	-95.9921	-82.0237
Restricted vs MCMC MI	-80.6227	-98.2792	-104.0975	-93.9618
Pearson Lawley vs FIML	-1.2349	-2.9321	-4.6865	-6.8466
Pearson Lawley vs MCMC MI	-0.7389	-2.5005	-4.3432	-6.5528
FIML vs MCMC MI	0.5112	0.4444	0.3505	0.2801

Table 8.14.: *T-tests for comparing the mean bias of Pearson Lawley, FIML and MI based on EM and MCMC algorithm under the two hurdle validity selection design with full information only on selection test z. The T values highlighted in green were significant with p-values of less than 0.0001*

	F values (r^u_{zy})			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs FIML	2.6202	4.4524	5.5166	5.2059
Restricted vs Pearson Lawley	2.1048	2.4160	1.9308	1.0153
Restricted vs MCMC MI	2.6336	4.4223	5.4796	5.1081
Pearson Lawley vs FIML	1.2448	1.8429	2.8571	5.1274
Pearson Lawley vs MCMC MI	1.2512	1.8305	2.8379	5.0310
FIML vs MCMC MI	1.0051	0.9932	0.9932	0.9812

	F values (r^u_{xy})			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs FIML	3.4041	5.3906	6.6541	5.4465
Restricted vs Pearson Lawley	3.1332	4.7786	5.4510	4.0068
Restricted vs MCMC MI	3.4565	5.3867	6.5928	5.3200
Pearson Lawley vs FIML	1.0865	1.1281	1.2207	1.3593
Pearson Lawley vs MCMC MI	1.1032	1.1273	1.2095	1.3277
FIML vs MCMC MI	1.0154	0.9993	0.9908	0.9768

Table 8.15.: *F-tests for comparing the mean bias of Pearson Lawley, FIML and MI based on EM and MCMC algorithm under the two hurdle validity selection design with full information only on selection test z. The F values highlighted in green and blue were significant with p-values of less than 0.0001 and less than 0.05 respectively*

Comparison	T values (r^{u}_{zy})			
	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs FIML	-74.9758	-99.1720	-111.2488	-110.7069
Restricted vs Pearson Lawley	-75.3533	-99.3610	-111.3232	-110.7125
Restricted vs MCMC MI	-74.6200	-98.7795	-110.9098	-110.0121
Pearson Lawley vs FIML	0.3863	0.1660	0.05846	0.0001
Pearson Lawley vs MCMC MI	0.9143	0.5757	0.4073	0.3353
FIML vs MCMC MI	0.5272	0.4095	0.3488	0.3352
Comparison	T values (r^{u}_{xy})			
	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs FIML	-0.9611	-3.8294	-6.4344	-9.2556
Restricted vs Pearson Lawley	-0.9841	-3.8630	-6.4530	-9.2610
Restricted vs MCMC MI	-0.4102	-3.3651	-6.0554	-8.9934
Pearson Lawley vs FIML	0.0225	0.0353	0.0207	0.0061
Pearson Lawley vs MCMC MI	0.5487	0.5241	0.4403	0.3025
FIML vs MCMC MI	0.5265	0.4888	0.4197	0.2964

Table 8.16.: *T-tests for comparing the mean bias of Pearson Lawley, FIML and MI based on EM and MCMC algorithm under the single hurdle concurrent validity selection design. The T values highlighted in green and blue were significant with p-values of less than 0.0001 and 0.05 respectively*

Comparison	F values (r^u_{zy})			
	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs FIML	2.6509	4.8601	6.6205	6.7841
Restricted vs Pearson Lawley	2.6477	4.8649	6.6254	6.7856
Restricted vs MCMC MI	2.6731	4.8616	6.6176	6.6801
Pearson Lawley vs FIML	1.0012	0.9990	0.9993	0.9998
Pearson Lawley vs MCMC MI	1.0096	0.9993	0.9988	0.9844
FIML vs MCMC MI	1.0084	1.0003	0.9996	0.9847
Comparison	F values (r^u_{xy})			
	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs FIML	0.8178	1.2440	1.6234	1.8980
Restricted vs Pearson Lawley	0.8158	1.2443	1.6240	1.8983
Restricted vs MCMC MI	0.8227	1.2409	1.6090	1.8820
Pearson Lawley vs FIML	1.0025	0.9998	0.9997	0.9998
Pearson Lawley vs MCMC MI	1.0059	0.9972	0.9908	0.9916
FIML vs MCMC MI	1.0154	0.9975	0.9911	0.9768

Table 8.17.: *F-tests for comparing the mean bias of Pearson Lawley, FIML and MI based on EM and MCMC algorithm under the single hurdle concurrent validity selection design. The F values highlighted in green and blue were significant with p-values of less than 0.0001 and less than 0.05 respectively*

	T values			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case II	-63.5478	-105.4215	-139.3924	-123.8945
Restricted vs EM MI	-63.7512	-105.5818	-139.3337	-123.3602
Restricted vs MCMC MI	-63.3062	-105.1000	-138.9025	-123.4102
Thorndike Case II vs EM MI	-0.2722	-0.2746	-0.1326	-0.0673
Thorndike Case II vs MCMC MI	0.3722	0.2877	0.3701	0.1917
EM MI vs MCMC MI	0.6444	0.5621	0.5024	0.1240
	F values			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Thorndike Case II vs Restricted	2.8616	3.0601	8.4093	5.9413
EM MI vs Restricted	2.8868	3.0834	8.4349	5.9508
MCMC MI vs Restricted	2.8244	3.0412	8.3611	5.9314
Thorndike Case II vsEM MI	0.9913	0.9924	0.9970	0.9984
Thorndike Case II vs MCMC MI	1.0132	1.0062	1.0058	1.0017
EM MI vs MCMC MI	1.0221	1.0139	1.0088	1.0033

Table 8.18.: *T and F-test comparison of the methods under direct range restriction with imputation based on the selection test PLAB I for bias and MSE respectively. The T and F values highlighted in green were significant with p-values of less than 0.0001*

	T values			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case III	-25.0714	-55.5768	-84.6708	-85.6403
Restricted vs EM MI	-1.7276	-17.2790	-22.7814	-28.2303
Restricted vs MCMC MI	1.2095	-15.3927	-21.3426	-27.1613
Thorndike Case III vs EM MI	24.3238	43.6981	68.4474	62.3999
Thorndike Case III vs MCMC MI	26.5447	45.3012	69.6068	63.2632
EM MI vs MCMC MI	3.0682	2.0458	1.4988	1.0648
	F values			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Thorndike Case III vs Restricted	3.1518	3.9990	6.7882	5.5899
Restricted vs EM MI	1.0947	1.1159	0.8439	0.7885
Restricted vs MCMC MI	1.1715	1.1561	0.8670	0.8012
Thorndike Case III vs EM MI	3.4503	4.4625	5.7284	4.4075
Thorndike Case III vs MCMC MI	3.6925	4.6233	5.8855	4.4786
EM MI vs MCMC MI	1.0702	1.0360	1.0274	1.0161

Table 8.19.: *T and F-test comparison of the methods under indirect range restriction with imputation based on the selection test PLAB I and predictor IELTS for bias and MSE respectively. The T and F values highlighted in green and blue were significant with p-values of less than 0.0001 and 0.05 respectively*

	T values			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs Thorndike Case III	-26.3831	-55.6237	-85.4958	-84.9945
Restricted vs EM MI	146.8842	123.4281	109.9061	70.2585
Restricted vs MCMC MI	148.0555	124.0224	110.2450	70.6167
Thorndike Case III vs EM MI	130.8214	154.3443	178.1188	147.2712
Thorndike Case III vs MCMC MI	131.5384	154.8014	178.3742	147.4748
EM MI vs MCMC MI	0.9604	0.8268	0.4452	0.5172
	F values			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Thorndike Case III vs Restricted	3.2488	3.8003	6.7318	5.5745
EM MI vs Restricted	3.7951	5.5800	4.2596	2.7048
MCMC MI vs Restricted	3.8345	5.6375	4.2905	2.7399
EM MI vs Thorndike Case III	1.1681	1.4683	0.6328	0.4852
MCMC MI vs Thorndike Case III	1.1803	1.4834	0.6374	0.4915
EM MI vs MCMC MI	0.9897	0.9898	0.9928	0.9872

Table 8.20.: *T and F-test comparison of the methods under indirect range restriction with imputation based on only the selection test ,PLAB I, for bias and MSE respectively. The T and F values highlighted in green were significant with p-values of less than 0.0001*

	T values ($r^{u}_{PlabI.PlabII}$)			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs FIML	-80.8803	-104.8327	-105.5194	-97.0922
Restricted vs Pearson Lawley	-47.8826	-51.9512	-36.2060	-0.6010
Restricted vs MCMC MI	-80.1155	-104.5022	-104.9398	-96.3985
Pearson Lawley vs FIML	-35.6206	-55.5585	-71.2664	-96.7440
Pearson Lawley vs MCMC MI	-34.6763	-55.1235	-70.7730	-96.0498
FIML vs MCMC MI	1.0169	0.5123	0.3288	0.2152
	T values $r^{u}_{Ielts.PlabII}$			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs FIML	-60.2092	-78.3333	-64.7073	-60.4831
Restricted vs Pearson Lawley	-48.1498	-56.3528	-31.8370	-6.2270
Restricted vs MCMC MI	-59.2882	-76.8310	-62.9910	-59.2762
Pearson Lawley vs FIML	-9.5809	-20.0061	-33.0083	-54.8124
Pearson Lawley vs MCMC MI	-8.1630	-18.4209	-31.3898	-53.6250
FIML vs MCMC MI	1.5446	1.6566	1.5110	0.8767

Table 8.21.: *T-test comparison of the methods under the two hurdle validity selection design with imputation based on the selection tests PLAB I and IELTS with full information on PLAB I only. The T values highlighted in green were significant with p-values of less than 0.0001*

Comparison	F values ($r^{u}_{PlabI.PlabII}$)			
	SR=0.2	SR=0.4	SR=0.6	SR=0.8
FIML vs Restricted	4.6262	6.0992	4.8262	3.9825
Pearson Lawley vs Restricted	1.8678	1.7497	1.0414	0.9836
MCMC MI vs Restricted	4.5042	6.0303	4.8061	3.9868
FIML vs Pearson Lawley	2.4768	3.4859	4.6342	4.0490
MCMC MI vs Pearson Lawley	2.4115	3.4465	4.6150	4.0535
FIML vs MCMC MI	1.0271	1.0114	1.0041	0.9989

Comparison	F values $r^{u}_{Ielts.PlabII}$			
	SR=0.2	SR=0.4	SR=0.6	SR=0.8
FIML vs Restricted	1.0308	1.3020	1.5467	2.2927
Pearson Lawley vs Restricted	1.0050	0.9495	0.8980	0.9558
MCMC MI vs Restricted	0.9731	1.2517	1.5114	2.2740
FIML vs Pearson Lawley	1.0256	1.3713	1.7223	2.3988
MCMC MI vs Pearson Lawley	0.9682	1.3183	1.6830	2.3792
FIML vs MCMC MI	1.0593	1.0402	1.0233	1.0082

Table 8.22.: *F-test comparison of the methods under the two hurdle selection design with imputation based on the selection tests PLAB I and IELTS with full information on PLAB I only. The F values highlighted in green were significant with p-values of less than 0.0001.*

	T values ($r^{u}_{PlabI.PlabII}$)			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs FIML	-74.6656	-96.0544	-95.2109	-82.0224
Restricted vs Pearson Lawley	-75.1295	-96.4055	-95.3765	-82.0898
Restricted vs EM MI	-75.4598	-96.7260	-95.3517	-81.7077
Restricted vs MCMC MI	-74.0126	-95.7134	-94.7110	-81.6422
FIML vs Pearson Lawley	-0.1934	-0.1857	-0.1106	-0.0532
FIML vs EM MI	-0.6331	-0.5636	0.2761	-0.0136
FIML vs MCMC MI	0.9120	0.4352	0.2300	0.0965
Pearson Lawley vs EM MI	-0.4410	-0.3784	-0.1658	0.0394
Pearson Lawley vs MCMC MI	1.1083	0.6217	0.3405	0.1495
EM MI vs MCMC MI	1.5481	0.9997	0.5053	0.1097
	T values ($r^{u}_{Ielts.PlabII}$)			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs FIML	-51.6805	-61.8408	-42.8452	-24.3451
Restricted vs Pearson Lawley	-52.2607	-61.8730	-42.7185	-24.2234
Restricted vs EM MI	-52.6816	-62.1804	-42.9026	-24.3406
Restricted vs MCMC MI	-50.7346	-60.2221	-41.4524	-23.2609
FIML vs Pearson Lawley	0.3922	0.0196	0.1385	0.1268
FIML vs EM MI-0.8932	-0.4262	-0.2566	-0.1906	
FIML vs MCMC MI	1.5030	1.7280	1.3031	0.9461
Pearson Lawley vs EM MI	-0.5030	-0.4862	-0.3946	-0.3164
Pearson Lawley vs MCMC MI	1.9030	1.7096	1.1652	0.8203
EM MI vs MCMC MI	2.4063	2.1917	1.5530	1.1277

Table 8.23.: *T-test comparison of the methods under the two hurdle selection design with imputation based on the selection tests PLAB I and IELTS with full information available on both selection tests. The T values highlighted in green and blue were significant with p-values of less than 0.0001 and 0.05 respectively*

	F values ($r^u_{PlabI.PlabII}$)			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
FIML vs Restricted	4.3971	5.5455	4.1505	3.0147
Pearson Lawley vs Restricted	4.3954	5.5571	4.1573	3.0174
EM MI vs Restricted	4.4490	5.6029	4.1807	3.0327
MCMC MI vs Restricted	4.2849	5.4916	4.1439	3.0233
FIML vs Pearson Lawley	1.0003	0.9979	0.9984	0.9991
FIML vs EM MI	0.9883	0.9898	0.9928	0.9941
FIML vs MCMC MI	1.0262	1.0098	1.0016	0.9972
Pearson Lawley vs EM MI	0.9879	0.9918	0.9944	0.9950
Pearson Lawley vs MCMC MI	1.0258	1.0119	01.0032	0.9980
EM MI vs MCMC MI	1.0383	1.0203	1.0089	1.0031

	F values ($r^u_{Ielts.PlabII}$)			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs FIML	1.0836	1.0851	1.0257	0.9805
Restricted vs Pearson Lawley	1.0895	1.0883	1.0286	0.9824
Restricted vs EM MI	1.0762	1.0697	1.0059	0.9503
Restricted vs MCMC MI	1.1447	1.1243	1.0365	0.9684
FIML vs Pearson Lawley	1.0053	1.0029	1.0029	1.0020
FIML vs EM MI	0.9931	0.9857	0.9807	0.9692
FIML vs MCMC MI	1.0564	1.0361	1.0105	0.9877
Pearson Lawley vs EM MI	0.9878	0.9829	0.9779	0.9673
Pearson Lawley vs MCMC MI	1.0507	1.0331	1.0077	0.9857
EM MI vs MCMC MI	1.0637	1.0511	1.0304	1.0190

Table 8.24.: *F-test comparison of the methods under the two hurdle validity selection design with imputation based on the selection tests PLAB I and IELTS with full information on available on both selection tests. The F values highlighted in green and blue were significant with p-values of less than 0.0001 and 0.05 respectively*

Comparison	T values ($r^{u}_{PlabI.PlabII}$)			
	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs FIML	-61.1037	-105.46587	-140.4319	-122.8577
Restricted vs Pearson Lawley	-61.8792	-106.2758	-140.8275	-123.0020
Restricted vs MCMC MI	-61.6173	-105.8963	-140.2370	-122.1896
Pearson Lawley vs FIML	0.7849	0.5248	0.2443	0.1122
Pearson Lawley vs MCMC MI	0.4696	0.3842	0.3707	0.3352
FIML vs MCMC MI	-0.3168	-0.1413	0.1264	0.2232
Comparison	T values ($r^{u}_{Ielts.PlabII}$)			
	SR=0.2	SR=0.4	SR=0.6	SR=0.8
Restricted vs FIML	-24.3461	-53.0948	-83.8366	-83.5515
Restricted vs Pearson Lawley	-24.5266	-53.4321	-84.0336	-83.6439
Restricted vs MCMC MI	-23.8194	-52.6930	-83.0405	-82.2470
Pearson Lawley vs FIML	0.2854	0.3350	0.1860	0.0924
Pearson Lawley vs MCMC MI	0.6980	0.6653	0.7297	0.7215
FIML vs MCMC MI	0.4146	0.3306	0.5441	0.6298

Table 8.25.: *T-test comparison of the methods under evaluation under the single hurdle validity selection design with imputation based on the selection test PLAB I and predictor IELTS. The T values highlighted in green were significant with p-values of less than 0.0001*

	F values ($r^{u}_{PlabI.PlabII}$)			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
FIML vs Restricted	2.9022	3.1801	9.0214	6.0366
Pearson Lawley vs Restricted	2.9598	3.2081	9.0526	6.0475
MCMC MI vs Restricted	2.9067	3.1799	9.0050	6.0287
Pearson Lawley vs FIML	1.0198	1.0088	1.0035	1.0018
Pearson Lawley vs MCMC MI	1.0183	1.0088	1.0053	1.0031
FIML vs MCMC MI	0.9985	1.0001	1.0019	1.0013
	F values ($r^{u}_{Ielts.PlabII}$)			
Comparison	SR=0.2	SR=0.4	SR=0.6	SR=0.8
FIML vs Restricted	3.2241	3.8868	6.7352	5.4521
Pearson Lawley vs Restricted	3.2817	3.9196	6.7592	5.4619
MCMC MI vs Restricted	3.2055	3.8630	6.6806	5.4275
Pearson Lawley vs FIML	1.0179	1.0084	1.0036	1.0018
Pearson Lawley vs MCMC MI	1.0238	1.0146	1.0118	1.0063
FIML vs MCMC MI	1.0058	1.0062	1.0082	1.0045

Table 8.26.: *F-test comparison of the methods under evaluation under the single hurdle validity selection design with imputation based on the selection test PLAB I and predictor IELTS. The F values highlighted in green were significant with p-values of less than 0.0001.*

8.7. The computation of Peer Competition Rescaling (PCR)

8.7.1. Guidelines for computing PCR

Given selection data, one needs to conduct Peer Competition Rescaling (PCR) following the guidelines enumerated below:-

1. **Identification of the predictor (used for selection) and the outcome of interest**

   Determine the predictor and outcome variables. The predictor should be scores from a selection test sat for by all the applicants whilst the outcome should be a local outcome. In the UK selection setting, the predictor would be a national predictor like the United Kingdom Clinical Aptitude Test (UKCAT)) whilst the outcome would be *knowledge-based* exams for year one of medical school training from different medical schools.

2. **Means and corresponding standard deviation for the predictor and outcome**

   Compute the mean and standard deviation for the predictor and outcome. Note that for the predictor, one should have one *mean* and *standard deviation* since the predictor scores would be obtained from a selection test sat for by all the applicants. For the outcome, the number of means and corresponding standard deviations would be equal to the number of local institutions for which entrant data is available. This is because each participating institution would have its own local exam which would differ from other local exams (from other institutions). Therefore, the means and standard deviations of the outcome would have to be computed separately by institution. In addition, if the predictor in question is a national selection test sat for by applicants in different testing periods then the number of means (and corresponding standard deviations) of the predictor would be equal to the number of different testing periods. For example, if the applicants sat for the UKCAT in 2006 and 2007, then two means and their corresponding standard deviations would be computed as was the case shown by equation 6.1.1.

3. **Standardisation of the predictor and outcome scores**

   Next, using the results obtained from step 2, standardisation of the predictor and the outcome may be obtained by taking their reported scores, subtracting their corresponding computed mean and dividing the result by their computed standard deviation. Note that standardisation of the predictor would utilise means and standard deviations obtained from data of all applicants whilst the outcome would utilise data only for entrants (as rejected applicants would not have any of their outcomes observed). The purpose of standardisation of the predictor and outcome is to enable the comparison of the scores on a common scale (mean of zero and variance of one).

4. **Computation of the Peer Rescaled outcomes**

   Following steps 1 to 3, the Peer Competition Rescaling (PCR) makes use of the entrant data. That is, the standardised predictor and outcome scores of those who were selected. The means and standard deviations of the standardised predictor scores are computed for the entrants by institution (for which outcome data is available). Thereafter, the peer rescaled scores are obtained by taking the entrants' standardised outcomes for the from step 3, adding to them, the mean entrant score of the standardised predictor scores and dividing the results by the standard deviation of the entrant scores (of the standardised predictor).

8.7.2. Data generation of selection data

In this section, selection data were generating in R software, with the view of demonstrating with the aid of a simulated example, the procedure of conducting PCR as described in 8.7.1. The R code presented may be copied and re-run in R software after the packages *MASS* and *plyr* are installed. For example installation of the R package *MASS* may be done by the R code install.packages("MASS"). In the R code, green and blue colours are used to highlight the user comments (non-executable R code) and R software key words (e.g. R functions) respectively. The R function set.seed() with an integer as an argument ensures that the same results are obtained every time the R-code is run (as long as that integer is retained, see example in section 8.7.2, line 4).

In the R code, the computation of PCR is demonstrated for a 100 applicants (section 8.7.2, line

9) who apply to two medical schools, with identity code “1” and “2” (section 8.7.2, line 20). For simplicity, it is assumed that an applicant can apply to only one of each medical school randomly with probability of applying to either being 50% (section 8.7.2, line 20). This is far from realistic (as applicants tend to apply to more than one medical school) but will suffice for the purpose of demonstration. Further, it is assumed that the selection ratio for both of the medical schools is 0.4 (that is, top 40% of the applicants are selected based on their predictor (UKCAT scores, 8.7.2, lines 49 to 59 and 62 to 73 for medical schools 1 and 2 respectively). The UKCAT scores (selection test) and knowledge-based exam outcome (criterion) were drawn from a standardised multivariate normal distribution (section 8.7.2, line 4 to 23, with mean and covariance matrix specified in line 7 ans 8 respectively). The proportion of total variability accounted for by the between variability due to correlated outcomes of entrants in a medical school was arbitrarily taken to be $\frac{5}{105} * 100 = 4.76\%$ (section 8.7.2, line 28).

```
1  ##################################################################
2  #Simulate data from multivariate normal distribution, pre-selection.
3  ##################################################################
4  set.seed(46464685) # This allows repeatability
5
6  # Here, specify the mean and covariance structure of the data.
7  mu.data=c(65,70)
8  sigma.data=matrix(c(50,30,30,50),2,2)
9  n.applicants=100 # how many applicants to be simulated
10
11 require(MASS)
12 data=mvrnorm(n =n.applicants, mu=mu.data, Sigma=sigma.data, tol = 1e-6,
       empirical = TRUE)
13 colnames(data)=c(”x”,”y”)
14
15 #####################################################################
16 # Now assume applicants apply to one of two medical schools, 1 or 2,
17 # with random probability of 0.5
18 #####################################################################
19 set.seed(48747477)
20 med.sch.allocation=data.frame(sample(x=1:2,size=nrow(data),replace=TRUE,
    prob=c(0.5,0.5)))

med.sch.data=cbind(med.sch.allocation ,data)
colnames(med.sch.data)=c("med_sch","x","y")

# Inorder for multi-level data to be created, assume that correlation,
# that is variability between  medical schools is 5/105.
# The figure 105 obtained by 5 to variances of predictor and outcome.
var.co=5

set.seed(474747454)
re=data.frame(rnorm(n=2,mean=0,sd=var.co))
med.sch.data$co=NULL
for (i in 1:nrow(med.sch.data))
{
med.sch.data$co[i]=ifelse(med.sch.data$med_sch[i]==1,re[1,1],re[2,1])
}
med.sch.data$knowldege=med.sch.data$y + med.sch.data$co

# Introduce unique id for every applicant
med.sch.data$app_id=1:nrow(med.sch.data)
med.sch.data2=med.sch.data[,c(6,1,2,5)]
colnames(med.sch.data2)=c("app_id","med_sch","ukcat","knowledge")
head(med.sch.data2)

###############################################################
# Simulate selection into medical schools
###############################################################
############## Medical School 1
data.med.sch.1=med.sch.data2[med.sch.data2$med_sch==1,]

# Taking top 40%, sort according to ukcat ascending
data.med.sch.1.sorted=data.med.sch.1[order(data.med.sch.1$ukcat),]
qnt60_1=quantile(data.med.sch.1.sorted$ukcat,0.6)

```

```
top40_med_1=data.med.sch.1.sorted
top40_med_1$knowledge_sel=NULL

# Create systematic missingness in knowledge scores for those rejected
top40_med_1$knowledge_sel=ifelse((top40_med_1$ukcat) >= qnt60_1,
    top40_med_1$knowledge, NA)

############## Medical School 2
data.med.sch.2=med.sch.data2[med.sch.data2$med_sch==2,]

# Taking top 40%,  sort according to ukcat ascending
data.med.sch.2.sorted=data.med.sch.2[order(data.med.sch.2$ukcat),]
qnt60_2=quantile(data.med.sch.2.sorted$ukcat,0.6)

top40_med_2=data.med.sch.2.sorted
top40_med_2$knowledge_sel=NULL

# Create systematic missingness in knowledge scores for those rejected
top40_med_2$knowledge_sel=ifelse((top40_med_2$ukcat) >= qnt60_2,
    top40_med_2$knowledge, NA)

################################################################
# Now collate selection data
################################################################
selection.d=rbind(top40_med_1,top40_med_2)

selection.d$mean_ukcat=mean(selection.d$ukcat)
selection.d$sd_ukcat=sd(selection.d$ukcat)

require(plyr)
knowldege.summary=ddply(selection.d, .(med_sch), summarize,
    mean_knowldeg_sel=mean(knowledge_sel, na.rm=TRUE),
sd_knowldeg_sel =sd(knowledge, na.rm=TRUE))

selection.d2=merge(selection.d,knowldege.summary,by=”med_sch”)
```

```

###################################################################
# Standardisation of the ukcat (predictor) and knowledge (outcome)
##################################################################
selection.d2$z_ukcat=((selection.d2$ukcat-selection.d2$mean_ukcat)/
                             selection.d2$sd_ukcat)

selection.d2$z_knowledge=((selection.d2$knowledge_sel-
selection.d2$mean_knowldeg_sel)/selection.d2$sd_knowldeg_sel)

selection.data=selection.d2[,c(2,1,3,10,5,11)]
```

Snapshots of the data generated are shown inside the frames that follow for medical school “1” and “2” respectively. These snapshots were obtained using the R function *head()* with the integers between the colon (:) specifying the rows of observations to display. For example “27:32” displays rows 27 to 32 of the simulated data. Notice that the right most column named “zknowledge” represents the standardised simulated outcome (step 3 in section 8.7.1) with “NA” denoting those applicants who were rejected and thus having no outcome observed. The proportion of applicants with outcome data (*knowledge-based* exam scores) was 0.4082 and 0.4117 for medical schools “1” and “2” respectively. This is in line with the selection ratio of 0.4 selected. Note also that from the snapshots, the approximate threshold for selection for the different medical schools “1” and “2” may be deduced by examining the predictor scores (ukcat) at the point where the first entrant appears. This is because the predictor scores are sorted in ascending order.

```
> # Applicants selected out of 52 selected
> length(na.omit(top40_med_1$knowledge_sel))
[1] 20
# Proportion selected
> length(na.omit(top40_med_1$knowledge_sel))/nrow(top40_med_1)
[1] 0.4081633
> round(head(selection.data[27:32,]),2)
app_id med_sch ukcat z_ukcat    knowledge_sel   z_knowledge
 11        1    65.11    0.02            NA            NA
 47        1    65.58    0.08            NA            NA
 48        1    65.83    0.12            NA            NA
 33        1    66.19    0.17          82.57          0.02
 51        1    66.52    0.22          65.51         -2.39
 64        1    66.81    0.26          88.69          0.89
```

```
> # Applicants selected out of 48 selected
> length(na.omit(top40_med_2$knowledge_sel))
[1] 21
# Proportion selected
> length(na.omit(top40_med_2$knowledge_sel))/nrow(top40_med_2)
[1] 0.4117647
app_id med_sch ukcat z_ukcat  knowledge_sel  z_knowledge
 6        2    65.58    0.08            NA            NA
24        2    66.06    0.15            NA            NA
85        2    66.31    0.18          68.67         -0.79
19        2    66.34    0.19          83.50          1.28
56        2    67.38    0.34          64.67         -1.35
45        2    67.39    0.34          70.74         -0.50
```

8.7.3. Computation of the peer rescaled outcomes

Following the data generated in section 8.7.2, the PCR was implemented on the simulated standardised *knowledge-based* outcome scores (“zknowledge”, see snapshots of data in section 8.7.2) by first computing the mean and standard deviation of the predictor score (in this example called “zukcat” see snapshots of data in double and single black frames) among the entrants for each medical school (“1” and “2” separately).

```
require(plyr)


ukcat_summary_pcr=ddply(selection.data, .(med_sch),
              summarize, ukcat_mean_pcr= mean(z_ukcat,
              na.rm=TRUE),
              ukcat_sd_pcr =sd(z_ukcat, na.rm=TRUE))


> ukcat_summary_pcr
med_sch ukcat_mean_pcr ukcat_sd_pcr
1       1    -0.06596744    1.0895305
2       2     0.06338048    0.9121646
```

The resulting mean and standard deviations for the different medical schools“1” and “2” are shown inside the triple black frame. Thereafter, the peer rescaled *knowledge-based* outcome scores were then obtained for each medical school by adding to the computed means of the predictor (ukcat) score and subsequently divided by the computed standard deviations as described by step 4 in section 8.7.1.

```
#################################################################
# Compute PCR
#################################################################
selection.data.merge=merge(selection.data,ukcat_summary_pcr,by="med_sch")
selection.data.merge$pcr=((selection.data.merge$z_knowledge+
6  selection.data.merge$ukcat_mean_pcr)/selection.data.merge$ukcat_sd_pcr)
7
8  # Exploratory data analysis
9  par(mfrow=c(1,2))
10 cl=rainbow(4)
11
12 boxplot(z_knowledge~med_sch,data=selection.data.merge,xlab=c(”Medical
       school”),names=c(”1”,”2”),ylim=c(−2,2),
13 main=c(”Standardised knowledge−based score”),col=c(cl[1],cl[2]))
14 boxplot(pcr~med_sch,data=selection.data.merge,
15              xlab=c(”Medical school”),names=c(”1”,”2”),ylim=c(−2,2),
16 main=c(”PCR standardised knowledge−based score”),col=c(cl[3],cl[4]))
```

Figure 8.2 shows the distribution of the *knowledge-based* exam outcome before and after peer competition re-scaling in each of the medical school produced by running the R code in lines 5 to 6 (in this section 8.7.3).

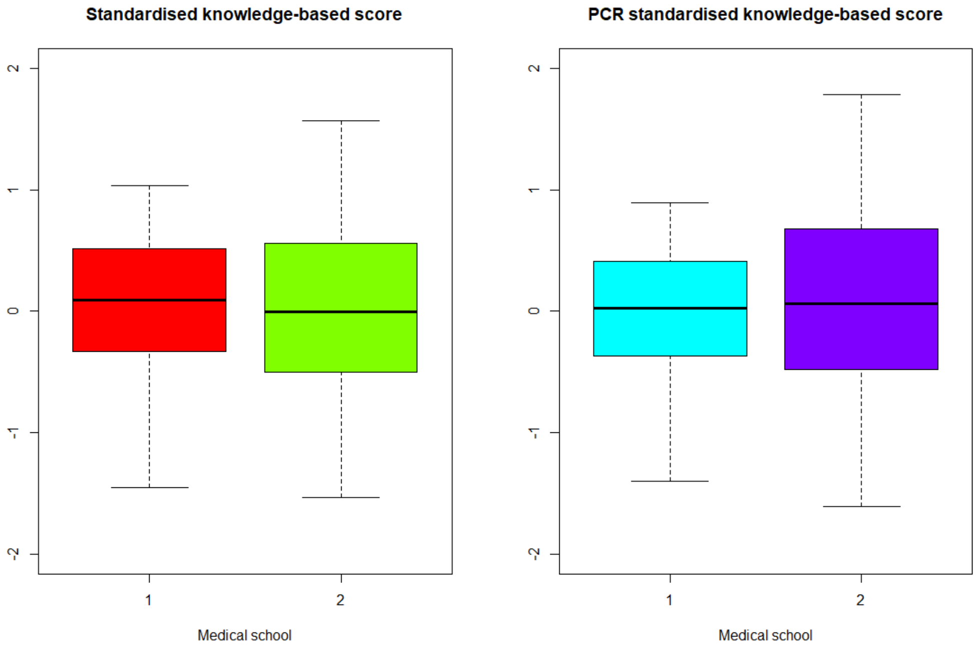


Figure 8.2.: *Distribution of simulated knowledge-based outcome scores for two hypothetical medical schools before and afterPeer Competition Rescaling (PCR)*

Note that, the observed distribution depends on the pre-specified mean and covariance structure of predictor and outcome (section 8.7.2, line 7 and 8), number of applicants (section 8.7.2, line 9), selection ratio (section 8.7.2,lines 53 and 66 for medical schools “1” and “2” respectively) and variability between medical schools (section 8.7.2, line 28). Therefore, to make a conclusion on the impact of Peer Competition Rescaling (PCR), these pre-specifications need to be evaluated in a simulation study that considers many samples.

Glossary

attenuated correlation artificial deflation of empirical correlation between two variables.

construct-level predictive validity predictive validity corrected for attenuation resulting from range restristion and/or measurement error.

improper posterior posterior distribution whose AUC is NOT equal to 1.

improper prior prior distribution whose AUC is NOT equal to 1.

locally uniform prior prior that is appropprimately constant on the interval where the likelihood is not (close to) zero, AUC is equal to 1.

measurement error inconsistencies in measurements associated to a test or instrument.

predictive validity association between a score in a selection measure (i.e. education attainment) and scores on a outcome measure (i.e undergraduate and postgraduate performance).

range restriction lower truncation in a variable due to incidental or direct selection .

reliability Proportion of variability of a variable devoid of measurement error.

Acronyms

AAMC Association of American Medical Colleges.

A-level General Certificate of Education Advanced Level.

AUC Area Under the Curve.

BHM Bayesian Hierarchical Model.

BMAT BioMedical Admissions Test.

DRR Direct Range Restriction.

EI Emotional Intelligence.

EM Expectation Maximisatiom.

EPM Educational Performance Measure.

EU European Union.

FIML Full Information Maximum Likelihood.

GAMSAT Graduate Medical School Admissions Test.

GCSE General Certificate of Secondary Education.

GMC General Medical Council.

GPA Grade Point Average.

HPAT-Ireland Health Professions Admission Test-Ireland.

IMGs International Medical Graduates.

IRR Indirect Range Restriction.

IRT Item Response Theory.

LMM (General) Linear Mixed Model.

LOCF Last Observation Carried Forward.

MAR Missing At Random.

MCAR Missing Completely At Random.

MCAT Medical College Admissions Test.

MCMC Markov Chain Monte Carlo.

MCSE Monte Carlo Standard Error.

MI Multiple Imputation.

MICE Multiple Imputation Chained Equation, also known as FCS (Full Conditional Specification).

ML Maximum Likelihood.

MLE Maximum Likelihood Estimates.

MMIs Multiple Mini Interviews.

MNAR Missing Not At Random.

MRCP UK Membership of the Royal Colleges of Physicians of the United Kingdom.

MSE Mean Square Error.

MSN Multivariate Skew Normal.

MVN Multivariate Normal.

NNR Number Needed to Reject.

OSCE Objective Structured Clinical Examination.

PCR Peer Competition Rescaling.

PLAB Professional and Linguistic Assessments Board.

PMM Pattern Mixture Model.

PQA Personal Qualities Assessment.

REML Restricted Maximum Likelihood.

RMSE Root Mean Square Error.

SeM Selection Model.

SEM Structural Equation Modeling.

SJTs Situation Judgment Tests.

SPM Shared Parameter Model.

UCAS Undergraduate Courses At University and College.

UK United Kingdom.

UK FPO United Kingdom Foundation Programme.

UKCAT United Kingdom Clinical Aptitude Test.

UKCATSEN United Kingdom Clinical Aptitude Test for Special Educational Needs.

UKMED United Kingdom Medical Education Database.

UMAT Undergraduate Medical and Health Sciences Admission Test.

USA United States of America.

USMLE United States Medical Licensing Examination.

WGEE Weighted Generalized Estimating Equations.

Bibliography

Abbiati, M., Baroffio, A., and Gerbase, M. W. (2016). “Personal profile of medical students selected through a knowledge-based exam only: are we missing suitable students?” In: *Medical education online* 21.

Abe, K., Evans, P., Austin, E. J., Suzuki, Y., Fujisaki, K., Niwa, M., and Aomatsu, M. (2013). “Expressing one’s feelings and listening to others increases emotional intelligence: a pilot study of Asian medical students”. In: *BMC medical education* 13.1, p. 1.

Adam, J., Bore, M., Childs, R., Dunn, J., Mckendree, J., Munro, D., and Powis, D. (2015). “Predictors of professional behaviour and academic outcomes in a UK medical school: A longitudinal cohort study”. In: *Medical Teacher* 0, pp. 1–13.

Adam, J., Bore, M., McKendree, J., Munro, D., and Powis, D. (2012). “Can personal qualities of medical students predict in-course examination success and professional behaviour? An exploratory prospective cohort study”. In: *BMC medical education* 12.1, p. 69.

Adam, J., Dowell, J., and Greatrix, R. (2011). “Use of UKCAT scores in student selection by UK medical schools, 2006-2010”. In: *BMC medical education* 11.1, p. 98.

Ahmed, H., Rhydderch, M., and Matthews, P. (2012). “Can knowledge tests and situational judgement tests predict selection centre performance?” In: *Medical education* 46.8, pp. 777–784.

Al Alwan, I., Al Kushi, M., Tamim, H., Magzoub, M., and Elzubeir, M. (2013). “Health sciences and medical college preadmission criteria and prediction of in-course academic performance: a longitudinal cohort study”. In: *Advances in Health Sciences Education* 18.3, pp. 427–438. ISSN: 1573-1677. DOI: 10.1007/s10459-012-9380-1. URL: http://dx.doi.org/10.1007/s10459-012-9380-1.

Albanese, M. A., Farrell, P., and Dottl, S. (2005a). “Statistical criteria for setting thresholds in medical school admissions”. In: *Advances in health sciences education* 10.2, pp. 89–103.

Albanese, M. A., Farrell, P., and Dottl, S. L. (2005b). “A comparison of statistical criteria for setting optimally discriminating MCAT and GPA thresholds in medical school admissions”. In: *Teaching and learning in Medicine* 17.2, pp. 149–158.

Albers, W. and Kallenberg, W. C. (1994). “A simple approximation to the bivariate normal distribution with large correlation coefficient”. In: *Journal of multivariate analysis* 49.1, pp. 87–96.

Albishri, J. A., Aly, S. M., and Alnemary, Y. (2012). “Admission criteria to Saudi medical schools”. In: *Saudi Med J* 33.11, pp. 1222–1226.

Alexander, R. A., Alliger, G. M., and Hanges, P. J. (1984). “Correcting for range restriction when the population variance is unknown”. In: *Applied Psychological Measurement* 8.4, pp. 431–437.

Alexander, R. A., Hanges, P. J., and Alliger, G. M. (1985). “Correcting for restriction of range in both X and Y when the unrestricted variances are unknown”. In: *Applied psychological measurement* 9.3, pp. 317–323.

Alhadlaq, A. M., Alshammari, O. F., Alsager, S. M., Neel, K. A. F., and Mohamed, A. G. (2015). “Ability of Admissions Criteria to Predict Early Academic Performance Among Students of Health Science Colleges at King Saud University, Saudi Arabia”. In: *Journal of dental education* 79.6, pp. 665–670.

Allen, N. L. and Dunbar, S. B. (1989). “Standard errors of correlations adjusted for incidental selection”. In: *ETS Research Report Series* 1989.2.

— (1990). “Standard errors of correlations adjusted for incidental selection”. In: *Applied Psychological Measurement* 14.1, pp. 83–94.

Alliger, G. M. (1987). “An equation to simplify correction of range restricted standard deviations and correlations when the population variance is unknown”. In: *Educational and psychological measurement* 47.3, pp. 615–616.

Altman, D. G. (1998). “Confidence intervals for the number needed to treat”. In: *BMJ: British Medical Journal* 317.7168, p. 1309.

Altman, D. G. and Bland, J. M. (1994). “Statistics Notes: Diagnostic tests 2: predictive values”. In: *Bmj* 309.6947, p. 102.

Andridge, R. R. and Little, R. J. (2010). “A review of hot deck imputation for survey non-response”. In: *International statistical review* 78.1, pp. 40–64.

Azur, M. J., Stuart, E. A., Frangakis, C., and Leaf, P. J. (2011). “Multiple imputation by chained equations: what is it and how does it work?” In: *International journal of methods in psychiatric research* 20.1, pp. 40–49.

Azzalini, A. (2018). *The R package `sn`: The Skew-Normal and Related Distributions such as the Skew-t (version 1.5-2).* Università di Padova, Italia. URL: http://azzalini.stat.unipd.it/SN.

Azzalini, A. and Capitanio, A. (1999). “Statistical applications of the multivariate skew normal distribution”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61.3, pp. 579–602.

Banerjee, A., Chitnis, U., Jadhav, S., Bhawalkar, J., and Chaudhury, S. (2009). “Hypothesis testing, type I and type II errors”. In: *Industrial psychiatry journal* 18.2, p. 127.

Baraldi, A. N. and Enders, C. K. (2010). “An introduction to modern missing data analyses”. In: *Journal of school psychology* 48.1, pp. 5–37.

Baser, O., Crown, W. H., and Pollicino, C. (2006). “Guidelines for selecting among different types of bootstraps”. In: *Current medical research and opinion* 22.4, pp. 799–808.

Bedeian, A. G., Day, D. V., and Kelloway, E. K. (1997). “Correcting for measurement error attenuation in structural equation models: Some important reminders”. In: *Educational and Psychological Measurement* 57.5, pp. 785–799.

Behseta, S., Berdyyeva, T., Olson, C. R., and Kass, R. E. (Jan. 2009). “Bayesian Correction for Attenuation of Correlation in Multi-Trial Spike Count Data”. In: *Journal of Neurophysiology*, pp. 90727.2008+.

Benbassat, J. and Baumal, R. (2007). “Uncertainties in the selection of applicants for medical school”. English. In: *Advances in Health Sciences Education* 12.4, pp. 509–521. ISSN: 1382-4996.

Bengt O. Muthén, L. K. M. and Asparouhov, T. (2016, pp 443-445). *Regression and Mediation analysis using MPLUS*.

Beunckens, C., Molenberghs, G., and Kenward, M. G. (2005). "Direct likelihood analysis versus simple forms of imputation for missing data in randomized clinical trials". In: *Clinical Trials* 2.5, pp. 379–386.

Bland, J. M. and Altman, D. G. (2000). "The odds ratio". In: *Bmj* 320.7247, p. 1468.

Bracht, G. H. and Glass, G. V. (1968). "The external validity of experiments". In: *American educational research journal* 5.4, pp. 437–474.

Briggs, D. C. (2004). "Causal inference and the Heckman model". In: *Journal of Educational and Behavioral Statistics* 29.4, pp. 397–420.

British Medical Association (2017). Widening participation into medicine. https://www.bma.org.uk/advice/career/studying-medicine/becoming-a-doctor/widening-participation, [Accessed: September 27, 2017].

Buehler, J. A. and Trainer, J. B. (1962). "Prediction of medical school performance and its relationship to achievement." In: *Academic Medicine* 37.1, pp. 10–18.

Burt, C. (1943). "Validating tests for personnel selection". In: *British Journal of Psychology* 34.1, pp. 1–19.

Burton, A., Altman, D. G., Royston, P., and Holder, R. L. (2006). "The design of simulation studies in medical statistics". In: *Statistics in medicine* 25.24, pp. 4279–4292.

Buuren, S. and Groothuis-Oudshoorn, K. (2011). "mice: Multivariate imputation by chained equations in R". In: *Journal of statistical software* 45.3.

Byrne, A. T., Arnett, R., Farrell, T., and Sreenan, S. (2014). "Comparison of performance in a four year graduate entry medical programme and a traditional five/six year programme". In: *BMC medical education* 14.1, p. 1.

Camponovo, L. (2015). "On the validity of the pairs bootstrap for lasso estimators". In: *Biometrika* 102.4, pp. 981–987.

Carlisle, D. M., Gardner, J. E., and Liu, H. (1998). "The entry of underrepresented minority students into US medical schools: an evaluation of recent trends." In: *American Journal of Public Health* 88.9, pp. 1314–1318.

Carpenter, J. R., Goldstein, H., and Rasbash, J. (2003). "A novel bootstrap procedure for assessing the relationship between class size and achievement". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 52.4, pp. 431–443.

Carpenter, J. and Bithell, J. (2000). "Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians". In: *Statistics in Medicine*.

Carpenter, J., Goldstein, H., and Rasbash, J. (1999). "A non-parametric bootstrap for multilevel models". In: *Multilevel modelling newsletter* 11.1, pp. 2–5.

Carr, S. E. (2009). "Emotional intelligence in medical students: does it correlate with selection measures?" In: *Medical education* 43.11, pp. 1069–1077.

Casey, P. M., Palmer, B. A., Thompson, G. B., Laack, T. A., Thomas, M. R., Hartz, M. F., Jensen, J. R., Sandefur, B. J., Hammack, J. E., Swanson, J. W., et al. (2016). "Predictors of medical school clerkship performance: a multispecialty longitudinal analysis of standardized examination scores and clinical assessments". In: *BMC medical education* 16.1, p. 128.

Catellier, D. J., Hannan, P. J., Murray, D. M., Addy, C. L., Conway, T. L., Yang, S., and Rice, J. C. (2005). "Imputation of missing data when measuring physical activity by accelerometry". In: *Medicine and science in sports and exercise* 37.11 Suppl, S555.

Centre for Evidence-Based Medicine (2017). Number Needed to Treat (NNT). http://www.cebm.net/blog/2014/03/03/number-needed-to-treat-nnt/, [Accessed: December 17, 2017].

Centre for Reviews and Dissemination (CRD) (2008). Systematic reviews:CRD's guidance for undertaking reviews in health care. https://www.york.ac.uk/media/crd/Systematic_Reviews.pdf, [Accessed: September 25, 2017].

Cerutti, B., Bernheim, L., and Van Gessel, E. (2013). "The predictive validity of the aptitude test for the performance of students starting a medical curriculum". In: *Swiss medical weekly* 143, w13872.

Charles, E. P. (2005). "The Correction for Attenuation Due to Measurement Error: Clarifying Concepts and Creating Confidence Sets." In: *Psychological Methods* 10.2, pp. 206–226. ISSN: 1082-989X.

Chatellier, G., Zapletal, E., Lemaitre, D., Menard, J., and Degoulet, P. (1996). “The number needed to treat: a clinically useful nomogram in its proper context”. In: *Bmj* 312.7028, pp. 426–429.

Chen, B., Yi, G. Y., and Cook, R. J. (2010). “Weighted generalized estimating functions for longitudinal response and covariate data that are missing at random”. In: *Journal of the American Statistical Association* 105.489, pp. 336–353.

Cherry, M. G., Fletcher, I., O’Sullivan, H., and Dornan, T. (2014). “Emotional intelligence in medical education: a critical review”. In: *Medical education* 48.5, pp. 468–478.

Cherry, M. G. (2014). “The emotional side of selecting for medicine”. In: *Medical education* 48.12, pp. 1143–1145.

Chew, B. H., Zain, A. M., and Hassan, F. (2013). “Emotional intelligence and academic performance in first and final year medical students: a cross-sectional study”. In: *BMC medical education* 13.1, p. 1.

Clarivate Analytics (2017). Web of Science. https://login.webofknowledge.com/, [Accessed: September 22, 2017].

Cleland, J., Dowell, J., McLachlan, J., Nicholson, S., and Patterson, F. (2012). *Identifying best practice in the selection of medical students (literature review and interview survey).*

Coates, H. (2008). “Establishing the criterion validity of the graduate medical school admissions test (GAMSAT)”. In: *Medical education* 42.10, pp. 999–1006.

Cook, R. J. and Sackett, D. L. (1995). “The number needed to treat: a clinically useful measure of treatment effect.” In: *BMJ: British Medical Journal* 310.6977, p. 452.

Coombs, W. T., Algina, J., and Oltman, D. O. (1996). “Univariate and multivariate omnibus hypothesis tests selected to control Type I error rates when population variances are not necessarily equal”. In: *Review of Educational Research* 66.2, pp. 137–179.

Costa, P., Alves, R., Neto, I., Marvao, P., Portela, M., and Costa, M. J. (2014). “Associations between medical student empathy and personality: a multi-institutional study”. In: *PloS one* 9.3, e89254.

Cote, J. A. and Greenberg, R. (1990). “Specifying Measurement Error in Structural Equation Models: Are Congeneric Measurement Models Appropriate?” In: *ACR North American Advances*.

Creemers, A., Hens, N., Aerts, M., Molenberghs, G., Verbeke, G., and Kenward, M. G. (2010). “A Sensitivity Analysis for Shared-Parameter Models for Incomplete Longitudinal Outcomes”. In: *Biometrical Journal* 52.1, pp. 111–125.

Crowley, P. H. (1992). “Resampling methods for computation-intensive data analysis in ecology and evolution”. In: *Annual Review of Ecology and Systematics* 23.1, pp. 405–447.

Culpepper, S. A. (2015). “An improved correction for range restricted correlations under extreme, monotonic quadratic nonlinearity and heteroscedasticity”. In: *Psychometrika*, pp. 1–15.

Dahlin, M., Söderberg, S., Holm, U., Nilsson, I., and Farnebo, L.-O. (2012). “Comparison of communication skills between medical students admitted after interviews or on academic merits”. In: *BMC Medical Education* 12.1, pp. 1–6. ISSN: 1472-6920. DOI: 10.1186/1472-6920-12-46. URL: http://dx.doi.org/10.1186/1472-6920-12-46.

David Millett (2016). Medical school offering places through clearing inundated with 1,825 calls. https://www.gponline.com/medical-school-offering-places-clearing-inundated-1825-calls/article/1406161, [Accessed: December 19, 2017].

Davidson, R. C. and Lewis, E. L. (1997). “Affirmative action and other special consideration admissions at the University of California, Davis, School of Medicine”. In: *JAMA* 278.14, pp. 1153–1158.

Davidson, R. and Flachaire, E. (2008). “The wild bootstrap, tamed at last”. In: *Journal of Econometrics* 146.1, pp. 162–169.

De Corte, W., Lievens, F., and Sackett, P. R. (2007). “Combining predictors to achieve optimal trade-offs between selection quality and adverse impact.” In: *Journal of Applied Psychology* 92.5, p. 1380.

— (2008). “Validity and adverse impact potential of predictor composite formation”. In: *International Journal of Selection and Assessment* 16.3, pp. 183–194.

De Corte, W., Sackett, P. R., and Lievens, F. (2011). "Designing Pareto-optimal selection systems: formalizing the decisions required for selection system development." In: *Journal of Applied Psychology* 96.5, p. 907.

De Corte, W., Sackett, P., and Lievens, F. (2010). "Selecting predictor subsets: Considering validity and adverse impact". In: *International Journal of Selection and Assessment* 18.3, pp. 260–270.

Derrick, B., Russ, B., Toher, D., and White, P. (2017). "Test Statistics for the Comparison of Means for Two Samples That Include Both Paired and Independent Observations". In: *Journal of Modern Applied Statistical Methods* 16.1, p. 9.

Diab, P., Flack, P., Mabuza, L., and Moolman, H. (2015). "Curriculum challenges faced by rural-origin health science students at South African medical schools". In: *African Journal of Health Professions Education* 7.1, pp. 51–54.

DiCiccio, T. J. and Efron, B. (1996). "Bootstrap confidence intervals". In: *Statistical science*, pp. 189–212.

DiCiccio, T. J. and Romano, J. P. (1988). "A review of bootstrap confidence intervals". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 338–354.

Dixon, P. M. (2002). "Bootstrap resampling". In: *Encyclopedia of environmetrics*.

Donders, A. R. T., Heijden, G. J. van der, Stijnen, T., and Moons, K. G. (2006). "A gentle introduction to imputation of missing values". In: *Journal of clinical epidemiology* 59.10, pp. 1087–1091.

Dong, Y. and Peng, C.-Y. J. (2013). "Principled missing data methods for researchers". In: *SpringerPlus* 2.1, p. 222.

Donnon, T., Paolucci, E. O., and Violato, C. (2007). "The predictive validity of the MCAT for medical school performance and medical board licensing examinations: a meta-analysis of the published research". In: *Academic Medicine* 82.1, pp. 100–106.

Duan, B. and Dunlap, W. P. (1997). "The accuracy of different methods for estimating the standard error of correlations corrected for range restriction". In: *Educational and Psychological Measurement* 57.2, pp. 254–265.

Dunbar, S. B. and Linn, R. L. (1991). *Performance Assessment for the Workplace*. Vol. II. Chap. Range Restriction Adjustments in the Prediction of Military Job Performance, pp. 127–157.

Dunlap, J. W. and Cureton, E. E. (1930). "The Correlation Corrected for Attenuation in One Variable and Its Standard Error". English. In: *The American Journal of Psychology* 42.3, pp. 405-407. ISSN: 00029556.

Dunleavy, D. M., Kroopnick, M. H., Dowd, K. W., Searcy, C. A., and Zhao, X. (2013). "The predictive validity of the MCAT exam in relation to academic performance through medical school: a national cohort study of 2001–2004 matriculants". In: *Academic Medicine* 88.5, pp. 666–671.

Durrant, G. B. et al. (2005). "Imputation methods for handling item-nonresponse in the social sciences: a methodological review". In: *ESRC National Centre for Research Methods and Southampton Statistical Sciences Research Institute. NCRM Methods Review Papers NCRM/002*.

Dustin Fife (2016). Correcting Biased Estimates Under Selection. https://cran.r-project.org/web/packages/selection/index.html, [Accessed: November 1, 2017].

Dvison, A., Hinkley, D. V., and Schechtman, E. (1986). "Efficient bootstrap simulation". In: *Biometrika* 73.3, pp. 555–566.

Edwards, D., Friedman, T., and Pearce, J. (2013). "Same admissions tools, different outcomes: a critical perspective on predictive validity in three undergraduate medical schools". In: *BMC Medical Education* 13.1, pp. 1–7. ISSN: 1472-6920. DOI: 10.1186/1472-6920-13-173. URL: http://dx.doi.org/10.1186/1472-6920-13-173.

Edwards, J. C., Johnson, E. K., and Molidor, J. B. (1996). "Admission to medical school: International perspectives". In: *Advances in Health Sciences Education* 1.1, pp. 3–16.

Eekhout, I., Boer, R. M. de, Twisk, J. W., Vet, H. C. de, and Heymans, M. W. (2012). "Missing data: a systematic review of how they are reported and handled". In: *Epidemiology* 23.5, pp. 729–732.

Efron, B. (1981). "Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods". In: *Biometrika* 68.3, pp. 589–599.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. SIAM.

— (1992). “Bootstrap methods: another look at the jackknife”. In: *Breakthroughs in statistics*. Springer, pp. 569–593.

Efron, B., Tibshirani, R., et al. (1986). “Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy”. In: *Statistical science* 1.1, pp. 54–75. URL: http://www.jstor.org/stable/2245500.

Elam, C. L. and Johnson, M. (1992). “Prediction of medical students’ academic performances: does the admission interview help?.” In: *Academic Medicine* 67.10, S28–30.

Emery, J. L. and Bell, J. F. (2009). “The predictive validity of the BioMedical Admissions Test for pre-clinical examination performance”. In: *Medical Education* 43.6, pp. 557–564. ISSN: 1365-2923. DOI: 10.1111/j.1365-2923.2009.03367.x. URL: http://dx.doi.org/10.1111/j.1365-2923.2009.03367.x.

Enders, C. K. (2001a). “The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data.” In: *Psychological methods* 6.4, p. 352.

— (2001b). “A primer on maximum likelihood algorithms available for use with missing data”. In: *Structural Equation Modeling* 8.1, pp. 128–141.

— (2001c). “The performance of the full information maximum likelihood estimator in multiple regression models with missing data”. In: *Educational and Psychological Measurement* 61.5, pp. 713–740.

— (2010). *Applied missing data analysis*. Guilford Press.

— (2011). “Analyzing longitudinal data with missing values.” In: *Rehabilitation Psychology* 56.4, p. 267.

Enders, C. K. and Bandalos, D. L. (2001). “The relative performance of full information maximum likelihood estimation for missing data in structural equation models”. In: *Structural equation modeling* 8.3, pp. 430–457.

Eric W Weisstein (2017). MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/CorrelationCoefficientBivariateNormalDistribution.html, [Accessed: December 13, 2017].

Eskander, A., Shandling, M., and Hanson, M. D. (2013). “Should the MCAT exam be used for medical school admissions in Canada?” In: *Academic Medicine* 88.5, pp. 572–580.

Eva, K. W. and Macala, C. (2014). “Multiple mini-interview test characteristics:‘tis better to ask candidates to recall than to imagine”. In: *Medical education* 48.6, pp. 604–613.

Eva, K. W., Reiter, H. I., Rosenfeld, J., Trinh, K., Wood, T. J., and Norman, G. R. (2012). “Association between a medical school admission process using the multiple mini-interview and national licensing examination scores”. In: *Jama* 308.21, pp. 2233–2240.

Fan, A., Tsai, T. C., Su, T.-P., Kosik, R. O., Morisky, D. E., Chen, C.-H., Shih, W.-J., and Lee, C.-H. (2010). “A longitudinal study of the impact of interviews on medical school admissions in Taiwan”. In: *Evaluation & the health professions* 33.2, pp. 140–163.

Farrokhi-Khajeh-Pasha, Y., Nedjat, S., Mohammadi, A., Rad, E. M., Majdzadeh, R., Monajemi, F., Jamali, E., and Yazdani, S. (2012). “The validity of Iran’s national university entrance examination (Konkoor) for predicting medical students’ academic performance”. In: *BMC medical education* 12.1, p. 1.

Federation of State Medical Boards (2017). United States Medical Licensing Examinations. http://www.usmle.org/, [Accessed: May 3, 2017].

Feng, X., He, X., and Hu, J. (2011). “Wild bootstrap for quantile regression”. In: *Biometrika* 98.4, pp. 995–999.

Ferguson, E., James, D., and Madeley, L. (2002). “Factors associated with success in medical school: systematic review of the literature”. In: *Bmj* 324.7343, pp. 952–957.

Ferguson, E., Semper, H., Yates, J., Fitzgerald, J. E., Skatova, A., and James, D. (2014). “The ‘dark side’and ‘bright side’of personality: When too much conscientiousness and too little anxiety are detrimental with respect to the acquisition of medical knowledge and skill”. In: *PloS one* 9.2, e88606.

Fife, D. A., Hunter, M. D., and Mendoza, J. L. (2016). “Estimating Unattenuated Correlations With Limited Information About Selection Variables Alternatives to Case IV”. In: *Organizational Research Methods*, p. 1094428115625323.

Fife, D. A., Mendoza, J. L., and Terry, R. (2013). "Revisiting Case IV: A reassessment of bias and standard errors of Case IV under range restriction". In: *British Journal of Mathematical and Statistical Psychology* 66.3, pp. 521–542.

Fisher, C. R. (2014). "A Pedagogic Demonstration of Attenuation of Correlation Due to Measurement Error". In: *Spreadsheets in Education (eJSiE)* 7.1, p. 4.

Flachaire, E. (2005). "Bootstrapping heteroskedastic regression models: wild bootstrap vs. pairs bootstrap". In: *Computational Statistics & Data Analysis* 49.2, pp. 361–376.

Fosdick, B. K. and Raftery, A. E. (2012). "Estimating the correlation in bivariate normal data with known variances and small sample sizes". In: *The American Statistician* 66.1, pp. 34–41.

Foundation Programme (2017). The UK Foundation Programme. http://www.foundationprogramme.nhs.uk/pages/home, [Accessed: May 3, 2017].

Fukui, Y., Noda, S., Okada, M., Mihara, N., Kawakami, Y., Bore, M., Munro, D., and Powis, D. (2014). "Trial use of the Personal Qualities Assessment (PQA) in the entrance examination of a Japanese medical university: similarities to the results in western countries". In: *Teaching and learning in medicine* 26.4, pp. 357–363.

Gad, A. M. and Darwish, N. M. (2013). "A shared parameter model for longitudinal data with missing values". In: *American journal of applied Mathematics and Statistics* 1.2, pp. 30–35.

Gao, S. (2004). "A shared random effect parameter approach for longitudinal dementia data with non-ignorable missing data". In: *Statistics in Medicine* 23.2, pp. 211–219.

Garces, L. M. and Mickey-Pabello, D. (2015). "Racial diversity in the medical profession: The impact of affirmative action bans on underrepresented student of color matriculation in medical schools". In: *The Journal of higher education* 86.2, pp. 264–294.

General Medical Council (1973). "Conference on Methods of Examination and Assessment". In: p. 5.

— (2015). Professional and Linguistic Assessments Board (PLAB) test. https://www.bma.org.uk/advice/career/studying-medicine/becoming-a-doctor/widening-participation, [Accessed: March 23, 2017].

General Medical Council (2017a). A widening participation programme helps local applicants enter further education in medicine. http://www.gmc-uk.org/education/28201.asp, [Accessed: September 27, 2017].

— (2017b). Medical Licensing Assessment (MLA). http://www.gmc-uk.org/education/28201.asp, [Accessed: December 19, 2017].

Genest, C., Rémillard, B., et al. (2008). "Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models". In: *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*. Vol. 44. 6. Institut Henri Poincaré, pp. 1096–1127.

Gleason, J. R. (1988). "Algorithms for balanced bootstrap simulations". In: *The American Statistician* 42.4, pp. 263–266.

Godwin, M., Ruhland, L., Casson, I., MacDonald, S., Delva, D., Birtwhistle, R., Lam, M., and Seguin, R. (2003). "Pragmatic controlled clinical trials in primary care: the struggle between external and internal validity". In: *BMC medical research methodology* 3.1, p. 28.

Good, P. (2013). *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media.

Google Scholar (2017). Google Scholar. https://scholar.google.co.uk/, [Accessed: September 22, 2017].

Graham, J. W. (2003). "Adding missing-data-relevant variables to FIML-based structural equation models". In: *Structural Equation Modeling* 10.1, pp. 80–100.

Green, B. N., Johnson, C. D., and McCarthy, K. (2003). "Predicting academic success in the first year of chiropractic college". In: *Journal of manipulative and physiological therapeutics* 26.1, pp. 40–46.

Greener, J. M. and Osburn, H. (1979). "An empirical study of the accuracy of corrections for restriction in range due to explicit selection". In: *Applied Psychological Measurement* 3.1, pp. 31–41.

— (1980). "Accuracy of corrections for restriction in range due to explicit selection in heteroscedastic and nonlinear distributions". In: *Educational and Psychological Measurement* 40.2, pp. 337–346.

Greenland, S. and Finkle, W. D. (1995). “A critical look at methods for handling missing covariates in epidemiologic regression analyses”. In: *American journal of epidemiology* 142.12, pp. 1255–1264.

Griffin, B., Yeomans, N., and Wilson, I. (2013). “Students coached for an admission test perform less well throughout a medical course”. In: *Internal medicine journal* 43.8, pp. 927–932.

Griffin, B. and Hu, W. (2015). “The interaction of socio-economic status and gender in widening participation in medicine”. In: *Medical education* 49.1, pp. 103–113.

Gross, A. L. (1990). “A maximum likelihood approach to test validation with missing and censored dependent variables”. In: *Psychometrika* 55.3, pp. 533–549.

Gross, A. L. and Fleischman, L. (1983). “Restriction of range corrections when both distribution and selection assumptions are violated”. In: *Applied Psychological Measurement* 7.2, pp. 227–237.

Gross, A. L. and Fleischman, L. E. (1987). “The correction for restriction of range and nonlinear regressions: An analytic study”. In: *Applied psychological measurement* 11.2, pp. 211–217.

Groves, M. A., Gordon, J., and Ryan, G. (2007). “Entry tests for graduate medical programs: is it time to re-think?” In: *Medical Journal of Australia* 186.3, p. 120.

Gupta, N., Nagpal, G., and Dhaliwal, U. (2013). “Student performance during the medical course: role of pre-admission eligibility and selection criteria.” In:

Gutowski, C. J., Thaker, N. G., Heinrich, G., and Fadem, B. (2010). “Current medical student interviewers add data to the evaluation of medical school applicants”. In: *Medical education online* 15.

Habersack, M., Dimai, H. P., Ithaler, D., and Reibnegger, G. (2015). “Time: an underestimated variable in minimizing the gender gap in medical college admission scores”. In: *Wiener klinische Wochenschrift* 127.7-8, pp. 241–249.

Hakstian, A. R., Schroeder, M. L., and Rogers, W. T. (1988). “Inferential procedures for correlation coefficients corrected for attenuation”. In: *Psychometrika* 53.1, pp. 27–43.

Hakstian, A. R., Schroeder, M. L., and Rogers, W. T. (1989). “Inferential theory for partially disattenuated correlation coefficients”. In: *Psychometrika* 54.3, pp. 397–407.

Harding, B., Tremblay, C., and Cousineau, D. (2014). “Standard errors: A review and evaluation of standard error estimators using Monte Carlo simulations”. In: *Quant Methods Psychol* 10, pp. 107–123.

Harris, B. H., Walsh, J. L., and Lammy, S. (2015). “UK medical selection: lottery or meritocracy?” In: *Clinical Medicine* 15.1, pp. 40–46.

Haukoos, J. S. and Newgard, C. D. (2007). “Advanced statistics: missing data in clinical research—part 1: an introduction and conceptual framework”. In: *Academic Emergency Medicine* 14.7, pp. 662–668.

Heckman, J. J. (Mar. 1977). *Sample Selection Bias As a Specification Error (with an Application to the Estimation of Labor Supply Functions)*. Working Paper 172. National Bureau of Economic Research. DOI: 10.3386/w0172. URL: http://www.nber.org/papers/w0172.

Hedeker, D. and Gibbons, R. D. (1997). “Application of random-effects pattern-mixture models for missing data in longitudinal studies.” In: *Psychological methods* 2.1, p. 64.

Heijden, G. J. van der, Donders, A. R. T., Stijnen, T., and Moons, K. G. (2006). “Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example”. In: *Journal of clinical epidemiology* 59.10, pp. 1102–1109.

Held, J. D. and Foley, P. P. (1994). “Explanations for accuracy of the general multivariate formulas in correcting for range restriction”. In: *Applied Psychological Measurement* 18.4, pp. 355–367.

Hesterberg, T., Moore, D. S., Monaghan, S., Clipson, A., and Epstein, R. (2005). “Bootstrap methods and permutation tests”. In: *Introduction to the Practice of Statistics* 5, pp. 1–70.

Hissbach, J. C., Klusmann, D., and Hampe, W. (2011). “Dimensionality and predictive validity of the HAM-Nat, a test of natural sciences for medical school admission”. In: *BMC medical education* 11.1, p. 1.

Holmes, D. (1990). “The robustness of the usual correction for restriction in range due to explicit selection”. In: *Psychometrika* 55.1, pp. 19–32.

Honaker, J., King, G., Blackwell, M., et al. (2011). “Amelia II: A program for missing data”. In: *Journal of statistical software* 45.7, pp. 1–47.

Houston, M., Osborne, M., and Rimmer, R. (2015). “Private schooling and admission to medicine: a case study using matched samples and causal mediation analysis”. In: *BMC medical education* 15.1, p. 1.

Hox, J. J. and Bechger, T. M. (2007). “An introduction to structural equation modeling”. In:

Huffcutt, A. I., Culbertson, S. S., and Weyhrauch, W. S. (2014). “Moving Forward Indirectly: Reanalyzing the validity of employment interviews with indirect range restriction methodology”. In: *International Journal of Selection and Assessment* 22.3, pp. 297–309.

Huitema, B. E. and Stein, C. R. (1993). “Validity of the GRE without restriction of range”. In: *Psychological Reports* 72.1, pp. 123–127.

Humphrey-Murto, S., Leddy, J. J., Wood, T. J., Puddester, D., and Moineau, G. (2014). “Does emotional intelligence at medical school admission predict future academic performance?” In: *Academic Medicine* 89.4, pp. 638–643.

Hunter, J. E., Schmidt, F. L., and Le, H. (2006). “Implications of Direct and Indirect Range Restriction for Meta-Analysis Methods and Findings”. In: *Journal of Applied Psychology* 91.3, pp. 594–612.

Husbands, A. and Dowell, J. (2013). “Predictive validity of the Dundee multiple mini-interview”. In: *Medical education* 47.7, pp. 717–725.

Husbands, A., Mathieson, A., Dowell, J., Cleland, J., and MacKenzie, R. (2014). “Predictive validity of the UK clinical aptitude test in the final years of medical school: a prospective cohort study”. In: *BMC Medical Education* 14.1, p. 88. ISSN: 1472-6920.

Husbands, A., Rodgerson, M. J., Dowell, J., and Patterson, F. (2015). “Evaluating the validity of an integrity-based situational judgement test for medical school admissions”. In: *BMC medical education* 15.1, p. 1.

Ibrahim, J. G. and Molenberghs, G. (2009). “Missing data methods in longitudinal studies: a review”. In: *Test* 18.1, pp. 1–43.

Imperial Centre for Endocrinology (2013). Educational Performance Measure (EPM) Framework. http://www.imperialendo.co.uk/sjt/EPM_framework.pdf, [Accessed: December 15, 2017].

InfluentialPoints (2017). Pearson's correlation coefficient. https://cran.r-project.org/web/packages/selection/index.html, [Accessed: October 5, 2017].

International English Language Testing System (2017). The IELTS scale. https://www.ielts.org/about-the-test/how-ielts-is-scored, [Accessed: March 23, 2017].

Iris Eekhout (2016). Missing data. http://www.iriseekhout.com/missing-data/missing-data-methods/imputation-methods/, [Accessed: July 30, 2016].

James, D., Yates, J., and Nicholson, S. (2010). "Comparison of A level and UKCAT performance in students applying to UK medical and dental schools in 2006: cohort study". In: *BMJ* 340. ISSN: 0959-8138.

Jason W Osborne (2015). Effect Sizes and the Disattenuation of Correlation and Regression Coefficients: Lessons from Educational Psychology. http://pareonline.net/getvn.asp?v=8&n=11, [Accessed: March 12, 2015].

Jerant, A., Fancher, T., Fenton, J. J., Fiscella, K., Sousa, F., Franks, P., and Henderson, M. (2015). "How Medical School Applicant Race, Ethnicity, and Socioeconomic Status Relate to Multiple Mini-Interview–Based Admissions Outcomes: Findings From One Medical School". In: *Academic Medicine* 90.12, pp. 1667–1674.

Johnson, H. G. (1944). "An empirical study of the influence of errors of measurement upon correlation". In: *The American Journal of Psychology* 57.4, pp. 521–536.

Julian, E. R. (2005). "Validity of the Medical College Admission Test for predicting medical school performance". In: *Academic Medicine* 80.10, pp. 910–917.

Kadengye, D. T., Ceulemans, E., and Van den Noortgate, W. (2014). "Direct likelihood analysis and multiple imputation for missing item scores in multilevel cross-classification educational data". In: *Applied Psychological Measurement* 38.1, pp. 61–80.

Kadengye, D. T., Cools, W., Ceulemans, E., and Van den Noortgate, W. (2012). "Simple imputation methods versus direct likelihood analysis for missing item scores in multilevel educational data". In: *Behavior research methods* 44.2, pp. 516–531.

Kadmon, G. and Kadmon, M. (2016). “Academic Performance of Students with the Highest and Mediocre School-leaving Grades: Does the Aptitude Test for Medical Studies (TMS) Balance Their Prognoses?” In: *GMS journal for medical education* 33.1.

Kankainen, A., Taskinen, S., and Oja, H. (2004). “On Mardia’s tests of multinormality”. In: *Theory and Applications of Recent Robust Methods*. Springer, pp. 153–164.

Katherine Sellgren (2016). Most students’ predicted A-levels ’wrong’. http://www.bbc.co.uk/news/education-38223432, [Accessed: December 19, 2017].

Kelly, M. E., Dowell, J., Husbands, A., Newell, J., Siun, O., Kropmans, T., Dunne, F. P., and Murphy, A. W. (2014). “The fairness, predictive validity and acceptability of multiple mini interview in an internationally diverse student population-a mixed methods study”. In: *BMC medical education* 14.1, p. 1.

Kelly, M. E. and O’Flynn, S. (2017). “The construct validity of HPAT-Ireland for the selection of medical students: Unresolved issues and future research implications”. In: *Advances in Health Sciences Education* 22.2, pp. 267–286.

Kelly, M. E., Regan, D., Dunne, F., Henn, P., Newell, J., and O’Flynn, S. (2013). “To what extent does the Health Professions Admission Test-Ireland predict performance in early undergraduate tests of communication and clinical skills?–An observational cohort study”. In: *BMC medical education* 13.1, p. 1.

Kennet-Cohen, T. and Bronner, S. (1998). *The Predictive Validity of the Components of the Process of Selection for Higher Education in Israel: A Correction for Sample-Selection Bias Using Heckman’s Method*. National Institute for Testing & Evaluation.

Kenward, M. G. and Molenberghs, G. (2009). “Last observation carried forward: a crystal ball?” In: *Journal of biopharmaceutical statistics* 19.5, pp. 872–888.

Khan, J. S., Mukhtar, O., and Tabasum, S. (2014). “Predictive validity of medical and dental colleges’ entrance test in Punjab: The way forward”. In: *JPMA. The Journal of the Pakistan Medical Association* 64.10, pp. 1132–1137.

Khan, J. S., Tabasum, S., and Mukhtar, O. (2013). “Comparison of pre-medical academic achievement, entrance test and aptitude test scores in admission selection process”. In: *J Pak Med Assoc* 63, pp. 552–557.

Kim Catcheside (2012). University admissions: increased AAB offers will lead to crisis at clearing. https://www.theguardian.com/higher-education-network/blog/2012/jul/23/university-clearing-crisis-aab-offers, [Accessed: December 19, 2017].

Kim, J. K. and Fuller, W. (2004). "Fractional hot deck imputation". In: *Biometrika* 91.3, pp. 559–578.

Kim, T., Chang, J.-Y., Myung, S. J., Chang, Y., Park, K. D., Park, W. B., and Shin, C. S. (2016). "Predictors of Undergraduate and Postgraduate Clinical Performance: A Longitudinal Cohort Study". In: *Journal of surgical education* 73.4, pp. 715–720.

King's College London (2017). Extended Medical Degree Programme (MBBS). https://www.kcl.ac.uk/study/undergraduate/courses/extended-medical-degree-programme-mbbs.aspx, [Accessed: September 27, 2017].

Knorr, M. and Hissbach, J. (2014). "Multiple mini-interviews: same concept, different approaches". In: *Medical education* 48.12, pp. 1157–1175.

Koczwara, A., Patterson, F., Zibarras, L., Kerrin, M., Irish, B., and Wilkinson, M. (2012). "Evaluating cognitive ability, knowledge tests and situational judgement tests for postgraduate selection". In: *Medical Education* 46.4, pp. 399–408.

Koehler, E., Brown, E., and Haneuse, S. J.-P. (2009). "On the assessment of Monte Carlo error in simulation-based statistical analyses". In: *The American Statistician* 63.2, pp. 155–162.

Korkmaz, S., Goksuluk, D., and Zararsiz, G. (2014). "MVN: An R Package for Assessing Multivariate Normality." In: *The R Journal* 6.2, pp. 151–162. URL: https://journal.r-project.org/archive/2014-2/korkmaz-goksuluk-zararsiz.pdf.

Kres, H. (1983). "The Mardia-Test for Multivariate Normality, Skewness, and Kurtosis: Tables by KV Mardia". In: *Statistical Tables for Multivariate Analysis*. Springer, pp. 420–431.

Kuncel, N. R., Kochevar, R. J., and Ones, D. S. (2014). "A meta-analysis of letters of recommendation in college and graduate admissions: Reasons for hope". In: *International Journal of Selection and Assessment* 22.1, pp. 101–107.

Lachin, J. M. (2016). "Fallacies of last observation carried forward analyses". In: *Clinical Trials* 13.2, pp. 161–168.

Laerd Statistics (2013a). Pearson's Correlation using Stata. https://statistics.laerd.com/stata-tutorials/pearsons-correlation-using-stata.php, [Accessed: October 5, 2017].

— (2013b). Pearson's Product-Moment Correlation using SPSS Statistics. https://statistics.laerd.com/spss-tutorials/pearsons-product-moment-correlation-using-spss-statistics.php, [Accessed: October 5, 2017].

Lahiri, S. N. (1999). "Theoretical comparisons of block bootstrap methods". In: *Annals of Statistics*, pp. 386–404.

Lahiri, S. N. (1993). "On the moving block bootstrap under long range dependence". In: *Statistics & Probability Letters* 18.5, pp. 405–413.

Lakhan, S. E. (2003). "Diversification of US medical schools via affirmative action implementation". In: *BMC medical education* 3.1, p. 6.

Lala, R., Wood, D., and Baker, S. (2013). "Validity of the UKCAT in Applicant Selection and Predicting Exam Performance in UK Dental Students". In: *Journal of dental education* 77.9, pp. 1159–1170.

Lanciano, T. and Curci, A. (2014). "Incremental validity of emotional intelligence ability in predicting academic achievement". In: *The American journal of psychology* 127.4, pp. 447–461.

Laurence, C. O., Zajac, I. T., Lorimer, M., Turnbull, D. A., and Sumner, K. E. (2013). "The impact of preparatory activities on medical school selection outcomes: a cross-sectional survey of applicants to the university of Adelaide medical school in 2007". In: *BMC medical education* 13.1, p. 1.

Le, H. and Schmidt, F. L. (2006). "Correcting for indirect range restriction in meta-analysis: testing a new meta-analytic procedure." In: *Psychological methods* 11.4, p. 416.

Leddy, J. J., Moineau, G., Puddester, D., Wood, T. J., and Humphrey-Murto, S. (2011). "Does an emotional intelligence test correlate with traditional measures used to determine medical school admission?" In: *Academic Medicine* 86.10, S39–S41.

Les Irwig Judy Irwig, L. T. and Sweet, M. (2008). *Smart health choices: making sense of health advice*. Hammersmith Press, London. Chap. 18: Relative risk, relative and absolute risk reduction, number needed to treat and confidence intervals.

Lesaffre, E. and Lawson, A. B. (2012). *Bayesian Biostatistics*. First Edition. John Wiley and Sons Ltd.

Li, J. C.-H. (2015). "Cohen's d Corrected for Case IV Range Restriction: A More Accurate Procedure for Evaluating Subgroup Differences in Organizational Research". In: *Personnel Psychology* 68.4, pp. 899–927.

Li, J. C.-h., Chan, W., and Cui, Y. (2011). "Bootstrap standard error and confidence intervals for the correlations corrected for indirect range restriction". In: *British Journal of Mathematical and Statistical Psychology* 64.3, pp. 367–387.

Lievens, F., Patterson, F., Corstjens, J., Martin, S., and Nicholson, S. (2016). "Widening access in selection using situational judgement tests: evidence from the UKCAT". In: *Medical education* 50.6, pp. 624–636.

Lin, H., McCulloch, C. E., and Rosenheck, R. A. (2004). "Latent pattern mixture models for informative intermittent missing data in longitudinal studies". In: *Biometrics* 60.2, pp. 295–305.

Lin, T. H. (2010). "A comparison of multiple imputation with EM algorithm and MCMC method for quality of life missing data". In: *Quality & quantity* 44.2, pp. 277–287.

Lindé, J. (2005). "Estimating New-Keynesian Phillips curves: A full information maximum likelihood approach". In: *Journal of Monetary Economics* 52.6, pp. 1135–1149.

Little, R. J. (1993). "Pattern-mixture models for multivariate incomplete data". In: *Journal of the American Statistical Association* 88.421, pp. 125–134.

— (1994). "A class of pattern-mixture models for normal incomplete data". In: *Biometrika* 81.3, pp. 471–483.

Little, R. J. and Wang, Y. (1996). "Pattern-mixture models for multivariate incomplete data with covariates". In: *Biometrics*, pp. 98–111.

Little, T. D., Jorgensen, T. D., Lang, K. M., and Moore, E. W. G. (2013). "On the joys of missing data". In: *Journal of pediatric psychology* 39.2, pp. 151–162.

Lomax, R. G. (1986). "The effect of measurement error in structural equation modeling". In: *The Journal of Experimental Education* 54.3, pp. 157–162.

Lovie, S. and Lovie, P. (2010). “Commentary: Charles Spearman and correlation: a commentary on ‘The proof and measurement of association between two things’”. In: *International journal of epidemiology* 39.5, pp. 1151–1153.

Lucid Software Inc (2018). Lucidchart. https://www.lucidchart.com/, [Accessed: June 29, 2018].

Lucieer, S. M., Stegers-Jager, K. M., Rikers, R. M., and Themmen, A. P. (2016). “Non-cognitive selected students do not outperform lottery-admitted students in the pre-clinical stage of medical school”. In: *Advances in Health Sciences Education* 21.1, pp. 51–61.

Lucinda Borrell (2017). Finding a place on a Medicine Clearing course for 2017. https://www.whatuni.com/advice/clearing/finding-a-place-on-a-medicine-clearing-course-for-2017/10838/, [Accessed: December 19, 2017].

Lynch Jr, J. G. (1982). “On the external validity of experiments in consumer research”. In: *Journal of consumer Research* 9.3, pp. 225–239.

Lynch, B., MacKenzie, R., Dowell, J., Cleland, J., and Prescott, G. (2009). “Does the UKCAT predict Year 1 performance in medical school”. In: *Medical education* 43.12, pp. 1203–1209.

MacKenzie, R., Cleland, J., Ayansina, D., and Nicholson, S. (2016). “Does the UKCAT predict performance on exit from medical school? A national cohort study”. In: *BMJ open* 6.10, e011313.

MacKenzie, R., Dowell, J., Ayansina, D., and Cleland, J. (2017). “Do personality traits assessed on medical school admission predict exit performance? A UK-wide longitudinal cohort study”. In: *Advances in Health Sciences Education* 22.2, pp. 365–385.

MacKinnon, J. G. (2006). “Bootstrap methods in econometrics”. In: *Economic Record* 82.s1, S2–S18.

Mahadevan, S. (1997). *Monte carlo simulation*, pp. 123–146.

Mammen, E. (1993). “Bootstrap and wild bootstrap for high dimensional linear models”. In: *The annals of statistics*, pp. 255–285.

Mankus, A. M., Boden, M. T., and Thompson, R. J. (2016). "Sources of variation in emotional awareness: Age, gender, and socioeconomic status". In: *Personality and individual differences* 89, pp. 28–33.

Manly, B. F. (2006). *Randomization, bootstrap and Monte Carlo methods in biology*. Vol. 70. CRC Press.

Matthew Pinchard (2015). With A-level Results Day almost upon us for another year, is clearing for medical school entry possible? https://www.gapmedics.com/uk/blog/2015/08/11/can-you-get-a-place-at-medical-school-through-clearing, [Accessed: December 19, 2017].

McGaghie, W. (2002). "Assessing readiness for medical education: Evolution of the medical college admission test". In: *JAMA* 288.9, pp. 1085–1090.

McManus, I., Dewberry, C., Nicholson, S., and Dowell, J. (2013). "The UKCAT-12 study: educational attainment, aptitude test performance, demographic and socio-economic contextual factors as predictors of first year outcome in a cross-sectional collaborative study of 12 UK medical schools". In: *BMC Medicine* 11.1, p. 244. ISSN: 1741-7015.

McManus, I., Dewberry, C., Nicholson, S., Dowell, J., Woolf, K., and Potts, H. (2013). "Construct-level predictive validity of educational attainment and intellectual aptitude tests in medical student selection: meta-regression of six UK longitudinal studies". In: *BMC Medicine* 11.1, p. 243. ISSN: 1741-7015.

McManus, I., Powis, D. A., Wakeford, R., Ferguson, E., James, D., and Richards, P. (2005). "Intellectual aptitude tests and A levels for selecting UK school leaver entrants for medical school". In: *BMJ* 331.7516, pp. 555–559. ISSN: 0959-8138.

McManus, I., Woolf, K., Dacre, J., Paice, E., and Dewberry, C. (2013). "The Academic Backbone: longitudinal continuities in educational achievement from secondary school and medical school to MRCP (UK) and the specialist register in UK medical students and doctors". In: *BMC medicine* 11.1, p. 1.

McNulty, J., Mackay, S., Lewis, S., Lane, S., and White, P. (2016). "An international study of emotional intelligence in first year radiography students: The relationship to age, gender and culture". In: *Radiography* 22.2, pp. 171–176.

Medic Portal (2017). Widening Participation. https://www.themedicportal.com/about-the-medic-portal/widening-participation/, [Accessed: September 27, 2017].

Medical School Council (2014). Selecting for Excellence Final Report. https://www.medschools.ac.uk/media/1203/selecting-for-excellence-final-report.pdf, [Accessed: September 22, 2017].

— (2017a). Educational Performance Measure. http://www.isfp.org.uk/PRE/Pages/Educational-Performance-Measure.aspx, [Accessed: May 3, 2017].

— (2017b). UK Medical Education Database. https://www.medschools.ac.uk/our-work/research/ukmed, [Accessed: December 13, 2017].

Mendoza, J. L. (1993). "Fisher transformations for correlations corrected for selection and missing data". In: *Psychometrika* 58.4, pp. 601–615.

Mendoza, J. L. and Mumford, M. (1987a). "Corrections for Attenuation and Range Restriction on the Predictor". English. In: *Journal of Educational Statistics* 12.3, pp. 282-293. ISSN: 03629791.

Mendoza, J. L., Bard, D. E., Mumford, M. D., and Ang, S. C. (2004). "Criterion-related validity in multiple-hurdle designs: Estimation and bias". In: *Organizational Research Methods* 7.4, pp. 418–441.

Mendoza, J. L. and Mumford, M. (1987b). "Corrections for attenuation and range restriction on the predictor". In: *Journal of Educational and Behavioral Statistics* 12.3, pp. 282–293.

Mercer, A., Crotty, B., Alldridge, L., Le, L., and Vele, V. (2015). "GAMSAT: A 10-year retrospective overview, with detailed analysis of candidates' performance in 2014". In: *BMC medical education* 15.1, p. 1.

Mercer, A. and Puddey, I. B. (2011). "Admission selection criteria as predictors of outcomes in an undergraduate medical course: A prospective study". In: *Medical teacher* 33.12, pp. 997–1004.

Miller, R. G. (1974). "The jackknife-a review". In: *Biometrika* 61.1, pp. 1–15.

Molenberghs, G. and Kenward, M. (2007). *Missing data in clinical studies*. Vol. 61. John Wiley & Sons. Chap. 1,7,9,14,33, pp. 5, 79, 107, 111, 175–182.

Molenberghs, G., Michiels, B., Kenward, M. G., and Diggle, P. J. (1998). “Monotone missing data and pattern-mixture models”. In: *Statistica Neerlandica* 52.2, pp. 153–161.

Molenberghs, G., Thijs, H., Jansen, I., Beunckens, C., Kenward, M. G., Mallinckrodt, C., and Carroll, R. J. (2004). “Analyzing incomplete longitudinal clinical trial data”. In: *Biostatistics* 5.3, pp. 445–464.

Molenberghs, G. and Verbeke, G. (2006). *Models for discrete longitudinal data*. Springer Science and Business Media. Chap. 9,8,13,16, pp. 200, 158, 160, 257, 298.

Molnar, F. J., Hutton, B., and Fergusson, D. (2008). “Does analysis using “last observation carried forward” introduce bias in dementia research?” In: *Canadian Medical Association Journal* 179.8, pp. 751–753.

Mooney, C. Z. (1997). *Monte carlo simulation*. Vol. 116. Sage Publications.

Moser, B. K. and Stevens, G. R. (1992). “Homogeneity of variance in the two-sample means test”. In: *The American Statistician* 46.1, pp. 19–21.

Muchinsky, P. M. (1996). “The correction for attenuation”. In: *Educational and psychological measurement* 56.1, pp. 63–75.

Mushtaq, J. and Ratneswaran, C. (2016). “Personality selection: An argument against the homogenisation of medical students”. In: *Medical teacher* 38.3, pp. 318–319.

Muthén, L. K. and Muthén, B. O. (2002). “How to use a Monte Carlo study to decide on sample size and determine power”. In: *Structural equation modeling* 9.4, pp. 599–620.

Muthu, V. (2003). “The number needed to treat: problemsdescribing non-significant results”. In: *Evidence-based mental health* 6.3, pp. 72–72.

Mwandigha, L. (2014). “Human Biomonitoring Project: Structural Equation Models”. Master Thesis. Hasselt University.

Mwandigha, L. M., Tiffin, P. A., Paton, L. W., Kasim, A. S., and Böhnke, J. R. (2018). “What is the effect of secondary (high) schooling on subsequent medical school performance? A national, UK-based, cohort study”. In: *BMJ open* 8.5, p. 10.

Myers, T. A. (2011). “Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data”. In: *Communication Methods and Measures* 5.4, pp. 297–310.

Naeem, N., Vleuten, C. van der, Muijtjens, A. M., Violato, C., Ali, S. M., Al-Faris, E. A., Hoogenboom, R., and Naeem, N. (2014). "Correlates of emotional intelligence: Results from a multi-institutional study among undergraduate medical students". In: *Medical teacher* 36.sup1, S30–S35.

National Institute of Standards and Technology (2003). e-Handbook of Statistical Methods: Maximum likelihood estimation. http://www.itl.nist.gov/div898/handbook/apr/section4/apr412.htm, [Accessed: September 9, 2017].

Naylor, S., Norris, M., and Williams, A. (2014). "Does ethnicity, gender or age of physiotherapy students affect performance in the final clinical placements? An exploratory study". In: *Physiotherapy* 100.1, pp. 9–13.

Nedjat, S., Bore, M., Majdzadeh, R., Rashidian, A., Munro, D., Powis, D., Karbakhsh, M., and Keshavarz, H. (2013). "Comparing the cognitive, personality and moral characteristics of high school and graduate medical entrants to the Tehran University of Medical Sciences in Iran". In: *Medical teacher* 35.12, e1632–e1637.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996). *Applied linear statistical models*. Vol. 4. Irwin Chicago. Chap. 4, pp. 165–168.

Newgard, C. D. and Haukoos, J. S. (2007). "Advanced statistics: missing data in clinical research—part 2: multiple imputation". In: *Academic Emergency Medicine* 14.7, pp. 669–678.

Nordman, D. J., Lahiri, S. N., and Fridley, B. L. (2007). "Optimal block size for variance estimation by a spatial block bootstrap method". In: *Sankhyā: The Indian Journal of Statistics*, pp. 468–493.

NOVA South Eastern University (2017). Criterion-related Validity. http://www.cps.nova.edu/~cpphelp/class/psy0507/critv.html, [Accessed: November 10, 2017].

Nuala Burgess (2017). Grammar schools: does selection help social mobility? https://blog.esrc.ac.uk/2017/04/19/grammar-schools-does-selection-help-mobility/, [Accessed: December 19, 2017].

O'Flynn, S., Fitzgerald, T., and Mills, A. (2013). "Modelling the impact of old and new mechanisms of entry and selection to medical school in Ireland: who gets in?" In: *Irish journal of medical science* 182.3, pp. 421–427.

O'Hare, L. and McGuinness, C. (2015). "The validity of critical thinking tests for predicting degree performance: A longitudinal study". In: *International Journal of Educational Research* 72, pp. 162–172.

Ogenler, O. and Selvi, H. (2014). "Variables Affecting Medical Faculty Students' Achievement: A Mersin University Sample". In: *Iranian Red Crescent Medical Journal* 16.3.

Oliver, T., Hecker, K., Hausdorf, P. A., and Conlon, P. (2014). "Validating MMI scores: are we measuring multiple attributes?" In: *Advances in Health Sciences Education* 19.3, pp. 379–392. ISSN: 1573-1677. DOI: 10.1007/s10459-013-9480-6. URL: http://dx.doi.org/10.1007/s10459-013-9480-6.

Olsson, U., Drasgow, F., and Dorans, N. (1982). "The polyserial correlation coefficient". English. In: *Psychometrika* 47.3, pp. 337–347. ISSN: 0033-3123.

Oluwasanjo, A., Wasser, T., and Alweis, R. (2015). "Correlation between MMI performance and OSCE performance - a pilot study". In: *Journal of Community Hospital Internal Medicine Perspectives* 5.3. URL: http://www.jchimp.net/index.php/jchimp/article/view/27808.

Padilla, M. A. and Veprinsky, A. (2014). "Bootstrapped Deattenuated Correlation Nonnormal Distributions". In: *Educational and Psychological Measurement*.

Paparoditis, E. and Politis, D. N. (2001). "Tapered block bootstrap". In: *Biometrika* 88.4, pp. 1105–1119.

Pardoe, I. and Weisberg, S. (2001). "An Introduction to bootstrap methods using Arc". In: *Unpublished Report available at www. stat. umn. edu/arc/bootmethREV. pdf.*

Patterson, F., Cousans, F., Edwards, H., Rosselli, A., Nicholson, S., and Wright, B. (2017). "The Predictive Validity of a Text-Based Situational Judgment Test in Undergraduate Medical and Dental School Admissions." In: *Academic medicine: journal of the Association of American Medical Colleges*.

Patterson, F., Knight, A., Dowell, J., Nicholson, S., Cousans, F., and Cleland, J. (2016). "How effective are selection methods in medical education? A systematic review". In: *Medical Education* 50.1, pp. 36–60.

Patterson, F., Rowett, E., Hale, R., Grant, M., Roberts, C., Cousans, F., and Martin, S. (2016). “The predictive validity of a situational judgement test and multiple-mini interview for entry into postgraduate training in Australia”. In: *BMC medical education* 16.1, p. 1.

Pau, A., Chen, Y. S., Lee, V. K. M., Sow, C. F., and De Alwis, R. (2016). “What does the multiple mini interview have to offer over the panel interview?” In: *Medical education online* 21.

Pau, A., Jeevaratnam, K., Chen, Y. S., Fall, A. A., Khoo, C., and Nadarajah, V. D. (2013). “The multiple mini-interview (MMI) for student selection in health professions training–a systematic review”. In: *Medical teacher* 35.12, pp. 1027–1041.

Pfaffel, A., Kollmayer, M., Schober, B., and Spiel, C. (2016). “A missing data approach to correct for direct and indirect range restrictions with a dichotomous criterion: A simulation study”. In: *PloS one* 11.3, e0152330.

Pfaffel, A., Schober, B., and Spiel, C. (2016). “A comparison of three approaches to correct for direct and indirect range restrictions: A simulation study”. In: *Practical Assessment, Research & Evaluation* 21.6, pp. 1–15.

Pfaffel, A. and Spiel, C. (2016). “Accuracy of Range Restriction Correction with Multiple Imputation in Small and Moderate Samples: A Simulation Study”. In: *Practical Assessment, Research & Evaluation* 21.10, p. 2.

Politis, D. N. and White, H. (2004). “Automatic block-length selection for the dependent bootstrap”. In: *Econometric Reviews* 23.1, pp. 53–70.

Poole, P., Shulruf, B., Rudland, J., and Wilkinson, T. (2012). “Comparison of UMAT scores and GPA in prediction of performance in medical school: a national study”. In: *Medical education* 46.2, pp. 163–171.

Poon, W.-Y. and Lee, S.-Y. (1987). “Maximum likelihood estimation of multivariate polyserial and polychoric correlation coefficients”. English. In: *Psychometrika* 52.3, pp. 409–430. ISSN: 0033-3123.

Preisser, J. S., Lohman, K. K., and Rathouz, P. J. (2002). “Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random”. In: *Statistics in medicine* 21.20, pp. 3035–3054.

Public Health Action Support Team (PHAST) (2017). Numbers needed to treat (NNTs) - calculation, interpretation, advantages and disadvantages. https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1a-epidemiology/nnts, [Accessed: November 7, 2017].

Puddey, I. B. and Mercer, A. (2014). "Predicting academic outcomes in an Australian graduate entry medical programme". In: *BMC Medical Education* 14.1, pp. 1–12. ISSN: 1472-6920. DOI: 10.1186/1472-6920-14-31. URL: http://dx.doi.org/10.1186/1472-6920-14-31.

Puddey, I. B. and Mercer, A. (2013). "Socio-economic predictors of performance in the Undergraduate Medicine and Health Sciences Admission Test (UMAT)". In: *BMC medical education* 13.1, p. 1.

Puddey, I. B., Mercer, A., Andrich, D., and Styles, I. (2014). "Practice effects in medical school entrance testing with the undergraduate medicine and health sciences admission test (UMAT)". In: *BMC medical education* 14.1, p. 1.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Raju, N. S. and Brand, P. A. (2003). "Determining the significance of correlations corrected for unreliability and range restriction". In: *Applied Psychological Measurement* 27.1, pp. 52–71.

Ranasinghe, P., Ellawela, A., and Gunatilake, S. B. (2012). "Non-cognitive characteristics predicting academic success among medical students in Sri Lanka". In: *BMC medical education* 12.1, p. 1.

Rao, J. N. and Shao, J. (1992). "Jackknife variance estimation with survey data under hot deck imputation". In: *Biometrika* 79.4, pp. 811–822.

Richard Adams (2017). Universities scramble to fill places, with Russell group still making offers. https://www.theguardian.com/education/2017/aug/18/fewer-uk-students-degree-courses-ucas-clearing, [Accessed: December 19, 2017].

Robert, C. P. (2004). *Monte carlo methods*. Wiley Online Library.

Rodgers, Joseph Lee (1999). "The bootstrap, the jackknife, and the randomization test: A sampling taxonomy". In: *Multivariate Behavioral Research* 34.4, pp. 441–456.

Rogers, W. T. (1976). “Jackknifing disattenuated correlations”. In: *Psychometrika* 41.1, pp. 121–133.

Rothwell, P. M. (2005). “External validity of randomised controlled trials:“to whom do the results of this trial apply?””. In: *The Lancet* 365.9453, pp. 82–93.

Roy, J. (2003). “Modeling longitudinal data with nonignorable dropouts using a latent dropout class model”. In: *Biometrics* 59.4, pp. 829–836.

Royston, P. and White, I. R. (2011). “Multiple imputation by chained equations (MICE): implementation in Stata”. In: *Journal of Statistical Software* 45.4, pp. 1–20.

Rubin, D. B. (1976). “Inference and missing data”. In: *Biometrika* 63.3, pp. 581–592.

— (1996). “Multiple imputation after 18+ years”. In: *Journal of the American statistical Association* 91.434, pp. 473–489.

— (2003). “Nested multiple imputation of NMES via partially incompatible MCMC”. In: *Statistica Neerlandica* 57.1, pp. 3–18.

Rubin, D. B. and Schenker, N. (1991). “Multiple imputation in health-are databases: An overview and some applications”. In: *Statistics in medicine* 10.4, pp. 585–598.

Al-Rukban, M. O., Munshi, F. M., Abdulghani, H. M., and Al-Hoqail, I. (2010). “The ability of the pre-admission criteria to predict performance in a Saudi medical school.” In: *Saudi medical journal* 31.5, pp. 560–564.

Ruxton, G. D. (2006). “The unequal variance t-test is an underused alternative to Student’s t-test and the Mann–Whitney U test”. In: *Behavioral Ecology* 17.4, pp. 688–690.

S Purcell (2007). Maximum Likelihood Estimation (MLE). http://statgen.iop.kcl.ac.uk/bgim/mle/sslike_3.html, [Accessed: September 9, 2017].

Sa, B., Baboolal, N., Williams, S., and Ramsewak, S. (2014). “Exploring emotional intelligence in a Caribbean medical school”. In: *The West Indian medical journal* 63.2, p. 159.

Sackett, P. R. and Yang, H. (2000). “Correction for range restriction: an expanded typology.” In: *Journal of Applied Psychology* 85.1, p. 112.

Saguil, A., Dong, T., Gingerich, R. J., Swygert, K., LaRochelle, J. S., Artino Jr, A. R., Cruess, D. F., and Durning, S. J. (2015). “Does the MCAT predict medical school and PGY-1 performance?” In: *Military medicine* 180.4S, pp. 4–11.

Saha, C. and Jones, M. P. (2009). “Bias in the last observation carried forward method under informative dropout”. In: *Journal of Statistical Planning and Inference* 139.2, pp. 246–255.

Salem, R. O., Al-Mously, N., AlFadil, S., and Baalash, A. (2016). “Pre-admission criteria and pre-clinical achievement: Can they predict medical students performance in the clinical phase?” In: *Medical Teacher* 38.sup1. PMID: 26984030, S26–S30. DOI: 10.3109/0142159X.2016.1142511. eprint: http://dx.doi.org/10.3109/0142159X.2016.1142511. URL: http://dx.doi.org/10.3109/0142159X.2016.1142511.

Sartania, N., McClure, J., Sweeting, H., and Browitt, A. (2014). “Predictive power of UKCAT and other pre-admission measures for performance in a medical school in Glasgow: a cohort study”. In: *BMC Medical Education* 14.1, p. 116. ISSN: 1472-6920.

Saupe, J. L. and Eimers, M. T. (2010). “Correcting correlations when predicting success in College: Presented at the 50th anniversary forum of the Association for Institutional Research”. In:

Schafer, J. L. (1999). “Multiple imputation: a primer”. In: *Statistical methods in medical research* 8.1, pp. 3–15.

Schafer, J. L. and Graham, J. W. (2002). “Missing data: our view of the state of the art.” In: *Psychological methods* 7.2, p. 147.

Schafer, J. L. and Olsen, M. K. (1998). “Multiple imputation for multivariate missing-data problems: A data analyst’s perspective”. In: *Multivariate behavioral research* 33.4, pp. 545–571.

Schmidt, F. L., Oh, I.-S., and Le, H. (2006). “Increasing the accuracy of corrections for range restriction: Implications for selection procedure validities and other research results”. In: *Personnel Psychology* 59.2, pp. 281–305.

Schripsema, N. R., Trigt, A. M. van, Wal, M. A. van der, and Cohen-Schotanus, J. (2016). “How Different Medical School Selection Processes Call upon Different Personality Characteristics”. In: *PloS one* 11.3, e0150645.

Sean Coughlan (2017). Poor students ’lose on grade predictions’. http://www.bbc.co.uk/news/education-42401660, [Accessed: December 19, 2017].

Sedgwick, P. (2013). “What is the number needed to treat (NNT)?” In: *Bmj* 347, f4605.

Shao, J. and Wang, H. (2002). “Sample correlation coefficients based on survey data under regression imputation”. In: *Journal of the American Statistical Association* 97.458, pp. 544–552.

Shehmar, M., Haldane, T., Price-Forbes, A., Macdougall, C., Fraser, I., Peterson, S., and Peile, E. (2010). “Comparing the performance of graduate-entry and school-leaver medical students”. In: *Medical education* 44.7, pp. 699–705.

Shulruf, B., Poole, P., Wang, G. Y., Rudland, J., and Wilkinson, T. (2012). “How well do selection tools predict performance later in a medical programme?” In: *Advances in Health Sciences Education* 17.5, pp. 615–626. ISSN: 1573-1677. DOI: 10.1007/s10459-011-9324-1. URL: http://dx.doi.org/10.1007/s10459-011-9324-1.

Silva, J. L. P. da, Colosimo, E. A., and Demarqui, F. N. (2015). “Doubly Robust-Based Generalized Estimating Equations for the Analysis of Longitudinal Ordinal Missing Data”. In: *arXiv preprint arXiv:1506.04451*.

Simmenroth-Nayda, A. and Görlich, Y. (2015). “Medical school admission test: advantages for students whose parents are medical doctors?” In: *BMC medical education* 15.1, p. 1.

Simpson, P. L., Scicluna, H. A., Jones, P. D., Cole, A. M., O’Sullivan, A. J., Harris, P. G., Velan, G., and McNeil, H. P. (2014). “Predictive validity of a new integrated selection process for medical school admission”. In: *BMC Medical Education* 14.1, pp. 1–10. ISSN: 1472-6920. DOI: 10.1186/1472-6920-14-86. URL: http://dx.doi.org/10.1186/1472-6920-14-86.

Siu, E. and Reiter, H. I. (2009). “Overview: what’s worked and what hasn’t as a guide towards predictive admissions tool development”. In: *Advances in Health Sciences Education* 14.5, p. 759.

Sjöberg, S., Sjöberg, A., Näswall, K., and Sverke, M. (2012). “Using individual differences to predict job performance: Correcting for direct and indirect restriction of range”. In: *Scandinavian journal of psychology* 53.4, pp. 368–373.

Sladek, R. M., Bond, M. J., Frost, L. K., and Prior, K. N. (2016). “Predicting success in medical school: a longitudinal study of common Australian student selection tools”. In: *BMC medical education* 16.1, p. 187.

Smith Hall (2005). More Correlation Coefficients. https://www.andrews.edu/~calkins/math/edrm611/edrm13.htm, [Accessed: April 26, 2017].

Sofia Lind (2016). Medicine degrees offered through clearing 'for first time ever'. http://www.pulsetoday.co.uk/your-practice/practice-topics/education/medicine-degrees-offered-through-clearing-for-first-time-ever/20032494.article, [Accessed: December 19, 2017].

Spearman, Charles (1904). "The proof and measurement of association between two things". In: *The American journal of psychology* 15.1, pp. 72–101.

Spiegelman, D. (2010). "Commentary: some remarks on the seminal 1904 paper of Charles Spearman 'The proof and measurement of association between two things'". In: *International journal of epidemiology* 39.5, pp. 1156–1159.

St George's University of London (2016). St George's University of London offers Medicine through Clearing for the first time. https://www.sgul.ac.uk/news/news-archive/medicine-clearing-first-time, [Accessed: December 19, 2017].

Stata (2016). Bootstrap — Bootstrap sampling and estimation. www.stata.com/manuals13/rbootstrap.pdf, [Accessed: July 20, 2017].

Statistics How To (2017). Criterion Validity: Definition, Types of Validity. http://www.statisticshowto.com/criterion-validity/, [Accessed: November 10, 2017].

Statsoft (2017). Structural Equation Modeling. http://www.statsoft.com/Textbook/Structural-Equation-Modeling, [Accessed: October 10, 2017].

Stef Van Buuren (2017). Multiple Imputation. http://www.stefvanbuuren.nl/mi/mi.html, [Accessed: October 19, 2017].

Stegers-Jager, K. M., Steyerberg, E. W., Cohen-Schotanus, J., and Themmen, A. P. (2012). "Ethnic disparities in undergraduate pre-clinical and clinical performance". In: *Medical education* 46.6, pp. 575–585.

Stegers-Jager, K. M., Steyerberg, E. W., Lucieer, S. M., and Themmen, A. P. (2015). "Ethnic and social disparities in performance on medical school selection criteria". In: *Medical education* 49.1, pp. 124–133.

Stegers-Jager, K. M., Themmen, A. P., Cohen-Schotanus, J., and Steyerberg, E. W. (2015). "Predicting performance: Relative importance of students' background and past performance". In: *Medical education* 49.9, pp. 933–945.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., and Carpenter, J. R. (2009). "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls". In: *Bmj* 338, b2393.

Steve Thornton (2017). UK Medical Education Database: an issue of assumed consent. http://www.clinmed.rcpjournal.org/content/17/2/191.1.full, [Accessed: December 12, 2017].

Steven, K., Dowell, J., Jackson, C., and Guthrie, B. (2016). "Fair access to medicine? Retrospective analysis of UK medical schools application data 2009-2012 using three measures of socioeconomic status". In: *BMC medical education* 16.1, p. 1.

Student (1908). "The probable error of a mean". In: *Biometrika*, pp. 1–25.

Štuka, Č., Martinková, P., Zvára, K., and Zvárová, J. (2012). "The prediction and probability for successful completion in medical study based on tests and pre-admission grades". In: *Stanisław Juszczyk*, p. 138.

Szumilas, M. (2010). "Explaining odds ratios". In: *Journal of the Canadian Academy of Child and Adolescent Psychiatry* 19.3, p. 227.

Taylor, N., Mehra, S., Elley, K., Patterson, F., and Cousans, F. (2016). "The value of situational judgement tests for assessing non-academic attributes in dental selection." In: *British dental journal* 220.11, p. 565.

Telegraph Reporters (2016). Only one in six A-level students is predicted the right grades by their teachers. https://www.telegraph.co.uk/education/2016/12/08/one-six-a-level-students-predicted-right-grades-teachers/, [Accessed: December 19, 2017].

Terregino, C. A., McConnell, M., and Reiter, H. I. (2015). "The effect of differential weighting of academics, experiences, and competencies measured by multiple mini interview (MMI) on race and ethnicity of cohorts accepted to one medical school". In: *Academic Medicine* 90.12, pp. 1651–1657.

Thai, H.-T., Mentré, F., Holford, N. H., Veyrat-Follet, C., and Comets, E. (2013). "A comparison of bootstrap approaches for estimating uncertainty of parameters in linear mixed-effects models". In: *Pharmaceutical statistics* 12.3, pp. 129–140.

Theresa Gillian, C. C. and Nelson, L. (2010). "Let SAS Do the Work : Correlation Crossroads". In: *SAS Global Forum: Stataistical Data Analysis*.

Thiele, T., Pope, D., Singleton, A., and Stanistreet, D. (2016). "Role of students' context in predicting academic performance at a medical school: a retrospective cohort study". In: *BMJ open* 6.3, e010169.

Thiele, T., Singleton, A., Pope, D., and Stanistreet, D. (2015). "Predicting students' academic performance based on school and socio-demographic characteristics". In: *Studies in Higher Education*, pp. 1–23.

Thorndike, R. L. (1949). *Personnel Selection: Test and Measurement Techniques*. Wiley.

Thornton, S. (2017). "UK Medical Education Database: an issue of assumed consent". In: *Clinical Medicine* 17.2, pp. 191–191.

Tiffin, P. A., Dowell, J. S., and McLachlan, J. C. (2012). "Widening access to UK medical education for under-represented socioeconomic groups: modelling the impact of the UKCAT in the 2009 cohort". In: *BMJ* 344, e1805.

Tiffin, P. A., McLachlan, J. C., Webster, L., and Nicholson, S. (2014). "Comparison of the sensitivity of the UKCAT and A Levels to sociodemographic characteristics: a national study". In: *BMC medical education* 14.1, p. 7.

Tiffin, P. A., Mwandigha, L. M., Paton, L. W., Hesselgreaves, H., McLachlan, J. C., Finn, G. M., and Kasim, A. S. (2016). "Predictive validity of the UKCAT for medical school undergraduate performance: a national prospective cohort study". In: *BMC medicine* 14.1, p. 140.

Tiffin, P. A. and Paton, L. W. (2017). Exploring the validity of the 2013 UKCAT SJT- prediction of undergraduate performance in the first year of medical school: Summary Version of Report. https://www.ukcat.ac.uk/media/1119/exploring-the-validity-of-the-2013-ukcat-sjt-prediction-of-ug-performance-in-1st-yr-of-med-school-summary-version-posted-27032017.pdf, [Accessed: May 3, 2017].

Tiffin, P. A., Paton, L. W., Mwandigha, L. M., McLachlan, J. C., and Illing, J. (2017). “Predicting fitness to practise events in international medical graduates who registered as UK doctors via the Professional and Linguistic Assessments Board (PLAB) system: a national cohort study”. In: *BMC Medicine*.

Till, H., Myford, C., and Dowell, J. (2012). “Improving Student Selection Using Multiple Mini-Interviews With Multifaceted Rasch Modeling.” In: *Academic medicine: journal of the Association of American Medical Colleges*.

Tiller, D., O’Mara, D., Rothnie, I., Dunn, S., Lee, L., and Roberts, C. (2013). “Internet-based multiple mini-interviews for candidate selection for graduate entry programmes”. In: *Medical education* 47.8, pp. 801–810.

Toomet, O., Henningsen, A., et al. (2008). “Sample selection models in R: Package sampleSelection”. In: *Journal of statistical software* 27.7, pp. 1–23.

Tsou, K.-I., Lin, C.-S., Cho, S.-L., Powis, D., Bore, M., Munro, D., Sze, D. M.-Y., Wu, H.-C., Hsieh, M.-S., and Lin, C.-H. (2013). “Using personal qualities assessment to measure the moral orientation and personal qualities of medical students in a non-western culture”. In: *Evaluation & the health professions* 36.2, pp. 174–190.

Turner, R. and Nicholson, S. (2011). “Can the UK Clinical Aptitude Test (UKCAT) select suitable candidates for interview?” In: *Medical education* 45.10, pp. 1041–1047.

UK Medical Education Database (2017). Welcome to UKMED. https://www.ukmed.ac.uk/, [Accessed: December 13, 2017].

UKCAT (2015). The UK Clinical Aptitude Test (UKCAT). http://www.ukcat.ac.uk/, [Accessed: March 15, 2015].

— (2017). The UK Clinical Aptitude Test (UKCAT) Official Guide. https://www.ukcat.ac.uk/media/1118/ukcat_guide_2017-web.pdf, [Accessed: September 28, 2017].

UKCAT Consortium (2017). The UK Clinical Aptitude Test (UKCAT) Dates and Fees. https://www.ukcat.ac.uk/ukcat-test/ukcat-dates-and-fees/, [Accessed: September 22, 2017].

University of Birmingham (2017a). Access to Birmingham (A2B). http://www.birmingham.ac.uk/undergraduate/preparing-for-university/teachers-advisors/reaching-regional/a2b.aspx, [Accessed: July 13, 2017].

University of Birmingham (2017b). Routes to the Professions. http://www.birmingham.ac.uk/undergraduate/preparing-for-university/teachers-advisors/reaching-regional/routes-professions.aspx, [Accessed: September 27, 2017].

University of Manchester (2017). Manchester Access Programme. http://www.manchester.ac.uk/study/undergraduate/aspiring-students/map/, [Accessed: September 27, 2017].

Valarie Blake (2012). Affirmative Action and Medical School Admissions. http://journalofethics.ama-assn.org/2012/12/hlaw1-1212.html, [Accessed: September 28, 2017].

Van de Ven, W. P. and Van Praag, B. M. (1981). "The demand for deductibles in private health insurance: A probit model with sample selection". In: *Journal of econometrics* 17.2, pp. 229–252.

Van Den Noortgate, W. and Onghena, P. (2005). "Parametric and nonparametric bootstrap methods for meta-analysis". In: *Behavior Research Methods* 37.1, pp. 11–22.

Van Iddekinge, C. H. and Ployhart, R. E. (2008). "Developments in the criterion-rlated validation of selection procedures: A critical review and recommendations for practice". In: *Personnel Psychology* 61.4, pp. 871–925.

Vance, C. (2009). "Marginal effects and significance testing with Heckman's sample selection model: a methodological note". In: *Applied Economics Letters* 16.14, pp. 1415–1419.

VanSusteren, T. J., Suter, E., Romrell, L. J., Lanier, L., and Hatch, R. L. (1999). "Do interviews really play an important role in the medical school selection decision?" In: *Teaching and Learning in Medicine* 11.2, pp. 66–74.

Veloski, J. J., Callahan, C. A., Xu, G., Hojat, M., and Nash, D. B. (2000). "Prediction of students' performances on licensing examinations using age, race, sex, undergraduate GPAs, and MCAT scores". In: *Academic Medicine* 75.10, S28–S30.

Verbeke, G. and Molenberghs, G. (2009). *Linear mixed models for longitudinal data*. Springer Science and Business Media. Chap. 5,6,8, pp. 41, 43, 69, 99.

Viswesvaran, C., Ones, D. S., Schmidt, F. L., Le, H., and Oh, I.-S. (2014). "Measurement error obfuscates scientific knowledge: Path to cumulative knowledge requires corrections for unreliability and psychometric meta-analyses". In: *Industrial and Organizational Psychology* 7.04, pp. 507–518.

Voet, H. van der (1994). “Comparing the predictive accuracy of models using a simple randomization test”. In: *Chemometrics and intelligent laboratory systems* 25.2, pp. 313–323.

Vonesh, E. F., Greene, T., and Schluchter, M. D. (2006). “Shared parameter models for the joint analysis of longitudinal data and event times”. In: *Statistics in medicine* 25.1, pp. 143–163.

Walther, B. A. and Moore, J. L. (2005). “The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance”. In: *Ecography* 28.6, pp. 815–829.

Wang, J. and Genton, M. G. (2006). “The multivariate skew-slash distribution”. In: *Journal of Statistical Planning and Inference* 136.1, pp. 209–220.

Wang, L. L. (2010). “Disattenuation of correlations due to fallible measurement”. In: *Newborn and Infant Nursing Reviews* 10.1, pp. 60–65.

Webb, N. M., Shavelson, R. J., and Haertel, E. H. (2006). “4 Reliability Coefficients and Generalizability Theory”. In: *Handbook of statistics* 26, pp. 81–124.

Wehrens, R., Putter, H., and Buydens, L. M. (2000). “The bootstrap: a tutorial”. In: *Chemometrics and intelligent laboratory systems* 54.1, pp. 35–52.

White, I. R., Royston, P., and Wood, A. M. (2011). “Multiple imputation using chained equations: issues and guidance for practice”. In: *Statistics in medicine* 30.4, pp. 377–399.

Wiberg, M. and Sundström, A. (2009). “A comparison of two approaches to correction of restriction of range in correlation analysis”. In: *Practical Assessment, Research & Evaluation* 14.5, p. 2.

Wilkinson, D., Casey, M. G., and Eley, D. S. (2014). “Removing the interview for medical school selection is associated with gender bias among enrolled students”. In: *Med J Aust* 200.2, pp. 96–99.

William M.K. Trochim (2008). Types of Reliability. https://www.socialresearchmethods.net/kb/reltypes.php, [Accessed: October 10, 2017].

Wilson, I., Griffin, B., Lampe, L., Eley, D., Corrigan, G., Kelly, B., and Stagg, P. (2013). “Variation in personality traits of medical students between schools of medicine”. In: *Medical teacher* 35.11, pp. 944–948.

Witt, P. L. and McGrain, P. (1985). “Comparing two sample means t tests”. In: *Physical therapy* 65.11, pp. 1730–1733.

Woolf, K., McManus, I. C., Potts, H. W., and Dacre, J. (2013). “The mediators of minority ethnic underperformance in final medical school examinations”. In: *British journal of educational psychology* 83.1, pp. 135–159.

Woolf, K., Potts, H. W., and McManus, I. (2011). “Ethnicity and academic performance in UK trained doctors and medical students: systematic review and meta-analysis”. In: *Bmj* 342, p. d901.

Wothke, W. (2000). “Longitudinal and multigroup modeling with missing data.” In:

Wouters, A., Bakker, A. H., Wijk, I. J. van, Croiset, G., and Kusurkar, R. A. (2014). “A qualitative analysis of statements on motivation of applicants for medical school”. In: *BMC medical education* 14.1, p. 200.

Wright, S. R. and Bradley, P. M. (2010). “Has the UK Clinical Aptitude Test improved medical student selection”. In: *Medical Education* 44.11, pp. 1069–1076. ISSN: 1365-2923.

Wyness, G. (2016). *Predicted grades: accuracy and impact*. University and College Union. URL: https://www.ucu.org.uk/media/8409/Predicted-grades-accuracy-and-impact-Dec-16/pdf/Predicted_grades_report_Dec2016.pdf.

Yates, J. and James, D. (2010). “The value of the UK Clinical Aptitude Test in predicting preclinical performance: a prospective cohort study at Nottingham Medical School”. In: *BMC medical education* 10.1, p. 55.

— (2013). “The UK clinical aptitude test and clinical course performance at Nottingham: a prospective cohort study”. In: *BMC medical education* 13.1, p. 1.

Yen, W., Hovey, R., Hodwitz, K., and Zhang, S. (2011). “An exploration of the relationship between emotional intelligence (EI) and the Multiple Mini-Interview (MMI)”. In: *Advances in health sciences education* 16.1, pp. 59–67.

Yoshimura, H., Kitazono, H., Fujitani, S., Machi, J., Saiki, T., Suzuki, Y., and Ponnamperuma, G. (2015). “Past-behavioural versus situational questions in a postgraduate admissions multiple mini-interview: a reliability and acceptability comparison”. In: *BMC medical education* 15.1, p. 1.

Zalewska, M., Niemiro, W., and Samoliński, B. (2010). “MCMC imputation in autologistic model”. In: *Monte Carlo Methods and Applications* 16.3-4, pp. 421–438.

Zhang, Z. (2016). “Missing data imputation: focusing on single imputation”. In: *Annals of translational medicine* 4.1.

Zhao, X., Oppler, S., Dunleavy, D., and Kroopnick, M. (2010). “Validity of four approaches of using repeaters’ MCAT scores in medical school admissions to predict USMLE Step 1 total scores”. In: *Academic Medicine* 85.10, S64–S67.

Zimmerman, D. W. and Williams, R. H. (1997). “Properties of the Spearman correction for attenuation for normal and realistic non-normal distributions”. In: *Applied Psychological Measurement* 21.3, pp. 253–270.