Developing Robust Statistical Scoring Methods for use in Child Assessment Tools

Gichuru Phillip Karanja, B.Sc, M.Sc.

Submitted for the degree of Doctor of Philosophy
at Lancaster University, UK
November 2017

Abstract

Timely and accurate diagnosis of developmental disability reduces its detrimental effect on children. Most of the current scoring methods do not appropriately remove the effect of age on development scores. This frustrates both disability status classification and comparison of scores across different child populations because their age dependent development profiles are usually quite different. Hence, the key objective of this research is to develop robust statistical scoring methods that appropriately correct for age using a) item by item age estimation methods that provide the expected age of achieving specific developmental milestones and b) overall score norms independent of the age effect using all the responses of a child to give one score across the entire domain for each child. Using data from 1,446 healthy and normally developing children (standard group) from the 2007 Malawi Development Assessment Tool (MDAT) study, a review of classical methods including generalised linear models, simple sum, Z-score, Log Age Ratio and Item Response Theory scoring methods in this child development context using binary responses only was carried out. While evaluating the pros and cons of each method, extensions to the current scoring methods using more flexible and robust methods including smoothing to reduce score variability are suggested. The results show that; a) the suggested generalised additive model extensions used for age estimation were more suited to deal with skewed item pass rate response distributions, b) smoothing of Z-scores was especially beneficial when variability in certain age groups is high due to low sample sizes, c) the more complex methods accounting for item response correlation or increase in item difficulty resulted in reliable and generalizable normative scores d) the extended overall scoring approaches were able to effectively correct for age achieving correlation coefficients of less than +0.25 between age and scores. The suggested overall scoring extensions improved the accuracy of detecting delayed development both in the disabled and even in the harder to classify malnourished children achieving sensitivity values of up to 98% and 85% respectively.

Acknowledgements

This thesis would not have been possible without the help, advice, support and facilitation of many people and institutions.

In particular, I would like to thank the Economic and Social Research Council (ESRC) for the funding that made my PhD work possible. I was also honoured to be a Kluge Fellow at the Library of Congress (LOC) in Washington D.C, U.S.A., after winning a three month AHRC fellowship award to carry out a systematic literature review to complement my work.

Both my supervisors Professor Gillian Lancaster and Dr. Andrew Titman offered excellent guidance that has made my Ph.D. experience very productive. Additionally, I would like to thank Dr. Melissa Gladstone who facilitated the use of the data used in this research as well as her contagious enthusiasm for child development research. A special thank you to the subject librarians both at Lancaster and the Library of Congress for their expert advice especially in sourcing materials to enrich the depth of reviews in this thesis. In a special way, I would like to thank Professor Brian Faragher for his motivation especially during tough times during my Ph.D. pursuit.

I would like to thank my family for all their love and encouragement. No words can express my gratitude to Mueni, for the relentless support and patience during the preparation and writing of this PhD. Finally, to God, who blessed me with the strength and perseverance to complete this PhD thesis. Thank you.

Declaration

This thesis is the result of my own work. The material contained in this thesis has not been presented, nor is it currently being presented, either wholly or in part for any other degree or qualification. The research was undertaken under the auspices of the Department of Maths and Statistics, Lancaster University between October 2011 and November 2017.

Signed……..

Gichuru Phillip Karanja

Dedication

I would like to dedicate this thesis to all the children of the world, more so those at the risk of having any developmental disabilities being diagnosed late.

Notation

The following are the general notations that will be used to formally describe the statistical model formulations used in this thesis.

Indices

- i for children, so that $i \in U_i$ where U_i is the set of items administered to the i^{th} child.
- j for items, so that $U_i = \{1, \dots, j\}$ where j refers to the number of administered items.
- n is the data sample size, so that $|U_i| = n_i$ is the number of items.

Data

$$y_{ij} = \begin{cases} 1 \; if \; child \; i \; passes \; the \; j^{th} item \\ 0 \; if \; child \; i \; fails \; the \; j^{th} \; item \end{cases}$$

Such that;

- p, is the probability of passing an item and lies in the interval $0 < p < 1$.
- p_i is the pass rate for passing the n items for the i^{th} child with a specific age, obtained from $Y_i, i = 1,2, \dots, n$ i.e. the binary responses to all the items the i^{th} child responds to.
- p_j is the pass rate of the j^{th} item obtained from $Y_j, i = 1,2, \dots, n$ i.e. the binary responses of all the n children to the j^{th} item.

Predictors

- X for item predictors (X_{ij}) or for predictors with fixed effects e.g. child age
- $\pi(X)$ is the probability of success i.e. the chance of getting an item correct.

Effects

- β for fixed effects of item predictors (β_j)
- θ for random effects of child predictors (θ_{ij})

Abbreviations

AIC – Akaike Information Criteria

BIC – Bayesian Information Criterion

CDC – Centre for Disease Control and Prevention

CTT – Classical Test Theory

FA – Factor Analysis

GAIC – Generalized Akaike information Criterion

GAM – Generalised Additive Model

GAMLSS – Generalised Additive Models for Location, Scale and Shape

GLM – Generalised Linear Model

GLMM – Generalised Linear Mixed Model

GM – Gross Motor

ICC – Item Characteristic Curves

IRT – Item Response Theory

MDAT – Malawi Development Assessment Tool

MM – Moderate Malnutrition

MUAC – Mid-Upper Arm Circumference

SCAM – Shape Constrained Additive Models

UN – United Nations

UNHCR – United Nations High Commissioner for Refugees

UNICEF – United Nations Children's Emergency Fund

WFP – World Food Program

WHO – World Health Organisation

Table of Contents

Abstract i

Acknowledgements ii

Declaration iii

Dedication iv

Notation v

Abbreviations vi

Table of Contents 1

PART I – INTRODUCTION 8

1. General Introduction 9

1.1. Introduction 9

1.1.1. What is the current child health burden and why is its accurate quantification relevant? 11

1.1.2. What is a standardised or norm referenced score and how is it used to classify development ability? 12

1.1.3. What are the positive, negative factors affecting and driving child developmental delay or disability? 16

1.1.4. What are the challenge(s) of child development research? 18

1.2. Collection of Assessment Response Data 19

1.3. Thesis contribution 21

1.4. Outline of thesis 24

2. Literature Review 26

2.1. Introduction 26

2.2. Why develop a tool using a translate or adaptation strategy, and how is its 'success' measured? 27

2.3. Statistical methods for developing tools for assessing child development 32

2.3.1. Design and content development stage 36

2.3.1.1. Reliability 39

2.3.1.2. Validity 39

2.3.1.3. Comparison between reliability and validity 40

2.3.2. Computing age estimates and overall ability scores 41

2.3.2.1. Item by item analysis 43

2.3.2.2. Overall scoring methods 47
2.3.3. Post hoc statistical analysis 59
2.3.3.1. Sensitivity analysis 59
2.3.3.2. Missingness in test item responses 61
2.3.3.3. Item to Total Correlation, Correlation within and between test item responses 63
2.3.3.4. Important Assessment Tool, Age Estimate and Score Properties 64
2.4. Summary of the systematic review 69
2.5. Summary 70
PART II – METHODS 72
3. The Malawi Development Assessment Tool 73
3.1. Introduction 73
3.2. The Malawi Development Assessment Tool study 74
3.2.1. The MDAT Study Population Description 74
3.2.2. The MDAT Item Development Process 78
3.2.3. The MDAT Tool Item Description and Characteristics 80
3.2.4. The MDAT Assessment Tool Kit 84
3.2.5. Recording an outcome using the MDAT tool 86
3.2.6. Item response recording into an analysis spreadsheet 89
3.3. Summary 90
4. Data and Exploratory Data Analysis (EDA) 91
4.1. Introduction 91
4.2. MDAT Data and External Validation 94
4.3. Exploratory Data Analysis (EDA) 96
4.3.1. Preparing to analyse the MDAT data 96
4.3.2. Exploring missing data 100
4.3.3. MDAT Study Characteristics 106
4.3.4. Exploring Item pass/failure rates 108
4.3.5. Item Correlation 111
4.4. Exploring Specific Item Characteristics 113
4.4.1. Item difficulty levels 114
4.4.2. Item Discrimination index 115
4.4.3. Item to raw total score correlations 118
4.4.4. Total raw score and age correlation 120

4.4.5. Empirical Item Characteristic Curves 122
4.5. Summary 126
5. Methods of Scoring Binary Assessment Data 127
5.1. Introduction 127
5.2. Scoring methods 127
5.2.1. Item by item analysis within each developmental domain 129
5.2.1.1. Generalized Linear Models – Logistic Regression (GLM) 130
5.2.1.2. Generalized Additive Models – (GAM) 133
5.2.1.3. Creating normal reference ranges for each item 136
5.2.1.4. Confidence intervals for fitted values 139
5.2.1.5. Item by item model checking and diagnostics 144
5.2.2. Creating an overall (total) score for a child using the entire (all) domain of items 149
5.2.2.1. Model based total scores using Simple Counts 153
5.2.2.2. Z-Score methods 156
5.2.2.3. Item Response Theory (IRT) scoring methods 159
5.2.2.4. Model selection and diagnostics for overall scoring methods 168
5.3. Methods for comparison of age estimate(s) and score characteristics 169
5.3.1. Comparison of age estimates from item by item analysis 174
5.3.2. Total score summary measures and distribution characteristics 175
5.3.3. Classification of developmental status: criterion validity 176
PART III – RESULTS 186
6. Results 187
6.1. Introduction 187
6.2. Item by item age estimation analysis 187
6.2.1. Comparison score characteristics of item by item analysis 196
6.2.2. Summary of item by item analysis 206
6.3. Overall scoring methods 206
6.3.1. Simple Sum Count Methods 212
6.3.1.1. Comparison score characteristics of Simple Scoring Methods 212
6.3.1.2. Sensitivity of simple scoring 217
6.3.1.3. Summary of simple scoring approach methods 220
6.3.2. Z-score methods 221
6.3.2.1. Comparison score characteristics of Z-Scoring Methods 221

6.3.2.2. Sensitivity of Z-scoring methods 227
6.3.2.3. Summary of the Z-score methods 231
6.3.3. Item Response Theory Overall Scoring Methods 231
6.3.3.1. Comparison of score characteristics of IRT Scoring Methods 232
6.3.3.2. Sensitivity of IRT scoring methods 241
6.3.3.3. Summary of Item Response Theory methods 245
PART IV – DISCUSSION 246
7. Discussion and Conclusions 247
7.1. Introduction 247
7.2. General Discussion 248
7.2.1. What are the current tool development, age estimation and scoring practices? 248
7.2.2. What is the current age estimation and scoring reporting practice? 250
7.2.3. What are the ideal properties of a typical assessment tool? 251
7.2.4. Why is item data quality important? 251
7.2.5. What are the important statistical implications? 252
7.2.5.1. Item by item methods 253
7.2.5.2. Overall scoring methods 253
7.2.6. What are the original contributions of this thesis and how are our suggested extensions more superior? 254
7.2.6.1. Item by item methods 255
7.2.6.2. Overall scoring methods 255
7.3. Limitations and future research 259
7.4. Concluding remarks 262
Bibliography 263
8. Appendix 279
8.1. Appendix A – Supplementary Material for Chapter 2 279
8.2. Appendix B – Supplementary Material for Chapter 3 280
8.3. Appendix C – Supplementary Material for Chapter 4 281
8.4. Appendix D – Supplementary Material for Chapter 5 282

List of Tables

Table 2.1: An overview of the types, descriptions, sources of bias and types of equivalence to be established in tool translation and adaptation ……..38

Table 2.2: Summary of problems of statistical methods used to compute age estimate and overall scores for norm creation using binary item response data……58

Table 3.1: A summary of MDAT item development process ……..78

Table 3.2: Gross motor item description list ……..83

Table 3.3: Item responses entered into spreadsheet……89

Table 4.1: A snap shot of Gross Motor MDAT data……99

Table 4.2: Case wise missing frequency summary in Gross Motor (GM) domain……..102

Table 4.3: Characteristics of cases with all 34 items missing in GM domain ……103

Table 4.4: Frequency counts of item pass/fail and missing rates in Gross Motor domain……..105

Table 4.5: Characteristics of Children in Normal, Disabled and Malnourished cohort samples……108

Table 4.6: Item responses entered into data spreadsheet ……..111

Table 4.7: Transposed Item responses entered into data spreadsheet ……112

Table 4.8: MDAT discrimination (D) item indices for the GM domain……117

Table 4.9: Item to total correlations in gross motor domain ……..120

Table 5.1: A summary of age estimation and overall scoring methods……173

Table 5.2: Types of errors in child development classification ……178

Table 6.1: Age estimates at the 25th and 90th percent probability of passing an item from GLM and GAM extention (SCAM) models for gross motor domain.……193

Table 6.2: Item by item model AIC value comparison in Gross Motor (GM) domain……198

Table 6.3: Age estimate Confidence Interval at the 25th and 90th % pass probability of items from GLM and SCAM models for gross motor domain ……201

Table 6.4: Score distribution summary statistics of the classical and extended overall scoring methods in the MDAT GM domain for the 3 sample data sets ……210

Table 6.5: Sensitivity and optimal score cut-off performance summary of the classical versus extended overall scoring methods in the MDAT GM domain ……211

Table 6.6: Means, standard deviations and the 2.5th and 97.5th percentile values of the Z-score (confidence intervals) for use in creating both unsmoothed (classical) and smoothed Z-scores as well as ability classification for the GM domain of the MDAT tool……227

Table 6.7: 1PL and 2PL IRT model parameter estimates for the MDAT tool in the normal sample……234

List of Figures

Figure 1.1: The normal distribution bell-shaped curve and its link to derived scores……15

Figure 2.1: The three stages of assessment tool development and normative score(s) computation; (a) design process-blue dotted box, (b) survey data collection and statistical analysis-red dotted box (focus of this thesis), (c) post-hoc analysis-green dotted box ……35

Figure 2.2: An extract from the Denver Developmental Screening Test for Sri Lankan Children (DDST-SL) showing gaps (red dotted box) and too much overlap (blue dotted box) in tool item age estimates……67

Figure 3.1: The MDAT Flow diagram of the recruitment of families and children for the MDAT study……75

Figure 3.2: The MDAT study data collection sites ……77

Figure 3.3: Stages in the creation of final MDAT tool ……80

Figure 3.4: MDAT gross motor tool chart ……82

Figure 3.5: Scoring child ability using MDAT chart ……88

Figure 4.1: A schematic flow diagram of the data exploring strategy of binary item responses before item by item age estimation and overall score computation ……93

Figure 4.2: Flow chart of Malawi Development Assessment Tool data ……95

Figure 4.3: Summary of item pass, fail and missing rates in gross motor domain……98

Figure 4.4 Back to back histograms items pass/failure rates against age for items that are ideal for babies (< 1 year old), toddlers (1 to < 3.5 years old) and pre-school age children (3.5 to < 7 years old) in the Gross motor domain ……110

Figure 4.5: Scatter plot of total sum raw score by age ……121

Figure 4.6: Empirical item characteristic curves of gross motor domain by age……125

Figure 5.1: Comparison of logistic and spline fit on MDAT item ……130

Figure 5.2: Creating normal reference ranges for each item under the GAM framework extension of the SCAM model ……138

Figure 5.3 GLM and isotonic regression fits for GM item 22 using aggregated data……147

Figure 6.1: A comparison of GLM and SCAM model fits for items ideal for infants (item 1), toddlers (item 17) and pre-school age children (item 34) in the gross motor domain……191

Figure 6.2: Display of model fit, confidence band around model fit, computation of age estimates and age estimate confidence band on single item data ……192

Figure 6.3 a): Normal reference age values for gross motor domain using Generalized Linear (Logistic) Model ……194

Figure 6.3 b): Normal reference age values for gross motor domain using Shape Constrained Additive (SCAM) Model ……195

Figure 6.4: SCAM model fit using relevant item data (age≤1 year) for GM item 1.....204

Figure 6.5: Centile plots on simple scores, Score density plots and Scatter plots for Quantile 50th percentile score and GAMLSS BB scores in GM domain for Normal, Disabled and Malnourished data……213

Figure 6.6: Percentile curves reference charts using GAMLSS regression for 0.025, 0.05, 0.25, 0.50, 0.75, 0.90 and 0.95 ……216

Figure 6.7: ROC curves: GAMLSS regression score method in GM domain……219

Figure 6.8: Mean and SD summaries of classical and Smoothed Z-scores, Z-score density plots and Z-score scatter plots with age in GM domain in normal, disabled and malnourished data ……..…..226

Figure 6.9a): ROC curves: Empirical Z- score method in GM domain……..229

Figure 6.9b): ROC curves: Smoothed Z-score method in GM domain……230

Figure 6.10: Density plots, scatter plots and ICC plots of 1PL IRT and 2PL IRT scores in GM domain for Normal, Disabled and Malnourished data ……..239

Figure 6.11: Density plots, score scatter plots, ICC plots of 1PL spline IRT scores in GM domain for Normal, Malnourished and Disabled data ……240

Figure 6.12a): ROC curves: 1PL score method in GM domain……242

Figure 6.12b): ROC curves: 2PL score method in GM domain ……243

Figure 6.12c): ROC curves: 1PL spline score method in GM domain ……244

PART I – INTRODUCTION

Chapter 1. General Introduction

Chapter 2. Literature Review

1. General Introduction

1.1. Introduction

The Convention on the Rights of the Child asserted the child's right to adequate conditions for overall development (United Nations, 1989). Therefore, the merit of early and accurate diagnosis of disability (or delayed development) is advantageous not only because a timely intervention means that there is a higher chance that one's health outcome is significantly improved if any form of morbidity exists, but also because this strategy makes better use of limited resources.

Various child health outcomes are used to assess or diagnose disability. A health outcome is a measure used to quantify and describe the effectiveness of a particular paediatric health care intervention. The implementation of internationally funded health intervention studies is critically dependent on viable tools to assess child development or enable early detection of any disability, and there is a dearth of such tools for use in children, particularly in Non-Western settings (Sabanathan, et al., 2015; Maulik & Darmstadt, 2007; Smit, et al., 2006). Further, crucial to developing suitable measurement tools and scoring methods for evaluating health interventions is the application of appropriate methodology and statistical techniques. This is both at the stages of the assessment tool creation and development as well as in the transformation of responses obtained from the tool to compute meaningful scores used to classify a child's development or disability status. Such tools need to be appropriate for the age, developmental status and physical capabilities of the child. Further issues arise in the suitability of tools when the child has mental and/or physical difficulties. Further, aside from age, other factors positively or negatively impact a child's health progression and are hence often strongly associated with computed development scores. This makes both disability status classification and comparison of scores across different child populations difficult because their age dependent development profiles are usually quite different.

Assessing child development with regard to ability is by concept a daunting undertaking because given the different modes of measurement often via cross sectional, case control and longitudinal studies, the actual outcome cannot be measured directly unlike physical growth or biological markers. The development indicator or outcome is thus measured indirectly by assessing other related or surrogate indicators. As we will soon appreciate, this makes the assessment process complex, lengthy, often requiring both large investments in terms of resources as well as very specialised expertise.

One methodological issue is where studies assessing developmental outcomes in resource-limited countries have tended to use Western assessment tools. Many are simply translated or adapted, with minimal validation before use. This approach is supposed to enable some comparison between groups. We make the point that this approach leads to misleading comparison conclusions as there is a lack of comparable outcome measures because these tools contain many items alien to children of a Non-Western culture. The Malawi Development Assessment Tool (MDAT) is a good example of how a western tool should be adapted and validated. Other tools have been created for children of a limited age range, have been based solely on urban children, or have excluded important domains of development such as language and social skills due to cultural differences. However, the MDAT (Gladstone, et al., 2008) research has highlighted how more culturally relevant items can and should be identified.

This literature review chapter briefly highlights the key issues that frustrate quantification of the current child heath burden to motivate the importance of having robust statistical methods applied in the tool/test development process from tool design to norm creation under selected subheadings. This will be in the form asking pertinent questions that should be constantly kept in mind as they guide the broader motivations of this research, such as outlining what exactly the current health burden is in Section 1.1.1, defining what a standardised or norm reference test is in Section 1.1.2, highlighting the factors that influence child development in Section 1.1.3 and outlining the problems and

challenges of child development assessment research in Section 1.1.4. These questions offer context and help focus the objectives of this thesis are given in Sections 1.3.

At the risk of being repetitive, the reader will constantly be reminded where this work aims to make its direct contribution specifically at the various stages of tool development where accurate assessment scores are required. To guarantee accurate assessment scores the implementation of robust statistical methods is utilised. This is an effort to differentiate our work and give the context needed to adequately set the stage for the root motivations for the specific research questions to be investigated and extended by this thesis while echoing the importance of accounting for cultural aspects. A detailed dissection of the tool development process will be outlined in Section 2.3 focusing on statistical methods and important issues that underpin age estimation and overall score quality. Having identified the research gap in the proposed research, the chapter will conclude by outlining the thesis contribution using proposed methods to deliver the specific research questions in Section 1.3 with the overall thesis outline in Section 1.4.

1.1.1. What is the current child health burden and why is its accurate quantification relevant?

Monitoring child mortality or morbidity that indicates a country's current child health burden is a subject of much interest to various stake holders such as the World Health Organisation (WHO,2011), the World Bank all over the world. Why? To us too, disability or development delay is a form of morbidity outcome and is an important component of chid development. Therefore, as is true with physical growth assessment (Feigelman, 2011), it is in only understanding and accurately quantifying the degree of this specific child health burden through paediatric health surveillance methods using mortality or morbidity can a country be best placed to implement informed and successful interventions to reduce the same burden.

'The state of the world's children; Celebrating 20 years of the convention on the rights of the child' (UNICEF, 2009) is an elaborate report that shows the general decline in child mortality everywhere that is a consequence of several reasons. One of the reasons that this report identifies to enhance tracking and monitoring child burden outcomes is the application of better tool development by using scientific translation or adaptation methodology. Thus more accurate assessment or diagnosis of delayed development are both required to simultaneously identify and focus timely interventions to the right children. As Grantham-McGregor, et al., 2007 state, this early diagnosis strategy has been shown to prevent or at least to a great extent reduce the detrimental effects of developmental delays on children. In accordance with this argument, this work makes its contribution by reviewing current assessment methodology at the point of age estimation of expected development milestones, scoring and norm creation to improve development or disability classification of children.

1.1.2. What is a standardised or norm referenced score and how is it used to classify development ability?

Literature (Cronback, 1970) often defines 'standardisation' as the process that encompasses the selection of tool items at the design stage, administration of the items or the actual solicitation of target outcomes from defined population, analysis of outcomes to computation of age-based norms. Tests used to assess child development are or should be norm referenced given their remit of diagnosing any development abnormality. In turn, norm-referenced tests are always standardised. Norm-referenced tests enable the comparison of a child's performance to the performance of another typically normal group; normative group. A 'norm' is therefore a performance measure of a normative group on a test which is used to assess or classify the performance of a specific child. Age has been shown to be a strong predictor of all aspects of child development. Other characteristics often mentioned are gender, nutrition status, disease status and social economic class (Walker, et al., 2007).

It is therefore important that the normative group data be a representative sample of children with similar characteristics as those to be tested.

Once the test items are administered to the representative normative group, the results are analysed to compute scores which are often expected to follow the normal distribution that is a bell-shaped curve having the range of scores or values on its horizontal x-axis and the corresponding percentage of number of children achieving a particular score on the vertical y-axis as shown in Figure 1.1 (Cronback, 1970; Glaser, 1963). Firstly, the bell-shape serves as assurance that the normative group is indeed representative, i.e. there is a large concentration of children in the middle (median or mean score) and a lesser number of children as one moves away from the middle in either direction. This is in line with the expectation that a representative sample of normally developing children will have most of their scores centred around the mean score value, and progressively fewer children will have their scores deviate from the mean score value to the progressively lesser or greater extents on the x-axis. The peak of the bell-shaped curve will coincide with the normal mean (average) or median (50^{th} percentile) expected score performance on the test. As explained by Stigler, (1982) a standard normal distribution such as the one shown in Figure 1.1 is the simplest case of a normal distribution having a mean of zero and variance of one. Secondly, when using a pre-specified cut-off value or score threshold, this bell-shaped property allows the classification of a child's ability status given a child's score's position being either; a) around the mean and within the confines of the pre-specified thresholds implying normal development b) to the left of the pre-specified threshold implying delayed development or disability c) to the right of the pre-specified threshold implying exceptional development. The classification (c) is usually of less concern as the primary objective of the assessment test is to detect disability or delayed development. Further, as we will see later in Section 5.3, the sensitivity of a test is also pegged on the value of the thresholds whose choice is predominantly motivated by research objectives being either screening tests or rigorous testing procedures and whether there are any factors influencing the sensitivity of the test.

The mean score described above is only a summary statistic, thus it represents the statistical average of a typical normal child's performance i.e. the 'norm'. For computation convenience, the score distribution can be transformed to conform to a standard normal distribution by adjusting the obtained scores using a suitable mean and variance value i.e. the score values from each child are put on a scale that has a mean of zero and variance of one as shown in Figure 1.1. Obviously, the score of an individual child will often not exactly correspond to the mean; it will most definitely deviate either positively (to the right) or negatively (to the left) of the mean. A natural question after computing a child's score is; 'what does their score say about their development status compared to the mean score of typically normally developing children of similar age?' To answer this question appropriately, we need to quantify the extent (or distance) a child's score deviates from the mean or 'norm'; this deviation or distance from the mean is called the standard deviation (SD). How 'small' or 'large' this deviation is on either side of the mean decides the development status of the child.

Using the bell-shaped property of scores from the normal representative sample, we can make the claim that it is expected that up to (a) 68.26% of all scores will fall within one standard deviation of the mean (34.13% above and 34.13% below the mean) (b) 95.44% of all scores will fall within two standard deviations of the mean (47.5% above and 47.5% below the mean) (c) 99.72% of all scores will fall within three standard deviations of the mean (49.85% above and 49.85% below the mean). Therefore, we see that a child's score value on a test may be understood in terms of its extent, distance or standard deviation from the mean. The position a child's score falls on the normal score scale distribution is indicative of the extent to which the score deviates from the average score for the standardisation sample i.e. the sample of normally developing children. This process not only quantifies the performance of a child, classifies their development status, it also has the added benefit of telling us the child's position within their development status age group. We note the use of the normal distribution features that serves to intuitively map what is naturally expected in the children's performance and also assists in development classification. There are instances where the normal

distribution cannot be achieved; therefore, other forms of expressing deviation from the 'norm' are used for skewed score distributions. Standardised scores (z-scores) that are reviewed in Section 2.3.2.2 are examples of the more frequently encountered transformed scores in standardised tests.

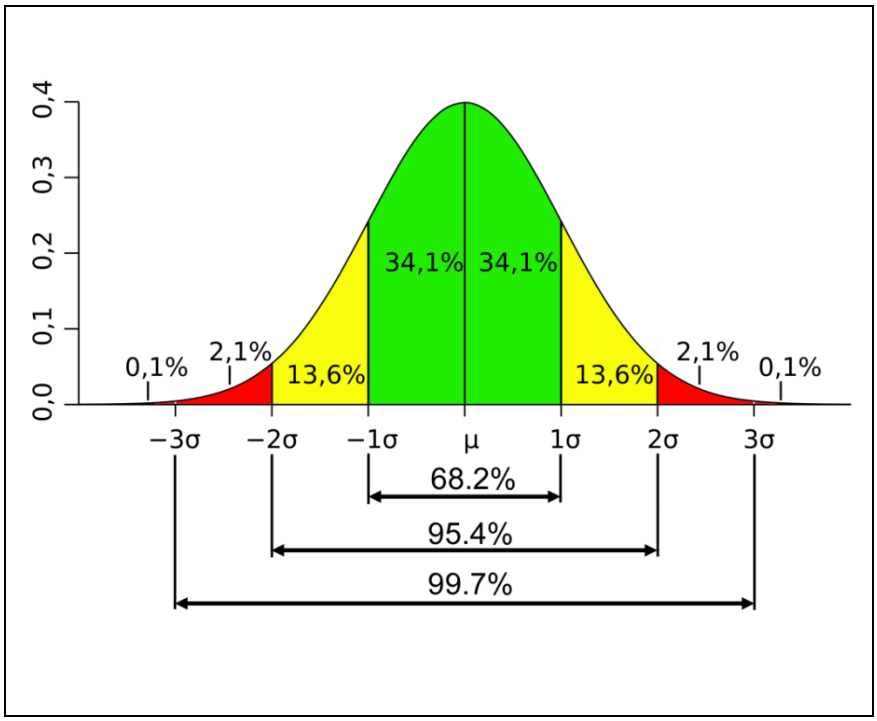


Figure 1.1: The normal distribution bell-shaped curve and its link to derived scores
*Image source: http://www.muelaner.com/wp-content/uploads/2013/07/Standard_deviation_diagram.png

The above paragraphs have explained the main idea and motivation for 'standardising' the raw scores that are computed form the raw item responses. This is the main focus of this thesis of formulating more suitable methods to statistically transform scores directly or indirectly into other kinds of more meaningful or appropriate scores to account for the relevant fundamental assessment features of item difficulty and the influence of age.

1.1.3. What are the positive, negative factors affecting and driving child developmental delay or disability?

It is a well-known fact that there are several factors that simultaneously influence child development both positively and negatively. Factors affecting child development negatively or that hamper it are referred to as risk factors, while factors that enhance or make child development flourish are called protective factors. Driving factors are factors that are predictive of child development or influence the pattern that a child's ability development trajectory will take. Age is one example of a strong driving or predictive factor. Depending on the outcome used to quantify development, driving factors are strongly associated with this outcome. Two other good examples of potential child development drivers or predictors are gender and socioeconomic status (Chin-Lun Hung, et al., 2015). As we will see later in Chapter six, many of the current scoring method techniques produce scores that are strongly associated with these factors especially age.

The main challenge is not only the lack of fully understanding these factors' interaction(s) with each other and ability development, but also the fact that their source or onset is often unknown (Klinnert, et al., 2001). Worse still, some factors affect the child right from conception with consequences presenting or manifesting much later in life i.e. post childhood (Pérez-García, et al., 2016). This complexity has been echoed by several other authors including Fenske, et al., (2013), Joffe, et al., (2012), Galea, et al., (2011), Jayasinghe, (2011), Smith, et al., (2001), and Krieger, (1994).

Also, the degree these factors amplify or hamper ability development differs depending on the cultural background and environment that the children are in. Usually, these factors are studied as prognostic factors using univariate analysis for example in the paper 'Long-term effects of early kwashiorkor compared with marasmus. III. Fine motor skills' by Galler, et al., 1987. They found that the presence of soft neurologic signs, indicated by low development scores measured six years earlier in the same children was significantly correlated with their current development scores performance, implying

that early malnutrition has effects on the nervous system function that are evident at 18 years of age. Just to amplify the relevance of our work, we note that the conclusion of such a study is dependent of the validity and accuracy of scores used to assess the fine motor skills. Further, even studies that have just looked at one biological or environmental factor's influence on an ability outcome still support the premise that child development is influenced by a multitude of factors (Galler, et al., 2013).

The world health report (WHO, 2007) highlights up to six risk factors namely malnutrition (underweight), inadequate stimulation, iodine, iron, zinc and vitamin A deficiency as well as a lack of breast feeding to be affecting at least 20-25 % of infants from the developing world. The Lancet series on risks for severely compromised development in young children in developing nations that reviewed the evidence linking compromised development with modifiable biological and psychosocial risks encountered by children from birth to five years (Grantham-McGregor, et al., 2007; Walker, et al., 2007; Engle, et al., 2000). This list of the potential risk factors of child development is clearly not exhaustive and many of the listed issues are also inter-related and still vary in their influence in various developing world environments.

To date the issue of risk factors influencing child development is studied in a non-holistic fashion (Syed, 2013.) mostly by focusing on children of a specific age range and thus design the study sample recruitment as such. There are many factors working in concert that influence child development. However, there is no study to our knowledge that has tried to assess the effect of several of these factors together. Most studies dissect the effects of various factors affecting the child or mother's disease status, social economic status, or social issues including mother's smoking status, mother's alcohol consumption or the effect of mother's incarceration on child development. We admit that fully understanding the influence patterns of these factors holistically is a daunting task and to some extent presents a barrier. However, without focusing on this negative realisation, we should not compromise the desire or motivation to understand this problem. Perhaps taking an additive approach to discovering and understanding this issue is the only way this 'puzzle' will be solved.

1.1.4. What are the challenge(s) of child development research?

The following are the general issues that render child development research difficult and further amplify its complexity. Some of the issues have already been highlighted by various authors including Geisinger, (1994). The various research study designs, development assessment and scoring strategies suggested to date to circumvent or address these issues are in one way or another affected by cultural aspects.

The primary challenge of child development research is how to measure 'delayed development' or 'disability', a latent construct, accurately at subject specific level as it cannot be measured directly. The strategy taken is to employ carefully designed assessment tools whose responses are converted to meaningful scores whose value is used to make an inference of the possibility of the presence of delayed development or disability. In this respect, the problem is twofold in the sense that the assessment process is in itself complex and its accuracy is simultaneously affected by diverse cultural issues that manifest differently in each setting. The later issue in turn makes the process of comparing the incidence or prevalence of delayed development or disability in child populations difficult. Therefore, we argue that only with a more robust scoring framework, can delayed development or disability be more sensitively quantified and various child populations with different characteristics be adjusted for to facilitate comparison.

Comparative studies that often need translation and adaptation of assessment tools are 'expensive' to fund, design, organise, conduct and monitor. The wish to accommodate different child populations of interest frustrates the comparative analyses of the information gathered due to disparity of these studies outcomes. This is further compounded by differences in: a) cultural practises that frustrate certain methods of health or development of assessment b) available infrastructure, c) differing law or ethical policy and legislation, d) effects of immigration and globalisation on children, d) differences in thought language and communication language ability in young children. Further, to facilitate

comparison of child development delay estimates, there is need for country or region specific comparative data to standardise data and correct for the factors highlighted in Section 1.1.3. This country or region specific comparative data is often difficult to find or does not exist for some developing world countries. Therefore be devising a framework that adjusts for country specific factors affecting child development, computed age estimates or scores can be fairly compared across countries or regions eliminating the need for comparative data.

This work agrees with the sentiments by both Gladstone, et al., (2008) and Grieve, (1992), that children need to be viewed within their own cultural context and when measuring cognitive abilities, the examiner needs to be sensitive to cultural variation. Cultural aspects can affect the management, practice and implementation of development assessment tools. For example, how the examiner perceives and interprets certain responses from the examinee (child) may be culturally driven. So too are the examinee perceptions on the examiner. In such instances adequate training of examiners and awareness of local cultural differences is recommended to avoid false conclusions. The work of Li & Karakowsky, (2001) titled 'Do we see eye-to-eye? Implications of cultural differences for cross-cultural management research and practice' discusses how observation is influenced by the cultural back drop of the observer. Other finer assessment strategies to overcome research challenges are outlined in the textbook 'Survey methods in multinational, multiregional and multicultural contexts' by Harkness, et al., (2010).

1.2. Collection of Assessment Response Data

The purpose of most research, be it a clinical trial, health services research, or an experimental study, is to demonstrate a relationship between an outcome of interest and one or more other variables or characteristics. The type and structure of an item response (Gaussian or non-Gaussian) determines the statistical analysis challenge in computing age estimates or scores to be used either for assessing tool equivalence or in development status classification. Univariate Gaussian item responses are

usually analysed using linear regression models which extend to the Linear Mixed Models (LMM) when data are correlated. Univariate non-Gaussian item responses are analysed using logistic regression, log-linear models, probit regression, etc., unified into the Generalized Linear Model (GLM) framework (Aitkin, 1999; Tacq, 1997; Liao, 1994; Aitkin, et al., 1989) and further extend to the Generalised Additive Models (GAM) and Generalised Linear Mixed Models (GLMM) in case or presence of correlation (McCulloch, et al., 2008; Dobson & Barnett, 2008). In the context of child assessment there is an extra third step in that the same methods that are used to develop scores or establish tool equivalence are again used to relate the scores to child characteristics (covariates) e.g. age. One objective may be to identify the causal effect of one or a set of variables on the computed score. Another objective may be the prediction of a child's future score value given their current score value as is done educational research to predict future performance trajectories. Alternatively, a regression or other model may provide meaningful summaries between the response variable (score) and the covariates. Section 2.3 will discuss the various stages of statistical analysis in more detail.

The statistical methods mentioned above may still perform poorly in some circumstances when for example assessment tools collect binary data whose item pass probability is skewed or the data collection process inadvertently introduces a form of bias through missing item responses as discussed in Section 3.2.5. It is the goal of this thesis to highlight and if possible address some of these issues by proposing alternative methodologies that are not only more flexible and robust but also easily implemented with available statistical software. This thesis therefore makes the specific contributions listed in Section 1.3 below of not only offering alternative scoring procedures in theory but also offering practical implementation solutions owing to handicaps data may present by frustrating the computation of the suggested scoring approaches using currently available statistical software. It is imperative that the entire data collection stage is well planned, diligently conducted and aspires to have the highest degree of quality so that the robust statistical methodology employed has the best chance of improving the quality of the scores to accurately classify child development status.

1.3. Thesis contribution

This section outlines the motivation for carrying out this research by answering the following pertinent questions; what is the relevance of this research? How do we intend to resolve some of the challenges outlined in Section 1.1.4 above or answer questions posed by our research objectives? What do we expect to find? And why is this the right time to carry out such research?

Given the challenges described in Section 1.1.4, Section 2.3 will identify that both the item by item analysis that computes age estimates and overall scoring methods may have various methodological concerns. Even if appropriate translation and adaptation methods have been applied to the tool, and all forms of bias eliminated to the extent of assuming that the right construct is being assessed, on the one hand, the age estimation may still be compromised if important model assumptions are not adhered to. These assumptions include: (a) assumptions made with respect to the underlying data structure and the fitted model; (b) the monotonicity assumption that ability increases with age, and therefore the computed age estimates should reflect this as the tool items are designed to increase in difficulty. On the other hand, to qualify as an appropriate method, the overall scoring approach has to be able to allow for; (i) differences in item difficulty; (ii) differences in the number of items administered to each child; (iii) the correlation between and within item responses; (iv) adjustment for the effect of age.

Now that we have identified what the statistical issues in age estimation and norm computation are, and why extending methodology in this regard is important, the broader aims of this thesis which are to address the statistical issues of concern both in the item by item and overall approaches can be defined. Further, in support of this research's motivation that is also shared by many authors including Gladstone, et al., (2008) and Grantham-McGregor, et al., (2007), this work also argues that high quality scores computed using robust methods will have significant benefit within research tools in the early identification of child disability or development problems and as an outcome measure in clinical trials.

By being able to suitably diagnose and assess children with disabilities via more accurate age estimates and norms, it will be a first step towards properly employing and directing scarce government resources. It is therefore expected that; (a) the suggested age estimation methods are able to appropriately deal with various item data structures that currently compromise the quality of age estimates and norms (b) the overall scoring approaches are more sensitive in detecting disability or development delay, or at the very least perform as good as the classical approaches but with the added advantage of being more generalisable by dealing with several sources of bias, or nuisance issues as outlined in Section 2.3.

The four research questions will be investigated using the following research strategies listed below;

1. What are the main statistical methodological issues to be addressed in the development of health assessment tools that have been devised for use in children?

   In line with the popular rhetoric that a cure to any malady must be based on diagnosing the problem, already this literature review while briefly discussing the problems of child development assessment and the tool development process, has identified and highlighted key problems in the tool development process that threaten the quality of tools developed. Because these problems in turn threaten the quality of the age estimates, scores and norms computed, robust statistical methods are required. We reiterate in assertion that accurate age estimates, scores and norms will improve the classification of developmental delay. A systematic review of developmental assessment tools will also be carried out to assess current translation and adaptation practise as well as current reporting practise in non-Western settings. Papers will be limited to those that tackle issues such as constructing measurement tools and the statistical techniques employed in developing such tools.
2. Can the performance of a developmental assessment tools be improved by using more robust item by item model fitting methods beyond GLM, such as the GAM framework?

Using the three MDAT datasets of normal, disabled and malnourished children, the quality, properties and performance of age estimates from the different item specific models will be compared especially with respect to accurate development status classification.

3. Can the overall scoring of tool items be more accurately and appropriately assessed by employing techniques that account for important issues within a child development context using an Item Response Theory framework or a model based framework such as the Generalized Additive Models for Location Scale and Shape (GAMLSS)? Further, can the suggested extensions of overall scoring methodology successfully adjust for the effect of age known to be not only strongly associated with the overall scores, but also frustrate development ability scores comparisons across various child populations?

   In a similar fashion to the second objective, using the MDAT normal, disabled and malnourished datasets, the quality of scores or norms produced using the above methods will be assessed. For the methods that account for the effect of age, the residual effect of age will be explicated by checking the strength of correlation and a scatter plot of the scores against age.

4. Does the sensitivity/specificity for diagnosing a child's degree of disability differ with respect to the overall scoring method adopted?

   We will compare the sensitivity/specificity of the classical versus suggested extensions of the various scoring methods. Further, the effect of age on sensitivity/specificity will also be assessed and the implications this has on the choice of appropriate cut-off thresholds used to classify the development status. This will be by comparing each of the scoring methods' sensitivity to correctly classify either a) normal versus children formally diagnosed with a neurological disability b) or normal versus children who are at a high risk of having delay in development or disability due to malnourishment and are therefore often harder to classify correctly.

Therefore, while to a large extent the extension of current age estimation and scoring methodology serves as the immediate objective of more accurate development classification and disability diagnosis, the work will also indirectly benefit other related patient care research trajectories such as quality of life assessment in primary care that are also reliant on assessment scores. Further, issues pertaining to the sensitivity to change assessment that is often frustrated by disparity in outcomes may benefit from this work. This work could not have come at a better time following the advancement of necessary statistical methodology to implement the suggested extensions supported by corresponding software technology and power.

1.4. Outline of thesis

The entire thesis is organised into three main parts. The first part referred to as the general introduction is made up of chapters one and two. The literature review has highlighted the issues and also identified the research gap in age estimation and scoring methods thus motivating our research. The second chapter reviews current statistical methods applied in tool development from tool design to norm creation to give the required context and also gives a summary of a systematic review charged with the task of collecting evidence of current translation and adaptation methods of tools adapted from western settings for use in developing countries. The output of the systematic review was the development of a quality assessment checklist to appraise literature reporting on the development of assessment tools and scoring methods. The second part of the thesis is the methods section. Within the methods section we describe the already developed MDAT tool that was used to collect the data used in this thesis in chapter three. This is then followed by an elaborate data description in the form of an exploratory data analysis in the fourth chapter. The fifth chapter is dedicated to formal descriptions of both classical item by item analysis methods and overall scoring methods pitted against our suggested extensions owing to the shortcomings of the former methods. This chapter also describes how we examine the sensitivity performance of classical approaches compared to our

extensions to showcase the suitability of our proposed robust statistical scoring methods. The third part of this thesis is the presentation of results and the discussion. The findings of the methods describing age estimate and overall score characteristics as well as the sensitivity analysis are presented in the sixth chapter of the thesis. The subsequent seventh chapter concludes the thesis giving an overall 'soft' discussion highlighting the key points of this thesis and specifically describing the implications of our findings, any identified limitations and the prospects for future research, respectively.

2. Literature Review

A Literature Review of Statistical Methodological Issues in Translation, Adaptation, Age Estimation, Scoring and Development of 'Norms' of Child Development Health Assessment Tools.

2.1. Introduction

Chapter one gave a brief background of the issues that frustrate child development assessment research over the entire process of the tool construction using the translation and adaptation of established child development assessment tools sourced from western countries. As was stated in Section 1.1, and which will soon become apparent, is the fact that the process of tool construction; from inception of a project, to developing a tool, to assessing development in children, to the creation of standard norms is quite an intricate process. Given the main focus of this thesis, we cannot cover all aspects that directly or indirectly underpin the quality of computed scores and standardisation process in their entirety. Instead, only methods deemed to directly affect the quality of scores and the development of standardised reference norms used for development status classification will be discussed in this thesis.

The take home message of this second chapter will be a recommended strategy of the 'best' translation, adaptation and scoring strategies from a developing world perspective. In the same spirit, this chapter will recommend a methodology for assessing the quality of reporting of this process. To review and take stock of the currently applied methodology used by researchers in this regard, and make a detailed assessment of the translation, adaptation, scoring and norm creation processes or methods that are so far being used in practice, a systematic review was conducted.

This second chapter reviews the importance of assessment tool development by translating and adapting already established tools in Section 2.2. Section 2.3 is a description of the statistical methods used in tool development, scoring and methods employed in the standardisation of scores. A summary

of the findings of the systematic review as well as the development of a quality assessment which was carried out in part during my internship at the Library of Congress in U.S.A Washington D.C. between November 2012 and January 2013 is given in Section 2.4. The chapter concludes by giving its recommendations in Section 2.5.

2.2. Why develop a tool using a translate or adaptation strategy, and how is its 'success' measured?

The above question is broken into four parts; firstly, the second paragraph will consider why it is preferable to translate and adapt a tool rather than develop a new tool by considering the benefits of the latter approach given our developing country and cultural research themes. Secondly we will outline the common translation and adaptation strategies. Thirdly, we will be highlight the importance these adapted tools by outlining how their derived scores are used to; a) either assess different types of population outcomes to facilitate child population comparisons or b) carry out further analysis to enhance tool quality. Finally, we will describe how best to assess whether appropriate and adequate translation or adaptation has been carried out again using computed scores from the same assessment tools.

To develop a tool to assess child development, one has two options; either create a completely new tool or translate and adapt a well-established tool to suit the new research context. In an ideal situation with unlimited research resources, then developing a new tool is usually the best approach. However, this is rarely the case thus one has to translate and adapt an existing tool to suit the new setting. In as much as this second approach is somewhat seen as inferior, it has the merits of not 'reinventing the wheel' and even deriving more benefit especially if issues that arise due to cultural differences that motivate the whole adaption process are appropriately addressed. Instead of dedicating substantial resources in developing a new tool, one can repurpose these resources on tasks that ensure the selected items are both reliable and valid, which we believe to be more important.

There are four common strategies used in the development of tools or scales and measures as outlined by Hubley & Zumbo, (2013);

- The rational–theoretical approach, in which an expert researcher uses both theory and intuition to develop (design) items for a test, is the most commonly used approach. In this case, expert opinion forms the basis for the development and selection of items.
- The empirical approach, in which items are selected if they can discriminate (differentiate) the group of interest from a control group.
- The projective approach, in which various stimuli (e.g. inkblots, pictures) are used to help individuals create their own drawing (e.g. draw a person) to project their own concerns, fears, attitudes, and beliefs onto their interpretation or drawing.
- Listing an over inclusive pool of relevant assessment items as described by Ozer & Reise, (1994). Here, one begins with a rough idea of an ability development construct and writes an over inclusive pool of possible items. Data are collected, and the analyses are used not just to refine the scale's psychometric properties, but also to iteratively generate new theories manifested in the final set of items about the nature of the construct.

A mix of the above strategies may be necessary depending the broad research teams' interests and resources. Project teams can use both pre-established recommended translation or adaptation and additional strategies of their choosing. We argue that the ideal translation and adaptation method of tool development is actually a hybrid of the above four strategies in that; first it avoids a waste of expert resource in designing new items that often end up only duplicate the existing ones, it is still reliant on expert input to decide on each included item's suitability especially where an item's cultural relevance is called into question, and is still reliant of statistical methodology to quantify and assess 'success' of the process. Further, the translation and adaptation approach allows for the continued refinement of the common already established items and measures as well as the testing of new ones.

Often western tools are adapted for use in developing countries (Gladstone, et al., 2008). Pfeifer, et al., (2011), Chang, (2001) and Hambleton & Kanjee, (1995) give an elaborate justification of the need and importance of adapting or translating tools and outlines the pitfalls of especially ignoring cultural aspects. We agree also that the process of translation or adaptation arises not only because of differences in language syntax, but also the need to maintain context of the tool in its new form. This makes the exercise quite complex as it is no longer a simple task of simply matching words, but an iterative process whose aim is to achieve a holistic transfer of both the source tool's language and context. Further, the process will almost often affect the design of the new translated or adapted tool. Even if two cultures are similar, still there is need to ensure that the set of norms are applicable in the new setting. Translation may also either increase or decrease difficulty. Usually the process of translation just concentrates on certain key words, some of which may be lost already in the translation process.

Besides having a viable and adequately adapted assessment tool to use in the new setting, the objectives of developing a tool using the translation and adaptation approach usually are either ;firstly, to compare the abilities of different types of children (e.g. schooled versus unschooled, diseased versus non-diseased), or children from different regions (countries) in multiregional surveys and secondly, to assess the impact of various disease or policy interventions on children and therefore use the 'ability' captured by the developed tool as an indicator or outcome measure that can further be statistically analysed cross-sectionally or longitudinally using various forms of trend analysis.

A further benefit of translating and adapting a tool is the potential of saving often scarce resources including time stemming from a developing world perspective. Also, note that because culture influences parenting methods that make child development trajectories to differ, it is questionable to adapt a tool without establishing its validity and reliability in the new setting. There is a competing argument by Tsai, et al., (2006) that one should only target the assessment of behaviours or

development indicators that are not influenced or dependent on culture. While this is a viable strategy, it implies leaving out many of the language and social development indicators that are dependent on culture. Successful transfer of the source tool's ability to assess the same constructs thus creating an adapted tool in the new environment will facilitate comparative study of multiple child populations. Given the current rate of population interaction and globalisation, tools that transcend many cultures are much needed to make fair comparisons to enable appropriate policy making decisions. While also contributing to the short list of comparable child health and development outcome indicators, such tools help in proper aggregation of analyses and scores reported at population level for evidence to inform universal disease prevention strategies for example.

To be able to use an established test or tool in a different environment with differences in language and culture, one needs to properly translate and adapt the tool to be able to test the same or similar themes and constructs as the source tool's settings i.e. the tool needs to possess the same psychometric properties in the new setting otherwise even if the translation was 'perfect', the interpretation from the resulting scores are not valid in the new context. However, 'complete or successful tool adaptation' is an impossible phenomenon to achieve and one can only aspire to get as close as possible to the source tool's constructs as highlighted by Greenfield, (1997).

Prior to the formal quantitative evaluation to test whether successful translation and adaptation of a tool has been achieved, there needs to be a specific cross-cultural adaptation process to allow using a western tool to be applied in a new setting. The WHO, (2012) recommends a four stage translation and adaptation process but we feel the work of Beaton & Guillemin, (2000) that suggests a six stage process is more suitable with the inclusion of a piloting stage. The latter's recommendation of including a piloting stage in our view is the back bone of the many translation and adaptation strategies commonly seen in literature. The six stages are composed of; (1) early translation where

expert translators translate a tool into a new language; (2) translation synthesis where at least two translations are merged and overseen by a mediator; (3) a back translation to the original language by independent experts; (4) an analysis of the first and third stage translations; (5) a thorough review of the translated and adapted versions to make final amendments; (6) a pre-test or pilot stage to test the new tool in new setting and resolve any unforeseen issues (tool or administration related) not previously identified. There are numerous variations to this intricate process driven by research objectives, types of tools or what aspects they measure given available resources.

The assessment and quantification of the degree of equivalence between the source and target tools via various statistical methodology helps validate any assumptions of similarity in construct between two populations differing in characteristics such as cultural practices. Equivalency between source and adapted tests/tools is mostly measured in terms of the degree of similarity between well-defined psychometric properties of constructs, represented by either age estimates or development scores, across the respective tested domains and across populations. This is then followed by viable expiations for observed differences as well as reasons for observed tool's good (indicating 'successful' adaptation) or poor (indicating 'unsuccessful' adaptation) performance in carrying out the assessment. The adaptation process may reveal that other aspects beyond tool content may have to be altered to suit the new setting. These changes may include altering the types of task responses, the scaling of the responses and even the tool administration.

There are two broad forms of psychometric properties that the translation and adaptation process aspires to achieve to prove satisfactory equivalent performance (Dunn, 1992). These are reliability and validity. Thus a good tool is indicated by higher values of these two summaries. 'Reliability' refers to the consistency of the measurement process several times i.e. can the process be reproduced with similar results multiple times. On the other hand, 'Validity' refers to the accuracy or closeness between a measurement tool's construct and the real world or truth; this is usually in terms of the tools ability

to accurately measure and correctly classify the development status of assessed children. There are several statistical analysis used to quantify both measures that use assessment scores that will be described in Sections 2.3.1.1 to 2.3.1.3. Therefore, we argue that with improved age estimate or scoring methods, better tool adaptation and assessment will be achieved.

As noted by Pfeifer, et al., 2011, the WHO (2012) recommends cultural adaptation and translation of existing tests with the immediate benefit of being faster and cheaper than developing a new tool. This was also put forth by Anastasi & Urbina, (1997). We note the argument of 'faster and cheaper' is made with the mind-set of the high child health burden (see Section 1.1.1) and lack of resource in developing countries. Although the WHO is yet to categorically come up with refined strategies or recommendations to go about improving the translation or adaption of established tools, they also attest to the fact that if both cross-cultural reliability and validity is achieved, it will allow a more comprehensive worldwide child development status data comparison.

2.3. Statistical methods for developing tools for assessing child development

We are now best placed to appreciate that; a) the process of developing health assessment tools is a lengthy process riddled with complex technical issues and b) both age estimates and scores that capture various child development outcomes are at the heart of this process. In this section we wish to dissect the tool development process from inception to the creation of norms that are used in development status classification. We situate that irrespective of the phase in the tool development process, the computation of appropriate and accurate age estimates or scores are needed; firstly, to do any reliability and validity analysis and secondly to compute the score norms that are eventually used for developmental status classification. Hence our work contributes at these two important tool development phases that rely on high quality age estimates or scores. We will outline the statistical methodology used to address various issues in the tool development process in each stage and

highlight the very crucial stage of age estimation and norm reference creation that this thesis hopes to directly contribute to.

We will discuss the statistical methodology used in tool development under three broad stages covering the entire process of the tool development as shown in Figure 2.1 below. There are numerous cycles of the tool development process depending on the research objectives and available resources including expertise, time and data. For example, after the design stage, a study could go straight into collecting assessment data, generating a score, carry out validation and reliability checks then compute the required age estimates or norms. Another more careful approach would be to first assess the preliminary reliability and face/content validity of the designed tool using pilot data and make any necessary tool design changes. Then the full data collection process (green dotted box in Figure 2.1) would be rolled out with much more confidence that the designed tool is appropriate for the target setting and its reliability and validity assessed accordingly on a larger sample. This second approach that we advocate is especially helpful in our given adaptation scenario. A pilot study (Lancaster, et al., 2004) goes a long way in unearthing several important issues unique to the new setting that may not be apparent at the design stage and even missed by the best tool development experts. However, including a pilot study would demand more resources.

The design or content development is the first stage at which the tasks or questions that assess the presence of target ability constructs are developed by translating and adapting already established tools. Whether or not we are using pilot data, the success of the design stage involves a psychometric investigation process to establish that adequate translation and adaptation has been achieved via various reliability and validity measures using a created ability score from pilot data. The relevant assessment data collected is used to compute either age estimates for defined mile stones or developmental norms. The final stage that involves the estimation and interpretation of scores could

also be a form of 'post hoc' analysis in which more rigorous reliability and validity testing could be undertaken to scale, link and create equivalence adjustment parameters (Dunn, 1992).

Not shown are the other equally important processes to ensure quality, monitoring, ethical concerns, documentation and iterative feedback that should be undertaken at every stage of tool development. A much finer tool development process is chronicled in the book titled 'A handbook of test development' by Downing & Haladyna, (2006). Lancaster, (2009) has reviewed the statistical issues in the assessment of health outcomes in children and highlighted problems of current practice at various stages of tool development. To facilitate harmonisation of the new tool in the new setting, these other development procedures also underpin the defined test standards and should share the centre stage besides the main translation and adaptation procedures. Previous work has defined these aspects to take place at designated stages in the development process but we argue that these other equally important processes should be interwoven within these three main stages; (a) design and content development stage, (b) data collection and main statistical analysis and (c) post-hoc statistical analysis.

The following Sections from 2.3.1 to 2.3.3 will dissect the three broad stages of tool development described above showing the use of statistical methodology. As will soon become clear, the successful achievement of each of the three stages' objectives are to a large extent pegged on the quality of computed age estimates or norms assuming that the other important tool development issues using a translation and adaptation strategy have been addressed scientifically.

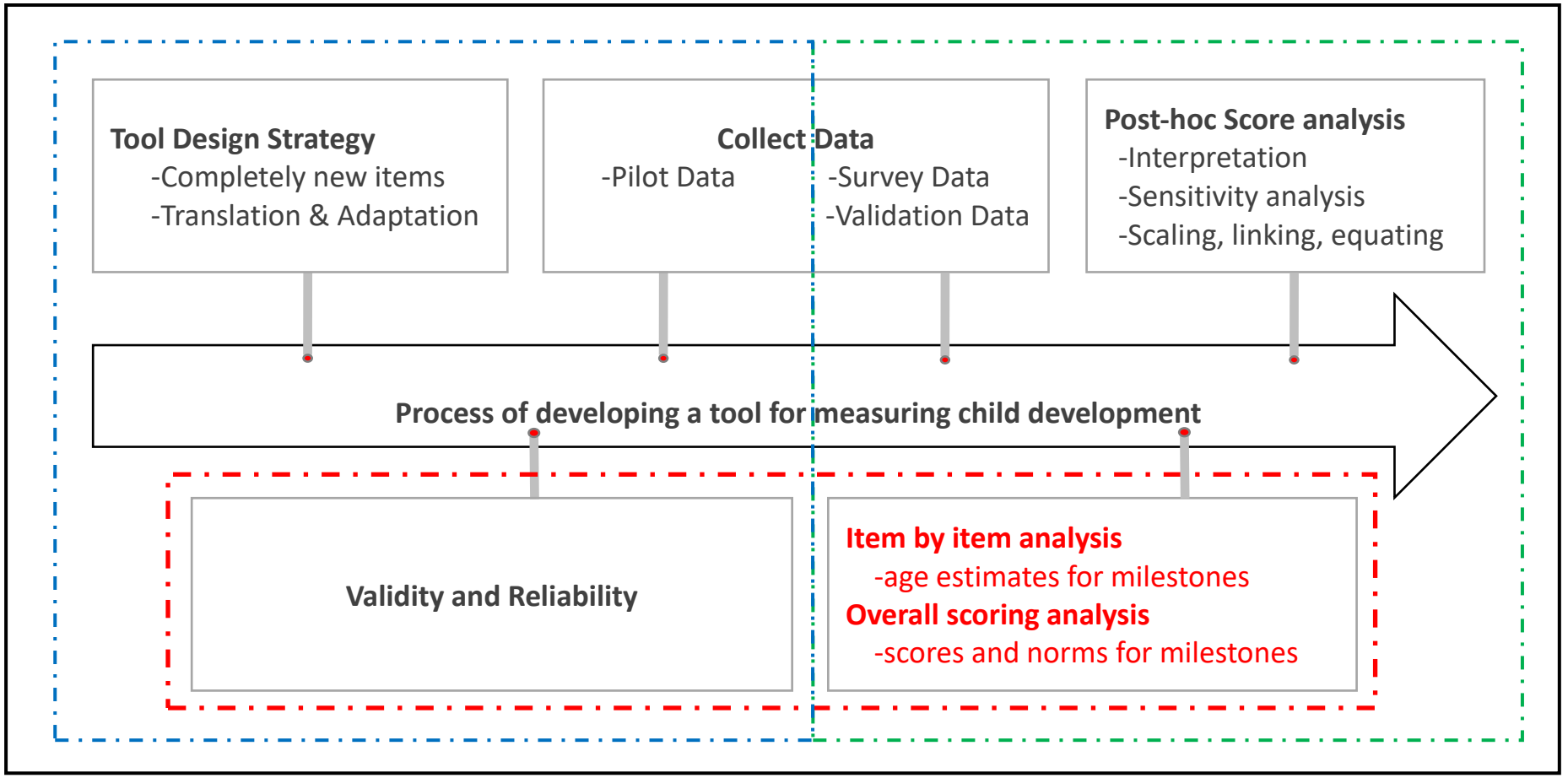


Figure 2.1: The three stages of assessment tool development and normative score(s) computation; (a) design process-blue dotted box, (b) survey data collection and statistical analysis-red dotted box (focus of this thesis), (c) post-hoc score analysis-green dotted box.

*This thesis contribution will be in the scoring of items after tool development (dotted red and green boxes).

2.3.1. Design and content development stage

The design and content development stage is tasked with the primary objective of tool content creation i.e. defining the specific items (tasks or questions) whose responses facilitate the assessment of the presence of the target latent ability construct. This content can be developed by translating and adapting already established tools and/or adding new items. Just like any other form of statistical analysis, the statistical analyses in tool development can be first fashioned as an exploratory process to investigate item data characteristics that advise the subsequent formal statistical analyses processes. There are various forms of bias summarised in Table 2.1 that could arise and can be detected and addressed by appropriate study designs and statistical methods. Shown on the fourth column of this table are the different forms of equivalence whose existence at an acceptable level has to be achieved. The existence of the different forms of equivalence cut across different forms of biases. Similarly, a statistical method may be used to detect one or a combination several forms of biases and to address one or several different forms of equivalence. The sole purpose of statistical methods used to establish tool equivalence is to assess whether the translation and adaptation was successful. Depending on the design of the study, the success of the tool design stage is measured by establishing acceptable recommended equivalence levels of different forms between the source and final tools via various reliability and validity statistical measures. Therefore, the equivalence analysis carried out is to ensure that all possible forms of bias that exist or were inadvertently introduced and challenge the success of the development process have been detected and addressed accordingly to prevent their manifestation in the age estimates or norms eventually computed.

The structure of data in reliability and validity studies often consists of two or more paired measurements differing in terms of examiner or examinee to assess different forms of biases. In the most common situation, there are two paired measurements on each subject or sample. In more complex designs, however, there may be more than two measurements on each sample, or different

samples may have different numbers of measurements. In this context, one could consider the number of measurements on each subject as the number of separate items of the questionnaire (in our case, the assessment tool). When considering the structure of the data, it is also important to consider the type of the variables and, if items are measured on a Likert scale or binary scale so that the appropriate statistical analysis can be planned accordingly as discussed in Section 1.2.

In our case we have binary items separated into four domains of gross motor, fine motor, language skills and social skills summarised by a continuous underlying latent score variable, representing a child's development ability. We will be focusing on; (a) reliability in terms of comparing different methods for creating age estimates and an ability score and (b) validity through comparisons of normal children with a group of formally diagnosed neurologically disabled children and a group with an extreme form of malnutrition that is known to be associated with or increase the risk of delayed development.

Table 2.1: An overview of the types, descriptions, sources of bias and types of equivalence to be established in tool translation and adaptation

Type of bias	Description	Sources of bias (examples)	Methods to address types of equivalence
Construct	An incomplete overlap of constructs in population groups	• Dissimilarity in the definitions of constructs across populations due to culture differences • Differential appropriateness of the behaviours associated with target construct	• **Construct equivalence**: Investigate and demonstrate that the source and final tool measure the same constructs across the different population groups using suitable forms of differential item functioning (DIF), dimensionality, differential test functioning (DTF), [†]various forms of factor analysis, correlation and agreement analysis • **Structural or functional equivalence**: Show that the source and final tool measure the same ability across different population groups using correlation and agreement analysis • **Measurement equivalence**: Show that the final tool measurement units, scales or surrogate outcomes facilitate the same across comparison across population groups • **Scalar or full score equivalence**: Show that the final tool has the same measurement units across comparison groups • Use various design and techniques like **matching, stratification** to ensure that possible confounding or interaction factors are addressed • Use various **adjusting, smoothing or robust** statistical techniques while computing age estimates or norms to deal with nuisance issues e.g. correlation, missingness and adjust for factors seen or known to influence ability
Item	The presence of anomalies of items due to inferior or inappropriate application of adaptation methods	• Poor item translation and adaptation • Item related nuisance factors induced by different population characteristics invoking additional traits or abilities	
Sample	The presence of nuisance issues e.g. correlation, missingness and factors arising from sample design and characteristics such as social economic status or disease	• Incomparability of samples caused by differences in sample characteristics e.g. age, social class, disease status or location • Nuisance issues compromising analysis	
Instrument	Tool features like measurement equipment, probing (tool kit) devices that are not related to target constructs that are influenced by population characteristics	• Differential familiarity with stimulus materials due to cultural differences	
Administration	Communication 'failure' between examiner and examinee	• Differential familiarity with stimulus materials due to cultural differences • Differential familiarity with response and styles procedures due to cultural differences • Differences in environmental administration conditions and methods due to population characteristics • Differences in tool administration logistical procedures due to available infrastructure	

[*]Adapted from Guidelines for Adapting Educational and Psychological Tests: A Progress Report by Hambleton, et al., (1994).
[†]Various forms of factor analysis-these include latent trait analysis, discriminant analysis and structural equation modelling.

2.3.1.1. Reliability

In general, reliability refers to the degree to which a measurement (age estimate or score) technique can be depended upon to secure consistent results over its repeated application as defined by Carmines & Zeller, (1979). In our research context reliability refers to the degree to which the final items produce the same consistent results as the source tool but in the new setting. There are up to six different forms of reliability measures that are applicable dependent on study design and research objectives. These are well described in numerous references e.g. in the book 'Health Measurement Scales: a practical guide to their development and use' by Streiner & Norman, (2008). It will be seen that various forms of reliability are designed to detect if a certain form of bias exists, thus the design of the study also has to support the form of reliability to be carried out.

In this project, the primary objective will be mainly to establish if our suggested scoring extensions are able to produce score values that are comparable to those produced by the classical methods. This will be in terms of the variability around item age estimate and score characteristics described in Section 5.3.

2.3.1.2. Validity

As defined by both Karras, (1997) and Engs, (1996), validity refers to the degree to which a study accurately reflects or assesses the specific concept that the researcher is attempting to measure. Validation has been emphasised in the design, sampling, measurement (referring data collection process) and analysis stages of research work. Within the context of the development of a child health assessment tool, validation refers to the degree to which a tool (through age estimate or score) measures a specific target construct related to ability. There are several types of validity depending on the different stages of the research, each charged with the remit of assessing a specific form of bias.

The several forms of validity measures are applicable dependent on study design, research priorities and objectives. Validity too detects if a certain form of bias exists and the design of the study has to support the form of validity to be carried out. Our work will be interested in criterion-related validity (using non-normal child samples), also referred to as instrumental validity. It applies to instruments that have been developed for use as indicators of a specific trait or behaviour; either now or in the future (to establish predictive validity). It checks how meaningful the research criteria are relative to other possible criteria by comparing it with another measure or procedure, which has been demonstrated to be valid (Karras, 1997). Therefore, we will assess whether our suggested scoring extensions' validity is as good as or superior to the classical scoring methods by correctly differentiating the development status between a sample of children with and without a known neurological disability or characteristics like malnutrition known to cause or be associated with delayed development as described in Section **Error! Reference source not found.**.

2.3.1.3. Comparison between reliability and validity

Reliability and validity are related and it may seem that these two quantities have only subtle semantic distinctions; but in fact they are quite different concepts. Many authors e.g. Wittchen, (1994) have attested to their relation and noted that researchers often inadvertently substitute various forms of the concepts with each other. In this child development context, if we consider the gross motor domain as the concept or construct of interest, then for each child assessed we are measuring their gross motor skill ability. Therefore, if we are measuring this construct accurately then we are indeed measuring their gross motor ability, and if not we are measuring something else (not gross motor skill ability).

We can therefore have three comparison result scenarios; firstly, we are inaccurately measuring gross motor ability but consistently i.e. we are consistently and systematically measuring the wrong ability construct for all respondents. This measure is reliable, but not valid (it is consistent but wrong). The

second scenario is a case where our measures are both inconsistent and very varied, i.e. not only do we not measure gross motor skills but our measures are very varied. In this second case our measure is neither reliable nor valid. The third scenario is a reliable and valid measure where we accurately and truly measure gross motor skills consistently. The third case is both highly reliable and valid as it yields consistent results in repeated application and it accurately assesses our target gross motor construct. Finally, it is not possible to have a measure that has low reliability and high validity i.e. you cannot really accurately measure the target construct of interest when its corresponding item response and consequently its age estimate or overall score fluctuates wildly. As outlined in Sections 2.3.1.1 and 2.3.1.2 above, we endeavour to show that the suggested statistical method extensions to compute age estimates and scores; a) are consistent (i.e. reliable) by assessing their similarity with age estimates and scores produced using classical methods b) are also valid by assessing their degree of sensitivity to correctly classify a child's development status in respective domains.

2.3.2. Computing age estimates and overall ability scores

This review would not be complete without mentioning research on the development of growth reference charts pioneered by the work of Cole, et al., 2012; Cole, 2008; Cole, et al., 2008; Cole, 2003; Cole, 1994a; Cole, 1994b; Cole & Green, 1992; Cole, 1998. Their work on statistical methods for assessing a child's anthropometric (physical body) measures spans a decade of ground breaking advances in the production of powerful graphical tools to assess aspects of child growth, especially growth chart construction using the Lambda-Mu-Sigma (LMS) method (Cole, 1990). The LMS method provides a flexible way of obtaining normalised growth centile standards by dealing with skewness that is usually present in the distribution of physical growth measurements e.g. height, weight, head and mid-upper arm circumference (MUAC) or skinfolds. It assumes that the data can be normalised by using a power transformation (e.g. Box-Cox, see Bickel & Doksum, 2011), which stretches one tail of the distribution and shrinks the other, thus removing the skewness of the growth measure response distribution. Cole's methods have since been duplicated widely in various countries e.g. the British

1990 growth reference, which was the official UK growth reference from 1996 to 2009 and the WHO Multicentre Growth Reference Study (MGRS) that was undertaken between 1997 and 2003 to generate new growth curves for assessing the growth and development for infants and young children around the world.

The collection of the item response assessment data is then followed by the statistical analysis stage i.e. the computation of either age estimates for defined mile stones and overall child ability scores. This is the stage at which this thesis hopes to make its direct contribution where it is envisaged that by extending current age estimate and norm creation methodology, the objective of accurate ability classification of development status will be better achieved with the added advantage of creating higher quality adapted tools. This is based on the fact that once a viable list of appropriate items (tool) has been obtained; the quality of the tool is further enhanced by using the item quality information derived from the more accurate age estimates or norms produced by applying these robust statistical methods. Therefore, in line with Cole's, et al's work that advanced statistical methods for physical measures using continuous data, ours will be to advance current statistical methods to produce charts for ability development classification using binary data.

Within the context of binary data, the scoring methods used can be broadly categorised into two groups according to whether one considered each item in a domain individually or all items simultaneously across the entire domain. To a large extent, the methods applied within these two broad analyses strategies, each have different purposes that complement each other, are dictated by the item data type and serve to self-check the tool development process. In this review we wish to mention their relevance in our work but also admit to the fact that it is beyond the scope of this review to give an in depth account of all their uses. As noted also by Sireci, (2005), besides the computation of age estimates or scores, the same statistical methods can also be used for item data exploratory purposes to assess item quality or to investigate problematic items due to a flawed language or

cultural adaptation process. This is through using simple methods based only on visual analysis as well as of more appropriate methods based on modern measurement theory. The Table 2.2 summarises the common problems of the various statistical methods used to compute age estimates for defined milestones and overall scoring methods for norm creation that will be discussed below in Sections 2.3.2.1 to 2.3.2.2. This approach of first identifying the specific classical methodology flaws therefore gives a glimpse of our thinking process of advancing methodology by building up on identified weaknesses of current statistical age estimation and overall scoring methods.

2.3.2.1. Item by item analysis

This work focuses on binary response data where the item by item analysis approach gives an expected age estimate for the child to pass a certain item to define a development milestone at a stipulated probability. This approach is currently used to compute age estimates used in the MDAT tool discussed in chapter three. As will soon be realised, the item by item analysis discussed in this section has two benefits; (i) Firstly, it can be viewed as a form of exploratory data analysis that informs the item quality. Therefore, as outlined by Langer, et al., (2007), Sireci, (2005), Dorans & Kulick, (1986), Angoff, (1982), Angoff & Ford, (1973), Angoff, (1972) and Mantel & Haenszel, (1959), this item by item analysis can be seen as a variation of Differential item(test) functioning (DIF) whose aim is to improve item quality and hence overall assessment tool quality. ii) Secondly, the findings of the item by item analysis have direct implications on the suitability of some of the overall scoring approaches used that are reviewed in Section 2.3.2.2. This is by utilising the robust features learned from the item by item modelling methods to either a) appropriately eliminate the age effect on the skewed item responses or overall scores for ability classification purposes or, b) smoothen overall raw scores for the purposes of eliminating high variability in summary measures used for score standardisation that threatens their quality.

a) Generalised Linear Models (GLM)

On account that the items considered in this thesis return a pass/ fail binary outcome and therefore are assumed to follow a binomial distribution, logistic regression is perhaps the most natural approach to measure the relationship between the categorical dependent variable (y) and one or more independent variable (x) using a logistic function. The binomial distribution (Ross, 1998) describes the distribution of the errors that equal the actual response, y minus the predicted response, y^*. Previously, the method has been discussed at length by Cox & Snell, (1969) and has been well received and applied in various research contexts as evidenced by the work of Agresti, (2002). As outlined in the data Section 1.2 the GLM modelling frame work has also been extensively researched and extended especially within longitudinal or clustered data scenarios to address the underlying correlation within outcomes leading to the generalized linear mixed models (GLMM) whose concepts are exhaustively discussed by both Stroup, (2012) and Dobson & Barnett, (2008).

However, just like the popular regression model, the GLM is based on several strict statistical assumptions including linearity in the logit, absence of multicollinearity, ratio of pass and fail cases, independence of errors made about the relation between the dependent (y) and independent variables (x), that have to be adhered to. For example, since the binomial distribution is also the assumed distribution for the conditional mean of the dichotomous outcome, it implies that the same probability is maintained across the range of predictor values. Adherence to this assumption is often tested by the normal z test (Siegal & Castellan, 1988) or may be taken to be robust as long as the sample is random; thus, observations are independent from each other. However, in this child development context we will see that the distribution of the pass probability of the dichotomous item responses is often skewed. Further, as we will see in Section 2.3.2.2 when generating an overall score, there is an underlying correlation owing to the multiple item responses from one child that renders the independence assumption invalid. However, this is not relevant in the item by item analysis.

b) Spline Models

Naturally, when a parametric modelling approach fails due to incompatibility of the assumed underlying assumptions, one can consider alternative flexible modelling approaches. The use of splines offers us such a framework to capture the relation of the binary outcome y and a covariate x. A spline simply makes no distributional assumptions about the covariate effect, but there are still distributional assumptions about the residuals (Greenland, 1995b). In general, splines are numeric piecewise-functions defined by polynomial functions. Because they possess a higher level of smoothness or flexibility, they are found to be a suitable alternative to the GLM framework in cases where the pass probability distribution is skewed. Using various interpolation methods, splines show case a higher level of smoothness at points where polynomials connect to each other. The points where polynomials connect are usually called knots.

Just like the GLM framework, the spline methodology framework too has received considerable interest in research owing to the flexibility rendered especially in dose response curves (Greenland, 1995c; May & Bigelow, 2005). A lot of research has gone into advancing the available types of splines, interpolation methods, maximisation methods to even advancement of necessary application software. There is a lot of literature in the form of peer reviewed papers and text books that adequately discuss the current spline theoretical concepts and applications, for example see 'Spline Models for Observational Data' by Wahba, (1990).

However, the choice of method used to define splines which is often either the degrees of freedom, number of knots and their positioning charged with the objective of improving the spline's model fit to the probability of passing an item given age, are to a large extent made subjectively or driven by research objectives. Beyond the flexibility accorded by splines, there is also the monotonicity assumption that has to be adhered to that is a corner stone property within this child development assessment context. In this work, both the sample size and the covariate of interest to make

adjustment for was fixed and varying the number and positioning of knots was found to improve model fit i.e. specifying knots allows more flexibility in the choice and number of the splines especially at younger ages due to the greater rate of developmental rate at these age ranges. Recall also that knots and degrees of freedom are equivalent if the knot points are equally spaced. However, due to the flexibility of the splines, the monotonicity assumption is not always adhered to. Therefore, unless a monotonic spline function can be defined the use of splines may not be viable in this research.

c) Generalised Additive Models (GAM)

Another alternative beyond generalised linear models is generalised additive models (GAMs), where instead of assuming a parametric or linear form of the covariates, one uses an unspecified smooth function estimated using various smoothing iterative procedures. This means combining a function for fitting additive models with the likelihood maximization to find an optimum smoothing function to capture the relation of the binary outcome y and a covariate x. GAMs that were originally extensively discussed by Hastie & Tibshirani, (1986) have recently been well researched and extended especially by Wood, (2006). GAMs are simply a generalised linear model with a linear predictor that depends linearly on unknown smooth functions of the predictor variables. It is an extension of additive models offering very flexible smooth functions to better relate a univariate response variable y to a predictor variable x. Wood's, (2008) work especially emphasises recent penalized regression spline approaches in GAMs for smoothing purposes as well as the mixed model extensions of these models.

However, this flexibility can still inadvertently lead to overfitting even if reasonable types and numbers of smoothing functions are specified. The modelling approach should adequately capture the fact that ability (captured by the item pass rate) increases with age. However, under the GAM framework, this flexibility is not always restricted to be monotonic, thus this approach is not appropriate to model pass rates under an ability measurement context. Further, as shown in the Figure A.1 in appendix A, the fitted model can lead to multiple age estimate solutions at pass probabilities of interest. It is for this

reasons that the classical GAM model will not be considered as a suitable model. Instead, the performance of alternative ways to guarantee monotonicity under the GAM framework or its extension, the SCAM model explained in Section 5.2.1.2 will be considered in comparison to the GLM models.

2.3.2.2. Overall scoring methods

As the name suggests, this approach seeks an 'overall' summary score by utilising 'all' the administered item responses per child in a particular tool domain simultaneously to compute a single aggregate index in the form of a raw score, standardised score or adjusted ability score. Why is this important? The results of a development test tool comprise a sequence of variables (items) whose outcome can be difficult to fully understand and interpret without reference to the original test materials and testing procedures. Also, the outcome of just one item may not be enough to classify the development status of a child. To facilitate interpretation, a child's score on each domain is therefore presented in the form of a transformed raw score, ability score or standardised score. The word 'transform' is used to imply making the score more meaningful, comparable or interpretable.

Obviously, as outlined by several authors but notably Cheung, et al., (2008), Drachler, et al., (2007), Jacobusse & van Buuren, (2007) and Jacobusse, et al., (2006), an appropriate use of the multitude of binary responses to score development status presents a challenge because; items in a child development context differ in difficulty, different children will respond to a different number of items and there is an underlying correlation or association between items caused by the fact that multiple items test the same construct and multiple responses are given by each respondent. Beyond the work of the authors mentioned above, apart from revisiting the merits and demerits of each method, we will offer more suitable extensions based on identified weaknesses, implement them and present their superiority in form of a performance comparison against the former classical methods.

Among others, Cheung et al (2008) pointed out that there is no consensus on the most appropriate method to score child development under the overall score approach, an issue that this work hopes to make a significant contribution to. Further, several of the already suggested advances in scoring methods build on the weaknesses of their predecessor inferior methods; therefore, only resolving a specific method's identified weaknesses. For example, the standardisation of the simple sum score (discussed in the Section 2.3.2.2b) below) is an advancement of the simple count. Therefore, it is still based on simple scores. The standardisation process is meant to mitigate a specific weakness of the simple score which is to eliminate the age effect only; hence there are still some inherent other weaknesses such as assuming uniform item difficulty of simple scores which will still carry on to the standardised score. Unlike previous work that discuss each overall scoring method in isolation or with different emphasis like overall quality of life, our work sets itself apart by pitting all the four classical overall score approaches against each other to facilitate their suitability comparison in this child development context.

a) Simple sum score

An overall simple sum score is created by simply adding the number of binary items passed (Connelly, 2013; Cheung, et al., 2008). This is a 'naive' and misleading method of creating an overall score because it not only ignores the fundamental fact that items differ in difficulty, it presents an even bigger challenge when it comes to using these sum scores to classify development status. The first problem of disregarding item difficulty in computation of the simple score makes it possible that the same score may be inadvertently used to classify two completely different children in terms of their ability status within the same category. Further, these naive scores are strongly associated with age which presents two problems; firstly, older children are likely to gain higher scores due to their greater experience, rather than their actual ability at the assessment time. Secondly, one unique cut-off threshold cannot be used to classify or determine the children's development status. Instead the use

of standardisation methods that are age category specific thresholds are defined and used for the classification process. The standardisation process also in itself presents a few challenges in as much as it attempts to mitigate the short comings of the simple score approach since the selection of age categories has to be appropriately chosen (Greenland, 1995a; Altman & Royston, 2006).

b) Standardisation of simple sum score

A popular statistical approach often used to alleviate the two fundamental flaws described in Section 2.3.2.2a) is to standardise the simple score according to age-specific data to produce various forms of transformed scores. Standardisation therefore serves two purposes; (a) to facilitate interpretation and comparison of raw scores in as far as development classification is concerned. This is because the nature of these score variables can be difficult to fully understand without reference to the original test materials and testing procedures. (b) To eliminate the effect of age because of the tendency of raw scores being strongly association with age. The most popular transformations used being z-scores and percentiles each of which has its proponents based on the methods history, strengths and weaknesses, ease of interpretation, relevance and wide use in different research contexts.

Wang & Chen, (2012) explained that compared to percentiles, Z-scores have a number of advantages; first, they are calculated based on the distribution of the reference population (mean and standard deviation), and thus reflect the reference distribution; second, as standardised quantities, they are comparable across ages, gender, and anthropometric measures; third, Z -scores can be analysed as a continuous variable in studies. In addition, they can quantify extreme growth status at both ends of the distribution. But Z -scores are not straightforward to explain to the public and are hard to use in clinical settings. Hence the growing support to the use of percentiles instead. This is due to the fact that Z-scores are developed based on different principles, data sets and have provided different cut points for the same anthropometric measures; they could, thus, provide different results. Despite being intuitively more understandable, indicative of expected prevalence, percentiles suffer from not

being comparable across different anthropometric measures, extreme values get lumped either to the lowest or highest percentile and are not suitable for assessing longitudinal growth status. Further as we will see in Section 5.2.2.2, the quality of Z-scores is only as good as the quality of the mean and standard deviations used to compute them. This is because summaries used for standardisation in form of look up tables available in test manuals (Elliot et al., 1996) run the risk of not only having sporadic variability, but also more importantly not being monotonically increasing with respect to age. Therefore, they require an appropriate form of smoothing to enhance the sensitivity of cut-off points used for development classification. The following sub-section will review the pros and cons of standard Z-scores.

Standard Z-scores

Simply, a standard score represents the degree to which a child's raw simple sum score deviates from the raw score mean of the 'normal standard sample'. The terms 'normal standard sample' are used literally to refer to the cohort of children living in a promotive environment that is conducive to their development, and have no known development abnormality. The 'normal standard sample' used in this research will be described in Section 3.2.1. The scores of this sample of 'normal' children will be used to characterise normal ability advancement in the other validation samples. The deviation from the mean score which interprets the raw score or tells us what the raw score implies given the normal sample raw scores is expressed in terms of a standard deviation (SD). It is therefore a measure of the distance between the group mean and the assessed child's score.

Z scores are based on the assumption that the mean is zero and the SD is 1. Z scores retain the original relationship between the raw scores. To calculate a Z score, an examiner needs to know the raw score of a child, the mean of the raw score from the standardisation group (M), and the corresponding SD of the normative sample. The formula for calculating the Z score is;

$$z = \frac{X - M}{SD}$$

For example, if a child's simple sum raw score (X) is 20, the mean (M) for raw scores from the standardisation sample is 18, and the standard deviation (SD) for that sample is 10, then the child's Z score is 0.2 (20 minus 18, divided by 10). This means that the child's score fell 0.2 SDs above the mean. If another child's score is 15, and both the mean and the SD are the same, then the child's Z score will be −0.3 (15 minus 18, divided by 10). This means that the child's score fell 0.3 SDs below the mean. As these examples show, if a child's score deviates to the negative side of the mean, then the Z score value will be negative implying delayed development if the classification cut-off threshold is arbitrarily set to be a Z score of 0; if it deviates to the positive side of the mean, the Z score value will be positive implying that they are developing normally. The Z score for a child who scores the same as the sample mean will be zero (no deviation from the mean). Note that it is clinically much more informative to know whether the child performed below, or above the mean for the standardisation group than to say that the child scored 20 or 15 in the example above.

c) Model based Scores

If we consider the simple sum raw score or another continuous outcome that is an overall simple raw sum score, we can use a model based approach to; i) Firstly, produce more appropriate score summaries i.e. smoothed raw scores to get scores that are less variable and monotonic with respect to age (Cole, 1994b; Carey, et al., 2004). ii) Secondly, the modelling framework allows us to adjust for the effect of age that is known to be strongly associated with raw scores and thereby facilitate comparison between children. ii) The model based framework will also enable creation of a confidence band around score estimates or summary measures that express the uncertainty needed to account for variations in test item performance that also varies with respect to age. In the following two sub-sections i) and ii), we will consider the quantile regression framework that has recently emerged as a suitable statistical approach in this child development context (Wei, et al., 2006) and the Generalised

Additive Models for Location, Scale and Shape (GAMLSS) regression framework. The contribution of this thesis is especially seen in the innovative use of the GAMLSS model that besides offering a framework to create confidence intervals also avails several alternative options to compute an overall score based on assumptions made on the item response (sum raw score) distribution.

i. Quantile regression

Petscher & Logan, (2014) notes that using a linear regression analysis would be the most obvious statistical method to model the relation between the continuous overall raw score and age. However, our motivation to explore the use of quantile regression in this project are two fold; a) Firstly, the regression approach only allows for an estimate of the mean (average) relation between the predictor(s) and outcome. Within this child development context, this would not completely meet our objective of being able to quantify how far above or below a child's performance is from the average normal population performance. Therefore, quantile regression is more suitable as it provides estimates of the relation between the predictor(s) and outcome across multiple points of the outcome's distribution. b) Secondly, as noted by Yu, et al., (2003), quantile regression continues to emerge as a comprehensive approach to the statistical analysis of linear and non-linear response models in many important application areas, such as medicine and survival analysis, financial and economic statistics and even environmental modelling. The application of quantile regression especially in the application of growth reference charts in children seems to have taken a life of its own.

Quantiles are especially widely used in preliminary medical diagnosis i.e. screening, to identify unusual children whose value of some particular measurement lies in the tail of the score reference distribution. Both Royston & Altman, (1994) and Cole & Green, (1992) note the need for quantile curves rather than a simple reference range arises when the measurement (and hence the reference range) is strongly dependent on a covariate such as age. Recently, Quantile regression has also been

applied within the multilevel modelling framework to investigate the risk factors influencing child development outcomes (Tzavidis, et al., 2016). The chosen quantiles are usually a symmetric subset of quantile values that are below and above the median quantile value. They are arbitrarily chosen based on research objectives. In this thesis we will explore the use of quantile regression to flexibly capture the non-linear relation between the overall raw score and age that usually entails use of an appropriate spline within the quantile regression framework. However, as noted in the spline section 2.3.2.1, this flexibility also puts the quantile regression approach at the risk of non-adherence to the monotonicity assumption and will therefore not be considered further.

ii. GAMLSS regression with Normal and Beta Binomial distributions

Rigby, et al., (2013) has accredited the focus of the idea of regression analysis where the response variable is allowed to vary according to explanatory variables (rather than just the mean or the variance) to the work of Kneib, (2013), that used other quantities aside from mean in regression. Apart from modelling the mean, modelling the distribution variance as a function of explanatory variables has also been discussed previously by Harvey, (1976) and Aitkin, (1987) for normal models and by Nelder, (1992) and Nelder & Pregibon, (1987), for exponential family models (using an extended quasi-likelihood, EQL, approach). Besides being non-linearly associated with age, (as shown in the scatter plot of the simple count raw score with age in Figure 4.5 in Section 4.4.4) the GAMLSS model framework allows us to appropriately model the shape of the simple sum score (response) distribution that varies according to the explanatory variable age.

Rigby & Stasinopoulos, (2005) explains the ideology that the GAMLSS model is based on; knowing the outcome distribution, the parameters and the mathematical structure of a model in real data situations is facilitated by a reasonable method to compare between different models and a way to check their assumptions. They go on to give the history that preceded its development following the development and use of the LMS method in centile estimation of Cole, (1988) and Cole & Green,

(1992) who were among the first to attempt to model skewness parametrically, as a function of age. Then Rigby & Stasinopoulos, (1996a; 1996b; 2006) introduced the Mean and Dispersion Additive models (MADAM) that used additive terms for the mean and dispersion. However, the likelihood maximisation used frustrated its comparison with other family models. They assume that the response (y) has a parametric distribution $y \sim D(\mu, \sigma, \nu, \tau)$, where μ and σ are usually location and scale parameters and ν and τ are usually shape parameters. Explanatory variables are introduced into the model through predictors that can be linear functions of the explanatory variables or can even take the form of the structured additive predictors (Rue & Held, 2005).

While there is now an enormous amount of literature on the GAMLSS model applications owing to the numerous continuous and discrete distributions the response variable can take, little if any has been written on its use in modelling child ability development outcomes. The GAMLSS approach can be considered in the realm of parametric models as a full distribution that is assumed for the response variable y. Our contribution will showcase the flexibility of the GAMLSS framework to; (a) model the overall probability of success of each child to pass administered items given their age by assuming that the item binary pass/fail responses follows a beta binomial (BB) distribution, (b) smoothen the overall score values that are strongly associated with age or indexes like Z-scores that have unstable variability. This will be by assuming a normal distribution that the overall score or index aspires to follow that facilitates threshold cut-off selection for delay or disability classification purposes, (c) provide a framework to create a confidence bound around score estimates that apart from quantifying variability around scores, serves to also create more sensitive score threshold cut-off values for development status classification.

d) Log Age Ratio score (LAR)

Initially proposed by Drachler, et al., (2007) and also discussed by Cheung, et al., (2008), the LAR approach takes into account the difficulty level of each item by first characterising the ideal 'ability

age' that a typical normal child should pass a given item and pits this estimate against the actual age estimate of a child using their responses in the form of a ratio. The fundamental purpose of the method is to enable the account for the increase in item difficulty of development tools. Therefore, a ratio value of less than one indicates a developmental delay while a ratio of greater than one indicates a normal child or an advance in development.

There are several drawbacks that frustrate this method; firstly, this ratio (score) cannot be estimated for children who either pass or fail all test items because it is difficult to estimate the item pass probability in these instances due to a lack of identifiability and separation within the GLM modelling framework (see Albert & Anderson, 1984; Lesaffre & Albert, 1989; Heinze & Schemper, 2002; Agresti, 2002, for further details on causes and remedies for nonidentifieability in GLM models). Secondly, as the classical LAR employs the GLM-logistic model reviewed in Section 2.3.2.1a), it is therefore likely that the computation of scores using this method is to a large extent contingent on the model fit performance of the logistic model. Put another way, the GLM model fit is likely to be poor for very easy or hard items and also very poor for very young or older children due to nonidentifieability issues. Therefore, in as far as the logistic model is a clincher to the suitability of LAR, unless more superior robust models can be substituted to the logistic model that are far more resilient to skewed item and subject specific item pass probabilities we will not be further considering LAR as a suitable overall scoring approach.

e) Item Response Theory score (IRT)

IRT models that have their roots from Classical Test Theory (CTT, Hedeker, et al., 1998; 2006; Hambleton & Russell, 1993; Lord & Novick, 1968). These models are now an established and well researched technique used for the statistical analysis of human assessment outcomes especially in education and social-health research contexts. There is a substantial amount of literature that chronicles both IRT application (Embreston & Reise, 2000) and theory (De Boeck & Wilson, 2004; van

der Linden & Hambleton, 1997). Within each application context, the IRT framework offers a vast application spectrum both in terms of measurement and explanation as they are used in the test development process to initially aid in item revision (design) and later to help understand why a test shows specific levels of reliability and validity using different item analysis indicators.

The measurement part is often geared towards item analysis and the explanation part is geared towards the person analysis. In fact, because the IRT framework connects the psychometrics and statistics field, such a broad method base to apply throughout the tool development process becomes available to the extent that there is now wide research in Bayesian IRT models (Sheng & Wikle, 2009; Kim & Bolt, 2007), IRT estimation methods (Beguin & Glas, 2001), IRT methods to investigate response dimensionality (Schmeiser & Welch, 2006) to IRT methods for detecting errors and suitability of items (Magis, et al., 2010; Zumbo, 2007). Simply, the IRT framework can be used to carry out almost all statistical requirements of the tool development process. Depending on the item response data type their varying complexity allows us to address a number of important issues that challenge the creation of overall scores including item difficulty, discrimination and different numbers of response items per child, hence our motivation of considering them as a suitable approach in this thesis. However, as is summarised in Table 2.2, the classical IRT models too have specific weaknesses that our suggested extensions will address.

Item analysis indicators include discrimination (a) parameters and difficulty (b) parameters that are examined across the range of the latent variable that are very relevant in this child development context. The item characteristic curve is the regression of the probability of endorsing the item (or, in achievement tests, the probability of getting the item correct) onto the latent variable score, which is the ability of the child variable we wish to quantify. The parametrisation of the various IRT models are intuitive. For example, we will see that the higher the difficulty parameter is, the more difficult that

item is. One can look at the difficulty of an item across the latent ability range, or one can compare difficulty across items at different points in the latent ability range.

We will first consider the one-parameter (1 PL) IRT model, where only the *b* parameter varies (one variant of the one-parameter model is called the Rasch model (Rasch, 1960). This model has been previously discussed and applied by Cheung, et al., (2008), Jacobusse & van Buuren, (2007) and Jacobusse, et al., (2006) in a child development context. Discrimination is identified by the slope of the item characteristic curve at its steepest point. The steeper the curve and the larger the parameter discrimination (*a*) is, the more discriminating the item. As the discrimination (*a*) parameter decreases, the curve gets flatter until there is virtually no change in probability across the latent ability continuum. Items with very low *a* values are not able to distinguish (discriminate) well among children with varying levels of latent ability. We would like to note that this approach does not standardise for age, but it has the advantage that the scores are on the same metric across age and can be compared across age. Under this approach, children who have the same number of passes will have the same score even if they pass different items with different level of difficulty. This may be misleading in this child development context not only because specific items test specific ability constructs, but also the items differ in difficulty. However, because a child development assessment exposes all children to the same set of items, we will consider the 1 PL IRT model in this work.

The two-parameter (2 PL) IRT model is a generalisation of the one-parameter model and allows both discrimination (*a*) and difficulty (*b*) parameters to vary in describing the items. A guessing parameter or lower asymptote can be added to the IRT model to form the three-parameter IRT model and is indicated by the *c* parameter. The third parameter reflects the probability of passing an item by guessing the correct answer (maybe due to experience) at the lowest level of ability. The higher the *c* parameter is, the greater the probability of guessing or selecting the correct answer, even though the latent ability is low. The *c* parameter is often used to model guessing on multiple-choice items and is

therefore not relevant in our research context as in as much as item responses are directed to the child, it is the examiner who decides whether the target ability construct exists.

Our work will take advantage of the IRT framework for computing an overall score and allow item difficulty, discrimination and different numbers of response items per child to change using the 1PL and 2PL IRT models. These 2 IRT models do not also correct for age either. Therefore, this work will therefore consider directly adjust for the age effect in the 1PL IRT model using a generalised linear mixed model framework approach.

Table 2.2: Summary of problems of current statistical methods used to compute age estimate and overall scores for norm creation using binary item response data

Approach	Method	Problems
Item by item analysis	Generalised Linear Models	• Adherence of model assumptions are often not met especially for'easy' items administered to older children or 'difficult' items administered to young children
	Spline Models	• Objectively selecting position and number of knots • Adherence to Monotonicity assumption
	Generalised Additive Models	• Adherence to Monotonicity assumption
Overall scoring	Simple count	• Assumes each item has similar difficulty • Ignores correlation between response items • Scores are strongly associated with age
	Standardisation methods	• Their validity is influenced by the overall score distribution that is often skewed • Ignores correlation between response items • Adherence to Monotonicity assumption of summary measures
	Model based methods: Quantile and [a]GAMLSS regression	• Shares the same limitations of the raw score e.g. assuming uniform item difficulty
	Log Age Ratio methods	• Performance is very reliant on the GLM model used to characterise item difficulty and subject specific overall probability
	Item Response Theory methods	• Lacks a suitable way of adjusting for the effect of age as both [b]1 PL and [c]2 PL IRT scores are strongly associated with age • May be very computer intensive as current software may experience convergence problems due to skewed item pass probabilities or too many response items

[a]Generalised additive models for location scale and shape
[b]One-parameter Item Response Theory model
[c]Two-parameter Item Response Theory model

2.3.3. Post hoc statistical analysis

We can now appreciate that the successful development of an assessment tool whose data are able to adequately classify development status or disability are hugely reliant on good age estimates or scores. Despite its complexity the tool development process is self-checking, hence its cyclic nature. Therefore, even after completion of the statistical analysis there is still an opportunity for further improvement of item quality to enable generalisability of both age estimates and reference norms; this warrants a post hoc analysis. The generalisability exercise encompasses the previously mentioned reliability and validation analysis in Sections 2.3.1.1 to 2.3.1.3 but tailored to be more rigorous with the availability of a larger survey data set to improve item quality but also enable the tool and norm references be used to assess different child populations.

There is also the scaling, linking and equivalence of scores analysis carried out at this last tool development stage that also serve the generalisability purpose but these are not covered in this thesis. The text books 'Statistical models for test equating, scaling, and linking' by von Davier, (2011) and 'Test Equating, Scaling, and Linking: Methods and Practices (Statistics for Social and Behavioural Sciences)' by Kolen & Brennan, (2014) have several dedicated chapters that emphasize the formal statistical principles of the theory and practice of equating, linking, and scaling scores. Our work will instead discuss four post hoc analyses in form of sensitivity analyses, investigating the extent and nature of item missingness, both between, within item correlation and item to total sum score correlation and checking a few age estimate and score distribution characteristics.

2.3.3.1. Sensitivity analysis

All this effort to enhance the current age estimation and norm creation methods is mainly to facilitate the appropriate comparison of different types of children with respect to their age, other demographic characteristics, culture and ability status. The Receiver Operating Characteristic Curve (ROC) as described by Pepe, et al., (2009) and Beck & Shultz, (1986) will be used to compare and contrast the sensitivity of the classical overall scoring methods versus the suggested extensions to the overall

scoring methods. The ROC summarises the performance of any binary classifier that can be summarised in a two by two cross tabulation (as shown in Table 5.2) for various test (score) cut-off thresholds. The ROC is found to be a more attractive tool to assess diagnostic accuracy unlike isolated measures of sensitivity and specificity (Zweig & Campbell, 1993).

We have seen that covariates especially age are strongly associated with the child's development process and therefore strongly correlated with raw ability scores. Several authors including Liu & Zhou, (2013) also advocate for covariate adjustment when they impact the magnitude or accuracy of the test under study. It is also possible that covariates are also correlated with the diagnostic testing procedure i.e. a given test, the chosen score cut-off threshold may be influenced by the subject's characteristics they are applied on and therefore in turn influence diagnostic classification (Janes, et al., 2009). For example, a certain score cut-off threshold may only be able to appropriately differentiate children of a very specific age category. This is usually a consequence of the type of test items included in the tool that are only relevant for a specific age range or type of children.

Hanley & McNeil, (1982) define and describe the use of the Receiver Operating Characteristic Area Under the Curve (ROCAUC) to compare sensitivity performance of classifiers from various disease contexts. We will use the ROCAUC for two purposes; firstly, in comparing ROC curves from the different overall scoring approaches. Secondly in comparing ROC curves drawn from the same scoring approach but groups of different aged children i.e. babies (< 1 year old), toddlers (1 to < 3.5 years old) and pre-school age children (3.5 to < 7 years old). The latter purpose is to assess if there is an age effect on the sensitivity of a scoring approach.

The pros and cons of using ROCAUC to summarise test sensitivity and specificity across various cut-off thresholds has been well summarised and argued by several authors including Hand, (2009), Fawcett, (2006) and Bradley, (1997). However, you will notice that each of the arguments is anchored on a specific research context or application. In the same spirit we argue that the suitability of ROCAUC in a child development assessment also depends on context and more specifically the behaviour of the

ROC curves over the entire threshold spectrum. If the classification objective is in a screening scenario, then ROCAUC may be suitable as interest lies in identification of 'suspect' delayed children. If the scenario is a formal diagnosis classification then in line with the demerits outlined in the literature for example when the ROC of different scoring methods or different aged children cross, then ROCAUC may not be suitable. As outlined in the Standards for Reporting of Diagnostic Accuracy (STARD, Bossuyt, et al., 2015) statement initiative, if the prevalence of a condition, its spectrum or study design influences the ROCAUC then the measure of diagnostic accuracy should be appropriately justified. Here, the objective of using ROCAUC will only be to highlight if there is indeed any difference in diagnostic accuracy between methods, and to what extent the child's age affects the difference in scoring method accuracy.

2.3.3.2. Missingness in test item responses

While carrying out a development assessment, an examiner is bound to face the challenge of not being able to discern with certainty whether a child can or cannot pass the test item either due to lack of ability, or the refusal to respond as consequence of another underlying issue. The missing data phenomenon is not a unique preserve of child development assessment surveys only. Because a particular attribute is measured repeatedly from the same child at one instance, or over a period of time, it is quite common to have child subjects with missing values/item responses for one or more items in the assessment tool. The underlying issue preventing an expert from discerning whether a child can pass an item could be anything from the child being too ill, the child being distracted by tool probing devices that are culturally foreign, to even being intimidated by or scared of the environment or examiner (Harzing, 2006). Rightfully, we also note that the degree of missingness could be an indicator of a poorly translated or adapted tool as well as a poor administration protocol.

Blom, et al., (2010) has asserted the importance of understanding bias introduced by non-response in the context of comparability of child language development multiregional studies. The type and extent of item response missingness often has a myriad of effects that negatively influence the validity of

inferences made on the computed age estimates or scores depending on the assumptions made about the reasons for the missing data during analysis. Dong & Peng, (2013), Tabachnick & Fidell, (2012) and Madow, et al., (1983) discuss various forms, patterns and extents of missing data giving various recommendations of dealing with each, while Kline, (2005) notes that the more sophisticated methods of dealing with missing data are useful if there is an identified systemic pattern and usually require specialised software and a sensitivity analysis to compare analyses with and without missing analyses.

The first issue in dealing with the missing data problem is to explore whether the missing data mechanism has distorted the observed data or introduced bias, and subsequently also compromising any inference to be drawn from the statistical analysis. Little & Rubin, (1991) formally distinguish between three missing data mechanisms that we briefly describe. As explained also by Briggs, et al., (2003) and Allison, (2001) data are said to be missing at random (MAR) if the mechanism resulting in its omission is independent of its (unobserved) value. If its omission is also independent of the observed values, then the missingness process is said to be missing completely at random (MCAR). In any other case the process is missing not at random (MNAR), i.e., the missingness process depends both on the observed and the unobserved values. Especially in this later missing data pattern, suitable methodology of dealing with the missingness should be considered.

Due to the concern on the negative impact of the missing mechanism will have on the quality of age estimates or norms and the drawn inferences, there is a vast amount of literature describing both rudimentary and advanced methods to handle missing data e.g. Nakai & Weiming, (2011) and Graham, (2009). These methods can broadly be classified into four groups: (a) Complete Case Analysis; where when some variables (items) are not observed for some of the units (children), one can omit these units with missing items responses from the analysis and only analyse the so-called ‘complete cases’ only. (b) Imputation-based methods; where one replaces each missing value with one or more than one imputed value and aggregates results of the different imputations. The goal is to combine the simplicity of imputation strategies, with unbiasedness in both point estimates and measures of

precision. For example, using confidentialised unit record data from the Household, Income, and Labour Dynamics in Australia (HILDA) survey, Watson & Wooden, (2012) discussed item non-response and the effect of various imputation strategies of analysis methods e.g. wage regression that are used as comparative indicators in cross-national surveys. (c) Weighting methods are a third approach based on the complete cases but now weighting them with the inverse of the probability that a case is observed as discussed for example by Seaman & White, (2013) and Zhao & Lipsitz, (1992). In this way cases with a low probability to be observed gain more influence in the analysis. (d) Finally, the use of fully model-based procedures also known as available case analysis; which rely on modelling the partially missing data using estimation methods such as maximum likelihood (Schafer & Graham, 2002). These have been extended as described in more recent literature using semi-parametric techniques to relax the missing assumptions made (Hens, 2005). This work will investigate the source and extent of missing item responses and whether any form of imputation method is necessary in Section 4.3.2.

2.3.3.3. Item to Total Correlation, Correlation within and between test item responses

Assuming that the different item responses from the same child are independent may not be entirely true for the following three reasons. Firstly, it is likely that the responses to different items by a given child will be correlated as these responses will all depend on their subject specific ability or their environment. Secondly, two adjacent tool's items may be related as they test the same underlying construct or ability milestone at differing difficulty degrees. Thus it is likely that a child will respond in a similar fashion to two or more items that are adjacent to each other in the tool due to item ordering as they are testing the same developmental construct. Thirdly, correlation may arise due to underlying population characteristics, or as a consequence of the study design leading to clustering of responses. For example, the target population may be a sample of children of the same age with a common underlying health condition or disease like malaria that is known to cause specific delay outcomes e.g.

onset of walking at one year. Hence it is likely that all responses testing the walking ability construct for one year old children will be lower than is expected (Bangirana, et al., 2014). The two overall scoring approaches using simple sum scores and Z-scores considered in Sections 5.2.2.1 to 5.2.2.2 ignore the correlation within a subject's item responses. Only the IRT framework considered in Section 5.2.2.3, makes the assumption of local independence i.e. the observed response items are conditionally independent of each other given an individual's score on the ability latent variable. This means that the latent variable explains why the observed item responses are related to each another as discussed by Lazarsfeld & Henry, (1968) and Henning, (1989).

The polychoric correlation coefficient (Drasgow, 1988) is usually used to compute the required correlation between items. Analogous to Pearson's or Spearman's correlation coefficients, polychoric correlation is a measure of association but for ordinal variables with the underlying assumption of a joint continuous distribution. It is a technique for estimating the correlation between two theorised normally distributed continuous latent variables from two observed ordinal variables. In this case the continuous latent variable is 'ability'. The method is frequently applied to assess correlation of personality test responses that have rating scales with a small number of response options. The smaller the number of response categories, the more correlation between continuous variables will tend to be attenuated. This method will also be used to assess and quantify the item to total correlation. In our work that involves binary data, the contribution of the item will be removed from the overall raw score before the item and overall correlation is computed as we will be mainly interested in assessing that the correlation is within an acceptable level which implies the suitability of that item. Ekstrom, (2011), Bonett & Price, (2005) and Lee, et al., (1995) have written extensively on this method of correlation calculation.

2.3.3.4. Important Assessment Tool, Age Estimate and Score Properties

This section highlights three key tool attributes that are important to bear in mind while choosing a source tool to adapt and which should carry over to the final adapted tool. These attributes discussed

in sections a) to c) below can quickly and easily be discerned from the tool's milestone charts of age estimate or score summary measures. Hence we note that with appropriate age estimation and scoring methods, we are not only able to better assess if these attributes are present in appropriate form in the new adapted tools, but we also are able to improve their quality.

a) Ordering of item age estimates and overlap

Section 2.3.2 has reviewed the statistical methods to convert binary item responses into meaningful reference age estimates or overall scores for ability classification. These age estimates or score are summarised in form of graphical milestone charts such as the one shown in Figure 3.5. We would like to note at this point that it is important and reasonable for the age estimate values to overlap to a reasonable degree for the following reasons;

1. If the age estimates of two adjacent items considerably overlap, it indicates that both items are in essence testing the same construct to the same degree of difficulty. Hence, both items may not be required unless they are used as substitutes of each other in the event that one of them cannot be assessed.
2. There should be no 'gaps' between age estimates at the percentile of interest between subsequent test items. The presence of a gap implies that the tool does not have an item that can test the ability trait over the length of the gap for age.
3. The age estimates for both the lowest and highest percentile of interest should always be monotonically increasing. This is to reflect the design of development assessment tools that are purposefully designed to increase in difficulty so as to differentiate specific items that a child can pass and those that they cannot in order to quantify their ability status.

The source and cause of the above characteristics if poor could be caused at the item/tool design stage where an item testing ability at that age is lacking, or could be a manifestation of variability of data causing a great shift in age estimates between subsequent items, or a weakness in the item by item

analysis that computes the age estimates as a consequence of inferior statistical item modelling methods. Such are the issues that we hope our suggested robust statistical methods will address.

Figure 2.2 below is a mile stone chart with age estimates for the 25th, 50th, 75th and 90th probability percentiles of passing an item from the work done by Wijedasa, (2012) in Sri Lanka using the Denver Developmental Screening tool. It presents an example where the above three attributes are lacking. This lowers the 'confidence' of using the tool in its current form to classify development status. The red dotted box highlights a gap in the personal social domain between the fourth and fifth items. Therefore, this chart cannot adequately assess the ability of a child who is four to six months old or be able to differentiate their ability appropriately. Also, notice the extent of overlap of subsequent items as shown in the blue dotted box. The rather significant overlap in the three items in the personal social domain suggests that these items could be testing the same ability construct for the ages 15 months to approximately 24 months.

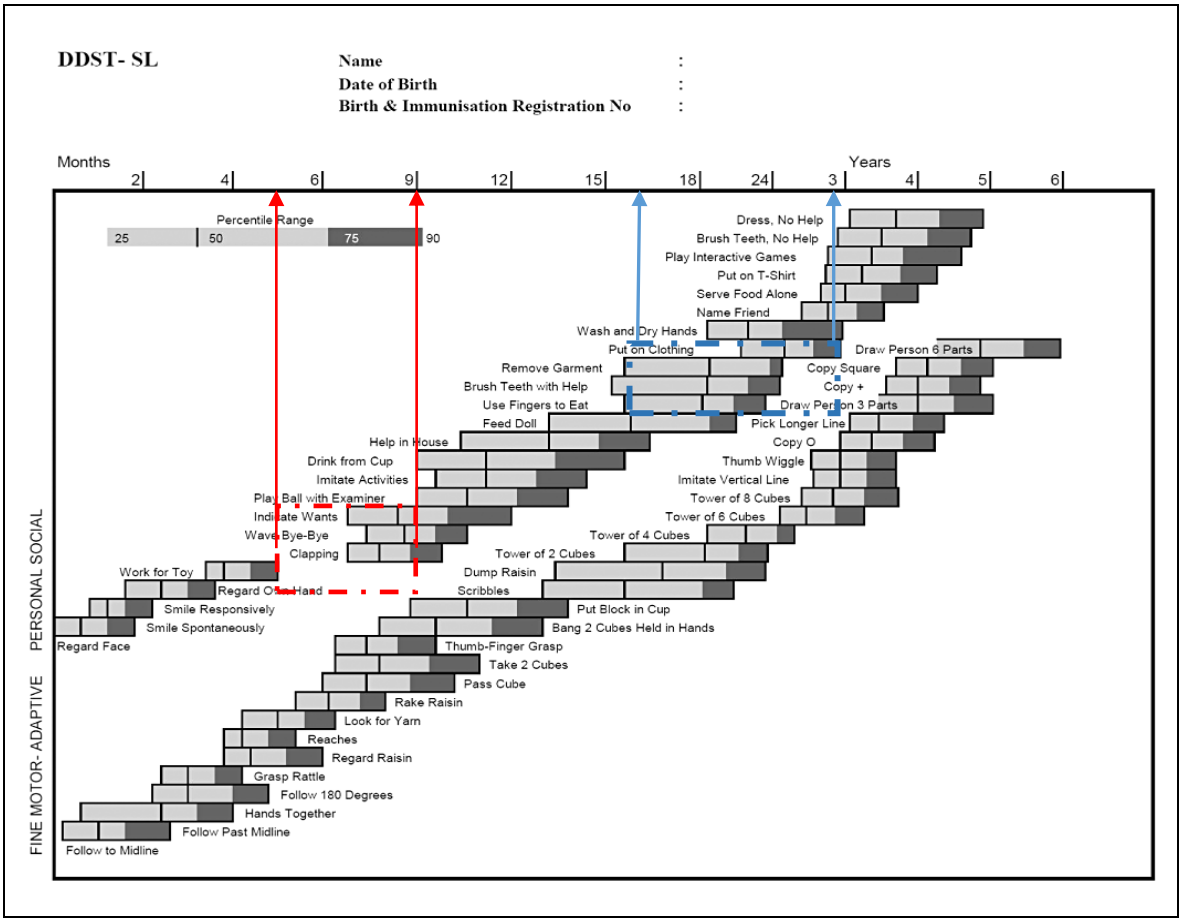


Figure 2.2: An extract from the Denver Developmental Screening Test for Sri Lankan Children (DDST-SL) showing gaps (red dotted box) and too much overlap (blue dotted box) in tool item age estimates. *Adapted from Wijedasa, (2012).

b) Evaluation of assessment Score Cut-off thresholds

Stopping rules in the context of developmental assessment tools refers to the criteria used to decide if a child can no longer pass subsequent items i.e. the child has reached their ability defining threshold. To the best of our knowledge this issue of identifying an objective criteria to stop testing a child has not been adequately discussed within a child development context.

The last item passed defines the child's ability milestone given their age which subsequently classifies their development status. This begs the question 'when do you stop testing?' A too conservative stopping criteria may be too stringent and therefore underestimate ability, and at the same time a less conservative approach may mislead or overestimate the true ability classification of a child. Often, the same stopping rules stipulated by the source tool are directly used in the new setting without any

consultation on their suitability in the new setting. For example, the stopping rule used for the MDAT tool described in Chapter three was directly adapted from the source tools used to compose the final tool. We wish to make the point that regardless of the stopping rule used, or whether an objective method has been used to select it, its degree of conservativeness has a direct bearing on the sensitivity performance of the tool. However, in this thesis we do not investigate this issue as this would require a comparison of two or more studies run in parallel, or a simulation study with exactly the same characteristics but different stopping rules and assess the impact of changes in stopping rules on the sensitivity of ability status classification.

c) Score Distribution and Monotonicity assumption

The univariate score distribution or with respect to age is key in informing the suitability of the threshold cut-off methods and values used for development status classification. This is because the cut-off method's appropriateness is pegged on adherence to the various computed score distribution assumptions. For example, we saw the relevance of the normal distribution in Section 1.1.2 above to give reassurance that the normal scores are from a representative sample and hence the normal distribution property can be utilised for disability status classification. Using several educational and psychological experiments, the work of Micerri, (1989), Cook, (1959) and Lord, (1955) have discussed departures from normality in educational test scores. They showed that the normality assumption is rarely met but is usually negatively skewed and platykurtic.

In this child development context, apart from score distribution adherence to underlying model assumptions, there are two other important implications; firstly, the score distribution influences the choice of overall score transformation as discussed in Section 2.3.2.2 above. Secondly, it is important to consider the score distribution with respect to age to advice on the most appropriate method to choose cut-off thresholds needed to classify the development status of children of different ages. This is where the monotonicity assumption becomes relevant since as ability trait levels increase (with increase in age), older children are expected to pass more items of achieve higher overall scores.

Therefore, if the scoring method does not correct for age, higher cut-off thresholds will be required to correctly classify the development status of older children with higher item pass rates.

Section 5.3 will explain in detail our two goals in as far as the relevance of the score distribution is concerned. The first is to check the adherence of the overall score to the normality assumption so as to advise on the most suitable method to choose cut-off thresholds. Secondly, depending of the scoring method used we will assess the distribution of scores with respect to age and check whether the age effect on scores has been eliminated as well as investigate if the monotonicity assumption has been adhered to. This is especially to augment the method used to choose and apply more sensitive cut-off thresholds on overall scores needed to classify the development status of different aged children.

2.4. Summary of the systematic review

To review and take stock of the currently applied methodology used by researchers and make a detailed assessment of the translation, adaptation, age estimation and overall scoring methods that are currently used in practice, a formal systematic review was conducted. A detailed account of the systematic review methodology, data extraction processes and explicit review findings are available in a separate report. This section summarises the key points revealed from the current state of methods employed in the development of child assessment tools and scoring in developing countries.

While some studies attested to the growing interest in the application of appropriate methods to culturally adapt assessment tools, several studies also declared the importance of or need for developing culturally adapted tests. To achieve this, adequate and appropriate application of translation, adaptation, scoring and statistical methodology in the assessment tool development are necessary. Following the realisation of the lack of uniformity in reporting of translation and adaptation studies and the statistical methods involved in age estimation, scoring and norm reference creation, a quality criteria checklist with 25 pertinent qualities to assess the quality of reporting of child

assessment tool development studies was developed. The developed quality assessment checklist scored and classified a retrieved reference as either poor or of minimal adherence, as having partial and moderate adherence, or extensive to excellent adherence to the application recommended scientific research methods in this regard.

Section 2.3.1 has discussed the most commonly mentioned and used statistical procedures involved the establishment of agreement between source and final tool developed. The analysis of agreement mostly took various forms of correlation to assess reliability especially the use of Cronbach's alpha. Few studies mentioned and used classical statistical procedures such as logistic regression, analysis of variance (ANOVA), analysis of covariance (regression, ANCOVA) or the more sophisticated scoring methods including Rasch models and IRT (item response theory). Even fewer studies adjusted for the effect of age on overall scores. This realisation offers the evidence for the importance and relevance of the work of this thesis.

The systematic review confirmed that while some studies attested to the growing interest in the application of appropriate methods to culturally adapt assessment tools, several studies also declared the importance of or need for developing culturally adapted tests using scientific methods. More importantly, very few studies carried out any analysis to assess scoring and for the creation of norms as well as to adjust for the effect of age.

2.5. Summary

Section 2.4 has summarised the key findings of the systematic review discussing both aspects that need more work and at the same time areas that have made tremendous advancement. In conclusion to the recommendations outlined above; firstly, the application of appropriate methodology of translation, adaptation and norm creation methods depends on primarily on available resources and research objectives. Secondly, the continued demand for both higher quality methodological application accompanied by both adequate and uniform reporting of the process of assessment tool

development studies will improve the current state and practice of this research. The quality assessment criteria checklist described in Section 2.4 pointed to pertinent qualities we should seek in a tool that we use as a source for adapting. Further, these qualities should also be carried on to the final tool that is the output of the translation or adaptation process. These can be summarised as follows;

- Look for a tool with a comprehensive manual describing the normative or standardisation sample, reliability and validity levels as well as its implementation procedures.
- Look for a recently adapted tool, that pairs well with your research objectives and expert capacity as well as being extensively researched and reviewed in literature. This will save you on important resources and avoid a false start in your research as most of the current translation adaptation or scoring issues you may be unaware of will most likely be already addressed and reported.

The important issues of methods involved in the development of child assessment tools and how exactly they influence the statistical methodology used in the creation of scores have been identified and discussed. We have gone further and given a summary of the collected hard evidence of current practice in the translation and adaptation as well as the statistical methods used in the computation of child development assessment scores. Therefore, we can now appreciate the potential gaps in this research area. The next chapter will describe the adapted tool that was used to collect the data used in this thesis to extend current scoring and norm creating methods.

PART II – METHODS

Chapter 3. The Malawi Development Assessment Tool (MDAT)

Chapter 4. Data and Exploratory Data Analysis (EDA)

Chapter 5. Methods of Scoring Binary Response Item Assessment Data

3. The Malawi Development Assessment Tool

3.1. Introduction

The objectives of this chapter are: a) to introduce and describe the assessment tool used in this project, and b) to qualify the fact that if one aspect of the tool is not of good quality, then the quality of every other subsequent stage leading to the creation of age estimates, overall scores or reference norms may be compromised. In particular, even with robust scoring methods, the sensitivity to classify the disability status of children may still be compromised.

The creation of the Malawi Development Assessment Tool (MDAT) was through an elaborate mixed method process involving both qualitative and quantitative strategies that is well documented in Gladstone, et al., (2008; 2010a; 2010b). The items in the MDAT tool were mainly taken and adapted from some of the popular established tools including the Denver Developmental Screening Test (DDST) II, Ages and Stages Questionnaire (ASQ II), Ages and Stages-Social Emotional (ASQ-SE) tool, Bayley Scale of Infant Development (BSID III) tool, Bayley Infant Neurodevelopmental Screener (BINS) tool and Griffiths Mental Development Scales tool.

Section 3.2 introduces and describes the MDAT study. Sections 3.2.1 to 3.2.6 describe aspects of the MDAT tool including the assessment tool kit with culturally appropriate probing props used while carrying out assessment, the elaborate process of recording responses from the MDAT tool up to the point of data entry ready for formal statistical analysis to compute age estimates and create reference norms. Section 3.3 concludes the chapter with some recommendations for selecting a high calibre tool to use as a source for adapting or qualities that a tool should have to ensure high quality data is collected. With an appreciation of the data collection instrument and its vital role, the stage will be adequately set for the reader to easily understand the data characteristics described in the fourth chapter.

3.2. The Malawi Development Assessment Tool study

This section describes the MDAT study (i.e. the country and regions where the study was carried out), the development process used to develop the MDAT tool items, item descriptions and item response recording on the MDAT assessment chart and data spreadsheet.

3.2.1. The MDAT Study Population Description

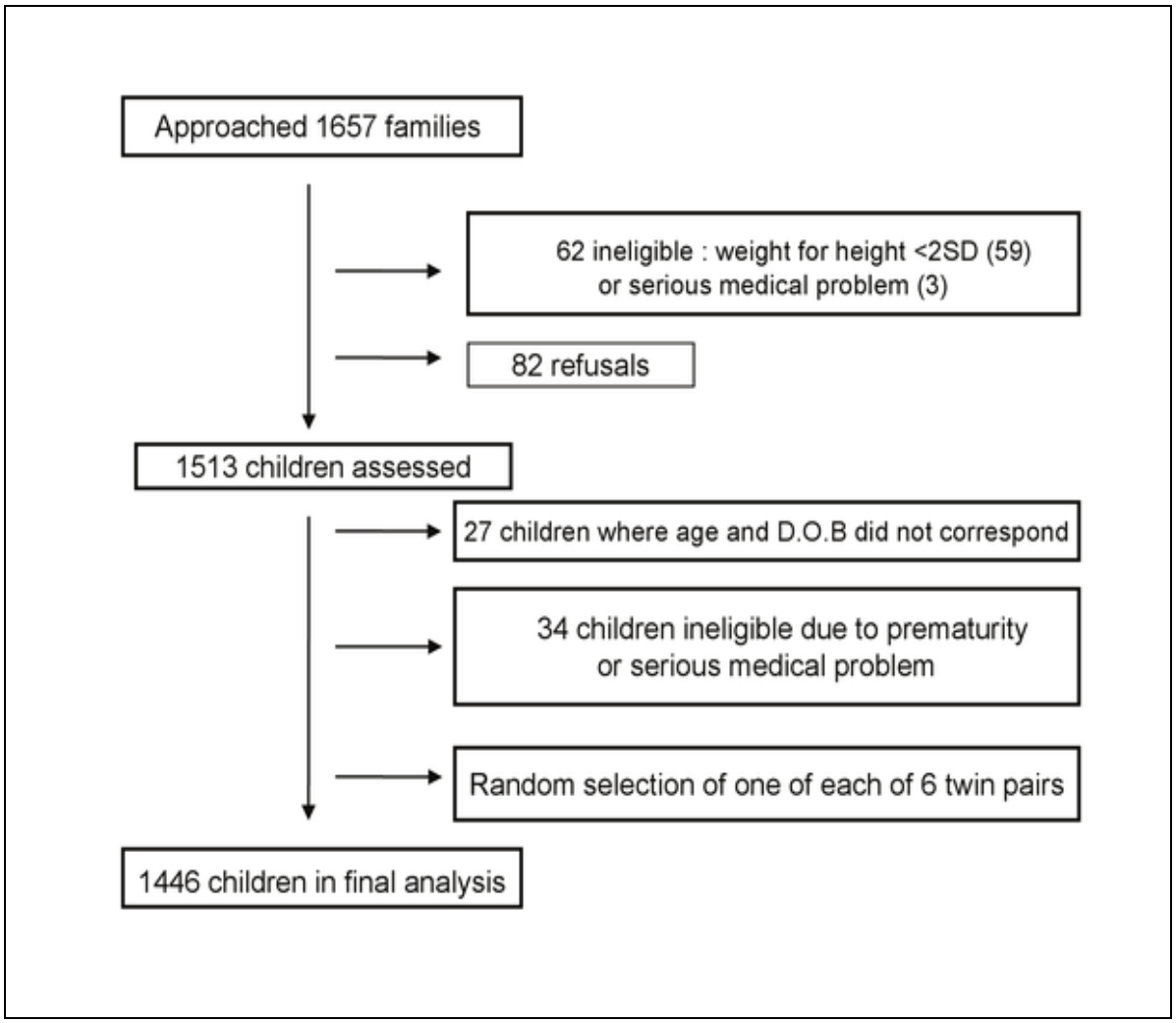
The Republic of Malawi is a landlocked country in Southeast Africa. It is bordered by Tanzania to the North, Zambia to the Northwest and Mozambique to the East, South and West. It is among the world's least developed countries according to the Human Development Index as included in a United Nations Development Programme's Human Development Report last released on March 2013 and compiled on the basis of estimates for the year 2012. The MDAT 2007 study population come from Blantyre, a district in the Southeast of Malawi as shown in Figure 3.2 below. The capital city, also called Blantyre, is in this district that covers an area of approximately 2,012 km^2 with a population of about 809,397 people.

To recruit children, one in every three mothers in clinic was requested to bring one child to their next appointment. A quota sampling technique similar to that used by Frankenburg, et al., (1992) where target numbers of children age between zero and seven years spanning 34 age groups were sought. Children's ages were determined from available birth data or from the 'health passport'. This is a document in form of a card that mothers in Malawi carry with them to all their health appointments (the document is used to record relevant child health data such as their birth dates, vaccination data or any growth measurements taken). Once enough children of a particular age range were recruited, no more children of that age range were invited to participate. Children of ages where there were inadequate numbers were targeted by requesting mothers to only bring children of those ages.

Healthy and normally developing children born to mothers attending post-natal clinics (one per family) without any known form of delayed development or disability between the ages of zero to seven years meeting the study inclusion criteria and who were receiving appropriate medical support were

included in the MDAT 2007 study. Those with significant malnutrition (weight for height Z score ≤ 2 using WHO (2011) criteria), significant medical problems, prematurity of 32 weeks or less (reported or measured on antenatal ultrasound), or significant neurodisability were excluded as the aim was to create a developmental assessment tool that identified children with developmental delay. Relevant sociodemographic characteristics were collected using similar questions as the 2005 Malawi Demographic Health Survey.

Figure 3.1: The MDAT Flow diagram of the recruitment of families and children for the MDAT study.



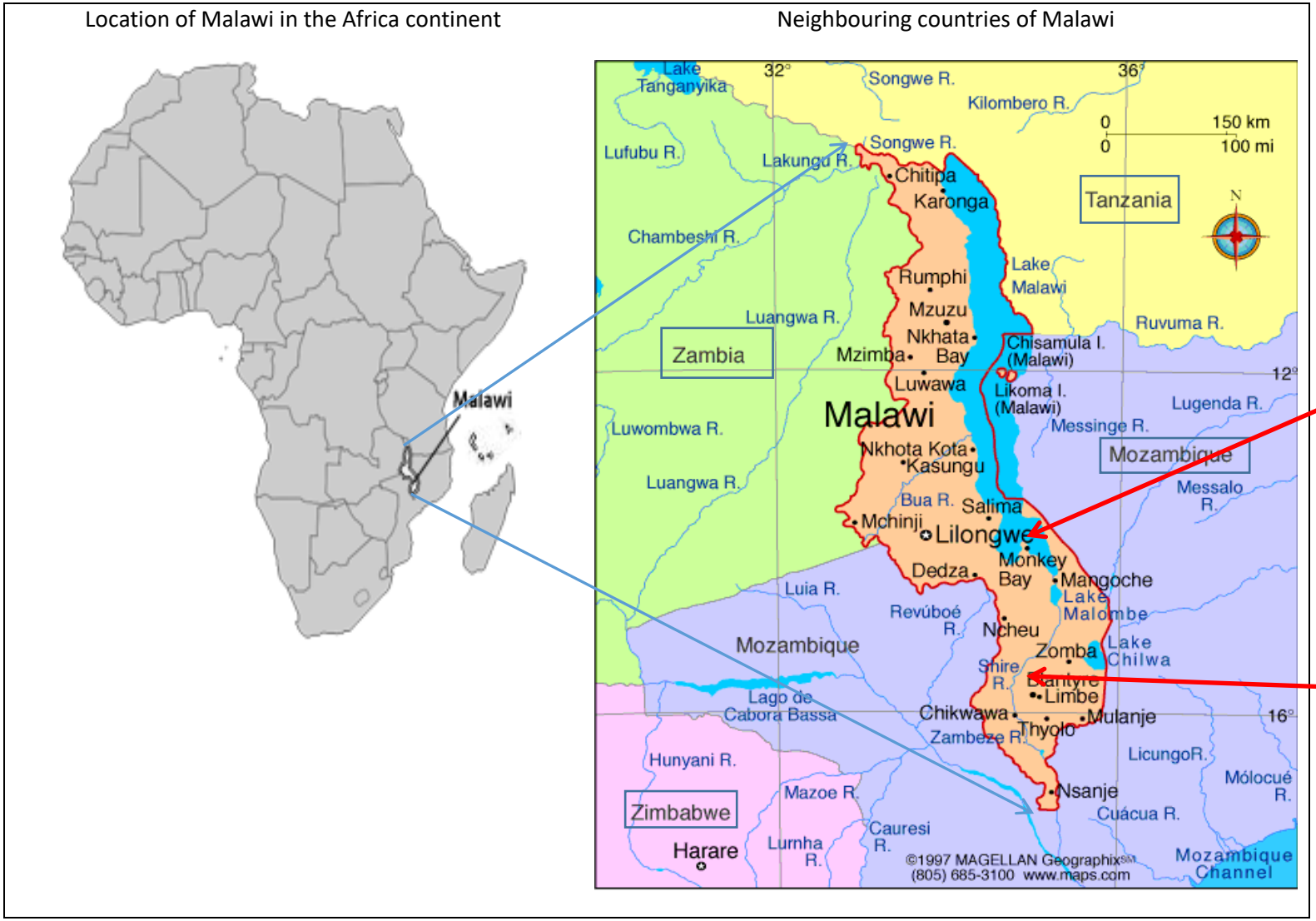
Source: Gladstone, et al., (2010a)

Therefore, as shown in Figure 3.1, to test the performance of the final MDAT draft III, a total of 1,513 children from four sites in the Southern region of Malawi were recruited and assessed. These were three rural and one semi-urban sites (Namitambo, Mikolongwe, Nguludi, and Bangwe), which were all

taking part in a larger antenatal trial. Assessments occurred over a year from June 2006 until July 2007 using a team of six well trained research midwives in local antenatal clinics in each of these areas.

There is an additional cohort of 120 children who were classified as malnourished according to the recommended WHO stunting and malnutrition criteria (2005) i.e. -2 SD below the norm for Height for Age Z scores from the median Height for Age of the reference population was recruited. However, for this study, very severely malnourished children (less than -2 SD weight for height Z score) were excluded. Another cohort of 80 children formally diagnosed with a known neurological disability by child experts was recruited from Cheshire and Moyo care centres that care homes for disabled children in Malawi. These two non-normal cohorts will be used for validation purposes of the suggested statistical scoring methods once the respective age estimates, overall scores and norms for developmental milestones have been developed using the standard normal child cohort data sample.

Figure 3.2: The MDAT study data collection sites.



3.2.2. The MDAT Item Development Process

The work of Gladstone, et al., (2008; 2010a; 2010b) has highlighted the importance of applying scientific methods to create a culturally appropriate tool by translating and adapting established western tools. In turn more robust statistical methods have a better chance to make more accurate age estimates, scores and norms for children from a different cultural setting. As was described in Section 2.3, the commonest tool development process is usually iterative where item quality can be improved at every stage of development. The intricate process of the tool development is adequately described in many handbooks e.g. 'A handbook of test development' (Downing & Haladyna, 2006). Table 3.1 shows a summary of the seven steps MDAT development process.

Table 3.1: A summary of MDAT item development process.

Phase	Stage	Activity
1.Tool development	1. Item selection	Tool background re-analysis of qualitative data to selected items to include
	2. Item review	Using expert steering groups, focus groups and methods such as †Delphi to review items
	3. Pre-testing	Small scale interviews to test and train administrators and resolve any identified issues
	4. Piloting	Piloting and resolving issues by consensus meetings
	5. Data collection, norm creation and reliability and validity testing	Resolve any pending or new realised issues before releasing final tool
2. Further quantitative (statistical) evaluation of psychometric properties of the tool items	1. Applying statistical methods to improve tool quality	Quantitative item reduction, psychometric evaluation and sensitivity analysis
	2. Final tool release	Collect item data and classify developmental status

†Delphi, see Linstone, (1975), Brown, (1968) and Norman, (1963)

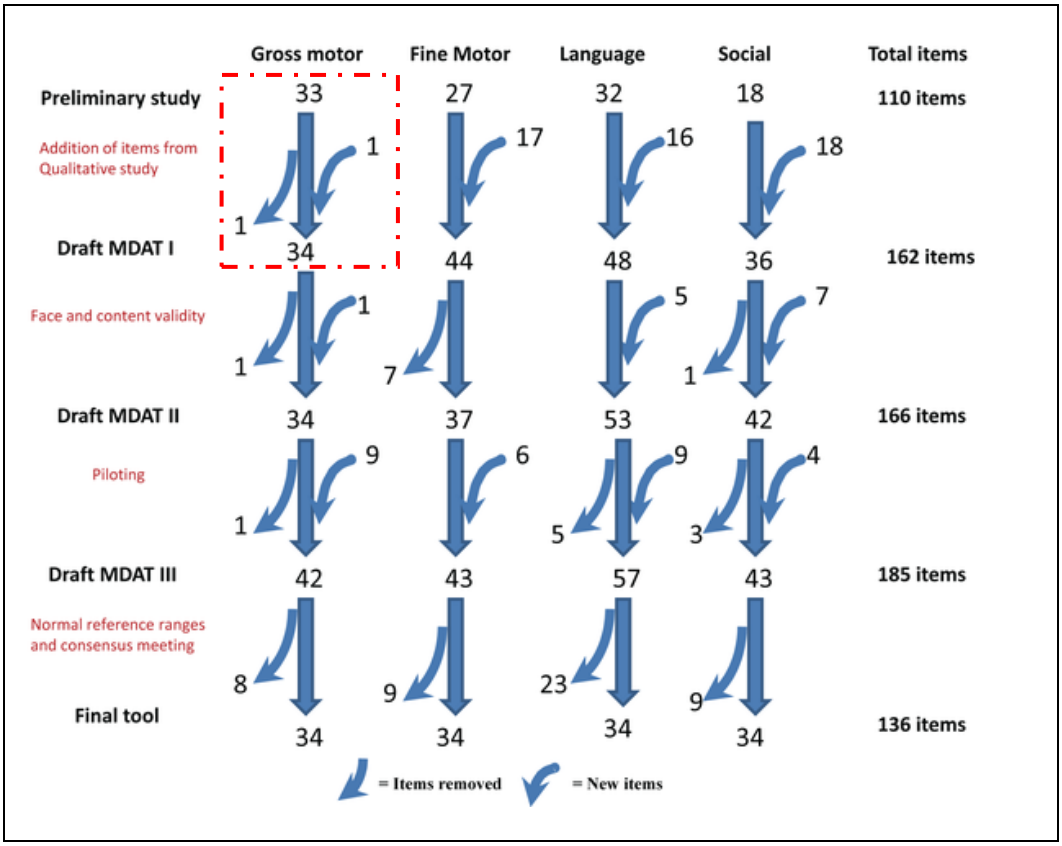
As described in detail in Gladstone, et al., (2010a), following preliminary and qualitative studies, the initial MDAT draft developmental assessment tool had 162 items in four domains of development. Figure 3.3 shows the stages in the creation of items included in the final MDAT tool where after face and content validity testing and piloting, the draft tool was expanded to 185 items. As described in Section3.2.1, 1446 normally developing children aged zero to seven years from rural Malawi were assessed. The performance of items was examined using logistic regression and reliability using kappa

statistics. All items were then considered at a consensus meeting and any items performing badly, those that were unnecessary or difficult to administer were removed, leaving 136 items in the final version of the MDAT.

This was then followed by various forms of reliability and validity analysis including the resolving of issues by consensus as recommended by Pope, et al., (2000). The MDAT tool was externally validated by comparing age-matched healthy normally developing children with those who had a potential to be delayed in ability development or disabled from the (neuro) disabled (n=80) and malnutritioned (n=120) cohorts. The remaining items in the MDAT had good reliability values of 94% to 100%. All included items used for scoring had kappa values that were greater than 0.4 for interobserver immediate, delayed, and intra-observer testing. A significant differences in overall mean scores (and individual domain scores) for children with neurodisabilities (35 versus 99 [p<0.001]) when compared to normal children was demonstrated, see Gladstone, et al., (2010a).

The initial analysis carried out during the development of the MDAT tool, using a pass/fail response technique similar to the Denver II, found that 3% of children with neurodisabilities passed in comparison to 82% of normal children, demonstrating good sensitivity (97%) and specificity (82%). Overall mean scores (computed by simply adding passed items, see Section 2.3.2.2) of children with malnutrition (weight for height <80%) were also significantly different from scores of normal controls (62.5 versus 77.4 [p<0.001]). Further, the scores in the separate domains, excluding social development, also differed between malnourished children and controls. In terms of pass/fail, 28% of malnourished children versus 94% of controls passed the test overall. By applying more robust scoring methods, this work hopes to significantly improve these initial sensitivity estimates.

Figure 3.3: Stages in the creation of final MDAT tool.



Source: Gladstone, et al., (2010a).
Note the error in arrow labelling in preliminary study of GM domain in red dotted box; two items should have been added and one item removed.

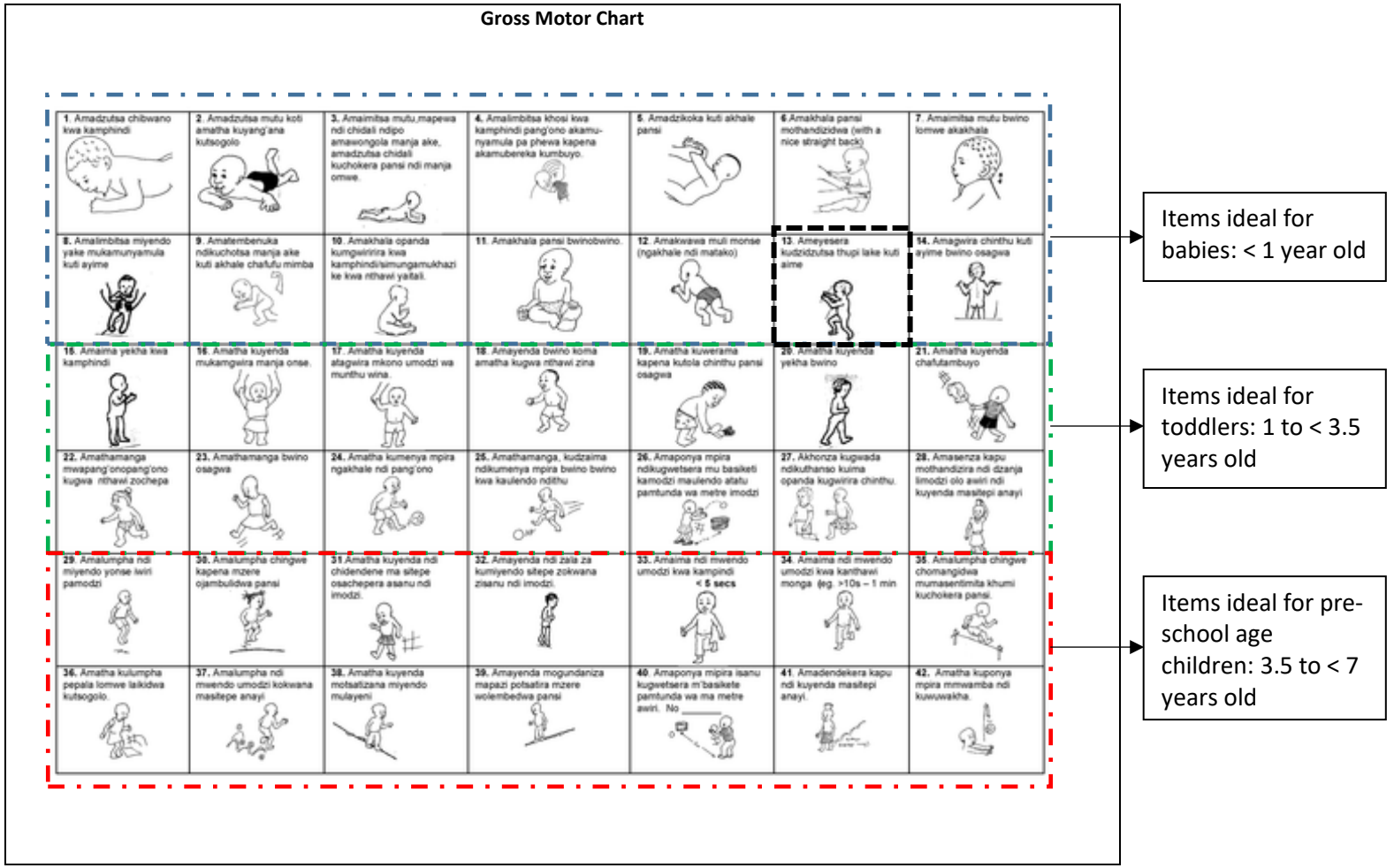
3.2.3. The MDAT Tool Item Description and Characteristics

The MDAT tool is composed of four domains (Gross motor, Fine motor, Language and Social skills) each with 34 items, designed to assess ability development in children who are zero to seven years of age and specifically developed for use in Malawi. Figure 3.4 shows the actual MDAT developmental questionnaire chart for the gross motor domain, showing items that are ideal for babies (< 1 year old), toddlers (1 to < 3.5 years old) and pre-school age children (3.5 to < 7 years old) in the blue, green and red dotted boxes respectively. The pictures for each item were developed by an artist to vividly portray what each item or task was evaluating as an aid during its administration. As explained in the MDAT manual, this form of the chart's structure allows its practical use at the place of examination during assessment.

In tools that measure achievement or ability, items are usually arranged in increasing order of difficulty. Since age has been shown to be strongly correlated with ability scores, then it is expected that as age increases, the dexterity of the child and therefore ability of the child should or is expected to increase monotonically. A logical way to measure ability development is thus to challenge it to the extent of neatly revealing a subject specific item beyond which a particular child can no longer convincingly pass any more administered tasks depending on the stipulated tool's assessment stopping rule. By using a score threshold computed from the number of items passed, the child's ability status can then be classified as either delayed, suspect delayed or normal by its position (value) in comparison to other normal children.

The MDAT items are also arranged with increasing difficulty as it is expected that a child's dexterity should increase with age. In an ability assessment context, this ascending ordering of items with respect to their difficulty also serves the purpose of exhuming confidence to encourage a child to deliver on various harder tasks given that they can deliver easy ones. An expert is advised to start administering easy items to continuously encourage, build a rapport with the child and simultaneously make them feel comfortable as the assessment proceeds to more difficult tasks i.e. the difficulty of items should be increasing monotonically. The MDAT manual recommends starting the administration of items with the Personal-Social domain, then the Fine Motor domain, the Language domain and finally the Gross Motor domain. This order allows for both examiner and child rapport establishment with items requiring less active participation of the child to be administered first and items requiring greater ability and confidence to be administered last. The scoring methods chapter will especially reflect how best to accommodate this item monotonicity feature in the computation of scores.

Figure 3.4: MDAT gross motor tool chart.



Items that are ideal for babies (< 1 year old), toddlers (1 to < 3.5 years old) and pre-school age children (3.5 to < 7 years old) in the blue, green and red dotted boxes respectively. Black dotted box shows starting item to assess a 1 year old child.

Table 3.2: Gross motor item description list

Type of item	Item #	Gross Motor domain item descriptions
Ideal items for infants <1 year old	**1**	GM1 - lifts chin up off floor for a few seconds
	2	GM2 - prone, head up to 90 degrees
	3	GM3 - holds head straight or erect for few seconds
	4	GM4 - pulls to sit with no head lag
	5	GM5 - lifts head, shoulders and chest when prone
	6	GM6 - bears weight on legs
	7	GM7 - sits with help
	8	GM8 - rolls over from back to front
	9	GM9 - sits without support for a period of time
	10	GM10 - sits by self well
	11	GM11 - crawls (in any way)
	12	GM12 - pulls self to stand
Ideal items for toddlers 1 to <3.5 years old	**13**	GM13 - able to stand if holding on to things
	14	GM14 - walks using both hands of somebody
	15	GM15 - walks with help - hand or furniture
	16	GM16 - walks but falls over at times
	17	GM17 - stoops and recovers
	18	GM18 - walks well
	19	GM19 - runs (basic running)
	20	GM20 - kicks a ball in any way/tries to kick ball
	21	GM21 - runs well (confidently) stopping and starting without falling
	22	GM22 - kneels and gets up without using hands
Ideal items for pre- School age children 3.5 to <7 years old	**23**	GM23 - throws a ball into a basket (at least one of 3 times) 1 metre away
	24	GM24 - runs, stops and is able to kick a ball some distance
	25	GM25 - jumps with feet together off ground
	26	GM26 - jumps over line/string on the ground
	27	GM27 - stands on 1 foot for less than 5 seconds
	28	GM28 - walks on heels 6 + (or more) steps
	29	GM29 - jumps over piece of paper
	30	GM30 - walks on tip toes 6 + (or more) steps
	31	GM31 - hops on one foot 4 steps
	32	GM32 - stands on 1 foot for a longer time
	33	GM33 - can throw ball in air and catch it with 2 hands
	34	GM34 - heel/toe walk precise one foot behind other along chalk line

Items in grey are examples of a tests assessing a child's walking ability (same construct) at different difficulty levels.
Items in blue are examples of items that returned count data but were dichotomised.
Item in red box shows the ideal starting item to assess a 1 year old child.

The MDAT also has items that returned counts of the number of times a child could repeatedly and successfully perform a given task. Such items were categorised into binary responses. For example, as highlighted in blue in Table 3.2 in items GM 28 and GM 30 in the gross motor domain that inquired how many steps a child could walk on heels or toes i.e. number of steps taken. This was broken into two questions/tasks where a child was expected to walk on heels at least six steps and another

subsequent question/task expecting the child to walk on toes at least six steps given age. The former task of walking on heels is easier than the latter, therefore it was expected that a younger child should have been able to do this better than walking on their toes. Each child would then be evaluated on the basis of whether they could perform the task or not, thus conforming to the binary response set up of assessed tasks. While this dichotomisation process was done purely for clinical interpretation convenience, to ease item the administration and interpretation, it may have led to loss of information (Royston, 2006). Worse still, it may have also introduced or increased the correlation between responses of these items. This is because the two new questions are asked in sequence as they both test the same construct but demand different ability levels.

Intuitively, from the description of various items shown in Table 3.2, it is also clear that some items are related or are testing the same construct even if to a differing degree. For example, items GM 6, GM 12, GM 13, GM 14, GM 15, GM 16, GM 18 and GM 19 that are highlighted in grey all assess a child's ability to use their legs and walk at different levels increasing in difficulty. Further, we note also that being able to deliver on some items depends on delivery of other items implying that not only do the items increase in difficulty, but there is dependence in their responses. This item dependency manifests itself as a correlation between item responses. For example, a child has to be able to stand comfortably before they can walk or even try to run properly. Thus it is expected that as a child advances in age, they are able to pass more complex items and hence the design of the items to increase in difficulty. The initial ordering of MDAT items was by consensus and after formal statistical analysis to produce age centiles, a final rearrangement was done. Section **Error! Reference source not found.** will highlight the fact that the suggested robust item by item age estimate methods may suggest different ordering of items.

3.2.4. The MDAT Assessment Tool Kit

The assessment tool kit refers to the equipment used in the assessment process. These include both machines operated by experts to measure height and weight such as weighing scales or 'slings' as well

as play items or toys to encourage play in children to help solicit the required responses from children. We note the distinction of the assessment tool kits and more specifically items of play found in western versus non-western settings. For example, as is shown in the Figure B.1 in Appendix B, the weighing scales found in western settings are likely to be more advanced having higher calibration and accuracy levels. In contrast, the weighing equipment found in non-western settings are likely to be very basic and have lower accuracy levels.

Gladstone, et al., (2010a) further explained that there are also culturally inappropriate question props that make up the assessment tool item kit e.g. 'prepares cereal' or 'plays board games/card games' are uncommon house chores or play activities for children in rural Malawi. Notably the 'pink doll' in the Denver Developmental Screening Test (DDST) source tool kit was found to be terrifying to most children during the piloting of the MDAT tool. Indeed Malawian children have seen a doll, but probably not a pink one. Thus many got scared making it difficult to carry out the assessment. Some of the naming questions in the Language section of the DDST or Denver II have pictures of objects that children, at least in the part of rural Malawi studied, have never seen before, such as a horse. Such alien play or probing items make it difficult for children to name them, especially as many children have also never seen pictorial representations of these objects in a book at their age.

Therefore, some items' administration style and kit may have to be altered for it to be suitable in a new setting. Instead locally available materials and props that children are familiar with should be used. Therefore, application of adapted tools with unadapted kits may still cause invidious reactions from children that will consequently lead to wrong ability status conclusions. In such instances local information and knowledge will offer feasible alternatives if the necessary equipment is not available or culturally suitable. We recommend also adequately describing such instances of unsuitable play or probing item kits. This reporting as was done for example by Pfeifer, et al., (2011), provides a source of valuable information for future researchers.

3.2.5. Recording an outcome using the MDAT tool

The assessment using the MDAT Draft III tool with 136 items took approximately 35 minutes in a quiet location that could even have been outdoors. Five to seven children were assessed daily by two to three research midwives trained on the MDAT use at two of the four different sites. Where possible, items were directly observed, but items were accepted by report if the mother was very clear that the child could do the item and there was no doubt when assessing associated areas of development. Items were scored as pass or fail, and if the child was uncooperative or unwell, items were scored as 'don't know'. Items were assessed until the child failed seven consecutive items as described in the MDAT manual. Aside from age, other relevant demographic details presumed to influence ability development including gender and social economic status are also collected.

If the child successfully performs the item, they 'pass' and a score of '1' is assigned. If the child does not successfully perform the item or the caregiver reports (when appropriate) that the child does not do the item, they 'fail' and a score of '0' is assigned. A 'no opportunity' outcome is often described when the child has not had the chance to perform the item due to restriction from the caregiver or for other reasons such as the item not being available to do or testing having to stop due to problems from the point of view of the examiner. A 'refusal' is where the child declines to attempt the item, again due to various reasons like fear, uncooperativeness or they are just too unwell on the day considering that recruitment in developing countries is often done in referral clinics or hospitals. In such instances, the parent can administer the item rather than the researcher as they are more familiar with the child. These items with no opportunity, refusal or no response are often recorded as a blank and analysed as missing data. One therefore needs to be as clear as possible when scoring items about whether the child is refusing because they cannot do the item or whether they are refusing because they do not want to do the item.

Figure 3.5 also shows how an examiner will score items for a child who is one year old. Recall that tasks 1 to 34 have been arranged in ascending order of difficulty given age. The yellow line at one year

corresponds to item 13 whether it is expected that a child of this age should pass. The examiner starts by asking easier questions to the left of the one year mark (yellow dotted line).and once it is apparent that the child can deliver (see tick (✓) marks on the chart) on these trivial tasks given their age, starts asking tasks to the right up to the point where the child can no longer respond (see cross (X) marks on the chart) given the stopping rules defined in MDAT manual. Items that are ideal or likely to be asked to babies (< 1 year old), toddlers (1 to < 3.5 years old) and pre-school age children (3.5 to < 7 years old) are highlighted in the blue, green and red dotted boxes respectively. We also note that in the first year from birth, the rate of development in each domain is quite rapid as can be seen by the steep initial gradient of percentile estimates in the first six months post-birth from the chart. Therefore, to be able to notice development differences, the scale for the first year has been magnified and shows changes on a monthly basis. In our methods chapter we will assess how well the item modelling approach captures this initial rapid rate of child development in the first few months after birth.

We would like to note that the recording of item responses is crucial as it is possible that the child may not able to carry out some tasks below item 13 and also may be able to still carry out tasks above the final item administered. Thus this item response recording may inadvertently introduce a form of bias or missingness pattern. However, while noting the importance of the item recording mechanism, we wish to note that this is a development assessment context so items are ordered in terms of difficulty. Thus if we are using a well-developed and validated assessment tool, it is reasonable to assume a non-response after reaching the child's final item means that a child has not yet developed the dexterity or skill to that level. Thus even if items beyond their threshold were administered they would still fail, and if they pass then it calls into question the soundness of the MDAT tool. Further, from a pragmatic point of view, considering that researchers want to save on time and assess as many children as possible, if a child walks into the examination room then it is reasonable to assume that they did at one time pass all preceding easier items testing their head support or crawling abilities.

Figure 3.5: Scoring child ability using MDAT chart.

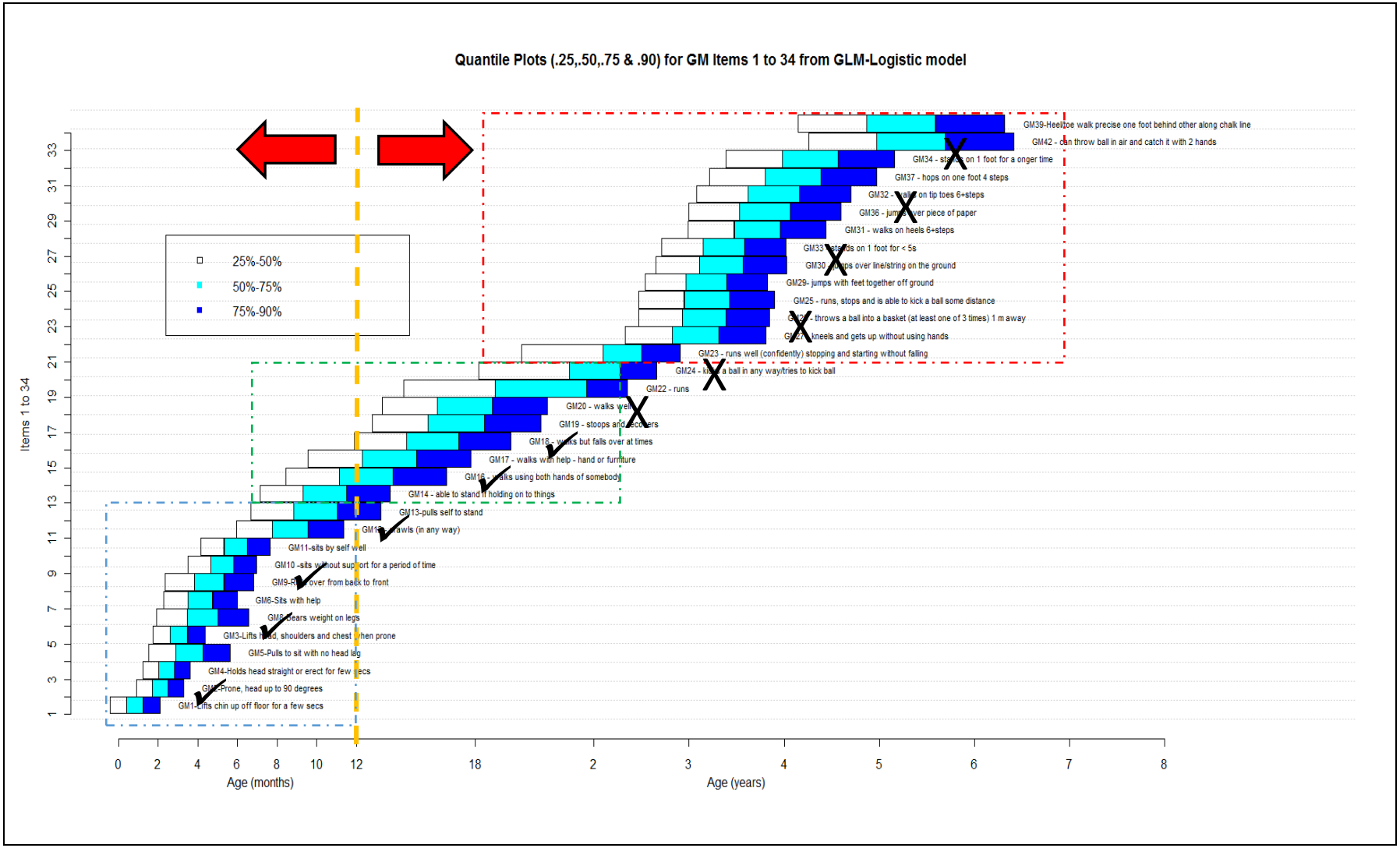


Chart used by Gladstone, et al., (2010a) but drawn using age estimates using 2007 MDAT survey data. Items that are ideal for babies (< 1 year old), toddlers (1 to < 3.5 years old) and pre-school age children (3.5 to < 7 years old) in the blue, green and red dotted boxes respectively. Yellow dotted line shows the starting assessment point for a 1 year old child.

3.2.6. Item response recording into an analysis spreadsheet

We have seen that either the chart in Figures 3.4 or 3.5 can be used to collect item responses following instructions in the MDAT manual. This section explains how these responses are represented in a typical data spreadsheet. The columns in the spreadsheet represent variables corresponding to tool items and other demographic variables while the rows represent responses for each child or cases in the study that was assessed by the expert. We will explain the process using a hypothetical example of item responses of a typical 1 year old child ID 001 shown in the Figure 3.5.

Table 3.3: Item responses entered into spreadsheet.

Child ID	Age	Other Demographic variable(s)	GM Item 1	•••	GM Item13	GM Item14	GM Item15	GM Item16		GM Item20	•••	GM Item33	GM Item34
001	1		1	…	1	1	1	1	…	0	…	0	NA
002	0.7		1		1	0	0	0	…	0	…	NA	0
003	0.1		1		0	0	0	0	…	0	…	0	0
004	1.5		1		1	1	1	1	…	1	…	0	0
005	3.6		1		1	1	1	1	…	1	…	0	NA
…	…	…	…	…	…	…	…	…	…	…	…	…	…
†n=1446	…	…	…	…	…	…	…	…	…	…	…	…	…

†n refers to the sample size of the MDAT study. Age is given in years

A one year old healthy normally developing child typically should be able to respond positively to all questions from task 1 to 13 (highlighted in red in Table 3.3) before starting to experience any difficulty. As is explained in the latest version of the MDAT manual, the expert starts by asking a few easier questions to the left of task 13. Once it is clear that the child can respond to these he/she proceeds on to administer questions 14, then 15 and so on until it is apparent that the child is struggling to respond to any further question. For this child, we see that they responded to all questions up to item 20 at which point the assessment stopped. Notice that all questions from 1 to 19 will have a pass response indicated by a 1, question 20 will have a fail indicated by a 0 and subsequent questions will not be administered. Child ID 002 is a slightly younger that child ID 001 thus he/she was only able to pass all items from item 1 to 13 only.

With this mode of administering items to children, a typical spreadsheet should have a general monotonic pattern of completed tasks to the left of their starting point, and any fails or missing items

should be on the right hand side of their starting point. How far below or how far above the normal threshold a child's positively responds defines the extent of delay or extent of advanced ability in comparison to normal children. As noted in Section 3.2.5 in the event some items are not administered and therefore missing, it should be clear that this was because the child had not achieved the required ability to pass them. If the examiner was unsure of the presence or lack of an ability trait, a repeat assessment at another time was organised. For items that are not administered to the left of the starting point of the assessment process as it is obvious a child can perform them, we argue that it is reasonable to assume these to be passes on the assumption that the MDAT tool is a scientifically developed and validated tool with a valid ordering of items in increasing difficulty. Further we note that in this research, no child failed more than seven items and then proceeded to pass any other subsequent items. This is because before the final assessment item was defined, as explained in the MDAT assessment guidelines (contained in the MDAT manual), it was asked up to seven times.

3.3. Summary

This chapter has described the MDAT assessment tool used in this project. Most importantly we have pointed out the potential bias that may be a result or the item response recording mechanism that has implications on the validity of both age estimate and scoring and analyses due to data quality. We recommend that the item response recording should be done very diligently invoking very reasonable assumptions even if any missing data can be explained against the backdrop of typical child development assessment particularity. The next chapter will discuss the exploratory of the collected binary item response assessment data. This is an important process as it investigates both underlying but important characteristics of the data that advise on the best statistical modelling approaches but also it can be used to flag any items that are not particularly useful in the development of the score norm and can thus be removed.

4. Data and Exploratory Data Analysis (EDA)

4.1. Introduction

This chapter discusses and presents the findings of the exploratory analysis (EDA) for the binary item response data. EDA refers to the process of analysing data to investigate and summarise its underlying distributional characteristics as well as item quality. The different types and tremendous merits derived from a well thought out, complete and rigorous EDA within an assessment tool context is to detect any anomalies in the performance of an item and advise on the most appropriate age estimating or scoring approaches.

The importance of the quality and role of the type of data cannot be over stated, and this is noted by several researchers for example by Altman, (1991). As we will soon realise data dictates the most appropriate and valid choice of the statistical analytical procedures to be used; its quality to a large extent has great bearing on the justification to use more robust statistical methods to deal with certain specific item characteristics so as to ensure that high accuracy of the computed scores is still achieved. Even though the eventual scoring methods used are complex and lean towards being more esoteric to justify their robustness, still the quality and performance of scores is underpinned by the quality of the data.

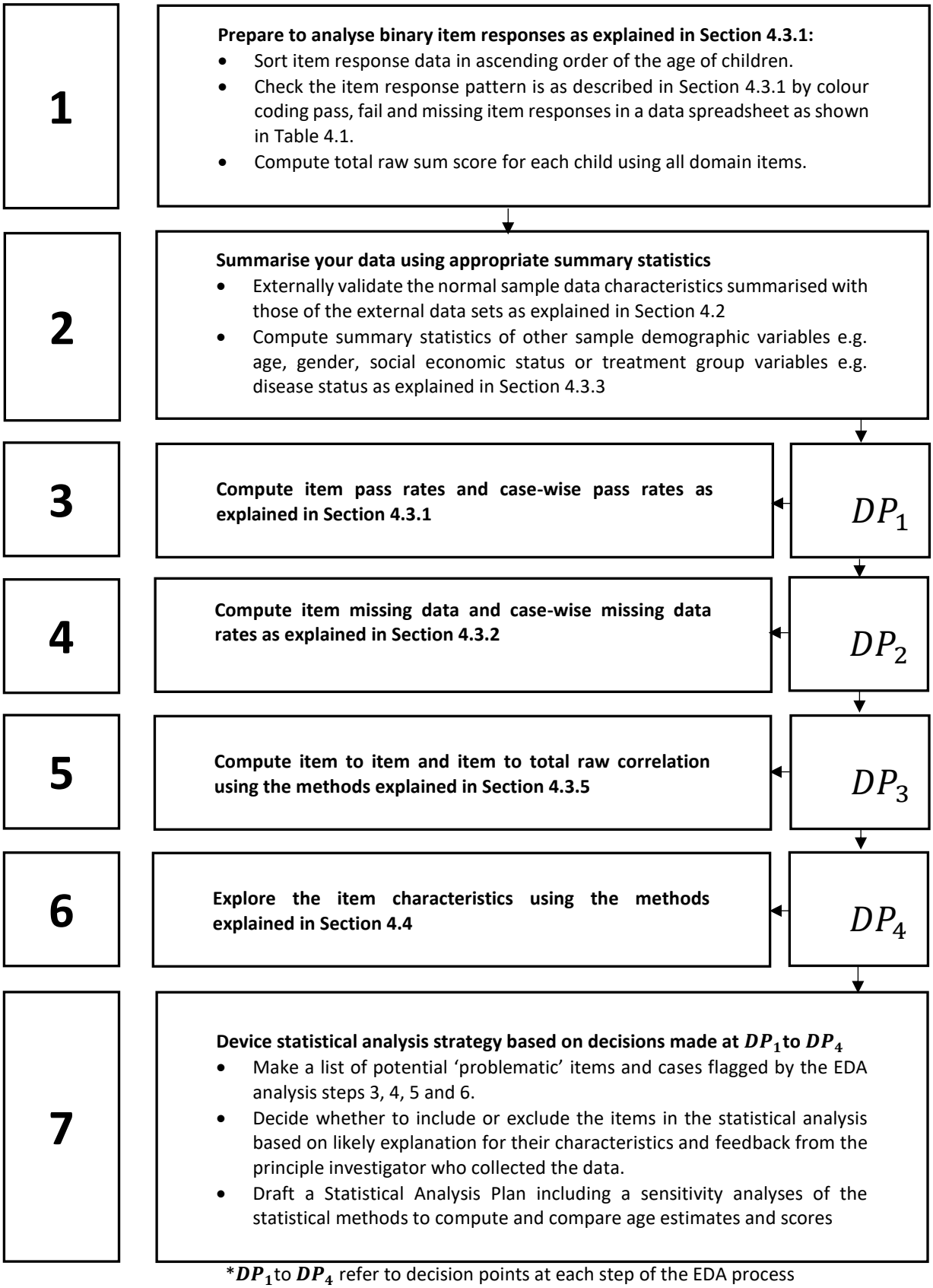
In accordance with the sentiments of Lawrence, et al., (1998), defining the study data characteristics is an integral part of posing and being able to deliver on research objectives using appropriate methods. Therefore, the aims of this chapter are two fold;

- Firstly, to describe the item data characteristics that will be used to apply the suggested scoring extensions as well as test their sensitivity.
- Secondly, to outline the importance of a thorough and well thought out exploratory data analysis of item assessment response data. While highlighting various concerns particular to child assessment data structure, fashion out a clear analysis plan that should be

followed to carry out EDA analysis in practice that also points to basis for using specific statistical modelling approaches. Figure 4.1 shows a schematic diagram of the data exploring strategy for binary item responses that will be used before the item by item age estimation and overall score computation.

Section 4.2 describes the representativeness of the MDAT data that will be used to carry out the formal age estimation and score computation in chapter five. Section 4.3 describes up to five preparation and exploratory analyses that we should carry out on binary item responses including examining the distribution of item pass/fail rates. Section 4.4 further highlights important, complementing and related item characteristics including difficulty, discrimination, investigating both item to raw total correlation as well as raw total and age correlation and the use of empirical item characteristic curves to characterise them. The chapter concludes with a summary in Section 4.5.

Figure 4.1: A schematic flow diagram of the data exploring strategy of binary item responses before item by item age estimation and overall score computation.



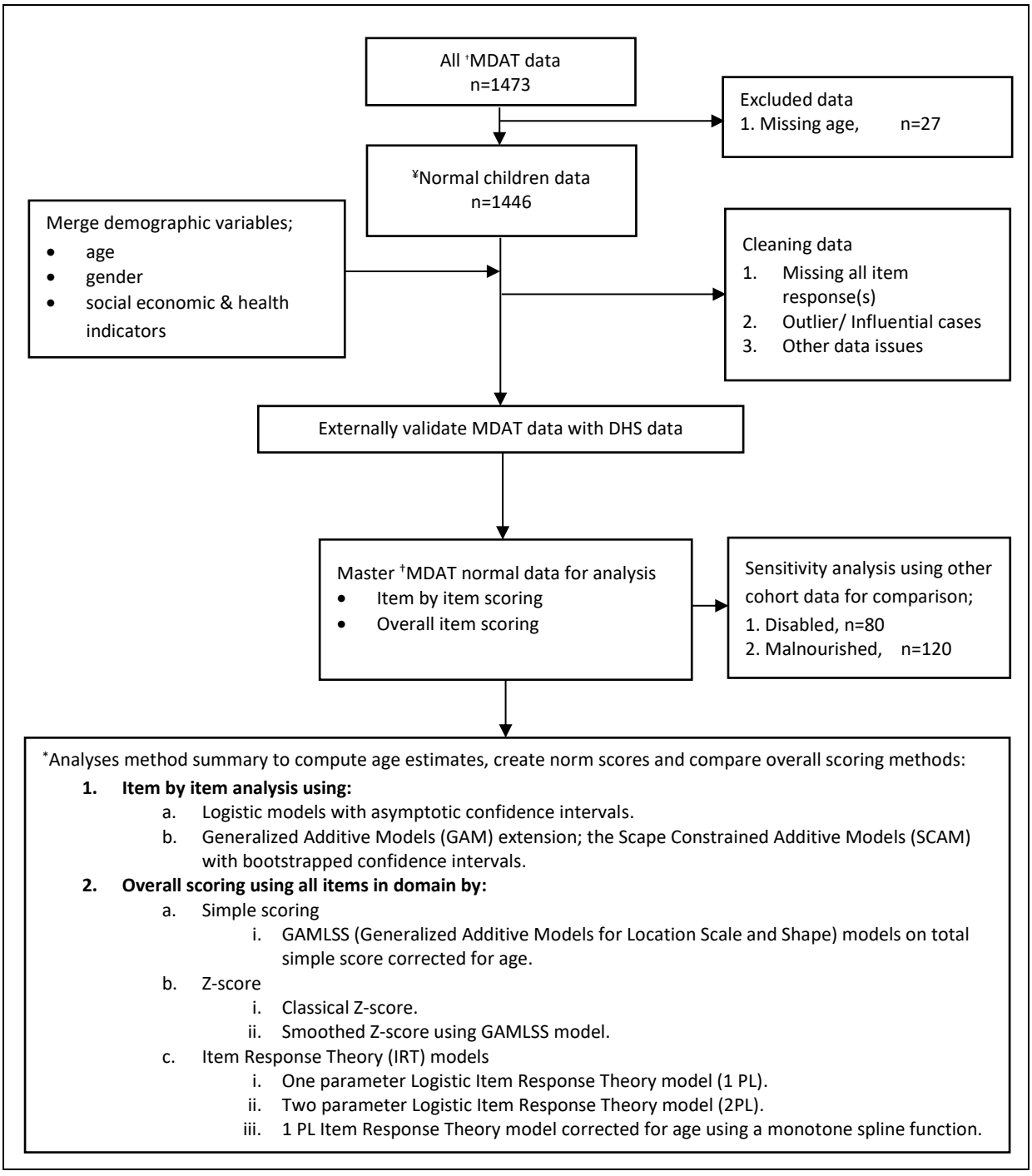
*DP_1 to DP_4 refer to decision points at each step of the EDA process

4.2. MDAT Data and External Validation

Section 3.2.1 described the binary item sample of data to use in this thesis was from a 'large' observational study that utilised a stratified quota sampling method to select a representative sample of healthy and normally developing children. The flow chart in Figure 4.2 gives a summary of the collected MDAT 2007 survey data and the eventual data used to develop and test performance of scores in this project after the data cleaning, resolving missing data item responses and external validation process.

In order to externally validate that the MDAT 2007 survey data was representative of the typical healthy and normal developing population of Malawian children aged zero to seven years; we requested the latest data collected from the Demographic and Health Surveys (DHS) for Malawi and compared the gender distribution, age distribution and wealth status of these data sets and the MDAT data. A detailed account of the external validation methodology and specific comparison findings are available in a separate report. In summary we found that although the DHS sample size of 2672 is almost twice that of the MDAT 2007 data used in this thesis, it is fairly comparable in gender and wealth status distributions. There is however a slight difference in age distribution between the DHS and MDAT 2007 data for the very young children in both samples. The MDAT has a higher proportion of very young children. This was attributed to the design of the MDAT study with a specific interest in the first few months of childhood where ability is known to increase at a high rate in comparison to later ages. However, given also that the age category limits on both data sets are not exactly similar, we conclude that the extent of this dissimilarity is severe and the MDAT 2007 survey data was indeed representative sample of the Malawi child population in the years 2007.

Figure 4.2: Flow chart of Malawi Development Assessment Tool data.



†MDAT data contains 34 items per domain.
¥Healthy children developing normally without any known form of disability or developmental disorder
*Scoring methods listed in data flow chart will be described in Chapter five.

4.3. Exploratory Data Analysis (EDA)

The Exploratory Data Analysis was performed to investigate the association: a) between item pass rates given the child age b) between overall scores and age. This was in order to advise on the best statistical modelling approaches that address various data issues especially the fact that the item outcomes of interest are binary, the necessity to address the complex correlated structure of the data across items and within items, and the fact that the outcome of interest is not only associated with the age of the child, but this association has to be captured and modelled in a unique fashion. Therefore, in line with the EDA strategy outlined above, this section is broken down into subsections each focusing on investigating a particular aspect of the MDAT item data structure.

4.3.1. Preparing to analyse the MDAT data

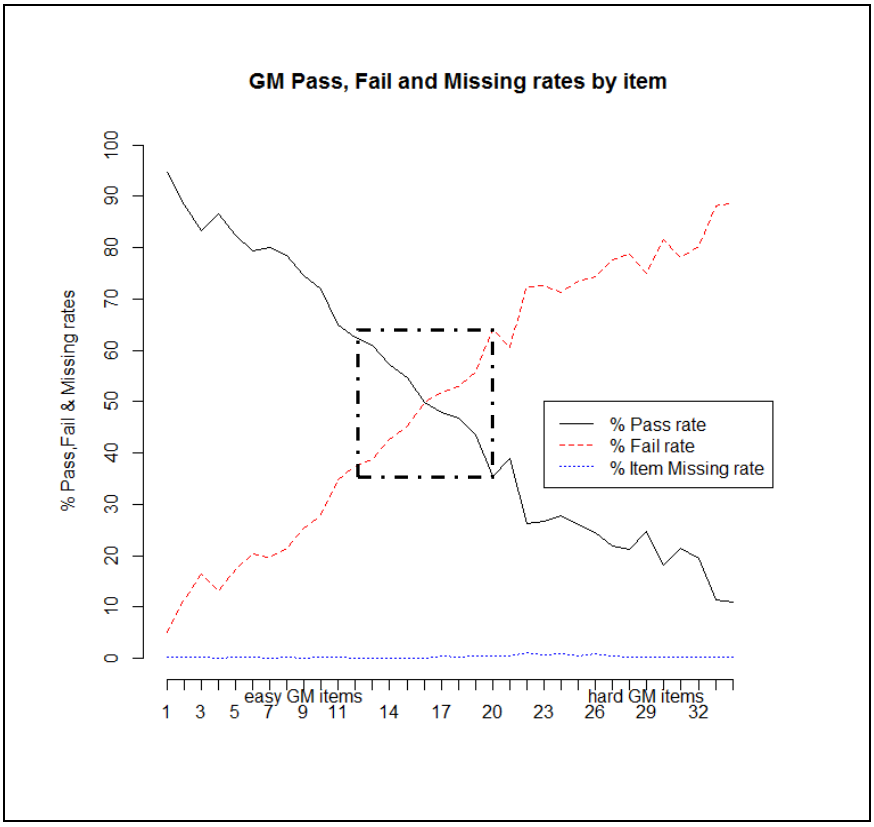
EDA is first tasked to ensure that all the item and demographic variables responses are valid. This involves carrying out a rigorous data cleaning and validation exercise to ensure correct and accurate response entries have been made. This is obviously important to maintain data integrity and contribute towards the aspired quality of age estimates, scores and norms.

All items responses had either a code 1 for a pass, code 0 for failed or a blank (NA) entry for an unadministered item. Further, one should check that only a few children should pass all 34 items per domain (as this would point to the assessment tool's inadequacy) and any child who had either a very low or high item pass rate given their age was acceptable. Such children with either too low or too high pass rates pointed to them being possible outliers. They were flagged and the principle investigator (PI) was consulted to confirm that the data entered for such a case was correct (i.e. valid) and acceptable. Cases with missing age were left out of the analysis. All age and nutritional variables were checked to be within acceptable ranges given the design of the survey. We were now confident that the MDAT data at hand had been properly coded, recorded and various summary statistics could now be produced to assess relevant distributional properties.

Even at this early stage, the increase in difficulty of items as they ascend from item 1 to 34 in each domain should be easily apparent from the data entry spread sheet described in Section 3.2.6. Table 4.1 shows hypothetical examples of item response patterns for babies (< 1 year old), toddlers (1 to < 3.5 years old) and pre-school age children (3.5 to < 7 years old) in the blue, green and red dotted boxes respectively. With the data sorted by ascending age, as expected, we see that; a) the number of items failed (highlighted in grey) decreases, b) the number of items passed increases c) therefore the total number of items passed increases with the increase in age, and the item pass rate reduces from item 1 towards item 34. This is as a consequence of the issue explained earlier that a development assessment tool's items are designed to increase in difficulty therefore it is expected that older children have a higher probability of passing more items.

Alternatively, one could plot the empirical item pass (black line), fail (red) or missing rates (blue) for each item as is shown in Figure 4.2. Therefore, the empirical item pass rate should gradually reduce from item 1 to 34 as the failure rate increases from item 1 to 34. This aspect should point to adequate ordering of items and echo the item response pattern seen in Table 4.1 when data is sorted by child age. The increase in pass or fail rate of the items should be subtle, not necessarily smooth, but should not have any abrupt changes or jumps in pass or failure rates which would suggest possible problems with the ordering of tool items, survey design or sample. Although merely plotting pass/failure rates maybe heuristic, the intension is to detect any anomaly in item ordering which is the corner stone of the usability of the collected item responses.

Figure 4.3: Summary of item pass, fail and missing rates in gross motor domain



*Very low (less than 2 %) missing data rates in gross motor item.
Rate = (Pass or Fail or Missing count/Total Normal Sample size) × 100 %.
Black dotted box shows the point where the item pass and fail rates are equal.

The missing rates of items should be fairly constant and preferably low in all items as shown in the Figure 4.3. The missing data rate for each item in the gross motor domain was quite low (<2 %) in general as shown by the almost horizontal blue line in Figure 4.3. However, the harder items from item GM 17 seem to have slightly higher missing rates. The black dashed box shows the point of intersection of the item pass and fail rates for the MDAT tool. This point represents the item in the tool that approximately the same number of children passed or failed this item. Ideally, in a sample of healthy normally developing children to be used for standardisation of scores to produce norms for ability classification, this intersection point should be one single point, and should point to an item that is positioned mid-way in the assessment tool. However, the ideal point or range at which the pass rate and failure rates should be equivalent is a likely important research question. However, this issue is beyond the scope of the objectives of this thesis. We envisage its position will probably depend on the design and objectives of the assessment tool.

Table 4.1: A snap shot of Gross Motor MDAT data

Child ID	Child Age	GM Item 1	GM Item 2	…	GM Item14	GM Item15	GM Item16	…	GM Item20	…	GM Item33	GM Item34	Total passed
0001	**0.20**	1	1	…	0	0	0	…	0	…	0	0	**4**
0002	**0.25**	1	1	…	1	0	0	…	0	…	0	0	**5**
0003	**0.55**	1	1	…	1	0	0	…	0	…	0	0	**6**
0004	**0.63**	1	1	…	1	0	0	…	0	…	0	0	**8**
0005	**0.84**	1	1	…	1	1	0	…	0	…	0	0	**8**
0006	**0.95**	1	1	…	1	1	0	..	0	…	0	0	**9**
…	**…**	…	…	…	…	…	…	…	…	…	…	…	**…**
0200	**1.00**	1	1	…	1	1	0	…	0	…	0	0	**10**
0201	**2.50**	1	1	…	1	1	0	…	0	…	0	0	**12**
0202	**2.70**	1	0	…	1	1	0	…	0	…	0	0	**17**
0800	**2.90**	1	1	…	1	1	1	…	0	…	0	0	**20**
0900	**3.05**	1	1	…	1	0	1	…	0	…	0	0	**18**
1000	**3.40**	1	1	…	1	1	1	…	0	…	0	0	**24**
…	**…**	…	…	…	…	…	…	…	…	…	…	…	**…**
1444	**5.00**	1	1	…	1	1	1	…	1	…	0	0	**27**
1445	**5.50**	1	1	…	1	1	1	…	1	…	1	0	**26**
1446	**6.00**	1	1	…	1	1	1	…	1	…	1	1	**28**
1444	**6.20**	1	1	…	1	0	1	…	1	…	1	0	**30**
1445	**6.50**	1	1	…	1	1	1	…	1	…	1	0	**34**
†n=1446	**6.97**	1	1	…	1	1	1	…	1	…	1	1	**33**
Pass rates		**100%**	**99%**	**…**	**70%**	**65%**	**50%**	**…**	**40%**	**…**	**25%**	**15%**	

†n, MDAT sample size=1446, Total Passed=Sum(Items passed per child), Pass rate = (Pass count/Total Normal Sample size) × 100 %.

Items response patterns for babies (< 1 year old), toddlers (1 to < 3.5 years old) and pre-school age children (3.5 to < 7 years old) in the blue, green and red dotted boxes respectively. Highlighted in grey are the items failed.

4.3.2. Exploring missing data

As discussed in Section 2.3.3.2, the missing item data pattern is an important aspect that needs to be checked in the EDA of assessment data. We noted that in real life, data from surveys, 'large' and complex regional or multiregional epidemiological studies, clinical trials or longitudinal studies, especially where a particular attribute is measured repeatedly from the same units at one instance, or over a period of time, it is quite common to have units or subjects with missing values/responses for one or more items. The type and extent of missingness will often have a myriad form of negative effects and influence on the inferences made on the analysis depending on the assumptions made about the reasons for the missing data at the time of analysis. Madow, et al., (1983) and Tabachnick & Fidell (2001) discuss various forms, patterns and extents of missing data giving various recommendations of dealing with each. Kline (2011) notes that the more sophisticated methods of dealing with missing data are useful if there is an identified systemic pattern which also requires a sensitivity analysis to compare analyses with and without missing data.

The first issue in dealing with the missing data problem is determining whether the missing data mechanism has distorted the observed data, and possibly the inferences. A child assessment context presents such a scenario where missing data is likely as ability is measured by the administration and measurement of several items at any given time. The objective of this section is therefore to explore missingness patterns in the MDAT 2007 data both case wise (per child) and item wise (per item) using simple graphical and tabular summary techniques with a view of establishing if the missing items are in any way different from those completed, i.e. investigate if there is any systematic form of missingness and then consider the best approach of dealing with it.

Section 3.2.5 highlighted that missing data is likely to occur in a child assessment scenario when the expert assessor is unable to discern with certainty whether a child has the ability that is under scrutiny. This can arise either because the item/question could not be administered due to various practical reasons or the item could not be reached due to lack of ability. Ideally, given the method detailed in

the MDAT manual of assessing a child, the missing data pattern should be monotonic and not intermittent. This is because no other task should be administered to the child once the threshold for stopping the task administration has been reached.

From a practical point of view an assessor may not complete entries for items not administered when the child reaches their final item, nor will they complete entries for items that it is obvious the child could pass that are to the left of the starting point. Obviously the examiner juggling between managing to keep a good rapport with the child and guardian during assessment may not have time to ask the entire set of items and simultaneously be always diligent to adequately complete the assessment chart. This is the purpose of the Section 4.3.1 to make sure that the former scenario's missing items are recoded to fails and the later scenario missing items are recoded to passes. As explained in Sections 3.2.5 to 3.2.6, in a child development assessment context, we argue that this is a safe assumption to make after thorough investigation and confirmation with the study P.I. of all instances with missing item responses. However, there may be still instances where a certain item could not be administered for logistical or practical reasons but these should be very isolated and few instances. We envisage that this later cause of missingness may not cause a problem as several items in a tool measure the same construct hence even if an item is not administered the trait of interest is still tested by another item even if to a differing degree. All other items prior to this threshold should indicate a pass or fail.

In an assessment context, the assessment of missingness both between and within items is equally important because of the nature of scoring computation may only affect one of the dimensions or both dimensions simultaneously. The missing data in the normal child sample was explored to check for any systematic patterns and possible causes discussed with the Principal Investigator of the MDAT study. This was especially in cases where there was a considerable amount of missing entries both between and within items. Colour coded profile schemes were prepared to flag any case with a 'high' rate missing item entries and missingness was explored and summarised.

In the MDAT data, it was found that most of the cases with missing item responses were indeed monotonic. In the few instances where it was not, this was attributed to situations where there were administration problems either stemming from tool equipment or the state of the child. Missing data patterns and rates were checked over items as well as across individual child profiles to assess their extent so as to advise the most feasible rudimentary approaches. The Table 4.2 shows a summary of the case wise frequency of the number of missing responses in the GM domain.

Table 4.2: Case wise missing frequency summary in Gross Motor (GM).

Number of items missing in GM domain	Total	†(%)
0	1374	(95.0 %)
1	54	(3.7 %)
2	9	(0.6 %)
3	3	(0.2 %)
4	2	(0.1 %)
5	2	(0.1 %)
6	2	(0.1 %)
7	0	(0.0 %)
34	2	(0.1 %)

†Percentage missing = (Number of items missing per case/Total Normal Sample size) × 100 %.

We see that the GM domain had 95.0 % complete data i.e. a pass or fail for all 34 items had been determined. Further we see that up to 54 (3.7 %) children had at least one missing item response and even fewer children had more than one item missing. However, there were two cases in the gross motor domain that had all 34 item responses missing. Their nutritional status variables seem to indicate that these two cases were borderline malnourished or stunted. Malnutrition and stunting have been shown to be a possible indicator of delayed development hence the possibility of the difficulty experienced in their assessment. The standard definitions of malnutrition and stunting are well defined by the WHO (2006). Consultation with the P.I. revealed that these two cases whose demographic characteristics has been summarised in Table 4.3 revealed that they in fact had a form of disability due to malnutrition hence were excluded from any further analysis that used the normal data as the standardisation sample to compute age estimates and norms.

There is also missing data with respect to child characteristics like age, gender wealth status and nutritional characteristics. The extent of missingness of these variables was summarized and it was

found that the extent of missingness of these variables was not extensive and will not interfere with the analysis. Given that most of the scoring methods and eventual classification of a child's ability are dependent on their age, if this variable is missing, then one cannot be able to score and hence classify their developmental status. The MDAT data had 27 cases with the missing age variable.

Table 4.3: Characteristics of cases with all 34 items missing in GM domain.

ID	Age (decimal years)	Gender	*Wealth status	¥WHOWAZ	$WHOHAZ
1178	2.42	Male	3 Middle	-0.98	0.11
1559	5.49	Female	1 Lowest	-2.54	-4.12

¥WHOWAZ-Weight for Age Z score, $WHOHAZ-Height for Age Z score, *Wealth Status-Social economic status.

A summary of the missing data situation in the normal standardisation sample in form of number of missing cases and percentage of missingness per item for each of the GM items is tabulated in the Table 4.4. For example, it was noticed that items 22, 23, 24 and 26 had the highest frequency of missing cases of 17 and 10 for items 22 and 23 respectively, and 14 missing cases for both items 24 and 26 in the gross motor domain. This relatively high missing rate in comparison to the rest of the items was attributed to the nature of the tasks that involved kneeling, throwing, running and jumping; these tasks do not only demand higher ability but also their administration and conclusive discernment of presence or absence of the ability being tested by examiner can a bit difficult.

It remains to be emphasised that a dropout process that is neither MCAR nor MAR cannot be ruled out i.e. MNAR. For the validity of the above proposed modifications, it is assumed that missingness is conditionally independent of the unobserved data, given the observed measurements. Unfortunately, it is very difficult to justify the assumptions of random dropout, especially when no information whatsoever is provided on the reasons for withdrawal. One way to check evidence for MAR as well as MCAR is to use a logistic dropout model with dropout as response and the genuine covariates including, the previous observations before dropout (*PREV)* and present observations *(PRES)*, as one of the explanatory variables. If the *PREV* has an effect on dropout then this provides some evidence (but does not proof) for the MAR process. However, if *PREV* is not significant the situation remains inconclusive. If both *PREV* and *PRES* do not have an effect on dropout, then some evidence of MCAR

can also be made. If on the contrary *PRES* has an effect on dropout then MNAR cannot be ruled out. As a way of conducting a small sensitivity analysis, different models are fitted under different assumptions, on the original data, the monotone-imputed and fully imputed data, as mentioned earlier. A more formal approach to the above tests is to couple a linear mixed-effects model for the measurements with a logistic model for dropout. This can be done under the framework of selection and pattern mixture models (Little & Rubin, 1991).

However, given our finding of the missing data pattern in the MDAT data, we conclude that missingness is very minimal and can safely be classified as MAR as it shows no systematic pattern to any item or covariate. Therefore, in its current nature we anticipate the present missingness will have minimal effect on any inference drawn from models built or on the quality of scores produced. In the methods section a weighting mechanism by administered items to compute total scores will be proposed and compared to a non-weighting scoring process. Therefore, while imputing the data of the few missing cases present can recover the missing item responses, the overall response uncertainty may be underestimated and possibly the correlation between items overestimated.

Table 4.4: Frequency counts of item pass/fail and missing rates in Gross Motor (GM) domain.

Type of item	GM Item(s)	Pass (1)	†Pass rate %	Sum, n	Missing	†Missing rate %
Ideal items for infants <1 year old	GM 1	1369	94.70	1442	4	0.30
	GM 2	1278	88.40	1442	4	0.30
	GM 3	1204	83.30	1443	3	0.20
	GM 4	1253	86.70	1444	2	0.10
	GM 5	1191	82.40	1441	5	0.30
	GM 6	1149	79.50	1443	3	0.20
	GM 7	1158	80.10	1444	2	0.10
	GM 8	1134	78.40	1443	3	0.20
	GM 9	1078	74.60	1444	2	0.10
	GM 10	1041	72.00	1443	3	0.20
	GM 11	940	65.00	1442	4	0.30
	GM 12	903	62.40	1444	2	0.10
Ideal items for toddlers 1 to <3.5 years old	GM 13	883	61.10	1444	2	0.10
	GM 14	827	57.20	1444	2	0.10
	GM 15	790	54.60	1444	2	0.10
	GM 16	722	49.90	1444	2	0.10
	GM 17	692	47.90	1440	6	0.40
	GM 18	678	46.90	1442	4	0.30
	GM 19	631	43.60	1438	8	0.60
	GM 20	512	35.40	1439	7	0.50
	GM 21	562	38.90	1438	8	0.60
	GM 22	381	26.30	1429	17	1.20
Ideal items for pre- School aged children 3.5 to <7 years old	GM 23	385	26.60	1436	10	0.70
	GM 24	402	27.80	1432	14	1.00
	GM 25	377	26.10	1439	7	0.50
	GM 26	356	24.60	1432	14	1.00
	GM 27	317	21.90	1439	7	0.50
	GM 28	306	21.20	1443	3	0.20
	GM 29	357	24.70	1442	4	0.30
	GM 30	263	18.20	1441	5	0.30
	GM 31	312	21.60	1442	4	0.30
	GM 32	283	19.60	1442	4	0.30
	GM 33	166	11.50	1441	5	0.30
	GM 34	160	11.10	1443	3	0.20

†Rate = (Pass or Missing count/Total Normal Sample size) × 100 %. Items in red text had relatively high missing rates. Items that are ideal for babies (< 1 year old), toddlers (1 to < 3.5 years old) and pre-school age children (3.5 to < 7 years old) in the blue, green and red dotted boxes respectively.

4.3.3. MDAT Study Characteristics

There were other demographic variables apart from age, including gender, nutritional and social economic status indexes that were recorded and that have been previously been shown to positively or negatively influence child ability development as discussed in Section 1.1.3. Knowledge of the extent and type of association of these variables on the child's ability expressed by the development score is important as the variables may be confounding or interacting with the probability of passing an item. Further, due to the lack of complete knowledge of these variables influence of ability, assessment studies are designed as such. For example most studies assessing child development will keep their child recruitment age below seven years to avoid the influence of gender differences due to the onset of earlier puberty in female children. Therefore, as recommended by Gladstone, et al., (2010a), it is best to develop a tool that is only dependent on age. Although this may practically mean that the tool is less sensitive in classifying developmental status especially for borderline ability delayed children, it allows the use of the tool across various groups of children. In this work we will not consider other variables apart from age in either the age estimate or overall score computation given the remit of this thesis of establishing a framework that can adjust for covariates. Further, for validation purposes, the two validation samples we had did not have all these other demographic variables. Various summary statistics with respect to these demographic variables have been summarised in the Tables 4.5 to describe the populations of the normal as well as the disabled and malnourished samples of children used to develop the MDAT tool, validate and compute the normative scores.

The most common form of descriptive statistics is the mean and variance (standard deviation) for continuous variables. The important aspect is to identify the underlying distribution of the variable(s) in order to recommend suitable statistical modelling approaches. From Table 4.5, we notice that the normal sample had a higher proportion of young children hence the age distribution in this cohort was slightly positively skewed. This was as a consequence of the MDAT study design described in Section

3.2.1 and the fact that it has been shown that the development rate is highest at this time in the life of a child. Therefore, the age category thresholds were chosen bearing this fact in mind while trying to have an equal frequency spread across all other age categories. The other demographic characteristics of gender, nutritional and social economic status were only available in the normal children sample. The nutritional indicators are fairly normally distributed and do not exhibit any anomaly to suggest that that the normal sample had an underlying nutritional problem. As would be expected there is the presence of both possibly underweight or stunted children and possibly overweight children in the normal sample. The gender distribution was fairly even and social economic status levels were fairly comparable with the highest proportions of 21.2 % each in the lowest and fourth levels.

Table 4.5: Characteristics of Children in Normal, Disabled and Malnourished cohort samples

Characteristics	Normal (*n* = 1446)	Disabled (*n* = 80)	Malnourished (*n* = 120)
Gender			
Male	689 (47.6 %)	43 (53.8 %)	63 (52.5%)
Female	717 (49.6 %)	34 (42.5 %)	56 (46.7%)
Unknown	40 (2.8 %)	3 (3.8 %)	1 (0.8 %)
Age summaries			
min	0	0.79	0.50
max	6.92	6.39	6.15
median	1.08	3.03	1.78
Mean (†s.d)	1.84 (1.85)	3.15 (1.43)	2.01 (1.02)
Age distribution			
0 < 4 months	355 (24.6 %)	0 (0.0 %)	0 (0.0 %)
4 < 7 months	134 (9.3 %)	0 (0.0 %)	1 (0.8 %)
7 < 12 months	198 (13.7 %)	2 (2.5 %)	13 (10.8 %)
1 < 1.5 years	156 (10.8 %)	9 (11.3 %)	28 (23.3 %)
1.5 < 2 years	120 (8.3 %)	8 (10.0 %)	30 (25.0 %)
2 < 3 years	130 (9.0 %)	20 (25.0 %)	31 (25.8 %)
3 < 5 years	209 (14.5 %)	30 (37.5 %)	15 (12.5 %)
5 < 7 years	144 (10.0 %)	11 (13.8 %)	2 (1.7 %)
*Wealth Status			
Lowest	306 (21.2 %)	-	-
Second	258 (17.8 %)	-	-
Middle	291 (20.1 %)	-	-
Fourth	306 (21.2 %)	-	-
Highest	285 (19.7 %)	-	-
¥WHOWAZ			
min	-5.79	-	-
max	8.12	-	-
median	-0.75	-	-
Mean (†s.d)	-0.63 (1.36)	-	-
$WHOHAZ			
min	-9.74	-	-
max	9.24	-	-
median	-1.56	-	-
Mean (†s.d)	-1.60 (1.60)	-	-

†s.d-Standard deviation, ¥WHOWAZ-Weight for Age Z score, $WHOHAZ-Height for Age Z score, *Wealth Status-Social economic status.

¥WHOWAZ, $WHOHAZ and *Wealth Status variables were not available in both disabled and malnourished samples.

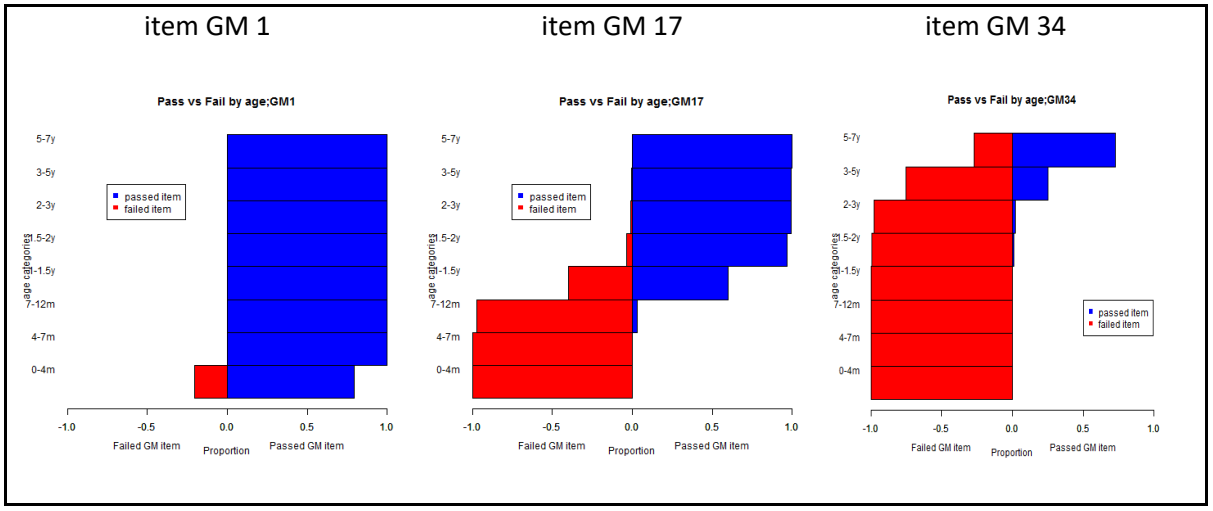
4.3.4. Exploring Item pass/failure rates

It may be not very helpful to simply assess distributional properties of variables univariately as it is their influence on the outcome of interest that is important in producing age estimates and scores. Therefore, it is more appropriate to assess the distribution properties of these variables with respect to the probability of passing an item. The mean and variance of dichotomous item variables can also be computed by multiplying π (proportion of children passing an item) by $1-\pi$ (proportion of

children failing an item) and the standard deviation is simply the square root of the variance got by the product of π(sample proportion for $\pi = 1$) by $1 - \pi$ respectively. This has thoroughly been described in various statistics and probability textbooks e.g. "A First Course in Probability, 5th edition" by Sheldon Ross, (1998). Again, the mean alone or with respect to other variables also gives an indication of distributional properties as well as some important item properties like difficulty levels that will be discussed in subsequent sections. The assessment of the distribution properties of these variables with respect to the probability of passing an item forms the preamble exploratory analysis required to advise the suitability of methods used in the item by item analyses to compute age estimates at pass probabilities of interest covered in Section 5.2.1.

The item pass/failure rate refers to the total number of children passing or failing an item divided by the total sample size or total number of children the item was administered to. Pass/failure rates per item in GM domain have been summarised in Table 4.4. Graphing item pass/failure rates with respect to child age was used to present the pass probability distribution with respect to age in the normal sample. Plots of histograms of the pass/failure rates in each domain were used to graphically evaluate the item pass rate distribution and confirm the expected item pass rate with increase in age. Figure 4.4 below show back to back histograms of the pass/failure rates against age for items 1, 17 and 34 that are ideal for babies (< 1 year old), toddlers (1 to < 3.5 years old) and pre-school age children (3.5 to < 7 years old) respectively in the gross motor domain only. The horizontal x axis shows the proportion of children who failed (in red) and passed (in blue) against the age category which is on the vertical y axis.

Figure 4.4: Back to back histograms items pass/failure rates against age for items that are ideal for babies (< 1 year old), toddlers (1 to < 3.5 years old) and pre-school age children (3.5 to < 7 years old) in the Gross motor domain.



From Figure 4.4 and pass rates in Table 4.4 we notice the evidence of the important underlying assumption of increased item pass probability with increase in age shown by increased item pass rates or reduction in failure rates as age increases. This aspect was evident across all four domains. This is a fundamental characteristic that the scoring method developed should adhere to and adequately or appropriately capture; thus the scores produced should be monotonically increasing with age.

Scatter diagrams of success proportions of items of were plotted against age to assess the item pass rate distribution and variability as age increased. While these scatter plots show the underlying pass distribution with respect to age, it also provides hints on how to adjust current methods to find more viable robust extensions to models when classical approaches fail to fit data appropriately. Item response data may often not conform or adhere to classical methods' underlying modelling assumptions due to various reasons stemming from the nature of primary item response, data collection process, quality of the assessment tool, design of the study and type of children sampled for example. The fitted item by item models shown in Figure 6.1 have been overlaid on a scatter plot of proportions of item pass rates given age to evaluate model fit by assessing the item pass rate distribution and variability as age increased.

We also checked the pass/fail rates within an item across all children. Having items with extreme pass rates or fail rates i.e. having all children pass or fail a given item is an indication of a poor tool item. None of the items in any of the four MDAT domains had 100 % or 0 % pass rates across cases. Generally, items that are ideal for infants will high pass rates and items that are ideal for pre-school aged children will have low pass rates as can be seen in the Table 4.4. We see that; a) the pass rates for items that are ideal for infants (< 1 year old) are high b) those of toddlers (1 to < 3.5 years old) are moderate c) while those of the pre-school aged children (3.5 to < 7 years old) are low. In these instances where the pass or fail rate is very high or low, some modelling assumptions may be violated and model fits are compromised due to identifiability issues that our extended robust methods will attempt to address in the item by item analysis described in Section 5.3.1.

4.3.5. Item Correlation

An assessment tool presents a complex correlation structure given that several tool items are administered to one child presenting a repeated measures scenario. Therefore, item correlation exists between the binary tool items given that one child responds to several questions/tasks and between children given the design of the study. The Tables 4.6 and 4.7 below give a clearer picture of these two sources of correlation.

Correlation between items

Correlation between items

Table 4.6: Item responses entered into data spreadsheet.

Child ID	GM Item 1	…	GM Item13	GM Item14	GM Item15	GM Item16	…	GM Item20	…	GM Item33	GM Item34
001	1	…	1	1	1	1	…	0	…	1	NA
002	1	…	1	1	1	1	…	0	…	NA	1
003	1		1	0	1	1	…	0	…	1	NA
004	1		1	1	1	0	…	1	…	1	1
…	…	…	…	…	…	…	…	…	…	…	…
†n=1446	…	…	…	…	…	…	…	…	…	…	…

†n refers to the child sample size of the study. Shaded in light grey is the correlation of item responses of item 15 and 16. NA-no item response.

The columns in the spreadsheet represent variables corresponding to tool items while the rows represent responses for each child or cases in the study that was assessed by the expert.

Correlation between children

Correlation between children

Table 4.7: Transposed Item responses entered into data spreadsheet.

Tool item	Child 1	Child 2	Child 3	Child 4	Child 5	Child 6	…	Child 200	…	Child 1445	Child 1446
GM 1	1	NA	1	1	1	1	…	0	…	1	1
GM 2	1	1	1	1	0	1	…	1	…	1	1
GM 3	1	1	1	0	0	0	…	0	…	1	1
GM 4	1	1	NA	1	NA	0	…	1	…	0	1
…	…	…	…	…	…	…	…	…	…	…	…
†n=34	…	…	…	…	…	…	…	…	…	…	…

†n refers to the total items in the tool. Shaded in light grey is the correlation of item responses of child ID 1445 and 1446. NA-no item response.

The columns in the spreadsheet represent each child's item responses considered as a variable while the rows represent tool items. Therefore, it is just a transposition of the previous data structure shown in Table 4.6 above.

The scoring methods employed generally assume the responses of different children to the same items are independent. However, this may not be entirely true for the following two reasons; a) Under the overall score domain scenario, it is likely that the responses to different items by a given child will be correlated as these responses will all depend on their ability or environment. b) Two tool's items may be correlated as they test the same underlying construct or ability milestone even if to differing degrees. Given that items in tools are designed to increase in difficulty, it means that is often the case that two adjacent items test the same, or almost the same construct. Thus it is likely that a child will respond in a similar fashion to items that are adjacent to each other in the tool due to item ordering, related or testing the same developmental aspect. Different modelling approaches considered in this thesis dealt with correlation in different ways.

As described in Section 2.3.3.3, the polychoric correlation coefficient was used to compute the required correlation between items was summarised in a correlation table. The correlation matrix for the within child correlation is not shown due to its large dimension. As expected, we can conclude the following from the correlation values;

- There was strong correlation between items that are close together
- There was strong correlation of item responses between children of similar age

4.4. Exploring Specific Item Characteristics

Following the preliminary item data checks described in Section 4.3, this section describes some specific item characteristics to investigate that are related to the pass/fail rates and are important in terms of the assumptions, ramifications and limitations or methods used to compute both age estimates and overall scores. Most of these item characteristics are intuitive and stem from classical test theory (CTT) as noted by both Embreston & Reise (2000) and Kline (2005) but are quite important to check at this early stage. This is because they help detect possible item problems and thus assess tool quality. If the item characteristics are poor and are left unchecked, they will definitely result in poor quality of scores and compromise the accuracy of development status classification.

At this stage while reflecting on these item characteristics, we wish to plant seed for the question 'which is the ideal age estimation or scoring approach?' We will attempt to answer this question by reflecting more on when and how either the item by item or the overall scoring approaches described in sections 5.3.1 and 5.3.2 are suited to deal with these item characteristics. It is then that a deeper appreciation of a detailed item response exploratory analysis will be appreciated given each scoring approaches' pros and cons. While it may later become apparent that we advocate for an overall score with reason, it should be obvious that building scores is a process, thus although this item characteristic exploration process seems laborious, it is bound to be very beneficial in the long run. As was noted in chapter one, one of the major contributions of this thesis is to devise a framework to be able to adjust for age in the computation of scores. Therefore, the following item characteristic exploration will to a large extent assess the influence of age in various respects of interest.

4.4.1. Item difficulty levels

As defined and discussed by Kline (2005), the difficulty level of an item refers to the number of passes in a dichotomous item expressed as a proportion of the total number of individuals (children) who undertook that item. This proportion of passes is usually referred to as the item's p value in psychology and education sectors. Therefore, to avoid confusion with the conventional p value that indicates the strength of an association in statistics, we will refer to the item pass rate as pr value in this thesis. As we have seen earlier in Section 4.3.4, high pass rates or high pr values will be seen in items that are ideal for very young children (babies) and low pass rates will be seen in items that are ideal for older children (pre-school age). Therefore, if an item has a high pass rate, it can be considered easy as almost all the children across the age spectrum will pass it, while an item with a low pass rate can be considered difficult as only a few who are likely to be older pass this item. Item difficulty ranges from zero to one, therefore items with extreme prvalues towards zero or one that correspond to either very easy or difficult items, meaning everyone tested passes or fails are not very useful in a tool. As we will see later in sections 4.5.2 and 4.5.3 and later in the item by item analysis in section 5.2.1, these items cannot adequately differentiate or discriminate between individuals. It can be shown that items with a 50% pass rate provide the highest levels of differentiation between subjects in samples in a binary response scenario. Therefore, items with a pr value of approximately 0.5 are more useful in differentiating assessed children. There are various recommendations for ideal pass rate ranges for various types of response scales. Table 4.4 showed a summary of the item pass rates (pr values) for the gross motor domain. It is evident that item difficulty increased from item 1 towards item 34.

It is important to note that if the tool items are strongly inter-correlated, this presents a problem. This is because even if an item has a pr value of 0.5 but is strongly correlated with other items, administration of the single item would have been enough to differentiate between children as highlighted in detail by Kline (2005). As mentioned earlier, for purposes of assessing developmental ability, most developmental tools are designed to increase in difficulty. In line with the above, it is

important to check that the pass rate or pr value increases with age. However, this only confirms that difficulty increase with age aspect has been appropriately inbuilt into the tool given that ability increases with age.

4.4.2. Item Discrimination index

A key aspect in the assessment of item quality that is related to item difficulty is the level of discrimination each included item can achieve. Within a child development context, the discrimination of a tool expresses the proportion of children able to pass a given item. The index of discrimination is a useful measure of item quality whenever the purpose of a test is to produce a spread of scores that can reflect or detect differences in child ability. Thus a tool item that is either too easy or too hard corresponding to high or low pass rates respectively will have low discrimination and should be replaced. This is so that distinctions or classifications may be made as far as their ability is concerned which is the purpose of most norm-referenced tests or tools.

Pass rates or pr values can be used to compute discrimination indexes that are often referred to as D values, see Kline (2005). Higher D values point to higher discriminating items. It can be shown that the highest D values are achieved by items with pr values of approximately 0.5. Therefore, tabulating D values of items at this early item exploratory stage with pin point 'poor' items initiating the appropriate rudimentary measures. There are various methods to compute D values for example the three step extreme group method described also by Kline (2005). To compute the item discrimination index, D, for the MDAT data, the following steps were followed;

- Compute the total raw sum score using all 34 items within a domain and then the 0.27 and 0.73 quantile values for total score. These quantile thresholds were chosen following recommendations documented in Kline (2005), but we note that they should be further guided by the pass/fail rate distribution for the respective domain i.e. approximately 50 % of the children should have a total score between the chosen lower and upper quantiles.

- Categorise data according to the lower and upper total score quantile thresholds and compute pr rates for each item for the lower and upper quantile groups i.e. for each item compute the pass rate for the children in lower and upper groups respectively.
- Compute D index which is the difference in pr rate obtained above between the lower group and upper group.

Discrimination indexes were computed for the GM domain items are tabulated in Table 4.8. From the table we see that the gross motor items 6 to 18 have very high (>0.70) discrimination indexes (highlighted in red). The values also confirm our previous findings shown in Table 4.4 that these items had relatively high pass rates and also that children with higher overall raw total test scores are more likely to pass these items than children with lower overall raw total test scores. Items 19 to 26 had very low discrimination index scores and no child in the lower level group managed to pass these items. The GM items 1 to 5 and 27 to 34 had medium D values. It is possible to have a negative discrimination D index score. Within a measuring ability context, this indicates a 'poor' item in that it implies those children with higher test scores, or those that had more ability, are not likely to pass this item, while those with lower test scores, or lower ability are likely to pass this item. This goes against what is expected naturally and therefore not intuitive in an ability assessment process. The normal standardisation sample MDAT data no item with a negative D score in the gross motor domain. This offers further reassurance that the MDAT is a scientifically developed assessment tool.

We see that item difficulty quantified by the pr value is related to discrimination D score. We would like to note that the total score upper or lower percentile threshold used will influence the value of the D score. Therefore, the quantile thresholds used should be appropriately motivated and reflect the distribution of item pass rates in your data. We recommend using the item pass rate distribution summarised in Table 4.4 to choose the appropriate thresholds to use in computing item D scores. Alternatively, the use of empirical item characteristic curves discussed in section 4.5.5 can be used to choose the appropriate thresholds to use in computing item D scores.

Table 4.8: MDAT discrimination (D) item indices for the GM domain.

GM domain				
Type of item	**Item #**	**pr -Level for Upper Group n=405 (28.00 %)**	**pr -Level for Lower Group n=395 (27.32 %)**	**$^{\dagger}D$**
Ideal items for infants <1 year old	**1**	1.00	0.82	0.18
	2	1.00	0.60	0.40
	3	1.00	0.41	0.59
	4	1.00	0.53	0.47
	5	1.00	0.38	0.62
	6	1.00	0.27	0.73
	7	1.00	0.29	0.71
	8	1.00	0.24	0.76
	9	1.00	0.10	0.90
	10	1.00	0.01	0.99
	11	1.00	0.00	0.99
	12	1.00	0.00	1.00
Ideal items for toddlers 1 to <3.5 years old	**13**	1.00	0.00	1.00
	14	1.00	0.00	1.00
	15	1.00	0.00	1.00
	16	1.00	0.00	1.00
	17	1.00	0.00	1.00
	18	1.00	0.00	1.00
	19	0.02	0.01	0.01
	20	0.02	0.01	0.01
	21	0.03	0.01	0.02
	22	0.06	0.00	0.06
Ideal items for pre- School aged children 3.5 to <7 years old	**23**	0.06	0.01	0.05
	24	0.05	0.00	0.05
	25	0.06	0.00	0.06
	26	0.08	0.00	0.08
	27	0.18	0.00	0.18
	28	0.22	0.00	0.22
	29	0.10	0.00	0.09
	30	0.33	0.00	0.32
	31	0.21	0.00	0.20
	32	0.28	0.00	0.27
	33	0.57	0.00	0.57
	34	0.59	0.00	0.59

pr-levels are the pass rates for each item

$^{\dagger}D$ Score-discrimination index. Highlighted in red dotted box are items with high D scores.

4.4.3. Item to raw total score correlations

It is important to assess how the responses to an item relate to the total raw aggregated score for all items. Given that item responses are binary, and the total raw score is continuous, their correlation is usually measured using a Pearson point-biserial item to total correlation coefficient as discussed in numerous references for example Sheskin (2011). As explained in Kline (2005), it is also important to correct the total score not to include the response of the item in question. For example, if we are computing the correlation of item 1 to the total raw score, then the item 1 binary responses should not be included in the total raw score.

Given that all of the responses in MDAT represent a true dichotomy (i.e. pass/fail), there is a vector of binary responses for each item and a continuous total score. Therefore, the Pearson point-biserial item to total correlation coefficient is the appropriate statistic to investigate item to total correlation. Formally, the Pearson point-biserial item to total correlation coefficient is given by the formulae;

$$R_{pbis} = \left[\frac{\overline{X_1} - \overline{X_0}}{\sigma_x}\right] \times \sqrt{\pi/1-\pi} \qquad 4.1$$

where; $\overline{X_1}$ and $\overline{X_0}$ denote the sample means of the X values corresponding to the first (item pass=1) and second (item fail=0) level of Y (item responses) respectively,

σ_x = the standard deviation of X,

π = the proportion of children who passed an item or whose response was 1,

$1-\pi$ = the proportion of children who failed an item or whose response was 0.

Table 4.9 summarises the Pearson point-biserial correlation coefficients for the gross motor domain items. In general, there were items with relatively low (item 1; highlighted in blue box), medium (items 2 to 8 and items 30, 33 and 34; highlighted in green box) to very high (items 9 to 29 and item 31 to 32 highlighted in red box) correlation with the score raw total. The Pearson point-biserial item to total

correlation coefficient possesses typical correlation coefficient characteristics. Firstly, the coefficient values range between ±1.0. Secondly, a high positive point-biserial value indicates that high performing children are likely to get the item right and those with low score are likely to fail the item which is intuitively expected. On the other hand, a low positive point-biserial value would indicate that children passing the item end up with a low overall score and those who fail the item tend to have an overall high score which would point to a possible item anomaly.

Therefore, the point-biserial correlation is also a form of an index of item discrimination as it shows us how well the item serves to discriminate between children with higher and lower levels of total raw scores i.e. the point-biserial correlation reflects the degree of relationship between passed (1), failed (0) and total raw sum test scores. Thus the point-biserial was positive if better children answered the item correctly more frequently than poorer children did, and it was negative if the opposite occurred. The value of a positive point-biserial discrimination index can range between 0 and 1 so that the closer the value was to 1, the better the item discrimination. The value of a negative point-biserial discrimination index can range between -1 and 0, but positive values were desirable. Item discrimination is greatly influenced by item difficulty. In general, from the item pass rate values in Table 4.4 and discrimination values in Table 4.8 we saw that; a) items with a difficulty (pass rate) close to either 0 or 1 in turn had a low discrimination index value close to 0 e.g. item GM 1 or GM 31 and b) item discrimination, reflected by high a point-biserial correlation was maximized when item difficulty (pass rate) was close to 0.5 e.g. item GM 12 to GM 18. It is recommended that items with point-biserial values of 0.20 and above should be considered to be desirable.

Table 4.9: Item to total correlations in gross motor domain.

Type of item	items	*Point biserial coefficient*
Ideal items for infants <1 year old	GM1	0.34
	GM2	0.52
	GM3	0.62
	GM4	0.57
	GM5	0.63
	GM6	0.68
	GM7	0.67
	GM8	0.69
	GM9	0.73
	GM10	0.74
	GM11	0.79
	GM12	0.80
Ideal items for toddlers 1 to <3.5 years old	GM13	0.82
	GM14	0.83
	GM15	0.84
	GM16	0.85
	GM17	0.85
	GM18	0.85
	GM19	0.83
	GM20	0.82
	GM21	0.83
	GM22	0.78
Ideal items for pre- School aged children 3.5 to <7 years old	GM23	0.80
	GM24	0.79
	GM25	0.80
	GM26	0.78
	GM27	0.75
	GM28	0.75
	GM29	0.79
	GM30	0.70
	GM31	0.75
	GM32	0.72
	GM33	0.56
	GM34	0.54

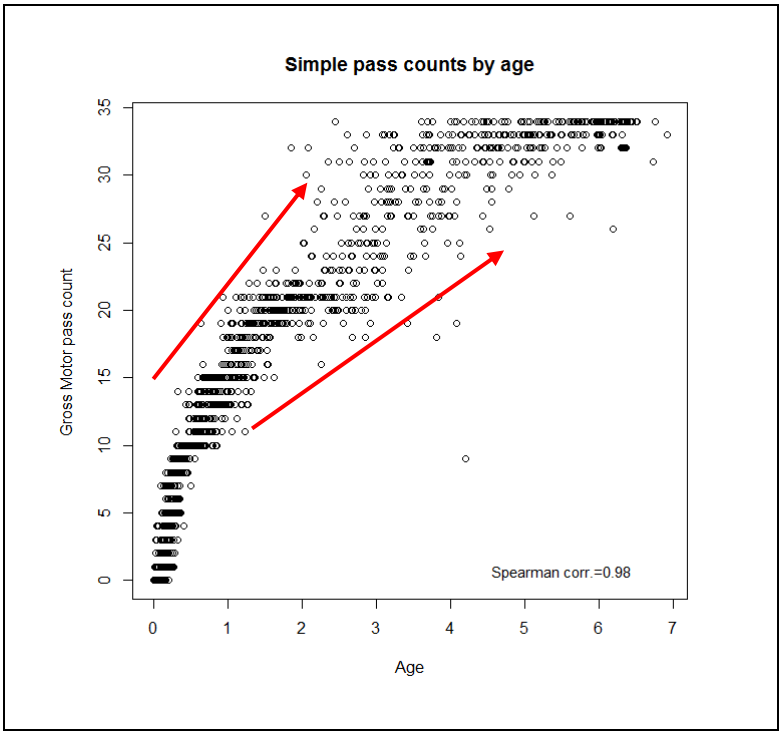
Highlighted in blue, green and red boxes are items with low, medium and high point-biserial correlation coefficients respectively.

4.4.4. Total raw score and age correlation

In Section 1.5 we explained one important objective of this thesis is to suggest extensions of overall scoring methodology successfully adjust for the effect of age known to be strongly associated with the overall scores. In section 4.3.4 we saw that the probability of passing an item increases with age. In this section we explore if indeed this relationship between age and an overall score like the raw total

score exists and describe its characteristics. The Figure 4.5 is a scatter plot of the total raw score of the normal children gross motor overall raw total scores against age.

Figure 4.5: Scatter plot of total sum raw score by age



Red arrows show the increase in overall total raw score variability as age increases.

From the scatter plot of overall raw total score and age, we see that;

- The raw total score has a strong curvilinear relation with age. The spearman correlation coefficient was 0.98 and significant ($p<0.00139$) indicating a strong correlation between the two variables.
- There is a noticeable rapid rate of development reflected by a high raw total score rate in the immediate first few months after birth. This fact was apparent from the scatter plots of item pass probability against age shown in Section 6.2. This is attributed to the fast development rate in first few months after birth but could be a consequence of the ceiling effect of respondents not being able to score above 34.
- There is also an increase in the total raw score variability as age increases as indicated by the red arrows in Figure 4.5. The increase in variability was attributed to the fact that the development milestones are very variably developed and defined as children get older.

Section 5.3.2 will discuss how both the classical overall scoring methods and the suggested extensions deal with the observed relation of age and the overall raw total score.

4.4.5. Empirical Item Characteristic Curves

As described in Theresa, (2005), a plot of a child's overall total raw score against pass rates summarises how an individual child performs on a single item against the overall performance level of the other children. This produces a curve typically called an item characteristic curve. This method was pioneered by research developed in the 1960's and work by the famous researcher Rasch, (1960). Such curves (see Figure C.1 of Appendix C for the GM domain) using pr values can give much more valuable insight into the underlying item response characteristics than the stand alone pr value or *D* index described in Sections 4.5.1 and 4.5.2. At the end of this section the reader will appreciate a) the use of empirical item characteristic curves to consolidate all the item specific response characteristics and their interrelationships as well as the suitability for perusing the Item Response Theory framework to compute overall scores discussed in Section 5.2.2.3.

These item curves are able to also simultaneously graphically show the rate increase of performance of a child with the corresponding performance on the items. Therefore, a steep monotonic increase will indicate high discrimination capability between children responding to that item i.e. it will be able to differentiate between children able to pass or fail this item. Also difficulty of an item is implied by the starting points of the curves. We will show this concept at this stage using the simple total scores and the IRT models described in Section 5.2.2.3 will validate any conclusions drawn at this point. Also the behaviour of overall total raw score with other important variables of interest such as age can be explored in a similar manner. Given that each domain had 34 items and the 1446 children that were assessed by the MDAT tool, the overall total raw score for each domain was computed and pr values computed for each item for children within every 10^{th} percentile up to the 99^{th} percentile. Each of the percentiles is then plotted against the respective pr level associated with that item for the respective percentile as described in Theresa, (2005).

Given the observation that most items ideal for toddlers and pre-school aged children had a plateau pr level for the initial percentile range and then change abruptly, meaning that they are very relevant at discriminating at these percentile values, we were prompted to plot pr levels are various age groups. This is further motivated by the fact that we are evaluating the MDAT that evaluated child ability that is known to be greatly influenced by child age. Therefore, the overall total raw score for each domain was computed and pr values computed for each item for children within the age categories of interest. Each of the item pr values is then plotted against the respective age category. More specifically the following steps were followed to draw the empirical item curves by item pr value of total score at age categories of interest;

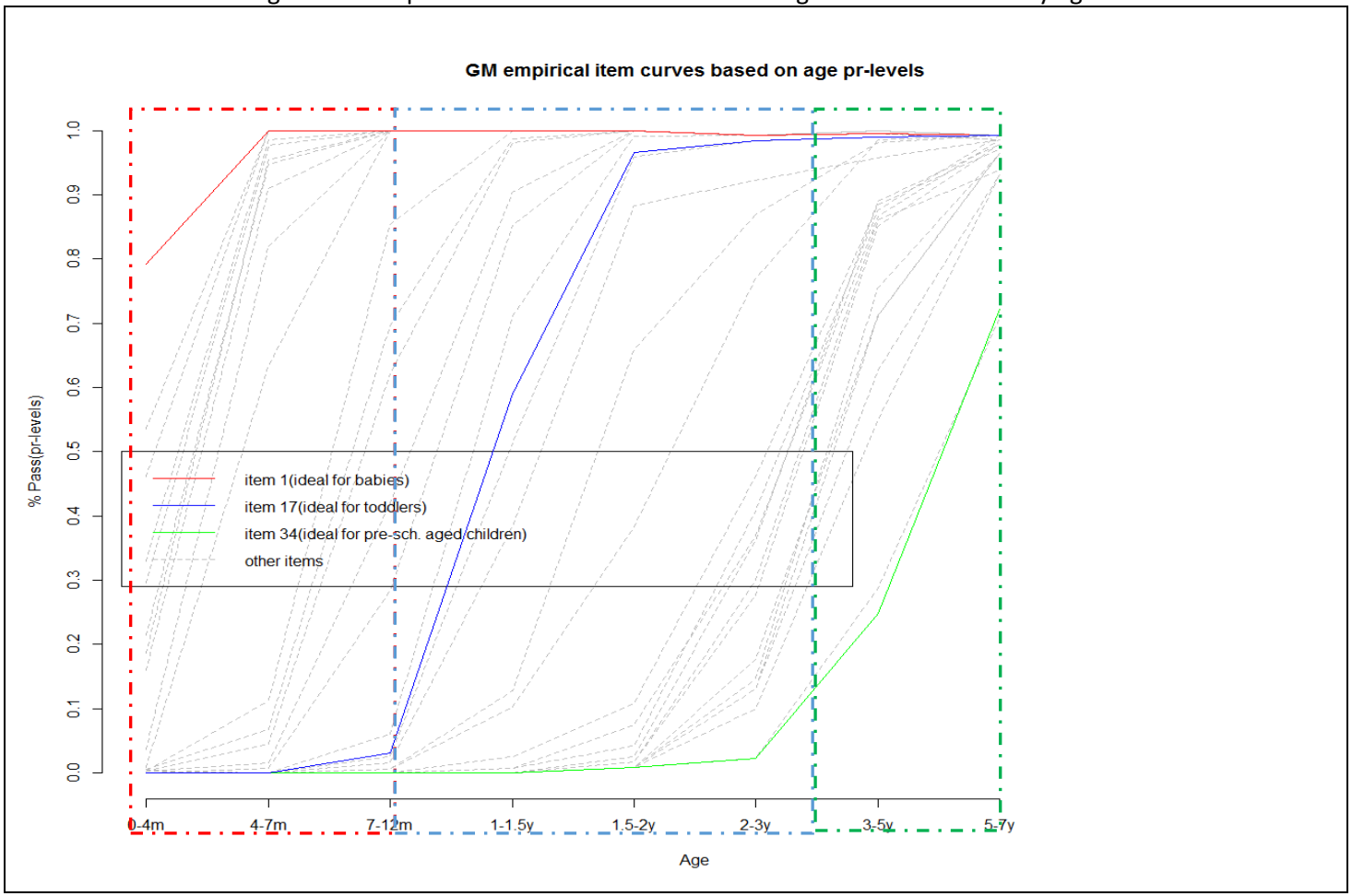
- Compute the overall total raw score for each child in each domain and create an 8 level indicator variable for each child that indicates which age category they belong to and create 8 separate data sets for each age category. Our eight age categories of interest were at 0 to less than 4 months, 4 to < 7 months, 7 to < 12 months, 1 year to < 1.5 years, 1.5 years to < 2 years, 2 years to < 3 years, 3 years to < 5 years and 5 years to < 7 years.
- Compute pass rates for all the 34 items in each of the 8 separate data sets above. The item pass rate for each item in each data set will be the count of children passing an item divided by the number (n) of children in that age category. Combine these items' pass rates in a 34 by 8 array i.e. 34 domain item pass rates or pr values in the 8 data sets.
- Make an empirical item curve for each item that is a plot of the item pass rates (pr values) over the eight age categories.

From the empirical item curves shown in the Figure 4.6, we see that pr levels for the item 1 (the red item curve) and the first 10 or so items (in red dotted box) were quite high and almost all children in the normal cohort were passing these items. This was seen as the very steep slope of item curves even at very low age values. This was an indication that these items were particularly useful at discriminating between the children who are very young babies (< 6 months) but offered no discrimination capability for toddlers or pre-school aged children. Item 17 that was shown as the blue

item curve and the other items in the blue dotted box showed almost no change in pr level until around 7 months where there was a sudden change in pr value. This was an indication that these items were particularly useful at discriminating between the children who were between 7 to 2 years old but offered no discrimination capability at other age categories. Similarly, item 34 that was shown as the green item curve and the other items in the green dotted box showed almost no change in pr level until after 2 years where there was a sudden change in pr value. This was an indication that these items were particularly only useful at discriminating between the children who are much older (pre-school age) but offer no discrimination capability at the other age groups. Also notice the clustering together of items that are ideal for babies (< 1 year old), toddlers (1 to < 3.5 years old) and pre-school aged children (3.5 to < 7 years old) which points to the strong correlation of items found in these clusters that assess similar constructs. The fact that in general the conclusions drawn from plotting the empirical item curves either using total score percentiles or age are similar or complement each other's conclusions, can be tied to the fact that these 3 variables are strongly associated i.e. as age increases, we expect the probability of passing an item to increase. Therefore, a higher pr level is expected, and this is in turn reflected by higher total score counts as children get older because of the inbuilt increase in item difficulty level.

At this point we wish to make the point that the use of empirical item curves is a more suitable method of exploring the suitability and likely performance of items as we know at which total score levels and ages they are relevant in terms of being able to differentiate or discriminate children across the raw score or age spectrums. The item curves by age echoed the findings already seen by the item curves by percentile total score. However, there is added benefit that now one can make conclusions on potential of discriminating power of an item at a given age. Further, it is now easier to view the clustering of items that are ideal for babies (< 1 year old), toddlers (1 to < 3.5 years old) and pre-school aged children (3.5 to < 7 years old) in the red, blue and green dotted boxes respectively. Therefore, a useful criteria to select items relevant for a given child age group could be selecting items with good or high discriminating potential at the given age groups.

Figure 4.6: Empirical item characteristic curves of gross motor domain by age.



pr-levels are the pass rates for each item

Items that are ideal for babies (< 1 year old), toddlers (1 to < 3.5 years old) and pre-school aged children (3.5 to < 7 years old) in the red, blue and green dotted boxes respectively.

4.5. Summary

Adequate item response (data) exploration ensures that 'poor' performing items are identified well in advance of the score computation process. This fourth chapter has discussed the initial or preliminary steps to undertake once the data has been collected; the exploratory of data process. We have highlighted the importance of this process as it advices us in any underlying mechanisms in the data and can be used to flag any items that are not particularly useful in the development of the score and can thus be removed or reviewed.

In each section of exploring item characteristics we have discussed important item response characteristics whose presence and extent is vital to advice on the most ideal statistical approaches to use to compute age estimates and overall scores. Further, our discussions in each section have also highlighted possible problematic items with extreme item response characteristics leading to a form of critical appraisal of items. However, in as much as these problematic items should be removed, their 'poor' item characteristics may be as a result of the study design and not necessarily a consequence of poor tool design. Further, it is these items with 'poor' item characteristics that we hope our suggested robust statistical methods will address given our contention that the MDAT assessment tool is a scientifically developed assessment tool. We have also seen that the item characteristics of interest were similar across all four MDAT domains. Therefore, deliberately, as the remit of this thesis is in extending scoring methodology, we will only present results for the gross motor domain as the conclusions drawn across the other domains are likely to be similar.

The following methods chapter will describe the current or classical methods used to derive age estimates and overall scores that are often dependent on age. We will further show the importance and explain our motivation for establishing a framework to correct or adjust for age that this chapter has shown to be strongly associated with the probability of passing an item.

5. Methods of Scoring Binary Assessment Data

5.1. Introduction

This chapter outlines the statistical methods used to either compute item by item age estimates for assessing age specific ability milestones or overall scores to classify development status for each child assessed using the MDAT tool. While a formal description of the different statistical approaches used, we will highlight the limitations of the classical methods which motivate the more robust methods suggested. It is assumed that the assessment tool is of high quality i.e. appropriate methods of translating, adapting as well as assessing reliability and validity, as set out in the second chapter, have already been applied to the MDAT tool.

Section 5.2 describes the scoring methods under two scenarios; (i) assessing items individually using a generalised linear model (GLM), an extension of the traditional generalized additive model (GAM) that uses the shape constrained additive model (SCAM) approach, and (ii) scoring across all items for a given child using a simple additive score as outcome for a model based scoring method, Z-scores and Item response theory methods. The chapter concludes with details of the methods used to compare both the score distribution characteristics and sensitivity in Section 5.3, which serves to objectively compare and contrast each age estimation and overall scoring method described.

The R statistical software and relevant packages were used to carry out the data manipulation, exploration and to implement the salient statistical analysis methods described in this fifth Chapter.

5.2. Scoring methods

After the preferred tool to carry out the developmental assessment has been identified (or translated and adapted) and the necessary data has been collected, then follows the intricate process of computing item specific age estimates or converting these data to normative scores to classify developmental status of children as either normal, or delayed.

As elucidated earlier in Chapter two, current scoring methods can be broadly categorised into two main groups; item-by-item analysis and overall scoring. Item-by-item analysis generates an age estimate for each item using a sample of healthy normal children, against which 'new' children can be assessed. Here the main interest is in estimating the ages at which a child from the population of 'healthy and normal' children has a probability, usually of 25%, 50%, 75% and 90%, of passing that item. The passing of an item is an indication that the child has achieved a specific developmental milestone. Hence the importance of choosing assessment items very carefully. This approach is currently used in the MDAT tool. Thus the focus of item by item analysis is principally to give age estimates at predefined probabilities, i.e. the expected age at which a healthy normally developing child should pass an item. The item-by-item analysis approach has the additional benefit of also checking the item quality in terms of its discriminatory ability. In the instance that a question or task has not been well administered due to cultural inappropriateness, its high pass rates variability manifests itself by poor model fits. This could be an indication that the item is not a suitable item to include in the assessment tool and needs to be substituted or removed.

The total score method considers all the tool item responses simultaneously across the entire domain. It therefore gives one score value (index or norm) that summarises the performance of the child over all items administered and hence captures their ability status. While this latter approach utilises all the item response details of a given child and provides a score combining all items for each domain for each assessed child, it does not use information on precisely which items a child could or could not complete. The total score gives the expected score in a given domain that a child of a given age should achieve comfortably if developing normally. Thus the focus is on the child specific score and its distance (above or below) the norm to classify the developmental status of the child. Each of the scoring methods will be explained in the following subsections highlighting their pros and cons as well as other important modelling issues that should be addressed. Our suggested extensions will be made pragmatically and while made on the basis of the cons of the current methods we will also consider their practical implementation using available software.

5.2.1. Item by item analysis within each developmental domain

In this scoring scenario, each item's responses in the standardisation sample for within a given domain is considered individually. Gladstone, et al., (2008) demonstrated the use of logistic regression models with decimal age as an explanatory variable to create age estimates that characterised a child's ability to pass individual items at defined age ranges during developmental assessment. In instances where there was a poor model fit, they considered the use of triple split spline regression as described in Greenland, (1995c), Pastor & Guallar, (1998), Smith, (1979) and Lemeshow & Hosmer, (1982). Their method assesses every binary item (question) in a given MDAT domain separately. Thus for example in the MDAT gross motor domain, they fitted 34 separate logistic models, each corresponding to one item in that domain.

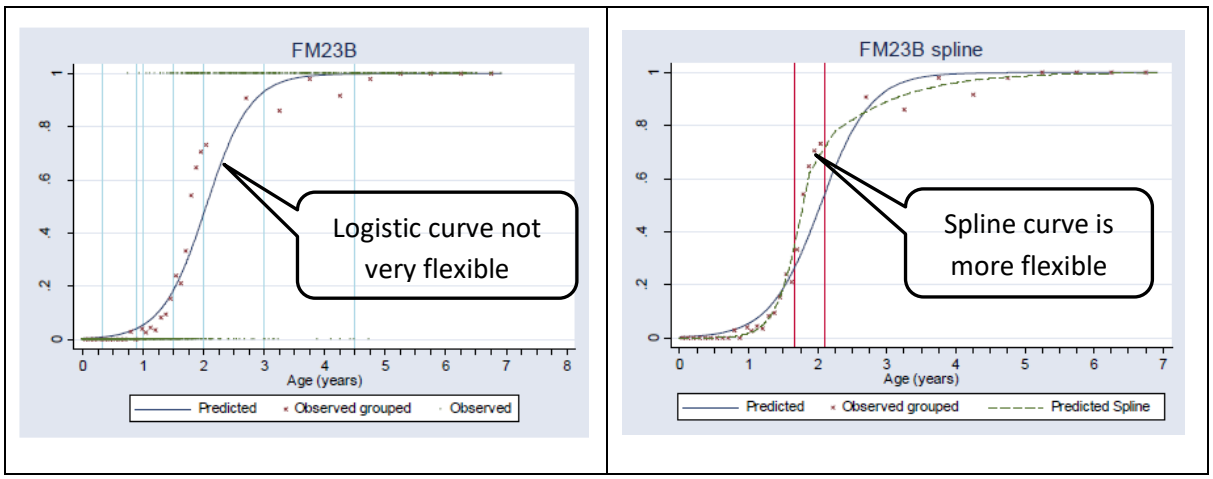
We therefore aim to investigate if the age estimates used to assess child ability development may be more accurately and appropriately estimated using more flexible statistical modelling methods. More appropriate model fits will be evidenced by improved model fits and consistency of estimates as described in Section 5.3.1. The application of the logistic model under the Generalized Linear Model framework is reviewed and the performance of suggested solutions in cases where the former did not fit the data well will be highlighted.

A standard logistic regression model might assume that the effect of age is linear with respect to the log-odds of passing the item or perhaps that it is linear in terms of the log of age. The logistic model may not fit well if the underlying logistic distributional assumptions of linearity in the logit for example are violated. It is possible to improve the fit of the model by allowing the pass rate to depend on a suitable spline function of age. However, the use of splines requires one to carefully choose the type, number as well as the location of knots objectively and can thus easily produce models that over fit data but in turn poorly fit new validation data.

As described in Section 2.3.2.1 b), a spline is basically a numeric function that possesses a high degree of flexibility or smoothness to allow curve fitting between two variables that have a non-linear

relationship. An example of an item where the logistic regression model did not fit well and a spline was used to improve model fit is shown in Figure 5.1 below from the work of Gladstone, et al., (2008).

Figure 5.1: Comparison of logistic and spline fit on MDAT item



*Sourced from Gladstone, et al., (2008).

In addition to logistic regression, which is a special case of a Generalized Linear Model, the following sections consider the application of Generalized Additive Model (Hastie & Tibshirani, 1986; Wood, 2006) framework to the MDAT data focusing especially on its extensions by the work of Pya & Wood, (2015) and Pya, (2012) involving Shape Constrained Additive Models. The extension under the generalized additive model framework offer a much more flexible and systematic approach to modelling the proportion of children successfully passing an item given age when classical methods like logistic regression fail to fit the data well. Therefore, we will be inquiring whether the benefit of achieving better item model fits using a more flexible methodology translates into achieving more accurate, improved quality and generalizable item age estimates to assess ability in children.

5.2.1.1. Generalized Linear Models – Logistic Regression (GLM)

The most widely used method for such binary and/or dichotomous outcomes is the logistic regression model. Logistic regression is part of a category of statistical models commonly referred to as generalized linear models (GLM). Within this framework, the response of interest Y (the dependent variable or outcome) is dichotomous, i.e. an outcome with only two possible levels. One is called the 'success/presence/pass' outcome (taking value 1) and the other is called the 'failure/absence/fail'

outcome (taking value 0) with probabilities π and $1-\pi$, respectively, Y as a Bernoulli random variable with parameter $E\{Y\} = \pi$ (Agresti, 2002). The simple logistic regression model can be written as;

$$Y_i = E\{Y_i\} + \varepsilon_i$$

Since the response variable returns a binary outcome, the relationship between the dependent and independent variable(s) is non-linear, hence standard linear regression models are not applicable. Further, several assumptions of classical linear regression (Neter, et al., 1996) are no longer viable; for example, the assumption of continuity (normality) of the response variable which is one of the fundamental requirements of linear models will not be appropriately adhered to and hence a relevant way to handle binary data is required. As the error term ε_i depends on the Bernoulli distribution of the response Y_i, it is preferred to state the simple logistic regression model as;

$$E\{Y_i\} = \pi_i = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}$$

where; Y_i are independent Bernoulli random variables with expected values,

$E\{Y_i\} = \pi_i$,

$\pi_i = P(Y = 1|X = x) = 1 - P(Y = 0|X = x)$,

X is an explanatory variable.

The logistic model formulates the logarithm of the odds of a success probability (the logit of the probability of success) as a linear function of one or a set of explanatory variables which can be of any kind; continuous, categorical, or the combination of these. If there is only a single explanatory variable, this model is known as a simple logistic regression model, otherwise the model is called a multiple logistic regression model. Formally, if we consider an explanatory variable denoted by X, then the equivalent linear form (linear predictor) of the logistic regression model is given by;

$$logit\big(\pi(X)\big) = log\left(\frac{\pi(X)}{1-\pi(X)}\right) = \beta_0 + \beta_1 X_i \qquad (5.1)$$

where; $\pi(X)$ is the probability of success (in this case it is the chance of getting an item correct),

β_0 corresponds to the intercept term,

β_1 is a fixed effect coefficient of the explanatory (age) variable.

Therefore the GLM model has the following basic structure;

$$g(\mu_i) = X_i\beta$$

where; $\mu_i = E\{Y_i\}$,

Y_i~exponential family distribution,

g is a smooth monotonic 'link function',

X_i is the i^{th} row of a model matrix, X ,

β is a vector of unknown parameters; In this case this will be one parameter for the age variable.

In Section 4.4.3, it was shown that the overall raw total sum score is positively correlated with pass rate and therefore ability. This clarifies the fact that with increasing age, ability is expected to increase. Therefore, the modelling approach should be able to adequately capture this feature by being monotonic i.e. it is entirely increasing. The GLM logistic model 5.1 above specifies that the log-odds of passing an item are described by a linear equation in age. Therefore, provided that the coefficient of age, β_1, is positive we are guaranteed to have increasing log-odds with age. If the log-odds are increasing this also implies the odds and probability must also be increasing with age. This means that the GLM model satisfy the required monotonicity assumption. The main limitation of the ordinary logistic regression is the underlying assumption of a linear effect of the covariates on the log-odds of passing an item. When this assumption is not adhered to, one cannot make 'safe' conclusions or draw reliable inference from the model in this child development context.

5.2.1.2. Generalized Additive Models – (GAM)

As outlined in the GLM Section 5.2.1.1, because modelling the non-linear relationship between the dependent and independent variable(s) presents a challenge with skewed response distributions, the GAM framework offers a more flexible approach. A closer look at the total raw sum score or pass rate relation with age reveals that it is somewhat curvilinear. There are changes in the gradient raw score or item pass rate as age values increase. Therefore, this warrants a more flexible approach to adequately capture this rather complex pass rate and age relation, hence our motivation of considering the GAM approach. Below is a brief formal definition of the GAM model.

If $Y_i, i = 1, \dots, n$, are independent observations of a response variable from a Bernoulli (or an exponential family) distribution and $x_{1i}, x_{2i}, \dots, x_{pi}$ are the explanatory variables, then the response is linked to an additive effect of the explanatory variables through a known link function (Hastie & Tibshirani 1986; 1990). The effect of the explanatory variables may be non-linear. This generalized additive model which is a generalized linear model with a linear predictor involving a sum of smooth functions of covariates can be written as follows;

$$g(\mu_i) = X_i^* \delta + f_1(x_{1i}) + f_1(x_{1i}) + f_1(x_{1i}) + \cdots (5.2a)$$

$$g(\mu_i) = X_i^* \delta + \sum_{j=1}^{p} f_j(x_{ji}) \qquad (5.2b)$$

where; $\mu_i = E\{Y_i\}$, and Y_i~exponential family distribution,

Y_i is the response variable,

$g(.)$ is a known smooth monotone log link function with respect to μ_i,

X_i^* is the i^{th} row of a model matrix for any strictly parametric effects or model components,

$\delta = (\delta_1, \delta_2, \dots \delta_{q0})^T$ are unknown vector parameters,

$f_j(x_{ji})$ are smooth unknown functions of the covariates $(x_{1i}, x_{2i}, \dots, x_{pi})$, written as a vector.

Since in our research the item responses Y_i, are binary, then the typical approach would be to take $g(.)$ to be the logit function. The $f_j(x_{ji})$ functions take the form of a simple coefficient, polynomials or fractional polynomials that may be specified parametrically, or semi to non-parametrically, to smooth functions that are estimated non-parametrically for example using a locally weighted mean. This flexibility allowing a non-parametric model specification and fit with relaxed assumptions provides better fits than the purely parametric models. Wood, (2008) explains the pros and cons of different approaches of avoiding overfitting using generalized cross-validation (Craven & Wahba, 1979), penalized likelihood estimation or Akaike Information Criterion AIC (Akaike, 1973) to optimally select smoothing functions. However, this flexibility can still inadvertently lead to overfitting even if reasonable types and numbers of smoothing functions are specified. Recall that the modelling approach should adequately capture the fact that ability (captured by the item pass rate) increases with age.

However, as noted in Section 2.3.2.1c) the GAM framework's flexibility is not restricted to be always monotonic, thus this approach may not always be appropriate to model pass rates under this ability measurement context i.e. the GAM model framework allows the log odds of passing an item through an unspecified function of X, $g(.)$ that is not necessarily increasing, and hence the pass probability is not always increasing with respect to X or age as expected. Indeed the pass probability may fail to increase with age due to measurement error stemming from a myriad of sources e.g. measurement error, but we insist on having a monotone transformation to also ensure that we can reverse the transformation to recover data values in the original scale. We acknowledge that the use of specific smooth monotonic functions or other link functions as described by Jiang, et al., (2014), Giampiero & Rosalba, (2010) and Aranda-Ordaz (1981) provide numerous avenues for flexibility and obtaining a monotonic function. But it would be still be difficult to not only prevent certain combinations of the coefficients from making the raw data process defined to be non-monotone, but also lack one practically usable unifying model framework as each item would potentially have its own link function aside from different smooth monotonic function on age. Instead, the following section will now

describe an extension of the GAM model that instead includes a unifying monotonic increasing constraint.

Shape Constrained Additive Models – SCAM

Within the generalised additive model framework outlined in Section 5.2.1.2 above, the effect of the explanatory variables $f_j(x_{ji})$ given in equation 5.2 is not restricted to be monotonic. Recall that one of the conclusions of the exploratory data analysis in Chapter four is that ability increases with age, thus a monotone increasing constraint should be implemented within the GAM model framework i.e. the model fitted to characterise the child development outcome, pass rate, has to increase with every unit increase in age. A detailed background of the SCAM model can be found in the work of Pya & Wood, (2015) and Pya, (2012). This work extended the GAM model within a binomial response framework using B-spline based functions to include a monotonic increasing constraint.

Formally, the SCAM model for a single covariate is defined as follows:

$$g(\mu_i) = f(x_{ji}), i = 1, \dots, n, \qquad (5.3)$$

where; $\mu_i = E(Y_i)$,

Y_i are independent response variables that follow a Bernoulli distribution,

x_i is a covariate e.g. age,

$f(x_i)$ is a smooth function that satisfies a monotonicity constraint,

$$f(x_k) \geq f(x_j) \; if \; x_k > x_j \qquad (5.4)$$

$g(.)$ is a known smooth monotone link function, e.g. $\log(\frac{x}{1-x})$.

We have seen that the GAM model cannot necessarily always guarantee a monotonically increasing function $g(.)$ across the entire age spectrum of interest. Beyond offering flexibility, the SCAM model has the added advantages of; a) ensuring efficiency provided that the true response function is

monotonic, the SCAM estimate will tend to have a smaller standard error at any given point than the GAM estimate, even though both estimators are consistent, b) ensuring adherence to the very important monotonicity assumption i.e. ability is expected to increase with age.

5.2.1.3. Creating normal reference ranges for each item

Once the preferred model has been fitted to the item binary data, it is used to obtain the age estimates corresponding to the 0.25, 0.50, 0.75 and 0.90 probabilities of success. These are the probabilities of success that we want to obtain the age estimates for but it should be noted that one can compute age estimates for any success probability of interest between 0 and 1. These age estimates make up the normal reference ranges for milestones defined in the MDAT tool. The age estimates at these probabilities characterise the development of the child by giving the typical age at which a child is expected to pass a given item.

GLM-Logistic framework

Under the generalized linear model (logistic) framework creating normal reference ranges for each item involved rewriting the GLM formulae defined above in Section 5.2.1.1 by making the unknown age estimate the subject of the equation 5.1 at the chosen percentile probability of interest and substituting into the equation the α (alpha) and β (beta) coefficients from the fitted model i.e. the intercept and slope values. For example to get the 95th percentile age estimate of the gross motor item 21 'Does the child run well (confidently) stopping and starting without falling?' the following formula was used;

$$x_{qj}, 95^{th}\text{percentile age estimate} = ((log\left(\frac{0.95}{1-0.95}\right)) - \hat{\propto}_{item\ 21})/\hat{\beta}_{item\ 21} \qquad (5.5)$$

where; x_{qj} is the unknown 95th percentile age estimate for the i^{th}child and the j^{th} item,

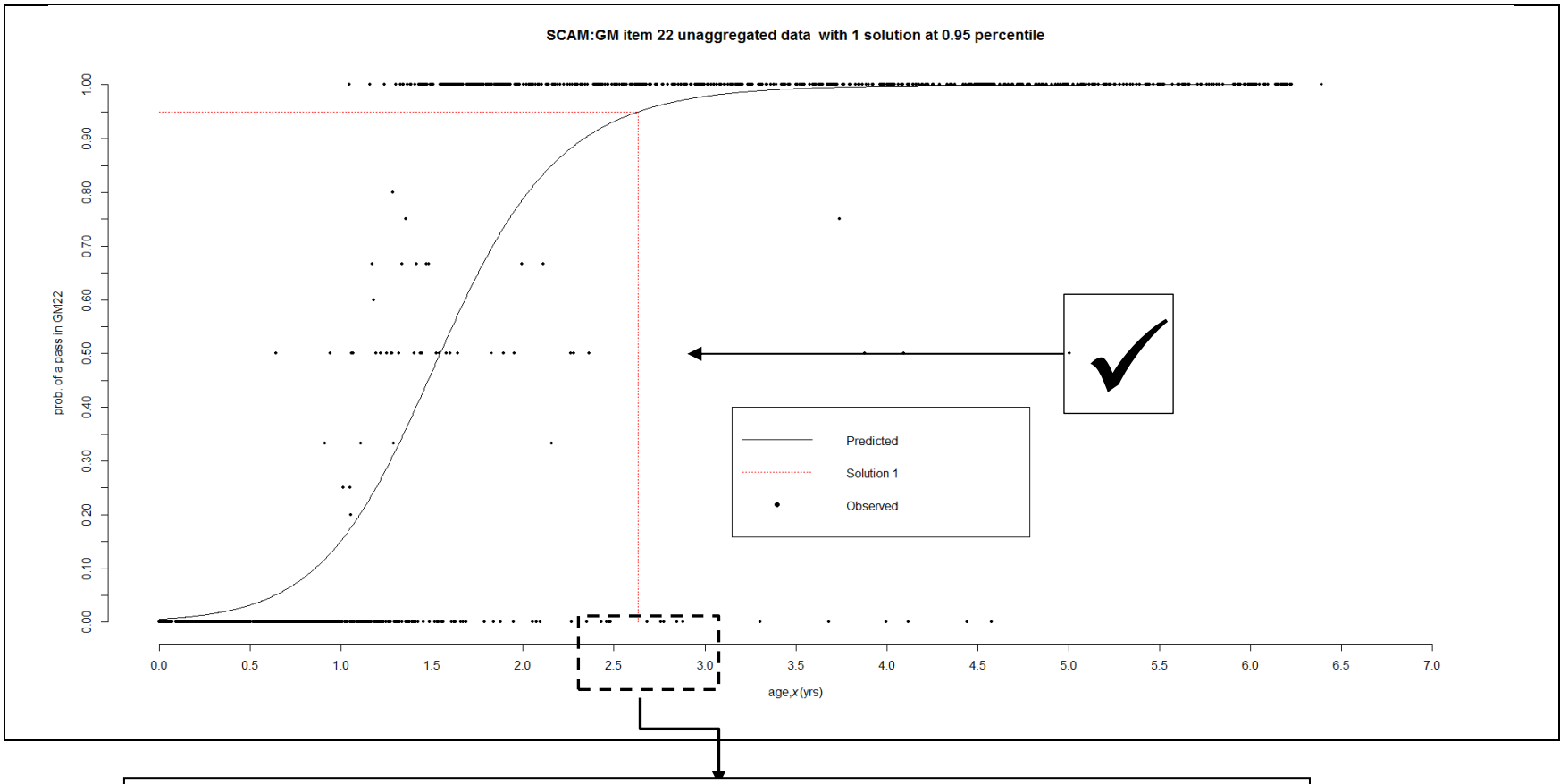
$\hat{\propto}_{item\ 21}$ and $\hat{\beta}_{item\ 21}$ are the fitted intercept and age (slope) coefficients respectively from the GLM model for the gross motor domain item 21.

GAM framework

Within the GAM model framework, because of the model formulations, analytically rewriting the model formulae to make the unknown age estimate the subject at the percentile probability of interest is may not be very straight forward to do within the GAM framework. Instead the age estimate values of interest can be found by numerically solving equations (5.2 and 5.3) with respect to age for a given value of $\pi(X)$, the item pass probability of interest. An approximate method to get the age estimate of interest is to obtain model predictions of the success probability for a fine grid of an age spectrum similar to the one in the standardisation normal sample. Using the fitted SCAM models we obtained age predictions for children ranging from 0 to 6 years of age. Then the age estimate of interest was found by selecting the closest predicted value from the model to the success percentile probability of interest. Therefore, if one wanted the 0.25 percentile probability age estimate for example, this is given by the corresponding age estimate from the model fitted that returns the smallest absolute difference with this probability of success from the fine grid value. Figure 5.2 graphically shows the process of obtaining the required age estimates for item 22 in the GM domain.

We would like to note that due to the flexibility availed within the GAM framework, it is possible to have more than one solution for the age estimates for a given success probability, the confidence intervals around the age estimate or both. This is because the GAM function is not monotonically restricted. As is shown in Figure A.1 in Appendix A, the model fit for a gross motor item 22 has three possible solutions for the 0.95 percentile age estimate. In such a case, one would take a conservative approach and consider taking the lesser of the age estimates (solution 1), as shown in the same figure. However, when the SCAM extension is used, this 'problem' does not arise as the model is constrained to be ever monotonically increasing across the entire age spectrum of interest (see Figure 5.2).

Figure 5.2: Creating normal reference ranges for each item under the GAM framework extension of the SCAM model.



Relevant age estimate(s) at pass percentiles of interest used to draw scoring chart shown in Figure 3.5.

5.2.1.4. Confidence intervals for fitted values

As is often good practice, when reporting estimates, we should also report their precision in the form of a confidence interval. We considered computing confidence intervals for the fitted values at specific pass probabilities of interest. The precision refers to a summary of the variability of the target value of interest that is given by the confidence interval, therefore a variance (or standard deviation) is required. Our confidence intervals therefore consist of a range of age values that act as good estimates of where the unknown true population age estimate of passing an item at the specified probability lies at a certain level of confidence. In this section we will briefly describe the construction of confidence intervals within the GLM and GAM model frameworks.

Asymptotic Confidence intervals for estimated age for GLM-logistic models

As outlined in Neter, et al., (1996) the logistic model response function can be represented as;

$$E\{Y\} = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)} = [1 + \exp(-\beta_0 - \beta_1 X]^{-1} \qquad (5.6)$$

If we let Y_i represent independent Bernoulli random variables corresponding to responses of our sample of children to a given item, then;

$$E\{Y_i\} = \pi_i = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} \qquad (5.7)$$

where as previously seen;

X are observations assumed to be known constants,

$E\{Y_i\}$ is viewed as the expected value or mean.

Often, the estimation of the probability π is for a single or several different sets of values of required predictor variables. In our research the main interest is in the probability of a child of a given age, and possibly also other characteristics like gender or social economic status, to be able to pass an item.

The vector of the levels of the X variables for which π is to be estimated can be estimated by a $p \times 1$ vector X_h. The mean response of interest can be estimated by π_h;

$$\pi_h = [1 + \exp(-\beta' X_h)]^{-1} \qquad (5.8)$$

The point estimator of π_h can be denoted by;

$$\hat{\pi}_h = [1 + \widehat{\exp(-b' X_h)}]^{-1} \qquad (5.9)$$

where; b is a $p \times 1$ vector of the estimated regression coefficients.

The confidence interval for π_h can be obtained in two stages. First calculate the confidence limits for the logit mean response π_h'. Secondly utilise the relation 5.8 to obtain confidence limits for the mean response π_h . Considering that $X = X_h$ we can write;

$$E\{Y_h\} = [1 + \exp(-\beta' X_h)]^{-1} \qquad (5.10)$$

and rewrite the expression 5.10 to convert limits for X_h' into confidence limits for π_h using the fact that $E\{Y_h\} = \pi_h$ and $\beta' X_h = \pi_h'$ as;

$$\pi_h = [1 + \exp(-X_h')]^{-1} \qquad (5.11)$$

Using the fact the point estimator of the logit mean response $X_h' = \beta' X_h$ is $\hat{\pi}_h' = b' X_h$ and also that $b' X_h = X_h'$ because it is scalar, then the estimated approximate variance of $\hat{\pi}_h' = b' X_h = X_h' b$ can be written as;

$$s^2\{\hat{\pi}_h'\} = s^2\{\hat{\pi}_h' b\} = X_h' s^2\,\{\mathrm{b}\} X_h \qquad (5.12)$$

where $s^2\{b\}$ is the estimated approximate variance covariance matrix of the regression coefficients when n, the sample size, is large. The approximate $1 - \alpha$ large sample lower and upper confidence limits for the logit mean response X_h' are then given by;

$$L = \hat{\pi}_h' - Z\left(1 - \frac{\alpha}{2}\right) s^2\{\hat{\pi}_h'\} \qquad (5.13)$$

$$U = \hat{\pi}'_h + Z\left(1 - \frac{\alpha}{2}\right) s^2\{\hat{\pi}'_h\} \qquad (5.14)$$

Now using the monotonic relation between π_h and π'_h shown in 5.13 and 5.14 above we can convert the lower and upper confidence limits for π'_h into approximate $1 - \alpha$ upper or lower confidence limits L^* and U^* for the mean response π_h using;

$$L^* = [1 + \exp(-L)]^{-1} \qquad (5.15)$$

$$U^* = [1 + \exp(-U)]^{-1} \qquad (5.16)$$

Bootstrap Confidence intervals for estimated age estimates for GAM models

The method described in the preceding section is available for evaluating the precision of estimated coefficients, fitted values (estimates) and predictions of new observations for logistic regression models in standard situations. However, in non-standard situations where important modelling assumptions such as non-constancy in error variance are violated, meaning that standard methods for evaluating the precision may not be available or may only be approximately available when the sample size is large, then employing more robust methods to assess precision are warranted.

We explain the bootstrap procedure as defined by both Efron & Tibshirani, (1993) and Efron, (1979) in terms of evaluating the precision of a fitted value for each of the fitted models. The procedure can be directly applied for any other estimate of interest e.g. model coefficients or new observation predictions. Consider fitting a model using an alternative method like SCAM and we obtain the fitted value for each age across the age spectrum of interest; we call this estimate b_1 and we intend to evaluate the precision of this value. As explained by Neter, et al., (1996) the bootstrap procedure calls for the selection from the observed sample data of a random sample of size n with replacement. As the sampling is done with replacement this implies that the bootstrap sample may contain duplicate data from the original sample and may also omit other data in the sample. Now the same model fitting procedure is used to compute the fitted value using the bootstrap sample leading to a new fitted value

b_1^*. This process is repeated iteratively for a large number of times and with each bootstrap random sample the fitted value is calculated. The estimated standard deviation of the bootstrap fitted values b_1^* normally denoted by $s^*\{b_1^*\}$, is an estimate of the sampling distribution of $b_1$and is therefore a measure of its variability or precision.

The confidence interval for b_1 is based on the $(\alpha/2)100$ and $(1-\alpha/2)100$ percentiles of the bootstrap distribution of b_1^* that are denoted by $b_1^*(\alpha/2)$ and $b_1^*(1-\alpha/2)$ respectively. We denote the distances of these percentiles from the target fitted value b_1 from the original sample by d_1 and d_2 and they are computed by;

$$d_1 = b_1 - b_1^*(\alpha/2) \qquad (5.15)$$

$$d_2 = b_1^*(1-\alpha/2) - b_1 \qquad (5.16)$$

Note that there are two distance values (d_1 and d_2) of these percentiles from the target fitted value b_1 because the confidence interval is not symmetric. The review by Forster, et al., (1996), explains why this may happen as odd ratios or the exponentiated regression coefficients are not distributed symmetrically. He also explains alternative exact methods obtaining variance of estimates in logistic models including an enumeration method and the Markov chain Monte Carlo method.

Then the approximate $1-\alpha$ bootstrap confidence interval for b_1 is given by;

$$b_1 - d_2 \leq b_1 \leq b_1 + d_1 \qquad (5.17)$$

Usually, the number of bootstrap samples to take to evaluate precision of the fitted value is dependent on special circumstances of each application. As discussed by Efron & Tibshirani, (1993) since we wanted to estimate the variability around the fitted value, and we had a sample size of 1,446, then up to 1000 bootstrap samples are adequate to ensure that the variability, $\mathrm{s}^*\{\mathrm{b}_1^*\}$, of the fitted value had reasonably stabilized before bootstrap re-sampling iteration process was terminated.

We implemented the bootstrapping procedure explained above by resampling the normal standardisation data 1000 times with replacement but refitting the same primary model for each of the respective item model(s) across the age spectrum at each of the resampling iterations while computing the fitted (predicted) values. In line with the work of Wood, et al., (2016) of controlling the automatic smoothing parameter selection that may be abitrary or change extensively in different iterations and items due to changes in sample data variability, our implementation ensures that exactly the same model smoothing parameters are fixed and used at each of the iterations to avoid the likelihood of too wide confidence intervals. Finally, this was followed by taking the 2.5^{th} and 97.5^{th} percentile values of the results that represent the 95% confidence band around the predicted model fit. We also would like to highlight that the approximate point wise confidence limits generated within the GAM framework are not constrained to be monotonic as the resampling is from non-monotonic fitted value estimates. However, using the bootstrapping process within the SCAM framework we ensured that the confidence limits produced were also monotonically constrained. Figure D.1 in Appendix D shows both the asymptotic and bootstrap 95% confidence intervals under GLM and GAM extension (SCAM) model frameworks respectively for the gross motor item 22.

A comprehensive summary of the different types and forms of bootstrapping procedures, their pros and cons and in what situations they should be used is summarised by Carpenter, et al., (2003) and Carpenter & Bithell, (2000). In this thesis, the bootstrapping method was applied to calculate an approximate confidence band around the predicted model fit under the generalised additive model framework. In chapter 2 we highlighted a lack of reporting of error margins of the normal age estimates at probabilities of interest. Under the GLM framework, creation of a confidence band is quite straight forward with already established methodology that is reliant on asymptotic assumptions and adequate sample sizes. However, when GLM assumptions are violated, as is often the case in the pass rate distributions in an assessment context, the resulting confidence intervals will also have questionable validity.

5.2.1.5. Item by item model checking and diagnostics

In the following subsections a) to e) we describe the model checking methods and strategies that were employed to assess the model fits of the normal sample pass rates (data) of items in the MDAT data under the item by item analysis scenario. The model diagnostic methods included a combination of graphical methods, summary statistics and formal tests that complement each other to assess adherence of residual error distributions to underlying model assumptions**.**

(a) Graphical methods to check model fit and residual error distributions

Scatter plots of the proportions of successes per item given age were plotted for each item and the model fit(s) overlaid and compared. This gave a visual indication of how 'good' or 'bad' the model fit was in comparison to the distribution of the empirical pass rates with respect to age.

As indicated by Olive, (2013), given that age was the only predictor, classical diagnostic plots under a GLM framework can be extended to generalised additive models with a few modifications for use in model fit diagnosis. However, the use of diagnostic methods under the GLM framework still have to be undertaken with caution to check that the various plots of the error terms from the respective models used to assess if the underlying assumptions such as constancy of variance and linearity were adhered to. There are other fundamental assumptions especially monotonicity in our research context that have to be adhered to as well. Scatter plots of the model residuals against age were used to check any indication of systematic patterns that indicated a problem with the fitted model.

(b) Isotonic regression using Pool Adjacent Violators Algorithm (PAVA)

An alternative method to assess the suitability of the model fitted to the item pass rate data is to estimate the effect of age non-parametrically using the Pool Adjacent Violators Algorithm (PAVA). This is a non-parametric monotonic regression method that is appropriate when the response variable (Y_i) is assumed to be increasing or decreasing with respect to one or more explanatory variables $(x_1, \dots, x_p)$. In our case the response variable, the pass rate of an item, is assumed to increase with

age. Several authors including Ghosh, (2007) have suggested extensions when there is more than one explanatory variable and also developed algorithms to facilitate a practical example of the PAVA algorithm. In this project however, we will be using the PAVA algorithm with only the one covariate, age.

Formally, let us consider a set of observations

$$\{(x_i, y_i), i = 1, \dots, n\},$$

The algorithm finds a set of fitted values

$\{z_i, i = 1, \dots, n\}$ such that the sum of squares given by

$$s = \sum_{i=1}^{n} (z_i - y_i)^2 \quad (5.22)$$

is minimised under the constraints induced by the partially ordered data for 2 subjects x_i and x_{i*}

$$z_i \leq z_j, \text{ if } x_i \leq x_j \text{for all i, j.}$$

The isotonic monotonic regression fit is then overlaid over the scatter plot of proportions of pass/success rates per item given age and compared against the respective GLM, GAM or SCAM model fits. Both the isotonic versus the GLM, or GAM or SCAM model fits should be reasonably comparable; indicating that the latter model is a good fit to the pass rate data. It is worth noting that the estimates derived from the above methods cannot be used directly as they may have very high and sporadic variability owing to a non-uniform sample sizes for all ages of children. Subsequently the confidence band around the fitted model would need to be smoothed to provide reliable predictive age estimates.

(c) Aggregated and un-aggregated data

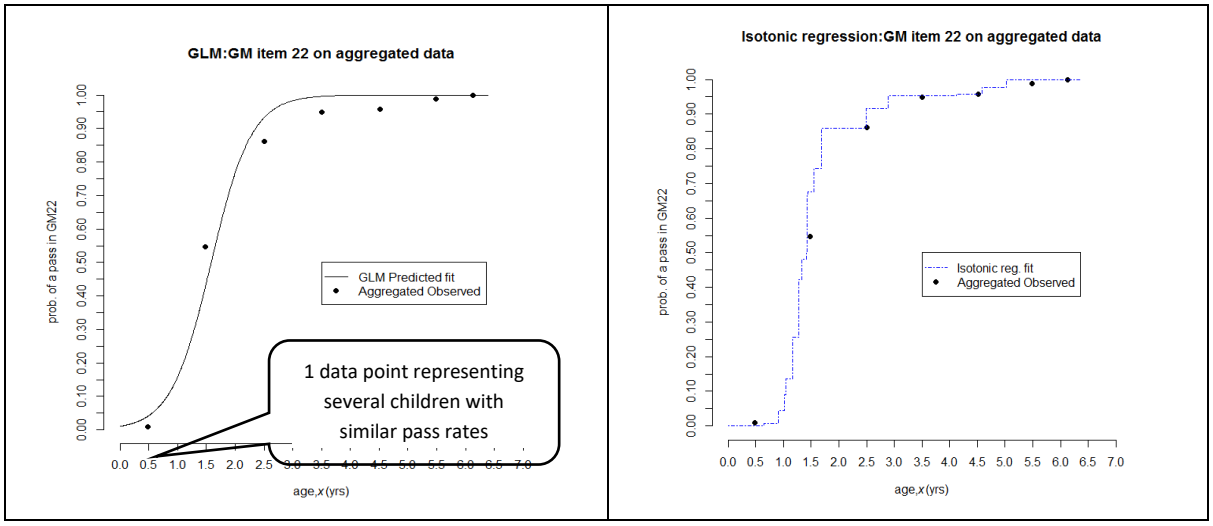
Binary data that is typically presented in the form of a proportion representing the number of children who passed an item divided by total number of children in the sample is referred to as the un-

aggregated data form. If these proportions are grouped for example by a certain age category then their total pass proportion can be presented as one single point. In this latter case, the data is said to be in an aggregated form. Aggregation is mostly done to either increase sample sizes within groups of interest or used to simplify the presented scatter plots that show the distribution of data over a given variable, see Figure 5.4. One can for example pool together all children who are below 6 months and compute the pass rate for a given item for these children. Thus, instead of presenting several data points for each age on the scatter plot of pass rates and age, there will be just a single point representing these children who are below 6 months as in shown in Figure 5.4 below. While this makes a plot simpler and less cluttered it has the disadvantage of masking some important distribution information.

Further, one can fit a model either using un-aggregated or aggregated data. Within the GLM framework that was explained in Section 5.2.1.1, similar results of model estimates are achieved by attaching frequency weights to the aggregated data. However, in the case of GAM models that are explained in Section 5.2.1.2, there is no obvious method to attach weights and therefore estimates of aggregated versus un-aggregated do not give exactly the same estimates. Another reason for using aggregated data is to get goodness of fit measures through residual deviance tests. However, our primary reason for using aggregated data is to be able to fit the isotonic regression described in section 5.2.1.5 part d) that is only possible using aggregated data.

In this analysis, all item by item scoring models were fitted using un-aggregated data as is shown in the various figures of model fits above as well as in the results chapter. The first panel of the Figure 5.4 below shows a GLM fitted on aggregated data while the second panel shows an example of the isotonic regression fit.

Figure 5.3: GLM and isotonic regression fits for GM item 22 using aggregated data



(d) Monotonicity in item by item analysis

The modelling of child development ability presents a situation where the probability of success of an item is expected to increase with age. Intuitively, this is an expected phenomenon in that as a child grows or develops; their ability is expected to advance or increase. Of course as we explained in chapter 1 there are many factors driving and affecting this process, and they differ from subject to subject, but the main fact is that if the child population is normal, their ability should always increase with age however minimal the increment rate i.e. the relationship between the probability of passing an item and the child's age has to have a monotonically increasing shape. Thus it is important that the chosen model adheres to this assumption so as to ensure the normal development scores produced reflect this aspect. This property has to be adhered to in all forms of scoring child development and also validated once scores are computed.

Under this scenario of item by item analysis, monotonicity was checked by ensuring that the increment of the probability of passing an item was always equal to or greater than 0 throughout the covariate age spectrum in accordance with the constraint defined in formula 5.22 above. As we will see in the results section as much as the classical GAM model framework often gave improved model fits, it did not always guarantee monotonicity which is a fundamental aspect to be adhered to in child

development scoring. Hence in the instance that the GAM model is not monotonically increasing the SCAM model that is an extension of the GAM framework offers a solution. This issue will be also be revisited again under the overall score creation scenario where a smoothing approaches to ensure the monotonicity of scores in this context will be suggested and implemented.

(e) Model comparison using Akaike Information Criterion (AIC)

As detailed in Agresti, (2002), the AIC is a good formal criterion to select the most parsimonious and 'best' model from a set of potential models. Aside from its simplicity, this model comparison method was mainly used because it can compare unnested and models from different families. We acknowledge that other potentially superior selection criteria including model fit significance tests and the Bayesian Information Criterion, (BIC; Schwarz, 1978) that is also likelihood-based and could have been used to guide the assessment of model fit. However, in our case our choice of model strategy had to first make a choice of the best fitting model within a given family by considering the type and number of knots, or the use of different degrees of freedome, or other link functions or other binary response transformations because the sample and number of covariates for each candidate model was fixed. Further, we observed that although in all instances only the covariate age was included in all versions of each family framework models, the other main challenge of getting an improved fit for items with either very low or high pass rates was due to identifiability issues that almost always frustrated any alternative model fit improvement strategy. The best model from the GLM given family was then compared with the best model from the competing GAM family bearing in mind the primary objective of computing more accurate and comparable age estimates.

Formally, the criterion selects the model that minimises the following function;

$$AIC = -2(maximized \log likelihood - number\ of\ parameters\ in\ model) \qquad (5.23)$$

Hence the models are assessed in terms of their deviance, but a model will be penalized for having too many parameters without any considerable improvement in the log likelihood value. Figure D.1 in

Appendix D shows the scatter plot of pass rates (black points) of gross motor item 22 with the GLM and SCAM model fits (black continuous line), respective confidence bands (red dotted line), the isotonic regression fit overlaid (blue dotted line) with the respective AIC values as an example.

5.2.2. Creating an overall (total) score for a child using the entire (all) domain of items

In the total score scenario, all items within a given domain for a given assessment tool are considered simultaneously. This second approach instead gives a single score to characterise a child's ability given age within a given domain. The benefit of combining all item responses within each domain to give one single score allows one to get a holistic 'picture' of a child's ability status. We will describe three main methods that include; a) a model based scoring method using Weighted Simple Counts, b) Z-score method and c) Item Response Theory (IRT) models discussed by Jacobusse, et al., (2006; 2007).

Most developmental tools' items are designed in the form of a series of tasks with increasing difficulty and are administered in sequence. However, at the end of the entire testing process, one is usually interested in summarising the ability level or developmental status of a given child. Therefore, instead of summarising the ability of a child with respect to individual tasks, a parent or assessor is likely to be more interested in knowing if the child is developing normally or whether they are delayed, and to what extent they are delayed compared to other healthy normally developing children. To be able to give a child's developmental status, being able to combine all scores within a given domain is important especially in this latent construct context. As noted by Cheung, et al., (2008), the use of all items or the multitude of binary data to score the developmental status of a child remains debatable. This is because there is yet to be a statistical framework that can address the fact that items in a child development context differ in difficulty, the item pass probability is dependent on age, different children will respond to a different number of items and that there is an underlying correlation or association between and within item responses simultaneously. Herein lays the motivation of this section to devise a suitable statistical framework that can address these 4 issues while computing an

overall score. We will first describe the GAMLSS model that will be used to extend some of the classical overall scoring methods.

The GAMLSS model

As described in Rigby, et al., (2005), Generalized Additive Models for Location, Scale and Shape (GAMLSS) are semi-parametric regression type models. They are parametric, in that they require a parametric distribution assumption for the response variable, and 'semi' in the sense that the modelling of the parameters of the distribution, are done as functions of explanatory variables which may involve using non-parametric smoothing functions. GAMLSS models were introduced by Rigby & Stasinopoulos, (2005) and Akantziliotou, et al., (2002) as a way of overcoming some of the limitations associated with the popular generalized linear models, GLM, and generalized additive models, GAM (see Hastie & Tibshirani 1990).

In the GAMLSS model, the exponential family distribution assumption for the response variable (y) is relaxed and replaced by a general distribution family, including highly skewed and/or kurtotic continuous and discrete distributions. The systematic part of the model is expanded to allow modelling not only of the mean (or location) but other parameters of the distribution of the response for the i^{th} subject, y_i, as linear and/or non-linear, parametric and/or additive non-parametric functions of explanatory variables and/or random effects. Hence the GAMLSS model is especially suited to modelling a response variable which does not follow an exponential family distribution, (e.g., leptokurtic or platykurtic and/or positive or negative skewed response data, or over dispersed counts) or which exhibit heterogeneity, (e.g., where the scale or shape of the distribution of the response variable changes with explanatory variables(s)).

The model assumes independent observations y_i for $i = 1,2,\dots,n$ with probability density function $f(y_i|\theta^i)$ conditional on $\theta^i = (\theta_{1i}, \theta_{2i}, \theta_{3i}, \theta_{4i}) = (\mu_i, \sigma_i, v_i, \tau_i)$ a vector of four distribution parameters, each of which can be a function applied to the explanatory variables. The first two

population distribution parameters are usually characterised as location and scale parameters, while the remaining two parameters are characterised as shape parameters, e.g., skewness and kurtosis parameters. In this thesis, the excess skewness and kurtosis was taken to be zero therefore only the location and scale parameters were used as the pass probability or simple sum score was seen to vary non-linearly with age and also increase in variability as age increased.

Formally, according to Rigby & Stasinopoulos, (2005) let $y^T = (y_1, y_2, \dots, y_n)$ be the n length vector of the response variable. Also for $k = 1,2$ let $g_k(.)$be known monotonic link functions relating the two distributional (location and scale) parameters to explanatory variables by the following functions;

$$g_1(\theta_1) = g_1(\mu) = \eta_1 = X_1\beta_1 + \sum_{j=1}^{J_1} h_{j1}(x_{j1}) \quad (5.24a)$$

$$g_2(\theta_2) = g_2(\sigma) = \eta_2 = X_2\beta_{k2} + \sum_{j=2}^{J_2} h_{j2}(x_{j2}) \quad (5.24b)$$

$\eta_k$is a vector of 'predictors' of length n, $\beta_k^T = (\beta_{1k}, \beta_{2k}, \dots, \beta_{J_k'k})$ is a parameter vector of length J_k' , X_k is a fixed known design matrix of order $n \times J_k'$, and h_{jk} is a smooth non-parametric function of explanatory variable x_{jk} with $j = 1,2, \dots J_k$. Therefore, we see that the GAMLSS framework models 5.24 allows one to also flexibly model parameters as linear functions of explanatory variables.

Smoothing scores using Generalized Additive Models for Location Scale and Shape (GAMLSS)

Once the scores to characterise child development have been created, the next step is to determine thresholds of the scores to classify development status. Classical approaches currently in use to create the classification thresholds or cut-offs used to decide whether a child is delayed or normal are reliant on whether the scores are normally distributed. As was discussed in Section 2.3.2.2 b) if the scores are normally distributed then various distances of deviation from the mean score of a given age category is considered as a viable (highly sensitive) thresholds.If the scores are skewed and not normally distributed, then the use of other measures of spread such as percentiles or percent of median can be considered. In line with this normality property is the fact that the scores should to be monotonically increasing with age. Section 5.3 describing the comparative methods will reveal in greater detail the

importance of score distributional properties of the different methods and their consequences on performance at different classification score cut-off thresholds in detecting disability or delay.

Typically, to summarise the computed scores, the means, standard deviations and percentiles of respective age categories of interest are reported. Owing to the possible underlying effects of sample sizes and/or study designs, some age categories may have extremely high or low variability which can result in non-monotonic mean summaries for age. For example, the presence of a borderline performing child in a given age group with a small sample size may lower the mean summary score to fall below that of the preceding age category. This will mean that there will be a lack of monotonicity of mean summary scores. Lack of monotonicity could also be as a result of the chosen age category cut-offs that in turn determine the number of children in a given age category. Further, even if monotonicity of these summary measures is achieved, the variability around them is likely not to be smooth or reasonably even across the entire age spectrum. This is an especially important feature that the summary of these scores should possess. This property will enable scores to be made more generalizable to allow their use to classify a different cohort of children for external validation purposes. This will in turn avoid high misclassification rates of the children's ability or development status in the new setting.

Flegal, (2013) argues that because of statistical variation in the reference sample, empirical percentile curves are generally irregular, and some type of smoothing over age should be applied. Further, as noted by Cole & Green (1992), the smoothing is partly for cosmetic reasons but more importantly because changes of the variable measured would be expected to be continuous. In line with these two sentiments we experienced the same issues in our research context and agree that too much and inconsistent variability may render the empirical summary measures of scores to be unreliable leading to wrong development status classification conclusions. Therefore, the use of GAMLSS models as described by Rigby, et al., (2006) was explored to ensure that; i) the smooth monotonicity of total scores or mean scores for age categories was maintained and ii) the variability of overall score

summaries across the entire age spectrum was not sporadic. The model based smoothing framework also allows for the creation of confidence intervals around scores and thus defines cut-off thresholds for scores to be used in development status classification.

5.2.2.1. Model based total scores using Simple Counts

The classical simple count (SC) method that combines all the binary responses for a given child by simply summing all the correct responses is used as the response to the model based scoring approach. While reflecting on the simple score's weaknesses we consider alternatives of enhancing this naïve approach by weighting the simple score by the number of administered items and then suggest a model based approach on the weighted total raw score using a GAMLSS model in to correct for age.

Simple Count and Weighted Simple Count methods

The simple count method can give misleading results for children with increasing age as the older a child is, the more items the child will be able to pass if they are developing normally. An additional problem occurs when not all items have been administered and there are missing data. We therefore suggest a weighted simple count method that weights the simple counts, i.e. the number of correct responses, by dividing this total by the number of questions answered to reflect the number of questions actually administered to the child. This simple (unweighted) count method has previously been used by Cheung, et al., (2008) but because they omitted to adjust their scores for children who were of similar age but answered different numbers of items or weight (with number of items administered to the child) then their scores did not correctly reflect the differing abilities of children of varying ages. The aim is to be able to differentiate children who have a similar simple score but were exposed to a different number of items for various reasons.

In general, building on the missing data terminology used by Rubin, (1987:1988) and Little & Rubin, (1987) for the i^{th} child in the study, we let Y_{ij} be the response for administered items $j = 1, \dots, n_i$.

Further we define an indicator R_{ij} for whether the j^{th} item was administered for the i^{th} child that is given by;

$$R_{ij} = \begin{cases} 1 \; if \; Y_{ij} \; is \; observed \; or \; administered \\ 0 \qquad\qquad\qquad\qquad otherwise \end{cases} \qquad (5.25)$$

Formally this weighted score can be written as;

$$Y_i^* = \frac{\sum R_{ij} * Y_{ij}}{\sum R_{ij}} \times 34 \qquad (5.26)$$

where; Y_i^* is the weighted simple count for the i^{th} child.

Recall that the MDAT tool has 34 items in each of its 4 domains; therefore, this total is used to weight the administered items. This method is able to differentiate children who were administered many more or less items for various practical reasons, such as differing levels of cooperativeness during assessment sessions. However, there is an implicit assumption that the missing items are of 'average' or 'moderate' difficulty. The total can be converted to a percentage of the total number of items in the testing tool or other summary measures for example means and standard deviations per age category.

For example if a child in the MDAT study passed the 1st, 3rd ,5th ,6th and 7th items, and yet only 8 out of 34 items in the gross motor domain were administered; then their simple score would be the sum of 5 passes while the weighted simple count score would be $(5/8) * 34 = 21.25$ using formulae 5.26. In this example we see that because only 8 out of 34 items were administered, the simple score of 5 was weighted up as few items were administered.

GAMLSS regression on Total passes and fails using Beta Binomial distribution

The testing process of the MDAT tool presents a typical example where there aren't a finite number of integers arising when the probability of success in each of a fixed or known number of Bernoulli trials is either unknown or random. As we saw in the previous item by item analysis, the probabilities of passing particular items are not the same. The beta-binomial distribution is the binomial distribution in which the probability of success at each trial is not fixed but random and follows a beta distribution. In this case the fixed number of Bernoulli trials refers to the 34 items that return a pass or fail with a certain probability given age. The beta distribution as described by Ross, (1998) can then be used to model a random phenomenon like ability whose set of possible values is within some finite interval $[a, b]$. By letting a denote the origin and taking $(\mathrm{b} - \mathrm{a})$ as a unit measurement, the finite interval can be transformed into the interval [0, 1]. In our case the finite interval or possible number of items that a child can pass given their age can range from 0 to 34. The number of items passed can be represented as a rate that can only take values between 0 and 1. In doing this, a Beta Binomial distribution allows there to be a more general mean-variance relationship than is the case for the binomial distribution.

Therefore, the Generalized Additive Models for Location Scale and Shape (GAMLSS) framework as described by Rigby, et al., (2005) was explored to provide a smooth functional estimate of the expected total score given age using the total passed and total failed for each child assuming a beta binomial distribution. Within this framework, confidence bounds can be computed. Total scores are obtained from a Beta Binomial (BB) GAMLSS model fitted to the number of passes and fails of each child and was defined as;

$$Y \sim Beta\ Binomial\ (n_i, p, \sigma) \qquad (5.32)$$

$$log\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 . age \qquad (5.33)$$

$$log[\sigma] = \eta_0 + \eta_1 . age \qquad (5.34)$$

where; Y_i, is a vector of sums of the binary random successes of j items administered to the i^{th} child,

n_i, is the number of items administered to the i^{th} child,

p, is the probability of passing an item and lies in the interval $0 < p < 1$,

σ, is the variance that is dependent on age,

$\beta_1(age)$ and $\eta_1(age)$ are monotonic spline functions of age.

Utilizing the GAMLSS flexibility of modelling multiple parameters of a distribution function, the GAMLSS model can be used to provide a smooth functional estimate of an overall outcome score; an expected mean or variance given any continuous quantity such as the simple sum total score or even a Z-score but assuming a normal distribution. Again within this framework the confidence bounds can be computed in a straight forward way. The smoothed mean and variance estimates obtained from the GAMLSS model fitted assuming a normal distribution are likely to be less variable and also better suited for developmental status classification as is explained in Section 5.2.2.2 b) below.

5.2.2.2. Z-Score methods

Perhaps the most widely used scoring approach is the Z–score method. It standardises the Simple Count Score (SC) described in Section 5.2.2.1 of each child using the mean and standard deviation corresponding to the age category of the child from a standard population of normal healthy children. If the reference population is the same population of children for which the z-scores are being calculated, then this is referred to as internal standardisation. One can also standardise the SC scores if separate means and standard deviations for age categories of interest are available from an external reference or standard population. This latter approach is referred to as external standardisation. The standardisation or normalization is designed to remove the effect of age on the scores produced. This section will consider the pros and cons of the classical Z-score approach and suggest one method to alleviate its disadvantages.

a) Empirical Z-Score method

Owing to the limitation of the simple count method highlighted in Section 5.2.2.1 we will instead standardise the Weighted Simple Count (Y_i^*) Score. Formally, the internal standardisation of the weighted simple score can be defined as;

$$Z\ score\ for\ i^{th} child\ = \frac{Y_i^* - \mu_i}{\sigma_i} \qquad (5.35)$$

where; Y_i^* is the weighted simple count of correct responses for items administered to the i^{th} child as defined in Section 5.2.2.1,

$\mu_{\ i}$ is the mean of the weighted simple counts of children in the age group this i^{th} child belongs to,

σ_i is the standard deviation of the weighted simple counts of children in the age group this i^{th} child belongs to.

Further, unless a very 'large' data set is available there will be likely too much variability between adjacent age categories because of differences in the number of children recruited per age. Therefore, using the empirical mean and standard deviation summary measures may cause the Z-scores computed not to be reliable for use to classify the developmental status of children. This is because the high variability of scores in some of the age categories also makes the means for each age category used to compute the Z-scores not to adhere to the monotonicity property discussed in Section 5.2.2. Thus instead of using the empirical mean and standard deviation of each age category to standardise weighted simple scores we suggest use of smoothed versions of these summary values to compute the Z-scores. The smoothed versions of the summary measures will have non-extreme variability and the property of monotonicity can be adhered to. The smoothing process is described in the following sub section.

b) Model based (GAMLSS) smoothed Z – Score

Several remedial measures have been suggested e.g. the Modified Z-Scores that are based on Median and Median Absolute Deviation and applied in various research contexts in psychology as described for example by Leys, et al., (2013).

In our case, recall from the EDA of pass rates discussed in Section 4.3.4 that the skewness in pass rates is likely to be a result of the recruitment process and therefore these remedial measures may still not remedy stability of the mean and variance. A consequence of this is that the mean and variance for certain age groups that are used to compute Z-scores may be sporadic. Further, the mean may not be monotonic with respect to age. This later issue has the potential to lead to misleading Z-score computations and threaten classification accuracy because certain older age groups' overall score means will have lower means than younger age groups. Therefore, by utilizing the GAMLSS flexibility of modelling multiple parameters of a distribution function, the model can be used to provide a smooth functional estimate of an expected mean and variance of any continuous quantity such as a total score given age but assuming a normal distribution. Again within this framework the confidence bounds can be computed.

Let $\mathrm{Y_i}$ represent the total number of passed items for i^{th} child. Then smoothed mean and variance estimates can be obtained from a Normal GAMLSS model fitted to a continuous score variable, the weighted simple count score, explained in Section 5.2.2.1 above can be fitted as;

$$\mathrm{Y_i^*} \sim \mathrm{Normal}\ (\mu(\mathrm{age_i}), \sigma^2(\mathrm{age_i})) \qquad (5.36)$$

where; Y_i^* is the weighted simple count score of the i^{th} child computed using the formulae 5.26,

$\mu(age_i)$ and $\sigma^2(age_i)$ denote the assumed mean and variance respectively that depend on the important covariate, age, using a GAMLSS model.

Given the non-linear relationship described in Section 4.4.3 between the overall total score and age, a B-spline function was used to capture this non-linear relationship.

The model based Z– score method standardises the weighted SC score defined in Section 5.2.2.1 of each child using a smoothed mean and standard deviation from a GAMLSS model of the age category that case belongs to. This involves fitting a GAMLSS model to the weighted simple count and using the fitted model to predict both the mean and variance values across the spectrum of ages of interest. The GAMLSS model estimate of mean ($\hat{\mu}(age_i)$) and standard deviation ($\hat{\sigma}(age_i)$) for a specific age group the i^{th} child belongs to are used to compute a smoothed Z-score. In other words, the model based predicted mean and variance are now used to standardise the weighted simple counts for each child. This has the benefit of having a less 'erratic' overall score variance, plus monotonicity of the mean score can now be adhered to by taking advantage of the flexibility afforded within the GAMLSS framework. We assumed a normal distribution, as this was the distribution the classical Z scores were are expected to adhere to for purposes of ability development classification explained in Section 1.1.2. We will refer to this method as the smoothed Z-score method.

Formally the model based internal standardisation to obtain the smoothed Z-scores can be presented as;

$$Z_i \text{ score for } i^{th}\text{child} = \frac{Y_i^* - \hat{\mu}(age_i)}{\hat{\sigma}(age_i)} \qquad (5.37)$$

where; Y_i^* is the weighted simple count of correct responses of items administered to the i^{th} child.

5.2.2.3. Item Response Theory (IRT) scoring methods

Several authors, for example Drachler, (2007), have applied IRT models to measure very specific child development outcomes in varied contexts. We have so far considered several total scoring methods and in the light of their disadvantages suggested and developed several extensions offering a framework for creating an overall score using all items for each child with the benefit of getting a holistic 'picture' of a child's developmental ability status. Additionally, our research pursues a framework that can simultaneously address and accommodate the often unaddressed issues stemming from the study design's multilevel structure, and type of outcome, as well as adjusting for

important covariates (e.g. age) as these underpin the quality and accuracy of the scores. However, even with the extensions suggested and developed so far, these methods still suffer a few limitations that are not fully or efficiently addressed; be they fundamental model assumption violations given our child development context, the fact that the extensions are still based on an inferior overall scoring approach (e.g. the developed model based scoring extensions discussed in Sections 5.2.2.1 and 5.2.2.1 were still based on the naïve simple sum scoring method), or issues arising from data structure or computation.

The fourth approach considered to create an overall score was the Item Response Theory (IRT) framework. The data collection tool, MDAT, and the scoring process used to collect our data were described in chapter 3. The data to be used to exemplify the superiority of our IRT extensions against the use of classical IRT models to create child development scores was described in chapter 4 of this thesis. The data qualities that are pertinent to modelling choices and the assumptions made in the form of an elaborate exploratory data analysis was also provided in chapter 4. The item by item analysis in Section 5.2.1 also plays a key role is advising extensions developed in this IRT framework.

Certain specifications of IRT models distinguish themselves from the classical alternatives seen in the previous sections for creating overall scores by not assuming that all items in the tool have equal difficulty. As will be explained in the subsequent sections, besides being able to combine all item responses, IRT showcases a variety of alternative generalisations to overcome the observed limitations of previously explored methods, for example while simultaneously allowing items to differ in difficulty, we may also allow ability to depend on the age of the child. It was especially because of this later generalisation that the next sections explore the use of IRT models to create overall scores for child development taking age into account in the modelling process.

IRT is based on establishing a model that specifies the probability of observing each response option to an item as a function of the target trait being measured by the assessment, which in our case is developmental ability. In testing situations where items are scored as correct or incorrect, IRT specifies

the probability of a correct response to an item as a function of ability. These models falling under the wider realm of latent trait models have now recently gained considerable momentum in their use especially in educational and developmental measurement. Basically, in the past decade, any application that involves combining response items of any form mostly advocates the use of IRT methodology to create overall scores to assess development. As was seen in our literature review, the subject of child developmental assessment tools has a long history, as does IRT. IRT has also evolved in its methodology and more so in its taxonomy. IRT or latent trait models provide a statistically-rich class of models for analysis of educational test and psychological scale data. In their simplest form the data are comprised of a sample of subjects responding to a dichotomous set of test or scale items. The items are a set of questions, tasks or observations that indicate presence or absence of a defined ability given a child's age. The main interest is in estimation of characteristics of the items and subjects.

An IRT model can be viewed as a mixed-effects regression model (Rijmen, et al., 2003) which enables its application to much more varied contexts including longitudinal studies especially in educational testing and psychological measurement. In this chapter, we describe IRT models with respective generalisations applicable to our research trajectory, and illustrate their application using cross sectional data. In particular, we will relate the IRT model to a mixed model framework and indicate how software can be used to estimate the IRT model parameters. Again, using data collected using the MDAT described in Chapter three, we describe how IRT models can be used to address key scoring questions in child development research.

Typically, in a sample with n subjects where each child responds to j item(s) at one occasion, though the amount of actual time that one occasion represents can vary for various reasons especially in a child development context. In this section of the methods chapter the basic IRT model for dichotomous items will be described and subsequent extensions developed to address weaknesses in the current methods will be suggested. Our example will be somewhat a traditional IRT illustration, in the sense that we will not examine responses to tests or validate questionnaire items, per se.

However, model fit performance is an indicator of quality of assessment tool items. Instead, as described more fully previously, our main interest is to produce and assess the quality of the score produced under the IRT framework against other methodologies and especially to determine if many of the ignored or unsatisfactorily addressed issues in the previous discussed methods can simultaneously be better addressed within the IRT framework.

Following the above introduction we hope to prove the prominence of the IRT framework in scoring child development in subsequent subsections. We will first describe 2 IRT models that have been applied in a similar child development context; a 1 PL model and its generalisation called the 2 PL model. This will be followed by a description of IRT models within a mixed model framework. The reasons for this will become much clearer as we expound on our developed extensions. While highlighting the limitations of these two classical IRT approaches and describing the formulation of our IRT extension, we fit our suggested IRT model extension using the 'back-bone' of the generalized mixed model framework (GLMM) implemented using the lme4 package in R.

The specification of IRT models

We will set notation under the IRT framework to be consistent with previous methods. Therefore, we will let Y_{ij} refer to the dichotomous response made by the i^{th} child ($i = 1,2, \dots, n$ children) to the j^{th} item ($j = 1,2, \dots, n_i$ items). The i^{th} child was measured or assessed on n_i items. In our illustration we have a total of 34 items per domain of which a child can answer all or some of them, therefore n depends on the number of items a child responded to i.e. we do not necessarily assume that all children are measured, assessed or responded to all tool items. In the following IRT model formulations, a correct or positive response to an item will be, in accordance with previous notation be denoted by $Y_{ij} = 1$ and an incorrect or negative response as $Y_{ij} = 0$.

An exception to this is in computerized adaptive testing where the items that a subject receives are selectively chosen from a large pool of potential items based on their sequential item responses. In

our illustration we will simply assume that there is a set of $n = 34$ items in total, but that not all our subjects ($N = 1446$) necessarily responded to all of these items for various reasons described in Section 2.3.3.2.

a) One Parameter Logistic model (1 PL)

The simplest and most popular IRT model is the one-parameter logistic (PL) model. The works of Jacobusse, et al., (2006; 2007) and Cheung et al., (2008) have used the one parameter model to score child development. Formally, the conditional probability that a child with a particular latent trait will correctly be able to answer an item with a specific difficulty can be presented as;

$$P(Y_{ij} = 1|\theta_i, \beta_j) = \frac{\exp\{\alpha(\theta_i - \beta_j)\}}{1 + \exp\{\alpha(\theta_i - \beta_j)\}} \qquad 5.43$$

where; Y_{ij} denotes the response made by the i^{th}child to the j^{th} item,

θ_i refers to the trait level of the i^{th} child where higher values reflect a higher level on the trait being measured by the items. The trait values are usually assumed to be normally distributed in the population of normal children with mean zero and variance one; $(N(0,1))$,

α refers to the difficulty of the j^{th} item. It determines the position of the logistic curve along the ability scale. The further the curve is to the right, the more difficult the item is and thus needs a higher level of ability,

β_j is the slope or the discriminating parameter which represents the degree to which the item response varies with ability θ_i , which is assumed to be 1 in this model.

In this 1 PL model, the discrimination parameter does not vary across items. This means all items are assumed to have the same slope and thus the parameter α does not carry the *j* subscript in the model formulation. In some representations of the 1 PL model the common slope parameter α does not explicitly appear in the model formulation i.e. rather than assuming that the trait levels θ_i are from a

standard normal distribution, $N(0,1)$, we assume a general normal distribution, $N(0,\sigma^2)$, you get an equivalent model for $\alpha = 1/\sigma$. Further from equation 5.43 above, we see that if a child's trait level $\theta_i$exceeds the difficulty of the item β_j then the probability of a correct response is greater than 0.5. Conversely, if $\theta_i < \beta_j$ the probability of a correct response is less than 0.5.

b) Two Parameter Logistic model (2 PL)

It has previously been shown that item difficulty per domain is not uniform across all items and increases with age in this context of child development. We saw this aspect in our exploratory data analysis where certain items demand a higher level or degree of ability that is assumed to be driven by age; thus older children are expected to have a higher ability and therefore will be expected to have a higher item pass rate. Therefore, the 1 PL model assumption that all items have equal discrimination may be inappropriate given the influence of age on a child's ability. Given that the MDAT tool items are designed to increase in difficulty, we argue that assuming equal discrimination of all items may be inappropriate as older children are likely to pass more items and therefore are likely to be poorly discriminated.

To relax this 1 PL assumption or incorporate this feature of changing item discrimination into the 1PL model, we allow the discrimination parameter α to vary across items. This generalisation of the 1 PL model results in the two parameter logistic (2PL) model such that the discrimination parameter α_j now carries the j subscript in the model formulation.Several authors, including Bock & Aitkin (1981), have described this generalisation. Now the conditional probability that a child with a particular latent trait level (assumed to be from a standard normal distribution, $N(0,1)$, will correctly answer an item with a specific difficulty can be presented as;

$$P(Y_{ij} = 1 | \theta_i, \beta_j) = \frac{\exp\{\alpha_j(\theta_i - \beta_j)\}}{1 + \exp\{\alpha_j(\theta_i - \beta_j)\}} \qquad 5.44$$

where; β_j refers to the discrimination parameter or slope of the j^{th} item,

Y_{ij} ,θ_i remain as previously defined in the 1 PL model above while α_j now carries the j subscript in the 2 PL model formulation.

IRT models have a long history and have been applied in various fields. There are numerous representations of model formulations each tailored to the specific application or researcher convenience. Despite the lack of a unifying model formulation, the underlying theory is exactly the same across all model formulations. At this point we would like to highlight a note by Bock & Aitkin (1981) of an alternative way to represent the 2 PL model;

$$P\left(Y_{ij}=1\middle|\theta_i\right)=\frac{1}{1+\exp\{-\left(c_j+\alpha_j\theta_j\right)\}} \qquad 5.45$$

where $c_j=-\alpha_j b_j$ is the item intercept parameter.

In a similar fashion the 1 PL 5.43 model can be written as;

$$P\left(Y_{ij}=1\middle|\theta_i\right)=\frac{1}{1+\exp\{-\left(c_j+\alpha\theta_j\right)\}} \qquad 5.46$$

with $c_j=-\alpha b_j$.

Under this formulation it is clearer to see these models as variants of the mixed effects logistic regression (Goldstein & Lewis, 1996) model that are briefly explained in the next section. Therefore, the 1 PL model 5.43 can be written as follows in terms of logit or log odds;

$$log\left[\frac{P\left(Y_{ij}=1\middle|\theta_i\right)}{1-P\left(Y_{ij}=1\middle|\theta_i\right)}\right]=c_j+\alpha\theta_j \qquad 5.47$$

As stated by Titman, et al., (2013) a limitation of these models is that by choosing a large or very small θ_i trait level parameter it is possible to find $P\left(Y_{ij}=1\middle|\theta_i\right)$ arbitrarily close to 0. Where there is a multiple choice option to select the correct response typically in aptitude tests, there is a possibility that a child could guess the correct answer. The 2 PL model can further be generalised to include a guessing parameter which can be represented as;

$$\left(Y_{ij}=1\middle|\theta_i,\beta_j\right)=\Upsilon_j+\left(1-\Upsilon_j\right)\frac{\exp\{\alpha_j(\theta_i-\beta_j)\}}{1+\exp\{\alpha_j(\theta_i-\beta_j)\}} \qquad 5.48$$

where; Υ_j represents a minimum success probability between 0 and 1 of a child giving a correct response to the j^{th} item i.e. guessing parameter,

Y_{ij} ,θ_i, and β_j remain as previously defined.

The model 5.48 is commonly referred to as the Three Parameter Item Response Theory Model (3 PL IRT). However, in this context of child development assessment, there is no possibility of a child guessing a task. This is because the assessment is carried out by an experienced practitioner whose response is based solely on what they observe.

c) Item Response Theory (IRT) Model Extensions

The rationale of considering the IRT model using GLMM framework stems from the fact that it is clear that IRT models have connections with the popular general class of models known as generalized linear mixed models (GLMMs; see McCulloch & Searle, 2008). Numerous authors including Rijmen, et al., (2003) present an informative overview of the bridge between IRT models, multilevel models, mixed models, and GLMMs. A more condensed review is given by Hedeker, (2006). GLMMs extend generalised linear models (GLMs) by inclusion of random effects, and are commonly used for analysis of correlated non-normal data. In an assessment context, every child responds to several items at one time or longitudinally over time, thus it is likely that their responses will be correlated. It is important to take into account this underlying correlation structure.

Given the limitation of the 1 PL model that constrains the discrimination parameter not to vary across items, that the 2 PL IRT model relaxes, we suggested and developed the following extensions that attempted to offer a complementary and more realistic modelling strategy under the IRT framework approach of measuring a child's ability. We will see that the mixed or multilevel model formulations easily allow incorporation of our extension features to correct for age.

One Parameter Item Response Theory (1 PL IRT) Monotonic Spline Model and its implementation using the GLMM framework approach

Setting it apart from the standard 1 PL model, because the ability of a child is assumed to increase non-linearly with age, we consider fitting an extension of the 1 PL model by adjusting for age using a spline function set on the backbone of the GLMM framework. Because of the importance of the monotonicity assumption, and the spline function selected will have to be monotonically increasing. The work of Geert & Geert, (2005) outline the various modelling options available when confronted with non-normal data within the generalized linear model framework.

The 1 PL IRT model assumes that the discrimination of each item is the same. Also, we have seen that it is plausible to assume that ability, quantitatively captured by a score, increases with age. Naturally, in accordance with these observations, as opposed to relaxing the discrimination assumption in line with the 2 PL model, we allow the discrimination parameter of each item to vary with age. Further, because this relationship of probability of passing an item and age is non-linear, we attempt to capture it using a spline function and thereby the effect of age is removed from the ability estimates. This 1 PL IRT spline model is still linear because it involves a linear combination of (known) spline based functions. Formally this 1 PL IRT model can be defined as follows using the mixed effects logistic regression model in terms of logit or log odds;

$$log\left[\frac{P\left(Y_{ij}=1\middle|\theta_i\right)}{1-P\left(Y_{ij}=1\middle|\theta_i\right)}\right]=\alpha_j+\beta_j x_i+\theta_i \qquad 5.74$$

where; Y_{ij} denotes the response made by the i^{th}child to the j^{th} item,

α refers to the difficulty of the j^{th} item,

$\beta_j(.)$ is a is a smooth function that satisfies a monotonicity constraint that characterises the discrimination of the j^{th} item against a child's age, x_i,

θ_i refers to the trait level of the i^{th} where higher values reflect a higher level on the ability trait being measured by the items.

The model 5.74 can be implemented using the lme4 package in R software used to fit both linear and generalised linear mixed-effects models but ensured that monotone spline functions which we need in order to adequately capture the non-linear relation of outcome and age. Higher positive score values reflect a higher trait level being measured by the items. In accordance with our research context the trait values are usually assumed to be normally distributed in the population of children with a mean zero and variance σ^2, $N(0, \sigma^2)$.

5.2.2.4. Model selection and diagnostics for overall scoring methods

(a) Model fit for GAMLSS models

To assess model fit in terms of overfitting and suitability of selected smoothing terms of the non-nested GAMLSS models used in both the total score and smoothing Z-score in Sections 5.2.2.1 and 5.2.2.2 respectively, the generalised Akaike information criterion (GAIC; Akaike, 1983) as described in Rigby, et al., (2005) was used. Further, in line with the primary objective of this research of developing robust and more sensitive statistical scoring methods, selected models' (developed using only the standard sample) ability to correctly classify delayed or disabled children was tested in both the disabled and malnourished samples as recommended by both Ripley (1996) and Hastie, et al., (2001).

(b) Goodness of Fit Statistics to assess suitability of IRT models

As described by Maydeu-Olivares, (2015), the goodness of fit (GOF) was used to describe how well the classical IRT models matched the set of observed data. Specifically this involved using *p*-values from the various goodness of fit indices and goodness of fit statistics to assess absolute model fit as well as piecewise model fit to identify sources and sections of misfit. Beyond determining whether each of the fitted IRT model could have generated the observed assessment data, the degree to which each model's scores met this research objectives distribution characteristics described in Section 5.3.2 so

as to be able to more sensitively detect delayed development in the study samples was also assessed. Assessment of the suitability of the IRT extension model was with regard to the fit of the model in the Generalized Linear Mixed Model (GLMM) framework. Thus as outlined in Section 5.2.2.4a) the use of the disabled and malnourished samples with a high percentage of children known to be delayed or disabled was used to validate each of IRT models developed using the standardisation normal sample only i.e. each if the IRT models developed using the standardisation normal sample were parsed on to both the disabled and malnourished sample data to check that predicted scores were indicative of the known delayed development status of these children.

5.3. Methods for comparison of age estimate(s) and score characteristics

We have so far reflected on various age estimation and overall scoring methods in Section 5.2 after lending ourselves to the highlights of important statistical issues that underpin appropriate assessment tool development in the preceding first two chapters of this thesis. Beyond this, a question that naturally arises is 'which is the best or most appropriate statistical approach to deal with highlighted classical method weaknesses?' The best approach avers to the most statistically sound method to compute the age estimate or score given all characteristic features of the study as well as those of the children assessed and more importantly the research objectives.

At the outset the decision to use a given age estimation or scoring method is primarily driven by two issues. Firstly, we must consider the research objective: Are we assessing performance of items in an item development process or are we actually developing assessment norms using an already established tool? Secondly, we must consider whether we are in a screening or a rigorous ability/disability diagnosis scenario which dictates the degree of accuracy the scoring method should achieve. An answer to this question(s) will be the main impetus towards making our scoring method recommendations adopted in common practice. 'The best approach or method' could mean several

things but in our context we take it to mean the simplest to compute but still be of the highest accuracy in terms of sensitivity and specificity of detecting delayed development without compromising on quality. Therefore, this section serves to convince any sceptics especially those influenced by the scoring method extensions of the preceding Section 5.2.

The claims argued by our motivations for suggested and developed scoring method extensions that are evidenced by our results findings need to be objectively justified to warrant their acceptance and assimilation as the new best scoring practise(s). Thus a comparison of the current scoring methods versus extensions will be carried out primarily to investigate score characteristics and test their robustness on various types of data given the highlighted limitations of the rudimentary or classical methods to produce more reliable and valid scores.

The age estimate or overall score comparison rationale mainly takes the form of assessing various aspects of both quality and performance especially with regard to accuracy of ability classification using the three data sets described in Section 4.2. The accuracy of classifying disabled children is expected to be higher than the accuracy of classifying malnourished children. This is because current scoring methods have not been shown to be very sensitive in detecting delayed development in malnourished children. Specifically, this will be by using the standard sample of healthy normal children;

- To fit item by item models and compare the quality of age estimates accuracy. The obtained age estimates that point to a specific number of items a normally developing child should pass can be used to assess to what extent the children in either the disabled or malnourished samples are able to pass the same items. It is expected that children in either the disabled or malnourished samples will not pass as many items for their age.
- To compute overall scores that will be used to compare overall scores from either the disabled or malnourished samples in order to classify their development status and therefore test each methods accuracy. Similarly, it is expected that due to disability or malnourishment, children in

these testing samples will on average have lower scores than children of similar age in the standard normal sample.

The next section outlines the dimensions of comparison of scores and general research questions given that the outputs and objectives of item by item analysis and overall total score analysis differ. In the subsequent sub-sections, we will describe the various methods that will be used to compare the characteristics and properties of scores objectively. We hope that this section will convince researchers of the robustness of our scoring extensions' to better mitigate the detrimental effects of poor or late development delay detection.

Dimensions of comparison of age estimates and scores

The item by item analysis output is to give age estimates at selected probabilities of interest that a child is likely to pass an item given their age, while total score analysis calculates an overall score for each child using all their administered items. The comparison of the performance of the item by item analysis or overall scoring methods described in this section will be with respect to their quality (i.e. distribution, consistency, validity) and accuracy (i.e. sensitivity) and the effect of age both on scores and sensitivity. More specifically this will be in terms of;

- Assessing whether the quality of estimates or scores differ with respect to the statistical method used and particular population characteristic(s)?
- Assessing the accuracy of the different modelling or scoring approaches; is the sensitivity across methods the same and to what extent does it differ according to statistical method?
- Assessing the extent the different overall scoring methods correct for age and its influence if any in the classification of a child's development status?

Table 5.1 below gives a summary of the scoring methods implemented in this thesis. Highlighted in grey are the suggested extensions we developed to the current scoring methods. Because the different approaches produce different outputs either age estimates or scores and therefore having

different estimate or score ranges, each method will be evaluated in terms of how better or worse its quality as well accuracy is for the classical versus its extended approach. For example, in discussing the Z-score method(s), we will explore the quality and accuracy of classical Z-scoring versus the proposed Z-score extensions as well as the other competing overall scoring methods.

Table 5.1: A summary of age estimation and overall scoring methods

Scenario	Statistical Approach	Classical method(s) and extension(s)
Item by item analysis	1. Generalised Linear Models (GLM)	• Logistic models with asymptotic confidence intervals
	2. Generalised Additive models (GAM)	• GAM extension (SCAM) with bootstrapped confidence intervals
Creating overall score using all administered items for each child	1. Model based scored based on Simple score	• GAMLSS model on total simple score corrected for age
	2. Z-score	• Classical Z-score
		• Smoothed Z-score using GAMLSS model
	3. Item Response Theory (IRT) Models	• One Parameter Logistic model (1 PL)
		• Two Parameter Logistic model (2 PL)
		• †1 PL IRT Spline Mixed Model

†Extension to the 1 PL IRT model allowing correction of age and implemented using a mixed model approach

Shaded in grey are the item by item and overall scoring extensions suggested and developed in this chapter

5.3.1. Comparison of age estimates from item by item analysis

This Section will outline how the quality of ability estimates will be evaluated in terms of their distribution, consistency (reliability), validity and if they differ according to various population characteristics such as age.

The GLM framework was used to estimate the age at the selected probability of interest that a child was likely to pass an item and was extended to include the use of GAM models. We have already discussed in Section 5.2.1.3 that on account of the classical GAM model's flexibility renders it not suitable within this ability measurement context and therefore we recommended the use of the SCAM extension as an alternative to the GLM approach to retain monotonicity. In this section we will assess the similarities of the age estimates from the 2 approaches at the different probabilities of interest. Further, we will compare the age estimates of the 2 methods using plots of each item against age estimate to check if a particular method consistently gives higher or lower age estimate values.

Owing to sample size and population characteristics that may result in skewed item response distributions, the confidence bound around the GLM models may be affected if various important assumptions are violated. These assumption violations may for example manifest as thinning at the tails of the logistic model, or appear as very narrow or too wide intervals that are not a true reflection of the actual variability in the data. Given that quoting a confidence bound for reported age estimates is important, we explained the use of bootstrapping as an alternative to classical asymptotic methods within the GAM model framework. We will therefore compare the behaviour of asymptotic versus bootstrap confidence intervals in terms of their width of the different methods of producing confidence bounds of the various age estimation approaches.

5.3.2. Total score summary measures and distribution characteristics

Distribution of scores

The distribution of scores used to classify developmental status is key in determining the ideal classification cut-off thresholds (see Lord, 1955; Cook, 1959; Micerri, 1989). In the second chapter we saw that the commonly used Z-score for example uses the distance or deviation from the mean score value to define score cut-off thresholds used for development status classification. This 'distance' is often expressed in terms of variability around the mean. Thus it is important that the distribution of scores is known.

The degree of adherence to the normal distribution was one of the ways of examining performance of the score produced. Score density plots, scatter plots of score with age and summary measures of mean, standard deviation, minimum and maximum (range) were used to assess the distribution of the total scores developed (see DeCarlo, 1997). The different plotting methods give a snapshot of the distribution of scores; however, the different methods magnify different aspects of the distribution characteristics that are of interest for our comparisons. The score density plots for example have the benefit of giving a general view of the distribution and indicate if there is any skewness. If overlaid, the density plots of scores from the different samples will help to assess the extent of overlap to advice on the most suitable cut-off thresholds to be used for development status classification. The densities also check if there are any systematic differences or similarities in spread of the scores like bimodality. Further, an objective confirmation of the score value adherence to the normality distribution was formally checked by the Shapiro-Wilks test of normality (Shapiro & Wilk, 1965).

Distribution of scores with respect to age

One of the primary objectives of this thesis was to create a framework for appropriately correcting for age of the child in the score computation. It is thus important to assess if our suggested extensions

achieve this and to what degree by assessing both the presence and strength of any correlation between scores with age or any systematic pattern of scores with respect to age.

Scatter plots of the age variable versus the score were produced; firstly, to assess the nature of the association between score and age variable, secondly to assess if the scoring method indeed corrected for age as a complement to the correlation findings and thirdly illustrate to what extent the suggested extension methods managed to correct for age. A lowess smooth curve (locally weighted scatter plot smoothing) as described by Cleveland & Devlin, (1988) was overlaid on the scatter plot between age and score to assess if it results in a horizontal line with a zero intercept implying that the scoring method was adequately correcting the effect of age. The Spearman correlation coefficient was used to objectively check the strength of the correlation between the scores with the age variable. We would like to also note that even if the computed correlation coefficient will be less than 0.5, thereby endorsing the scoring method as having adequately corrected for age, it is likely that a significant *p*-values will be reported. This is bound to happen especially in the normal children sample because of the large sample size (n=1446).

5.3.3. Classification of developmental status: criterion validity

The consistency of scores to classify child development status produced under the different methods will be compared. This will be to specifically assess if all the scoring methods manage to correctly classify a child known to be normal or delayed. We hypothesize that the severely delayed or disabled children will consistently be classified as such using the different scoring methods and have some similar underlying characteristic(s) that we will attempt to identify. We hope that the more robust scoring extensions will be able to correctly classify children with borderline (moderate) ability especially in the malnourished sample that are often harder to correctly classify.

Further, in line with the sensitivity analysis objectives, we will also compare inclusion of these characteristic variables in the sensitivity comparison method in the hope of; a) showing superiority of certain scoring methods by quantifying the misclassification error rates of the different scoring

methods under our extended framework of correcting for important variables such as age, b) investigating if age for example also influences the sensitivity of a scoring method i.e. exploring if different classification cut-off thresholds are required to maximise the sensitivity of differently aged children.

In our research context, criterion validity also called concrete validity refers to the extent to which a measure is related to an outcome; in this case normal or delayed/disabled. Criterion validity can be viewed as concurrent or predictive validity. The former refers to a comparison between the measure in question and an outcome assessed at the same time. The later compares the measure in question with an outcome assessed at a later time. Even if somewhat similar, in the book 'Validity in Educational and Psychological Assessment' by the authors (Newton & Shaw, 2014) caution that it is best to keep the two types of validity separate unless a suitable rational to view them as similar exists. Therefore, bearing the above in mind we will be interested in investigating the criterion validity of the scores from various methods i.e. assessing how well the scores computed using both classical and extended methods are able to correctly classify the developmental status of children known to be normal, malnourished or disabled.

To correctly classify the development status of a child will to a large extent depend on the scores ability to capture the target underlying trait, developmental ability, given the threshold (or cut-off) used to demarcate status. Therefore, Receiver Operating Characteristic (ROC) curves as described by many authors including Pepe, (2009) will be produced for each method to evaluate different score threshold choices under the item by item framework and total score frameworks respectively.

We consider the problem of classifying developmental status of a child similar to the problem of classifying individuals into one of two groups; diseased or non-diseased persons. In our classification context these two groups will be delayed or disabled and normal. Obviously, due to test (model) shortcomings, the process is bound to have some degree of misclassification error. Thus justifying error or accuracy evaluation and hence the motivation to pursue research aimed at the identification

of the more robust methods with less status misclassification error rates. The test here refers to both the scoring method and threshold used in classifying developmental status. The most appropriate method of estimating the accuracy measure of a test depends on the scores' data type, its characteristics and assumptions made about the distribution of these test scores.

Typically, while evaluating scores used to classify developmental status, you are likely to find yourself in one of two scenarios; either a scenario where there is a gold standard available i.e. the true developmental status of the child is known without error or the more practical scenario where there is no gold standard i.e. you are not certain of the true developmental status of the child. In this project we do know the true developmental status which is taken to be normal for the large standardisation normal sample and disabled for both the disabled and malnourished cohorts. What differs is the extent of normality or disability or delay that is reflected by the magnitude of the score. See Table 5.2 below showing the two types of misclassification errors of false positives or negatives.

Table 5.2: Types of errors in child development classification

	True development status (Gold Standard) or Test 2			
Test 1		Negative (T-)	Positive (T+)	Total
	Negative (T-)	True positive (T-,T-)	False negative, FN, (T-,T+)	
	Positive (T+)	False positive, FP, (T+,T-)	True negative (T+, T+)	NP
	Total			NT

To compute misclassification rate, we let:

- T+ = the test is positive (indicating that the development delay or disability is present)
- T- = the test is negative (indicating that the development delay or disability is absent)
- NP = Number with the disease in observations.
- NT = Total number of observations i.e. diseased and non-diseased.

The mis-classification error rate (MR) was computed as;

$$MR\ \% = \left[\frac{FP + FN}{NT}\right] \times 100\% \qquad 5.84$$

where; FP, FN and NT are as defined in Table 5.2. The misclassification error rate (MR) was used to compare the performance of the classical scoring methods versus proposed extensions.

Mis-classification error assessment

A mis-classification error is typically defined as either the number of false positives or false negatives. However, in other contexts certain types of misclassification errors may have higher 'costs' than others. As was explained in the score distribution Section **Error! Reference source not found.** above, it is important to advice on the appropriate diagnostic cut-off threshold to use for classification.

The main motivation to assess error rates stemmed from the fact that misclassifications may not occur symmetrically or often a more accurate classification method is needed for some classes or groups of children than others for reasons unrelated to the actual scoring method or relative class sizes. For example, it may be harder to classify development status of very young children using a particular scoring method i.e. certain scoring methods may be more sensitive for certain age groups of children. Therefore, beyond recommending a scoring method as being superior, we will also be able to recommend the age group the scoring method is likely to perform well in terms of development status classification.

Regardless of their relative frequency in the population, carriers of a disease or children of a given development status are more accurately predicted than carriers of the disease who differ by certain characteristics. Put another way, depending on the development delay context, either false negatives or positives maybe more severe in certain child groups than others. If one assumes that little is lost, in terms of 'cost', in avoiding one type of error in comparison to the other, the definition of the mis-classification error can be redefined accordingly. 'Cost' in this research context refers to the danger posed by for example classifying a normal child as delayed and subjecting them to subsequent unnecessary testing, or classifying a disabled child to be normal yet they are not, leading to the effect

of disability becoming worse. This 'cost' of mis-classification depends on the research context and is amplified further by factors like the prevalence of disease or disability.

However, our objectives here are to; i) Firstly, to assess the distribution of mis-classification error rates and check if they occur at the same rate. ii) Secondly, we wish to propose a more informed strategy to select thresholds that will optimize both types of errors in the instance where there is bias or imbalance in their occurrence across various score thresholds. Our target is to have a low MR of about 5% and a very high sensitivity greater than 80 % in differentiating between the normal and non-normal samples.

The receiver operating characteristic (ROC) curve estimation

As described by Beck, (1986) and Pepe, (2009) the ROC curve summarises the performance of any binary classifier that can be summarised in a cross tabulation as shown in Table 5.2 for various test (score) thresholds. The ROC is found to be a more attractive tool to assess diagnostic accuracy unlike isolated measures of sensitivity and specificity (Zweig, et al., 1993). The classification of a child as delayed/disabled or normal is assumed to be determined by an underlying diagnostic variable Y. Once Y exceeds a certain threshold c ($Y \geq c$), the child is classified as delayed or having the condition of interest (positive), otherwise if Y is less than the threshold c ($Y < c$), the child is classified as normal (negative).

Formally, we let the cumulative distribution function (CDF) of a (non) delayed child be denoted as $F_1(F_0)$,then the ROC-curve is a plot of $F_1(c)$ (true positive rate) on the y-axis against the corresponding $1 - F_0(c)$ (true negative rate) on the x-axis for varying values of c over the support of Y. If Y is continuous, the ROC-curve can be written as:

$$R(p|F_1, F_0) = 1 - F_0\left(F_1^{-1}(1-p)\right) \qquad 5.85$$

where; p is the probability of being in one of the 2 categories, $0 \leq p \leq 1$.

Based on literature recommendations, we too aspire to have a low False Negative Rate (FNR, < 5%). We specify that a cut-off value, c, for the score is suitable only if the FNR is not more than 0.05. Once a suitable cut-off value is found, we then estimate the specificity at this cut off and the corresponding score value on the original scoring method's scale.

Assessment of Covariate Effects on ROC Analysis or Misclassification Error

We have seen that covariates like age are strongly associated with the child's development process and therefore also strongly correlated with ability scores. Several authors such as Liu, (2013) advocate for covariate adjustment when they also impact on the magnitude or accuracy of the test under study. This is because it is also possible that covariates are also correlated with the diagnostic testing procedure. That is a given test or chosen score threshold may be influenced by the subjects' characteristics they are applied to and therefore in turn influence diagnostic classification. For example, a certain score threshold may only be able to appropriately differentiate children of a very specific age category. This maybe be due to the fact that the assessment tool was designed for a specific age group of children. It may be also possible for the covariates to be both related to the disease or process being assessed and the diagnostic testing procedure. The former case has been checked and addressed by our scoring extensions.

If it is possible that covariates also influence the discriminatory ability of the test, this should be explored and the extent of their influence on status diagnosis assessed. The work of Janes, et al., (2009) for example describes three methods of using covariate information; first, to use covariates to adjust classification markers which in our case refers to the scores, secondly, for factors that affect discrimination they describe ways of modelling ROC curves as functions of covariates, and thirdly, when factors contribute to discrimination, they suggest methods of combining covariate information to markers. However, in this project while our primary objective is to explore the effect of important covariates (age) on the classification process and consider their inclusion in the score computation, we also assess their influence on sensitivity performance.

The Receiver Operating Characteristic Area Under the Curve (ROC AUC)

Hanley, (1982) defines and describes the ROC AUC from another disease context. In our research the ROC AUC represents the likelihood or probability that the test or assessment tool will rank two children as delayed or normal in the correct order over all possible cut-off score thresholds. In other words, the ROCAUC can be defined as the area between the graph of the function defined in formulae 5.85 and the x-axis i.e. assuming that normal children will have higher ability scores, the AUC represents the chance that a randomly chosen child who is delayed will be correctly ranked below a randomly selected normal child. A perfect or optimal threshold cut-off of a test should have 100% sensitivity with zero false-positives (100 % specificity) across all possible thresholds. This point lies at the extreme top left-hand corner of the ROC plot resulting in AUC = 1·0. Such tests don't exist in reality, and we expect some failure or error in attempts to separate normal and abnormal children. A straight line connecting the extreme bottom-left (sensitivity, FPR: 0,0) and top-right (1,1) corners (the 'chance diagonal') describes a test with no discrimination i.e. the AUC = 0.5. Given that the ROC curve is a summary of the sensitivity and specificity of a given scoring method over a range of thresholds, the AUC provides an objective method to compare either; different ROC curves produced under different scoring methods or the same scoring method producing ROC curves under different scenarios, for example comparing two ROC curves drawn from the same scoring approach but using different child characteristics e.g. different age categories.

The pros and cons of using ROCAUC to summarise test sensitivity and specificity across various thresholds have been well summarised and argued by several authors such as Fawcett, (2006) and Hand, (2009). However, you will notice that each of the arguments is anchored on a specific research context or application. In the same spirit we argue that the suitability of the use of the ROCAUC in this child development assessment research also depends on context and more specifically the behaviour of the ROC curves over the entire threshold spectrum. If the classification objective is in a screening scenario, then ROCAUC may be suitable as interest lies in identification of 'suspect' delayed children.

If the scenario is a formal diagnosis classification then in line with the demerits outlined in literature for example when the ROC's being compared cross, then ROCAUC may not be a suitable method of comparison. This is because one can only say that a certain method of scoring is better/worse or is only suitable for a certain age group of children for only a specific threshold range, and not over the entire threshold range. Further, as outlined in the STARD statement (Bossuyt, et al., 2015) if the prevalence of condition, its spectrum or study design influences the ROCAUC then the measure of diagnostic accuracy should be appropriately justified. Here, since the objective was to only highlight if there is indeed a difference in diagnostic accuracy between scoring methods, and to identify if age also affects the difference in accuracy, the ROCAUC was used.

ROC within the total score framework

We note that it may not be advisable to consider the sensitivity and specificity of just a single item to decide the developmental status of a child as there are several items in a tool that test the same construct. Instead the assessment of item by item accuracy should be viewed from an item selection or development process perspective where one's objective is to find the best performing items in terms of their quality to adequately discriminate between children.

In this sub-section we describe the comparison of the mis-classification error of the overall scoring methods described in Section 5.2.2 using ROC. In order to compute the mis-classification error across all possible thresholds needed to plot the ROC curve, the following eight steps were followed;

1. Create an indicator variable 'True status' to indicate the actual development status for each of the children in the three data sets as; 0 for normal, 1 for disabled and 2 for malnourished. This variable will be taken to be the 'gold standard' variable to indicate the true status of the children i.e. the children in the standardisation normal cohort are normal and those in the other Disabled and Malnourished are deemed non-normal. Since it is of interest to compare

the performance of our scoring methods at classifying children of different characteristics, we will have two comparison scenarios; normal versus disabled and normal versus malnourished.

2. Parse the scoring computation model from the standardisation normal data to each child in the Normal, Disabled and Malnourished data and obtain their overall scores. Repeat this depending on the overall scoring method i.e. use estimates that characterise normal development derived from the standardisation normal sample to compute scores for the disabled and malnourished samples.
3. Create a variable 'Threshold' that is a vector containing 99 different quantile values of the score values from the normal data for each scoring method. These quantile values define the various cut-off thresholds to be considered and the one resulting in the lowest mis-classification error will be selected. For the scoring methods that do not adjust for age, the 'Threshold' values are defined by the predicted 2.5 % quantile scores value from the fitted GAMLSS model using the normal sample scores for the age range of interest.
4. Create three arrays with dimensions of n by 99, where n is the sample size of each sample data and 99 refers to the quantile thresholds defined in step three.
5. Create an indicator variable 'Model status' where, given a specific threshold, a child is deemed delayed/disabled if their score from each scoring method is less than or equal to (</=) the defined threshold defined in the step four above. The child is deemed normal if their score for the given method is greater than (>) the defined threshold defined in the step four.
6. Repeat step five iteratively creating the respective 'Model status' indicator variables and write the result on the arrays created in step five. This step should be repeated for all 99 thresholds of interest, all three data sets and for all overall scoring methods considered.
7. Append the three data sets; normal, disabled and malnourished data sets including the 'Model status' variable created in step six.

8. Using a cross tabulation of the ‘True status’ and ‘Model status’ variables, compute the mis-classification errors as described in equation 5.84 above for each defined ‘Threshold’ in step three, make the ROC plot and compute the area under the curve for comparison.

The following results chapter is the output of this methods chapter. While contributing to the current scoring discourse its main aim is to further reinforce the robustness argument pitched by our motivations for the proposed extensions of the age estimation and overall scoring methods.

PART III – RESULTS

Chapter 6. Results

6. Results

6.1. Introduction

This chapter presents the findings of the item by item age estimation and overall scoring methods. The performance of both age estimate and overall scores are compared in terms of their characteristics as well as their sensitivity to detect children with disability or delayed development as described in Section 5.3. These results of the item by item analysis are presented in Section **Error! Reference source not found.** and overall scoring methods presented in Section **Error! Reference source not found.** respectively, with the comparisons between each classical method versus extended method presented within either of the analyses frameworks. As the primary objective of this thesis is to extend age estimation and overall scoring methodology, and the EDA concluded that similar item characteristics exist in the other three assessed domains, only results of the gross motor domain will be presented.

6.2. Item by item age estimation analysis

The main objective is to assess if the age estimates obtained from the item by item analysis may be more accurately estimated using more flexible and robust modelling methods. As a starting point we considered the GLM model (Gladstone, et al., 2008) and then moved on to use the more flexible GAM framework. To adequately deliver on the main objective of this chapter, we will highlight issues that the various modelling approaches should address at the outset. By utilising appropriate plots or tables capturing and summarising the relevant issues pertinent to the quality of characteristics of the required age estimates, we outline how our suggestions remedy the issues raised under the following sub headings; age estimate distribution, model fit, monotonicity, confidence interval, data issues, computation or implementation problems and item ordering. Our motivation for using this approach to present our results and discussion is because many of the issues are related, influence each other

in concert and our suggested remedies may partially or completely address them either in isolation or simultaneously.

Presentation format of item by item analyses results

Following the key research issues initially prompted by the literature review, confirmed by the EDA findings, in addition to the pros and cons of the age estimation methods and extensions highlighted in Section 5.2.1, we will now discuss these same issues using vivid examples from the results. Specifically, the comparison of age estimates from the two item by item modelling approaches will assess how the following issues were addressed;

a) There is an increase in variability of pass probabilities as age increases. Which of the two approaches considered is able to capture this aspect more appropriately? How adequately each of the modelling approaches' confidence bands captures this aspect will be compared.
b) There is a noticeable rapid rate of development reflected by a high item pass rate in the immediate first few months after birth. Therefore, which of the two item modelling approaches are able to flexibly capture and model this aspect more appropriately? We will assess the suitability of the model at the first few months after birth.
c) Does the 'length' between age estimates at pass probabilities of interest differ as age increases across all the two modelling approaches?

   The difference could be attributed to either; the modelling approach's ability to deal with the change in variability with respect to age or the fact that there were more young children sampled than older ones resulting in more variability in older children.

   Which method gives the most appropriate distance between respective percentiles with respect to the age variable or are the lengths between percentiles consistent across methods?

   As was explained in the literature review, lengths between percentiles of interest between adjacent items is important for two reasons; i) Firstly, it cannot be too short to the point that there is no overlap of age estimates at probabilities of interest as this means that there is an age range

without a relevant item to characterise development. ii) Secondly, it cannot be too long such that there is too much or considerable overlap between the age estimates at percentiles of interest as this implies there is no discriminative difference between the two items. Thus, one of the two items may not be necessary as both items may be assessing exactly the same development construct. We will look at plots of age estimates between adjacent percentile values of a given modelling approach to answer this question.

d) We noted in our literature review that the ordering of items with respect to difficulty is done first using expert knowledge at the tool design stage and then ordered according to the age estimate at percentile of interest e.g. the 90^{th} percentile using pilot data. This was done in the work of Gladstone, et al., (2008) to come up with the current ordering of MDAT items. We will be interested to know the extent to which the ordering of items changes depending on the modelling approach used to compute age estimates at probabilities of interest.

Ultimately, we endeavour to address the question 'which approach is ideal in addressing all the above four issues and consequently leading to more accurate age estimates?' To address these specific issues listed above in line with our broader research objectives, we will pitch the presentation of results using the following three purposefully chosen item response scenarios; i) Firstly, an item that is ideal for infants (< 1 year old) and therefore has a high pass rate as almost all the children easily pass it. ii) Secondly, an item that is ideal for toddlers (1 to <3.5 years old) and therefore has an average (moderate) pass rate. iii) Thirdly, an item that is ideal for older pre-school aged children (3.5 to <7 years old) and therefore has a low pass rate as many of the infants and toddlers cannot pass it. These age categories have been chosen using guidance from expert opinion given their relevance in having very low, high or evenly distributed pass rates to facilitate comparison of the different modelling approaches.

Figure 6.1 below shows the model fit results of the generalised linear logistic model (GLM) and the shape constrained additive model (SCAM) fits for the items 1, 17 and 34 that are ideal for infants,

toddlers and older toddlers (pre-school age) respectively for the MDAT Gross Motor domain. These are scatter plots of the observed probabilities of success against age represented by black dots. Overlaid on each of the scatter plots are the model fits represented by a black line and the dotted red lines represent the confidence interval band around the model fit. The blue line shows the non-parametric isotonic regression fit on aggregated data whose computation was explained in Section 5.2.1.5 b). Beyond the model fit given the item pass probability distribution with respect to age, the isotonic regression fit assists to graphically evaluate the suitability of the fitted model. Figure 6.2 shows graphically how the 25^{th}, 50^{th}, 75^{th} and 90^{th} percentile age estimates were computed by rearranging the various model formulae explained in Section 5.2.1.3. Figure 6.2 also shows the graphical representation of confidence interval values around the age estimate which are summarised in Table 6.3 for both GLM and SCAM models.

Figures 6.3a) to 6.3b) shows the Normal reference age values for 25th to 50th (white box), 50th to 75th (cyan box) and 75th to 90th (dark blue box) percentile plots for the 34 items in the MDAT gross motor domain using the GLM and SCAM models. Note that the scale in the graph is not linear; the initial items that are ideal for infants use a monthly scale in order to show more detail in the developmental milestones between items. The rate of child development in the first year after birth is very rapid thus if the centiles were represented on the same scale, some detail would not be very apparent. Using the 90th % probability age estimates we see that gross motor items 1 to 12 are ideal for infants who are less than 1 year old, items 13 to 22 are ideal for toddlers who are 1 to less than 3.5 years old and items 23 to 34 are ideal for pre-school age children who are 3.5 to less than 7 years old. The work of Gladstone, et al., (2010) advised the ordering of the items using the 90% pass probability age estimate value in Figures 6.3a) to 6.3b).

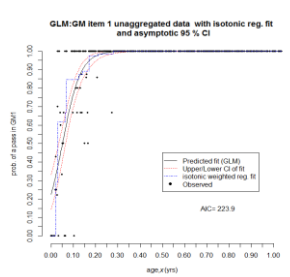
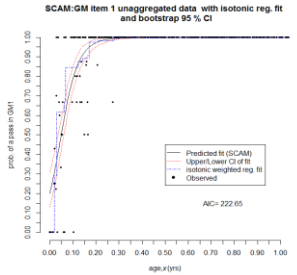
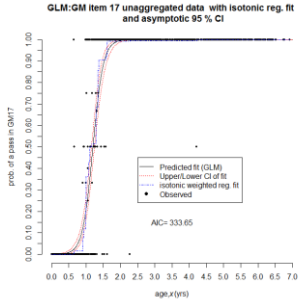
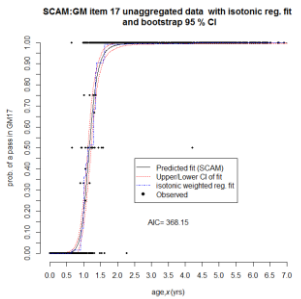
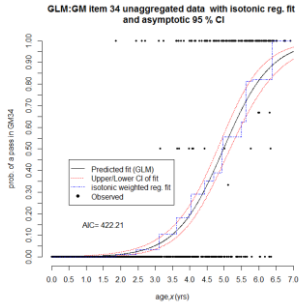
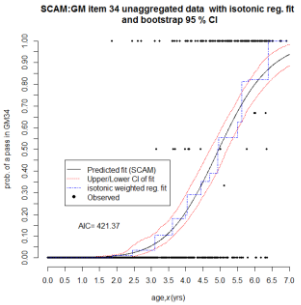
Type of items	GLM	SCAM
GM item 1: Ideal for infants (<1 year old)		
Gm item 17: Ideal for toddlers (1 to <3.5 years old)		
GM item 34: Ideal for pre-school age children (3.5 to <7 years old)		

Figure 6.1: A comparison of GLM and SCAM model fits for items ideal for infants (item 1), toddlers (item 17) and pre-school age children (item34) in the GM domain.

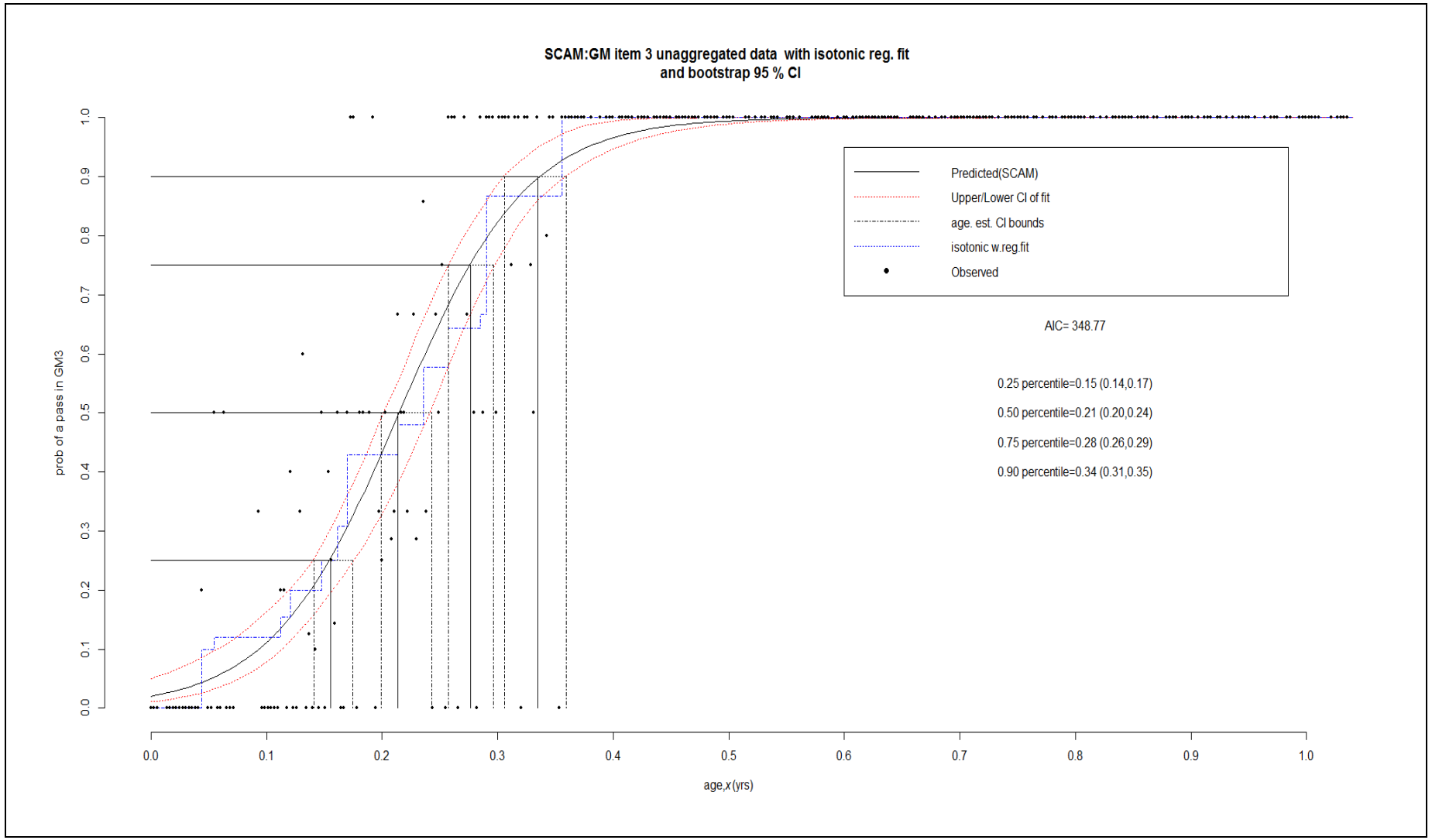


Figure 6.2: Display of model fit, confidence band around model fit, computation of age estimates and age estimate confidence band on single item data.

Table 6.1: Age estimates at the 25th and 90th percent probability of passing an item from GLM and GAM extension (SCAM) models for gross motor domain.

Type of item	Item(s) label	25th % probability		90th % probability	
		GLM	SCAM	GLM	SCAM
Ideal items for infants <1 year old	GM1 - Lifts chin up off floor for a few secs.	0.01	0.02	0.17	0.14
	GM2 - Prone, head up to 90 degrees	0.09	0.09	0.27	0.24
	GM3 - Holds head straight or erect for few secs.	0.15	0.09	0.30	0.26
	GM4- Pulls to sit with no head lag	0.12	0.17	0.47	0.33
	GM5 - Lifts head, shoulders and chest when prone	0.18	0.15	0.36	0.34
	GM6 - Bears weight on legs	0.22	0.21	0.55	0.38
	GM7 - Sits with help	0.21	0.22	0.50	0.38
	GM8 - Rolls over from back to front	0.23	0.23	0.57	0.42
	GM9 - Sits without support for a period of time	0.32	0.31	0.58	0.47
	GM10 - Sits by self well	0.35	0.37	0.64	0.52
	GM11 - Crawls (in any way)	0.51	0.53	0.95	0.79
	GM12 - Pulls self to stand	0.57	0.57	1.10	0.98
Ideal items for toddlers 1 to <3.5 years old	GM13 - Able to stand if holding on to things	0.60	0.61	1.14	1.02
	GM14 - Walks using both hands of somebody	0.70	0.69	1.38	1.34
	GM15 - Walks with help - hand or furniture	0.79	0.78	1.48	1.45
	GM16 - Walks but falls over at times	0.97	0.98	1.65	1.60
	GM17 - Stoops and recovers	1.04	1.06	1.78	1.71
	GM18 - Walks well	1.08	1.10	1.81	1.73
	GM19 - Runs	1.15	1.17	2.35	2.33
	GM20 - Kicks a ball in any way/tries to kick ball	1.63	1.41	2.66	3.07
	GM21 - Runs well (confidently) stopping and starting without falling	1.42	1.54	2.90	3.12
	GM22 - Kneels and gets up without using hands	2.35	2.20	3.81	3.96
Ideal items for pre- School aged children 3.5 to <7 years old	GM23 - Throws a ball into a basket (at least one of 3 times) 1 m away	2.37	2.34	3.85	4.04
	GM24 - Runs, stops and is able to kick a ball some distance	2.21	2.40	3.90	3.91
	GM25 - Jumps with feet together off ground	2.45	2.40	3.83	3.95
	GM26 - Jumps over line/string on the ground	2.56	2.54	4.03	4.14
	GM27 - Stands on 1 foot for < 5 seconds	2.90	2.66	4.01	4.08
	GM28 - Walks on heels 6+steps	2.99	2.91	4.44	4.55
	GM29 - Jumps over piece of paper	2.62	2.84	4.59	4.71
	GM30 - Walks on tip toes 6+steps	3.30	2.96	4.70	4.89
	GM31 - Hops on one foot 4 steps	2.91	3.06	4.97	5.19
	GM32 - Stands on 1 foot for a longer time	3.14	3.32	5.16	5.33
	GM33 - Can throw ball in air and catch it with 2 hands	4.13	4.20	6.42	6.00
	GM34 - Heel/toe walk precise one foot behind other along chalk line	4.22	4.00	6.32	6.00

*Highlighted in different shades of grey are instances where the age estimates in respective models across items are not monotonically increasing.

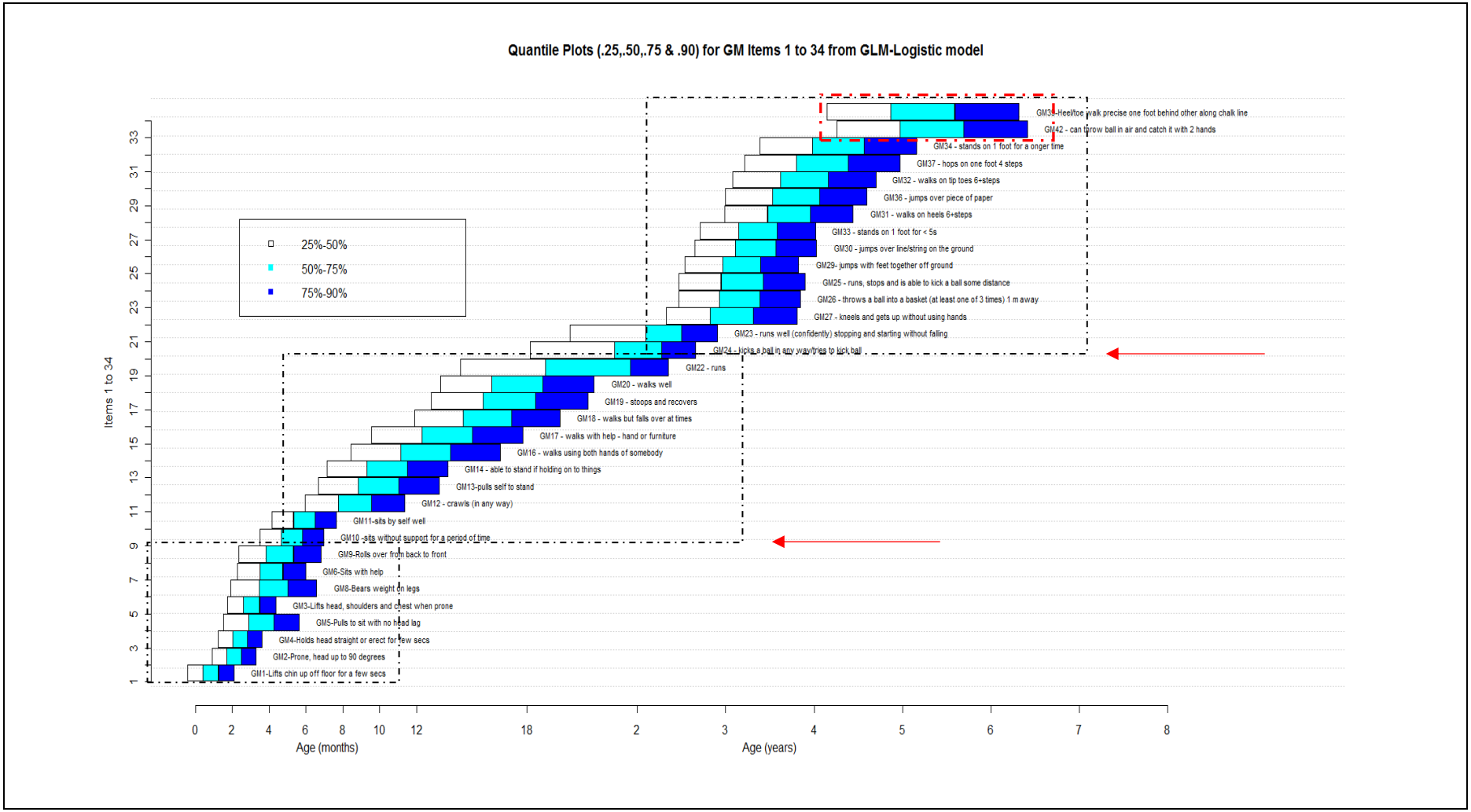


Figure 6.3 a): Normal reference age values for gross motor domain using Generalized Linear (Logistic) Model.

*Red arrows and black dotted boxes demarcate items that are ideal for infants (< 1 year old), item that are ideal for toddlers (1 to <3.5 years old) and items that are ideal for older pre-school aged children (3.5 to <7 years old). The red dotted box shows an example where 2 consecutive items are not appropriately ordered using the GLM model age estimates.

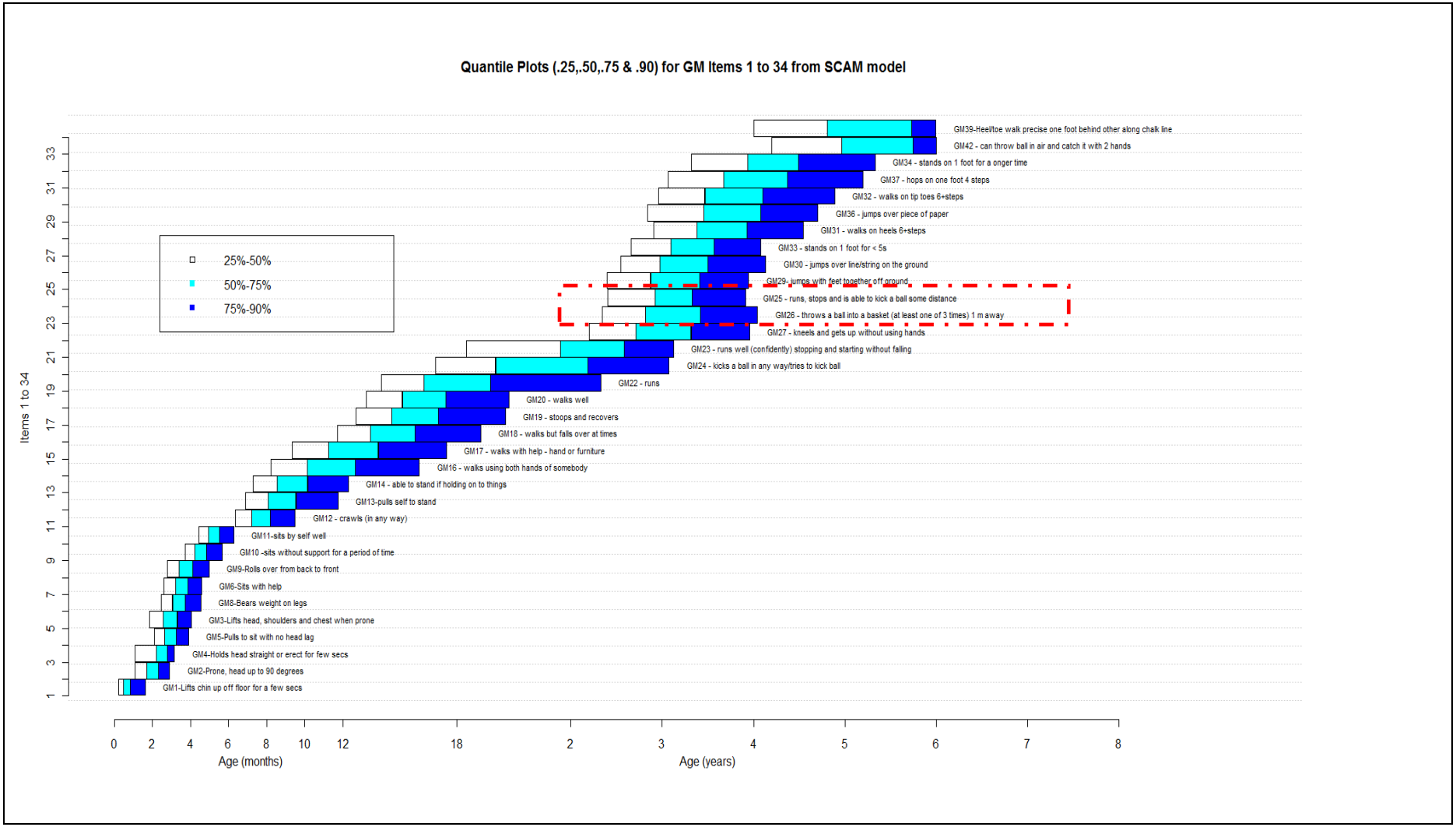


Figure 6.3 b): Normal reference age values for gross motor domain using Shape Constrained Additive (SCAM) Model.

*The red dotted box shows an example of a conflict in the ordering of items previously ordered using the GLM model age estimates presented in Figure 6.3a) above.

6.2.1. Comparison score characteristics of item by item analysis

This section compares several characteristics of the age estimates derived from the two item by item model analyses under the following sub-headings to justify the suitability of our proposed more robust modelling approaches.

Distributional aspects and assessment of item by item model fits

The distribution of pass rates for the respective items within each domain clarified that in general, with increasing age, the probability of passing a given item increases (see Figure 6.1 showing the scatter plots of the proportion of children passing an item against age). Therefore, items that are ideal for infants (< 1 year old) tended to have most of the children across the age spectrum passing while items for older children, toddlers (1 to < 3.5 years old) and pre-school aged (3.5 to < 7 years old) tended to have more children failing the items. It is clear that the GLM modelling approach seemed to struggle to adequately fit the data especially in the items ideal for the very young infants and older pre-school age groups. For example as shown in Figure 6.1, unlike the SCAM model fit, the GLM model fit for item 1 does not adequately overlay the isotonic regression fit over the entire age spectrum.

We assessed model fit graphically by plotting the observed pass probabilities and overlaying the model fits in the different item by item model approaches. Further we fitted a non-parametric isotonic model that does not make any distributional assumptions. Notice that the SCAM fit in all the three types of item scenarios is able to almost always mirror the non-parametric model fit. Further confirmation of an improved model fit is given by the reduced(less) AIC value even if subtle summarised in Table 6.3. The second row in Figure 6.1 shows a poor fit in the GLM case for item 17. GAM presents a more flexible framework to fit the data better as is evidenced graphically, the reduced AIC value. However, this flexibility within the GAM framework also presents a complication discussed in Section 5.2.1.5f) with regard to monotonicity which is discussed further in the sub-section below.

Therefore, from the model fit graphs and age estimates summaries we can conclude that in items where the distribution of pass rates was fairly even across the age spectrum or those designed for older toddlers, the GLM produced comparable estimates compared to the more complex SCAM approach as is highlighted by the AIC value summaries in the red dotted box on Table 6.2. Notice that at best for most of the items the SCAM outperformed the classical GLM modelling approach and at the very least performed as well as the GLM modelling approach. Actually the GLM can be considered to be a special case of a GAM model that has a linear term only. Even where the model fits are fairly comparable, we advocate for using the robust approach due to model assumptions that rely on pass rate distributions and therefore dictate the validity of derived age estimates from each modelling approach.

Table 6.2: Item by item model AIC value comparison in Gross Motor (GM) domain

Type of item	GM Item(s)	Model AIC values		
		[t]Item pass rates	GLM	SCAM
Ideal items for infants (< 1 year old)	1	0.95	223.90	222.65
	2	0.89	311.76	311.76
	3	0.83	348.77	348.77
	4	0.87	322.01	320.92
	5	0.83	276.52	276.52
	6	0.80	255.76	255.76
	7	0.80	272.09	272.09
	8	0.79	275.23	275.23
	9	0.75	187.22	187.22
	10	0.72	141.77	141.73
	11	0.65	316.52	247.25
	12	0.63	394.87	383.53
Ideal items for toddlers (1 to <3.5 years old)	13	0.61	393.22	379.86
	14	0.57	441.45	440.04
	15	0.55	403.03	401.59
	16	0.50	319.78	317.97
	17	0.48	333.65	368.15
	18	0.47	279.96	278.44
	19	0.44	448.66	409.14
	20	0.36	439.06	419.53
	21	0.39	423.75	396.20
	22	0.27	397.47	385.56
Ideal items for pre-school age children (3.5 to <7 years old)	23	0.27	378.52	370.25
	24	0.28	420.07	410.65
	25	0.26	357.55	339.95
	26	0.25	350.75	338.87
	27	0.22	343.02	338.87
	28	0.21	378.64	371.09
	29	0.25	338.01	334.62
	30	0.18	401.65	399.86
	31	0.22	404.00	392.78
	32	0.20	416.24	409.73
	33	0.12	448.37	439.36
	34	0.11	422.21	421.37

[t]Item pass rates in MDAT normal sample.
GM-Gross Motor.
GLM-Generalised Linear Model.
SCAM-Shape Constrained Additive Model.
Shaded in light grey are the model fit AIC values for items 10, 17, 19, 32 and 33 where only the relevant response item data for a specific age range for given item were used to fit both models to facilitate fair comparison between the GLM and SCAm models
Shaded in dark grey are the model fit AIC values for the preferred SCAM model.
Red dotted box shows items where the distribution of pass rates was fairly even across the age spectrum.

Confidence intervals/band around model fit and age estimates

In the methods chapter Sections 5.2.1.3 and 5.2.1.4 we outlined the process of the computation of age estimates and associated confidence bands around the age estimates at the probabilities of interest respectively in both the GLM and GAM model frameworks. The age estimates at the probabilities of interest have been summarised in Table 6.1. From this table we see that the age estimates from the two methods are fairly comparable. These model age estimates have been used to make the reference charts presented in Figures 6.3a) to 6.3b) that are used by experts during the child development assessment process. However, notice that in some cases, highlighted in grey that the age estimate at a particular probability of interest for two adjacent items decreases. This suggests a problem with the ordering of items with respect to this probability. A remedial measure to the above issue is to reorder the items according to the highest probability of interest. In this case it will involve ordering the items according to the 90^{th} % probability age estimate as this will often be the age estimate value of interest.

Within the GLM framework, there are certain fundamental model assumptions that should be adhered to in order to create valid asymptotic confidence bands around age estimates. Therefore, where there is a poor GLM fit for example due to a skewed item response pass probability distribution then this means that the confidence bands around age estimates also have poor coverage. This is evidenced by the 'thinning' of the confidence band seen in Figure 6.1 especially around the model fit tails. The GAM model frameworks also have methodology and corresponding assumptions for creating confidence bands around the model fits. The validity of these confidence bands would still be frustrated by similar assumption issues experienced within the GLM framework. Bearing this in mind, for the implemented model extensions in the GAM framework, we used the bootstrap procedure that was explained in Section 5.2.1.4 to create valid confidence intervals around model estimates. These bootstrap confidence bands appear to better capture the variability around the model fits more appropriately especially around the tails.

Table 6.3 shows the confidence bands around the age estimates. We see that both the GLM and SCAM modelling approaches gave fairly comparable values especially for items where there is an even distribution of item pass probabilities typically found in items that are ideal for toddlers that are highlighted in the red dotted box. We also noticed that the confidence intervals around age estimates tended to get wider from item 20 onwards for both model approaches at the two pass probabilities of interest presented. This is attributed to the fact that as age increases the pass probability variability also increased and is reflected in the width of the confidence band. The shaded grey box shows an example of an unrealistic confidence band for item 1 and is possibly a consequence of a poor GLM model fit.

Table 6.3: Age estimate Confidence Interval at the 25th and 90th % pass probability of items from GLM and SCAM models for gross motor domain.

Type if item	Item(s) label	25th % probability		90th % probability	
		GLM	SCAM	GLM	SCAM
Ideal items for infants (< 1 year old)	GM1 - Lifts chin up off floor for a few secs.	[-0.02, 0.03]	[0.00, 0.02]	[0.12,0.16]	[0.12, 0.17]
	GM2 - Prone, head up to 90 degrees	[0.08, 0.11]	[0.08, 0.11]	[0.22, 0.26]	[0.22, 0.27]
	GM3 - Holds head straight or erect for few secs.	[0.14, 0.17]	[0.14, 0.17]	[0.31, 0.36]	[0.30, 0.35]
	GM4 - Pulls to sit with no head lag	[0.10, 0.13]	[0.10, 0.14]	[0.25, 0.30]	[0.25, 0.28]
	GM5 - Lifts head, shoulders and chest when prone	[0.16, 0.19]	[0.16, 0.19]	[0.29, 0.34]	[0.29, 0.33]
	GM6 - Bears weight on legs	[0.21, 0.24]	[0.21, 0.24]	[0.34, 0.39]	[0.33, 0.38]
	GM7 - Sits with help	[0.19, 0.22]	[0.20, 0.23]	[0.33, 0.38]	[0.33, 0.38]
	GM8 - Rolls over from back to front	[0.22, 0.25]	[0.22, 0.25]	[0.37, 0.43]	[0.37, 0.42]
	GM9 - Sits without support for a period of time	[0.30, 0.33]	[0.31, 0.33]	[0.42, 0.48]	[0.43, 0.48]
	GM10 - Sits by self well	[0.33, 0.38]	[0.35, 0.40]	[0.54, 0.61]	[0.47, 0.51]
	GM11 - Crawls (in any way)	[0.49, 0.54]	[0.50, 0.55]	[0.77, 0.85]	[0.71, 0.83]
	GM12 - Pulls self to stand	[0.54, 0.60]	[0.55, 0.61]	[0.89, 0.99]	[0.85, 0.93]
Ideal items for toddlers (1 to <3.5 years old)	GM13 - Able to stand if holding on to things	[0.57, 0.64]	[0.59, 0.65]	[0.93, 1.04]	[0.89, 0.99]
	GM14 - Walks using both hands of somebody	[0.66, 0.74]	[0.63, 0.72]	[1.12, 1.25]	[1.09, 1.27]
	GM15 - Walks with help - hand or furniture	[0.75, 0.82]	[0.71, 0.82]	[1.20, 1.33]	[1.66, 1.33]
	GM16 - Walks but falls over at times	[0.93, 1.01]	[0.92, 1.01]	[1.34, 1.47]	[1.31, 1.48]
	GM17 - Stoops and recovers	[0.99, 1.08]	[0.98, 1.07]	[1.46, 1.60]	[1.44, 1.59]
	GM18 - Walks well	[1.04, 1.13]	[1.01, 1.13]	[1.50, 1.65]	[1.47, 1.59]
	GM19 - Runs	[1.09, 1.21]	[1.11, 1.21]	[1.96, 2.19]	[1.63, 1.99]
	GM20 - Kicks a ball in any way/tries to kick ball	[1.55, 1.71]	[1.32, 1.45]	[2.53, 2.82]	[2.08, 2.88]
	GM21 - Runs well (confidently) stopping and starting without falling	[1.35, 1.48]	[1.41, 1.60]	[2.16, 2.40]	[2.61, 3.01]
Ideal items for pre- School age children (3.5 to < 7 years old)	GM22 - Kneels and gets up without using hands	[2.24, 2.47]	[2.00, 2.38]	[3.56, 3.93]	[3.51, 4.07]
	GM23 - Throws a ball into a basket (at least one of 3 times) 1 m away	[2.26, 2.49]	[2.10, 2.41]	[3.53, 3.89]	[3.60, 4.32]
	GM24 - Runs, stops and is able to kick a ball some distance	[2.10, 2.31]	[1.95, 2.21]	[3.40, 3.76]	[3.47, 4.15]
	GM25 - Jumps with feet together off ground	[2.34, 2.57]	[2.09, 2.48]	[3.57, 3.92]	[3.54, 4.06]
	GM26 - Jumps over line/string on the ground	[2.44, 2.67]	[2.28, 2.62]	[3.69, 4.05]	[3.65, 4.21]
	GM27 - Stands on 1 foot for < 5 seconds	[2.77, 3.02]	[2.66, 2.95]	[4.09, 4.47]	[4.12, 4.61]
	GM28 - Walks on heels 6+steps	[2.86, 3.12]	[2.72, 3.02]	[4.33, 4.73]	[4.35, 4.96]
	GM29 - Jumps over piece of paper	[2.50, 2.73]	[2.36, 2.72]	[3.71, 4.07]	[3.72, 4.27]
	GM30 - Walks on tip toes 6+steps	[3.16, 3.45]	[2.98, 3.45]	[4.82, 5.27]	[4.78, 5.66]
	GM31 - Hops on one foot 4 steps	[2.78, 3.05]	[2.49, 2.92]	[4.32, 4.73]	[4.29, 4.86]
	GM32 - Stands on 1 foot for a longer time	[3.00, 3.28]	[2.69, 3.21]	[4.66, 5.10]	[4.75, 5.60]
	GM33 - Can throw ball in air and catch it with 2 hands	[3.95, 4.31]	[3.65, 4.13]	[6.18, 6.84]	[5.52, 6.70]
	GM34 - Heel/toe walk precise one foot behind other along chalk line	[4.05, 4.40]	[3.84, 4.36]	[6.17, 6.80]	[5.90, 7.03]

*Confidence interval values given to the nearest 2 decimal places.
The shaded grey box shows an example of an unrealistic confidence band for item 1 using the GLM model.
Highlighted in the red dotted box shows the GLM and SCAM modelling approaches gave fairly comparable values especially for items where there is an even distribution of item pass probabilities typically found in items that are ideal for toddlers.

Data issues

There were several children recorded to have a decimal age value of 0. Technically, a child cannot have a 'zero' value for age and therefore this caused computational issues in the item by item analysis. At this stage these children were omitted from the analysis. A remedial measure could be to add a 'small' age value, e.g. corresponding to half a day, to only these child cases or all cases in an effort to utilise their respective item responses.

Monotonicity

One of the most important underlying assumptions with regard to measuring child ability is that as age increases, it is expected that the item pass probability that reflects ability should increase. We have seen that in an attempt to fit a model to the data structure, by using a more flexible method like the classical GAM, monotonicity is not always assured. The SCAM framework offers similar fit flexibility features to the classical GAM but assures that the fit is monotonically increasing with respect to age. This in turn solves the issue of multiple solutions for age estimates at pass probabilities of interest. Adherence to monotonicity of estimates was checked as is described in Chapter 5 and it was found that only the GLM and SCAM model frameworks assure this important assumption in this child development context is always adhered to.

Computation or Software Package Problems

Items designed for infants or older pre-school age children often had convergence problems especially in the SCAM modelling approaches. The classical GLM is some instances produced unrealistic confidence intervals especially when there was no differentiation of pass/fail rates i.e. where almost all children either passed or failed the item. This is because items designed for infants will have most of the children over the age spectrum of interest in the standardisation normal sample passing them. Even if the GLM procedure may still converge or give model estimates, we note that these estimates are not correct. For example, as is highlighted in grey in Table 6.4, the GLM model resulted in a

negative lower age estimate confidence bound at the 25th percentile for the item 1. A way to prevent unrealistic estimates is to restrict the predicted age to be within the acceptable range of non-negative values.

The SCAM procedure did not run smoothly in these situations either; often failing or taking considerably too long to converge or giving unrealistic age estimates. Again a way to prevent the unrealistic age estimates is to restrict the age estimates not to go over a certain age value as is highlighted in red in Table 6.1. This problem is not present with more evenly distributed data with respect to pass patterns across the age spectrum. This was seen especially while trying to predict age estimates outside the learning data range and while bootstrapping to get confidence intervals. The current SCAM procedure cannot comfortably predict probabilities outside the sample data age range. Thus we could only safely predict probabilities of children whose ages were less than or equal to the oldest child in the MDAT data, which was 6.39 years. Given the above findings that may frustrate the computational process during model fitting we recommend the following remedial measures;

- For the SCAM model framework, one could;
    - Only use the relevant item data over a given age range that the item is designed for to get the needed age estimates at the probabilities of interest. For example, we know that item 1 in the GM domain is only relevant for very young infants who are ≤ 1 year old (see the x axis age range indicated by the red arrow in Figure 6.4); therefore we can fit the SCAM model only using the data for children that are less than 1 year as is shown in Figure 6.4 below.
    - Alternatively use a random sample of the item data to fit the required model.
- Adjust the model specification options like the type of spline used, number of knots or convergence tolerance level options used to fit the SCAM model
- Ensure that the competing GLM model is fitted using similar data and model specification options to facilitate fare comparison of each model's fit.

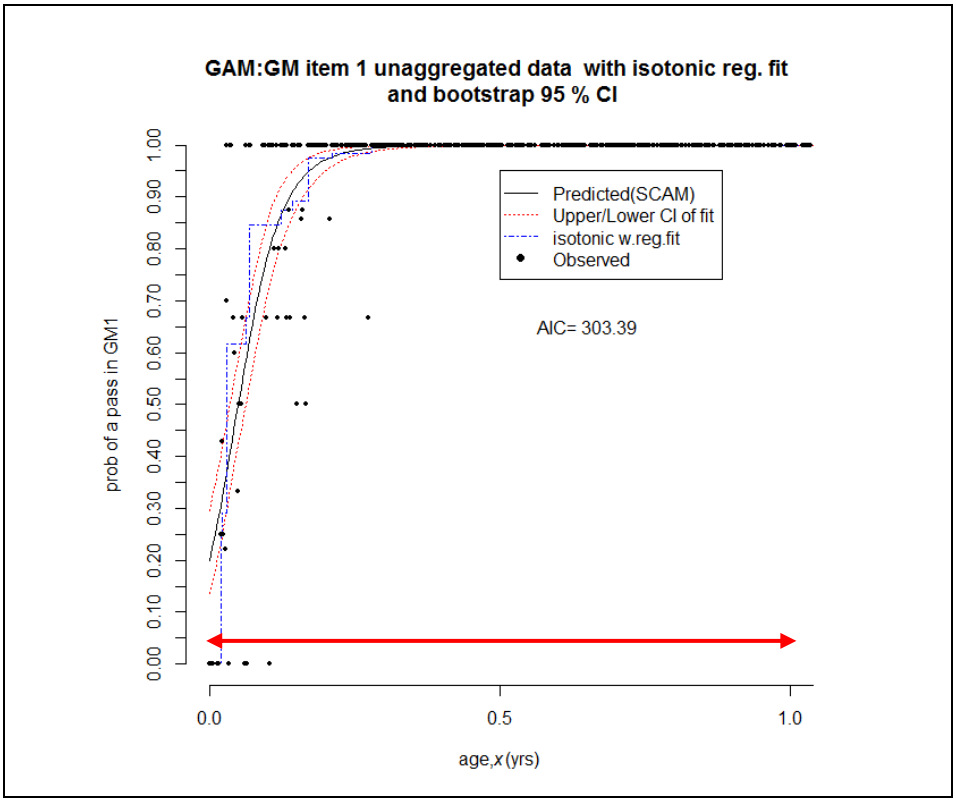


Figure 6.4: SCAM model fit using only relevant item data (age≤1 year) for gross motor domain item 1.

*Red arrow shows that only children who are ≤ 1 year were be used to fit the model.

Ordering of items on assessment chart

In order to appropriately assess child development using assessment charts such as those presented in Figure 6.3a) above, items are designed to be ordered to increase in difficulty. In turn, it is expected that as age increases, the probability of passing an item also increases if a child is developing normally. The probability to pass an item increases with respect to age implies an increase in ability. Therefore, items in an assessment tool should be ordered in ascending order with respect to age to reflect this. If the items are not correctly ordered, it may be difficult to assess a child's ability status as there may be 'easier' items that the child can pass and they are not administered as there are some preceding 'harder' items that are not appropriate for their age. The ordering of items in the work of Gladstone, et al., (2008; 2010a) used the 90% pass probability age estimate value of each item to order the items in each domain with respect to increasing age. This ordering was maintained for the charts presented in Figures 6.3a) to 6.3b).

However, the age estimate values shown in Table 6.1 are different for the two modelling approaches considered. This means that the ordering of items will change depending on the modelling approach used to compute the age estimates and which of the pass probability age estimates is used to order items. But, we also note that the ordering of items may also be influenced by the data used to fit the item by item models. Because the age estimates at the pass probabilities of interest should increase with age to enable the appropriate use of the charts to classify ability, we recommend that; i) Firstly, the primary reason for change in item order should be adequately motivated. ii) Secondly, the highest pass probability of interest should be used to order the items. iii) Thirdly, in the instance that there is a drastic change in item ordering, then expert opinion should be sought to confirm that the ordering of items is appropriate to assess child development.

For example, Figure 6.3a) shows an assessment chart that uses the age estimates from the GLM model using the MDAT 2007 data but keeping the Gladstone, et al., (2010a) order of these items. The ordering of items highlighted in the dashed red box should be altered as is confirmed by the grey shaded GLM age estimate values in Table 6.1. We see that if the 90% pass probability age estimate value from the SCAM model is used for ordering, the item order changes for the items 25, 26, 27, 29, 30 and 33.

Another issue related to ordering of items is the extent of overlap of age estimates between items ordered according to increasing difficulty. As was previously explained in the model assessment sub-section above it is clear that the MDAT items can be grouped into items ideal for infants, toddlers and pre-school aged children, as is shown by the red arrows in Figure 6.3a). As was highlighted in the literature review chapter Section 2.3.3.4 a), it is advisable to have some degree of overlap between age estimates of items, as if there is any gap it means that no item is appropriate to assess or characterise ability at that specific age.

6.2.2. Summary of item by item analysis

Our results have highlighted several key issues that are important in fitting item by item models to obtain development age estimates and we have gone on to show or suggest methods to circumvent various item modelling issues. In summary, instances where the item pass probability is not evenly distributed or where the log-odds of passing an item don't increase linearly with age will often result in the standard logistic regression model performing poorly. Once fundamental modelling assumptions are violated and pass probability rates are extreme, then more flexible and robust SCAM models should be used as these address most model requirements as discussed above.

Using the insight gathered from the individual item characteristics and more so how the two different modelling approaches characterise pass probabilities, the next section presents the results of the current overall scoring methods and attempts to devise suitable methods to compute an overall score for each child using all domain item responses while simultaneously accounting for age.

6.3. Overall scoring methods

As noted earlier, the overall scoring approach considers all items within a domain of an assessment tool simultaneously to give a single score to characterise a child's ability given age. The benefit of combining all item responses within each domain is to give one single score thus allowing one to get a holistic 'picture' of the child's ability status. As was explained in Section 5.3 and summarised in Table 5.1, we will consider three scoring approaches that are currently in use.

Again, the main objective here is to assess if the scoring of items using all items may be more accurately and appropriately assessed using more flexible and robust statistical methods and adjust for age. Some of the overall scoring methods are a combination of item by item analysis methods. Therefore, using lessons learned from the item by item analysis in Section **Error! Reference source not found.** above, we, for example, considered the Generalized Additive Model framework which was seen to be more flexible and robust to extend some of the current overall scoring approaches.

Presentation format of overall scoring methods

Before delving into the results and discussion reporting format that we will adopt for the overall scoring approach, there are a few important facts to consider; unlike item by item analysis the output of the overall scoring approach are overall summary scores at ages of interest. At the outset, we would like to note that a particular scoring approach may not address or only partially address certain scoring issues of interest. For example, it is not expected that any of the simple sum scoring methods will adjust for age, and thus they are expected to show a systemic strong association with age. The suggested extensions will at the very least address correcting for age and maybe a few other issues that improve their quality and use. Therefore, the discussion scope and extent will change depending on the number of issues that our developed extensions purport to address.

Further, between the different overall scoring realms, the score scales differ, and therefore the interpretation and use in turn will differ. For ease of reporting results and comparison of the various methods as was explained in the comparison Section 5.3, we will also group the results of scoring methods into three scoring families of; simple score methods, Z-score methods and Item Response Theory methods. Therefore, the discussion and comparison points of the results will be kept within each method. However, the main outlook of this project is to give the best approach that is robust and simultaneously able to detect a child's ability development delay or disability.

In a similar fashion as was used in reporting the item by item analysis we highlight general issues that the various overall scoring approaches raise at the outset. Then, using various statistical summaries and plots that will be described, discuss the issues raised and more so how our developed extensions remedy them under the following sub headings; score distribution, model fit, confidence interval, data issues, outliers, monotonicity and computation problems. Given the pros and cons of the current scoring methods, we will present more specific questions within each family that the suggested scoring method hope to address.

The specific research questions we intend to address under the overall scoring approaches are;

a) What is the distribution of scores produced from the various methods and what is the distribution of scores produced from the various methods with respect to age?

   These two aspects are important in devising the best approach to use for determining clinical suitable and sensitive threshold cut-off values to classify child development status.

b) Does the scoring method have a framework to adequately correct or adjust for age and other covariates? And does the scoring approach have a framework to create a confidence interval around score estimates?

c) Does the correlation of scores change with respect to scoring method and age?

d) Does the sensitivity of scores change between methods with respect to cut-offs and important covariates like age?

e) Which method is able to adequately deal with potentially hard to classify observations with no obvious disability or delayed development using the malnourished sample data?

f) Which of the scoring approaches is able to account for;

   i. Differences in the number of items administered to children

   ii. The underlying correlation within and between a child's responses

   iii. Allowing the difficulty across items to change

Each scoring method's ability to address the above specific questions will be assessed under the broader themes of; score distribution and quality, score distribution with respect to age, correlation agreement of scores with age and score sensitivity that will be compared and discussed within each of the methods' subsections. Table 6.4 summarises the distributional characteristics of the scores from the different scoring methods showing the scores' mean, standard deviation (SD), score range, correlation of score with age and sample size for each scoring method for the gross motor domain in the normal, disabled and malnourished sample groups of the MDAT 2007 data. It gives us a glimpse of the score characteristics of the classical methods against the developed extensions in this thesis that are discussed in more detail in subsequent sections.

We note that of main concern are the characteristics of distribution of scores mainly in the normal sample. The assessments of the score characteristics in the disabled and malnourished samples are mainly to validate the scoring methods, test the performance of scoring methods and inform us on other pertinent issues such as the appropriate threshold cut-off values to use to correctly classify child development status. To complement the summary statistics presented in Table 6.4, we have plotted density plots and scatter plots that give further insight into score characteristics and catalyse the discussion of the distribution of scores within each scoring method family to further assess their quality and performance. The criterion validity explained in Section **Error! Reference source not found.** is also assessed in each of the samples in the form of a sensitivity analysis where; a) the rate of false positives (FP) in the normal sample should be relatively low (about 5%) b) the rate of true positives (TP) in both the disabled and malnourished samples should be relatively high (> 50%). We anticipate that the true positive detection rate should be higher in the disabled sample than in the malnourished sample. This is because being malnourished does not necessarily imply disability.

As explained in Section **Error! Reference source not found.**, a scatter plot of the score values on the y axis against child age on the x axis will be plotted for each method with a loess curve overlaid. Apart from checking if there is any systematic pattern of scores with respect to age, the scatter plot also points to other important issues such as the child cases that are potentially hard to classify due to their borderline scores with respect to selected threshold cut-offs. Both the density plots and scatter plots allow us to assess the sensitivity of the scores to distinguish normal from disabled or malnourished children. A summary of the various scoring methods sensitivity performance is given in Table 6.5 which gives the ROCAUC for each scoring method to facilitate comparison, and also assesses the impact of age on sensitivity and subsequently prompts the finding of an optimal threshold or cut-off point that achieves the highest sensitivity.

Table 6.4: Score distribution summary statistics of the classical and extended overall scoring methods in the MDAT GM domain for the 3 sample data sets.

Summary Statistics	Simple Score Methods	Z-Score Methods		Item Response Theory Methods		
	GAMLSS Regression	Classical Z-Score	Smoothed Z-Score	Classical 1 PL	Classical 2 PL	1 PL Spline
Normal Data						
Mean	25.54	~0.00	-0.01	~0.00	~0.00	-0.06
†SD	8.61	0.99	1.00	1.20	1.21	0.82
Shapiro Test, W (*p*-value)	0.85 (<0.001)	0.97 (<0.001)	0.97 (<0.001)	0.91 (<0.001)	0.91 (<0.001)	0.98 (<0.001)
Minimum	1.39	-5.94	-7.25	-2.30	-2.25	-6.51
Maximum	33. 72	4.01	4.44	1.43	1.46	2.96
False Positives (FP %)	37 (2.56%)	28 (1.94%)	39 (2.70%)	47 (3.25%)	42 (2.91%)	37 (2.56%)
¥Corr. age *(p-value)*	0.21 (<0.001)	0.08 (0.002)	0.05 (0.04)	0.97 (<0.001)	0.97 (<0.001)	0.02 (<0.001)
n	1444	1444	1444	1444	1444	1444
Disabled Data						
Mean	8.88	-6.86	-5.75	-0.84	-0.84	-5.01
†SD	2.02	4.92	2.96	0.96	0.95	1.72
Shapiro Test, W (*p*-value)	0.95 (0.002)	0.83 (<0.001)	0.93 (<0.001)	0.96 (0.02)	0.96 (0.01)	0.92 (<0.001)
Minimum	4.66	-24.98	-15.38	-2.30	-2.25	-7.91
Maximum	11.87	-0.58	-0.56	1.43	1.46	-0.62
Sensitivity (TP %)	80 (100.00%)	76 (95.00%)	74 (92.50%)	76 (95.00%)	76 (95.00%)	75 (93.75%)
¥Corr. age *(p-value)*	-0.03 (0.79)	-0.19 (0.09)	-0.33 (0.003)	0.21 (0.06)	0.22 (0.05)	-0.01 (0.93)
n	80	80	80	80	80	80
Malnourished Data						
Mean	18.55	-1.95	-1.48	0.20	0.17	-1.66
†SD	3.80	2.48	1.55	0.67	0.67	1.53
Shapiro Test, W (*p*-value)	0.81 (<0.001)	0.80 (<0.001)	0.88 (<0.001)	0.93 (<0.001)	0.95 (~0.001)	0.88 (<0.001)
Minimum	7.50	-12.12	-8.90	-2.30	-2.25	-8.57
Maximum	21.83	2.32	0.91	1.43	1.46	0.57
Sensitivity (TP %)	80 (66.67%)	42 (35.00%)	36 (30.00%)	46 (38.33%)	40 (33.33%)	75 (62.50%)
¥Corr. age *(p-value)*	-0.01 (0.88)	0.03 (0.74)	-0.07 (0.43)	0.66 (<0.001)	0.66 (<0.001)	-0.13 (0.16)
n	120	120	120	120	120	120

Values are reported to nearest 2 decimal places, Values in brackets are the respective *p*-values for the Shapiro-Wilks test of normality and the Spearman's correlation coefficient, †SD-Standard Deviation, ¥Spearman correlation of scores with age in normal sample, TP-True positive (Proportion (%) of disabled children in sample), Highlighted in grey are the preferred overall scoring extensions developed in this thesis, Highlighted in red dotted box are the scoring methods that adequately correct for age.

Table 6.5: Sensitivity and optimal score cut-off performance summary of the classical versus extended overall scoring methods in the MDAT GM domain.

Overall Scoring Approach	Method	MDAT Sample Data	ROC AUC			
			Overall (0-<7 years)	Infants (<1 year)	Toddlers (1-3.5 years)	Pre-sch. age (3.5-<7 years)
Simple Sum Count	GAMLSS Regression Score	Normal vs Disabled	0.93	0.92	0.93	0.93
		Normal vs Malnourished	0.77	0.75	0.77	0.77
Z-Score	Classical Z-Score	Normal vs Disabled	0.98	0.94	0.98	0.99
		Normal vs Malnourished	0.81	0.82	0.81	0.88
	Smoothed Z-Score	Normal vs Disabled	0.98	0.95	0.98	0.99
		Normal vs Malnourished	0.80	0.82	0.79	0.90
Item Response Theory (IRT)	1 PL IRT	Normal vs Disabled	0.98	0.95	0.98	0.97
		Normal vs Malnourished	0.81	0.84	0.81	0.80
	2 PL IRT	Normal vs Disabled	0.98	0.94	0.98	0.97
		Normal vs Malnourished	0.80	0.81	0.81	0.77
	1 PL Spline IRT	Normal vs Disabled	0.98	0.96	0.98	0.97
		Normal vs Malnourished	0.83	0.85	0.84	0.77

Highlighted in grey are the overall scoring extensions developed in this thesis, Highlighted in red dotted box are the scoring methods that were found to be highly sensitive.

6.3.1. Simple Sum Count Methods

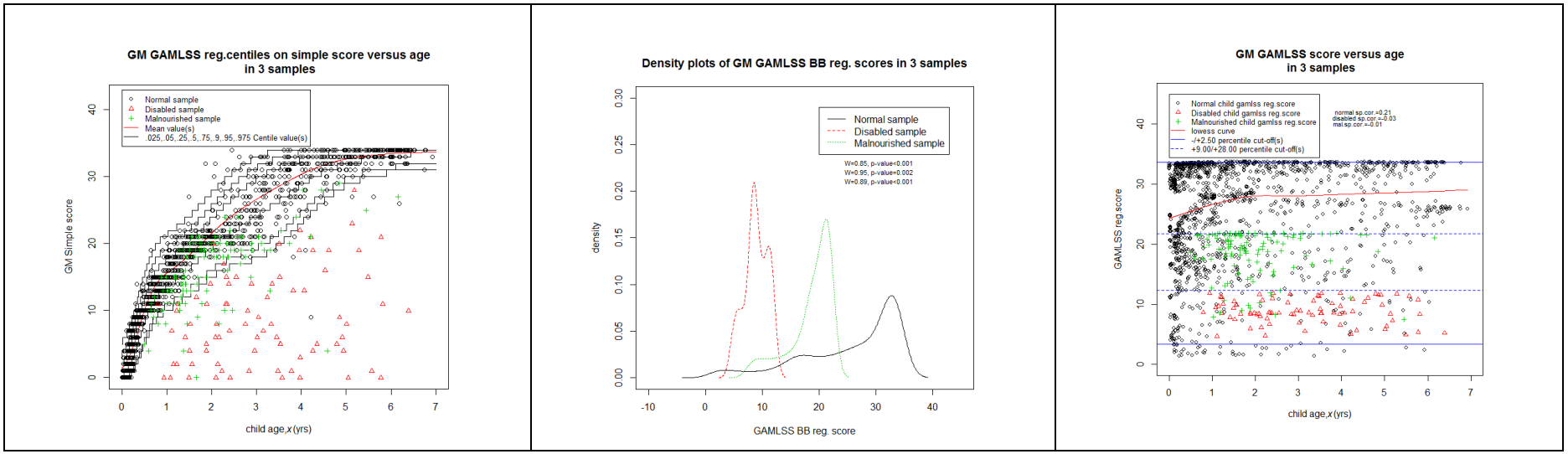
Under the simple sum count scoring approaches, we remedy the identified weaknesses by; a) first weighting the simple score by the actual number of administered items and then b) utilising the flexibility of the GAMLSS regression model described in Section 5.2.2.1 to correct for the effect of age. The following three sections describe the simple count methods in terms of score characteristics, correlation with age and sensitivity. While bringing out the weakness of the simple count method, we highlight the benefit of the suggested extensions.

6.3.1.1. Comparison score characteristics of Simple Scoring Methods

GAMLSS model scores

Section 5.2.2.1 outlined the process by which a GAMLSS Beta Binomial model is fitted on the number of passes out of the total administered MDAT items per child while adjusting for the non-linear relationship between the number of passes and age using a more flexible monotonic B-spline function to obtain expected scores at ages of interest. The second column on Table 6.4 shows the summary statistics of this scoring method that suggest a highly skewed distribution especially for the normal and malnourished samples. The skewness is confirmed by the density plots shown in second panel of Figure 6.5 with the disabled sample having a multimodal distribution. This could be an indication that there may be different groups in this sample depending on the extent of the formally diagnosed neuro-disability. A formal test of adherence to normality shows that the score distributions from all the three data samples are not normally distributed.

The third panel of Figure 6.5 shows a scatter plot of GAMLSS (50^{th} percentile) score values and age. We see that the previously strong non-linear association of the simple score with age is absent. The systemic pattern of scores and increase in sparsity as age increases no longer exists in the GAMLSS scores. These findings are confirmed by a weak correlation coefficient of 0.21 and a horizontal loess smooth curve (red line) overlaid on the scatter plot.



W is the Shapiro-Wilk test for Normal, Disabled and Malnourished scores

Figure 6.5: Centile plots on simple scores used as cut-off thresholds, Score density plots and Scatter plots for GAMLSS BB model scores in GM domain.

The following points can be drawn from both the model based GAMLSS regression scoring approach;

- These method does not directly adjust for age of the child in the score computation. Instead while still using the naïve simple score that is strongly associated with age it attempts to correct for age using the GAMLSS model frameworks. Therefore, even if this extension is able to mitigate the effect of age on scores, it is possible that some remnants of weaknesses of the simple score still carry over to these model based scores. These weaknesses are;
    - Because the response of the GAMLSS BB regression is a naïve sum of passed items that assumes each item has similar difficulty it still regards each item to have similar difficulty.
    - Further, both the GAMLSS BB regression model does not allow the investigation of dependence between items.
- However, this model based method offer a framework to get appropriate scores at percentiles of interest while accounting for the non-linear relationship between the total score and age. While doing so we see that;
    - If the simple scoring method is used for assessment, then the GAMLSS BB model framework can be used to create suitable age adjusted cut-off thresholds. These are shown in the first panel of the Figure 6.5.
    - We also notice that while the model based scores that have been adjusted for age, thus allow the use of one overall cut-off or threshold (shown in blue continuous line) to classify development status, owing to the skewed nature of the score distributions and subsequent overlap of scores seen in the density plots in the second panels of Figure 6.5, an alternative threshold should be sought to achieve high sensitivity. These alternative optimal thresholds shown in blue dotted lines in the third panel of Figure 6.5 prove to be more sensitive in differentiating between normal versus disabled and normal versus malnourished children.

To be able to use the model based approach to classify child development status using the simple score, one could plot percentile values across the age spectrum as shown in the Figure 6.6. The figure

shows a percentile curve reference chart using the GAMLSS BB regression for 0.025, 0.05, 0.25, 0.50, 0.75, 0.90, 0.95 and 0.975. Shown in the blue dotted lines are the expected overall sum score counts at the eight percentiles of interest that can be used for a child's development status given their overall raw sum simple score. For example, if we consider a two year old child who had an overall raw simple score of 25 using the MDAT tool as shown in the figure by the red dotted line; according to the GAMLSS regression percentile curve reference chart this child's gross motor development status would be classified to be on the 0.90 percentile.

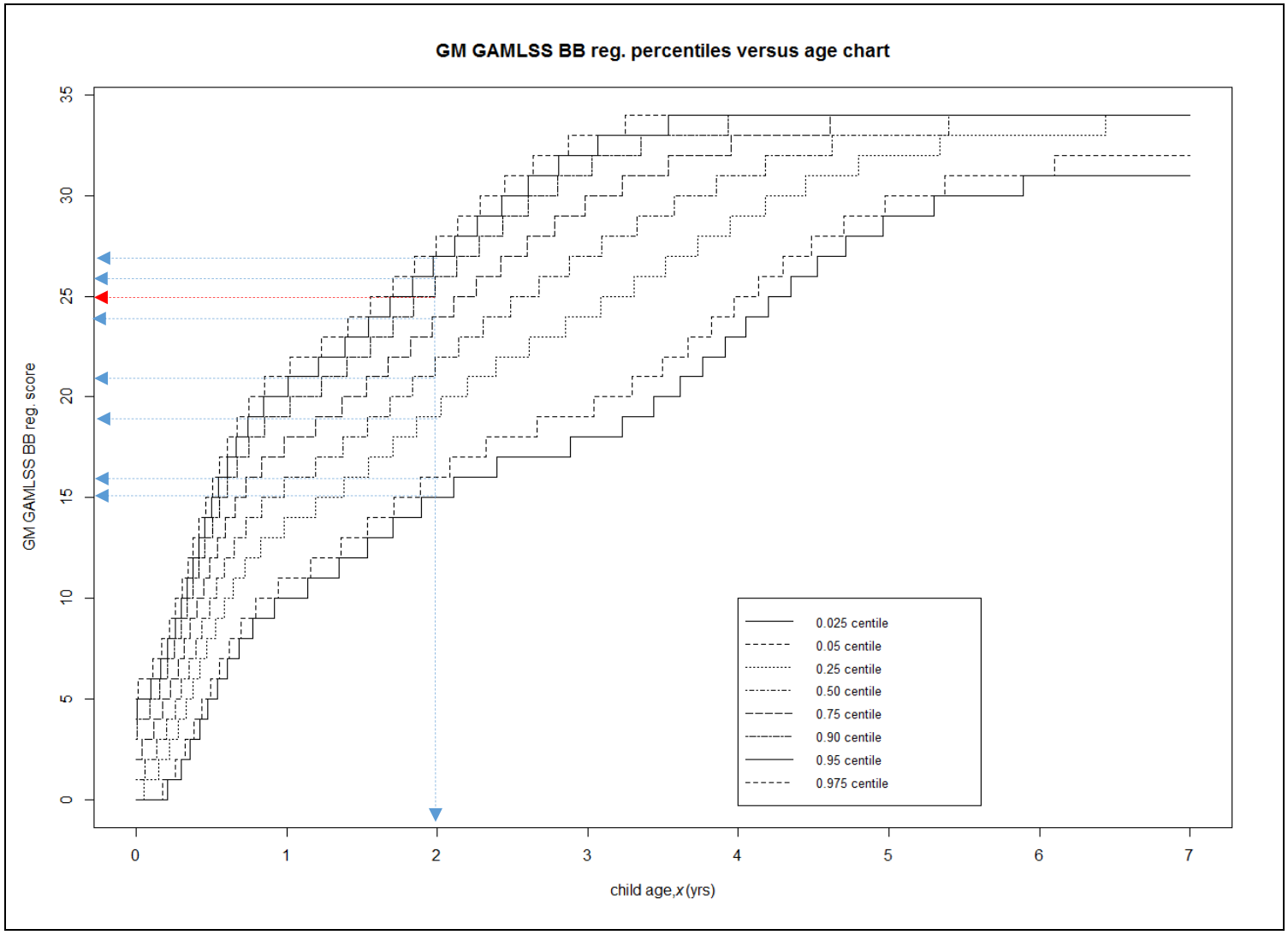


Figure 6.6: Percentile curves reference charts using GAMLSS regression for 0.025, 0.05, 0.25, 0.50, 0.75, 0.90 and 0.95.
Shown in the blue dotted lines are the expected overall sum score counts at the 8 percentiles of interest (the red dotted line shows 0.50 the expected overall sum score count).

6.3.1.2. Sensitivity of simple scoring

Ultimately the 'best' scoring approach is the one that while being 'simple' to implement or compute, is able to accurately detect or identify delayed or disabled children. To facilitate sensitivity comparison across the simple scoring methods described in Section **Error! Reference source not found.** were used.

The summary measures of False Positive rate in the normal sample that should be low and True Positive rate in the non-normal samples that should be high give a snap shot of the scoring methods ability to correctly classify different groups of children. Table 6.5 has summarised the performance of the various scoring methods and compared sensitivity in terms of; firstly, their sensitivity between normal versus disabled samples, and normal versus malnourished samples. Secondly, given our interest in correcting for age that has been shown to be strongly associated with simple score counts, also assess if age also influences the sensitivity of the score.

At the outset we note that given the degree of overlap seen by the density plot and scatter plots of the model based scores using simple scores in Figure 6.5 we expect that sensitivity will be higher when attempting to classify the normal versus disabled samples compared to classifying the normal versus malnourished samples. This expectation is confirmed by the ROCAUC values in Table 6.5 that are consistently higher for the normal versus disabled samples in comparison to the normal versus malnourished samples.

Finer detail with regard to sensitivity performance is shown by the ROC curves for each of the three methods in Figures 6.7 below. We would like to highlight the benefit of the model based scoring approaches' cut-off values to classify the naïve weighted scores. Obviously if one specific threshold that does not account for the effect of age is used as a cut-off for this method, the sensitivity will be very poor owing to the strong association of the scores with age. However, as seen from the respective ROC curves below, using the model based thresholds as cut-off is more suitable if the simple or weighted score is used as the scoring methods. Further, although we can now use one cut-off to distinguish the normal versus disabled or malnourished cases as the effect of age has been eliminated,

due to the skewness of the model based scores shown in the density plots in Figure 6.5 the choice of threshold needs to be adjusted slightly. As is shown in the scatter plot in the third panel of Figure 6.5, it is not until the cut-off threshold reaches at least the 10^{th} percentile that one is able to detect any disabled cases.

Effects of age on sensitivity

To evaluate the effect of age on sensitivity, we grouped our sample data into the three age groups of; infants who are less than 1 year, toddlers who are 1 to less than 3.5 years and pre-school age children aged 3.5 to less than 7 years, and then carried out the sensitivity analysis described in Section **Error! Reference source not found.** for each age group. The respective ROC AUC summary values that are used to make comparisons are also summarised in Table 6.5.

For the simple score method, it seems that this scoring approach seems more sensitive for the toddler and pre-school age children that both report the same slightly higher ROCAUC value of 0.93 for the normal and disabled children. However, as was highlighted in our literature review on some of the pitfalls of using the ROCAUC, we notice that the respective ROC curves for these age groups shown in Figures 6.7 cross at certain cut-off values. Therefore, merely having a lower ROCAUC value for a certain age group does not necessarily mean that the scoring method is poorer for this age group. It may just mean that the scoring method is not appropriate for children within a specific age spectrum. For example, in Table 6.5 we see that for normal versus disabled samples the ROCAUC for pre-school age children is slightly higher at 0.93 as opposed to the ROC AUC value for infants which is 0.92. However, in Figure 6.7 the ROC curve between the normal and disabled samples for infants and pre-school age children cross at several cut-off values. Therefore, for certain cut-off values the sensitivity performance of the simple scoring approach is higher for pre-school age than infants and in other cases it is not.

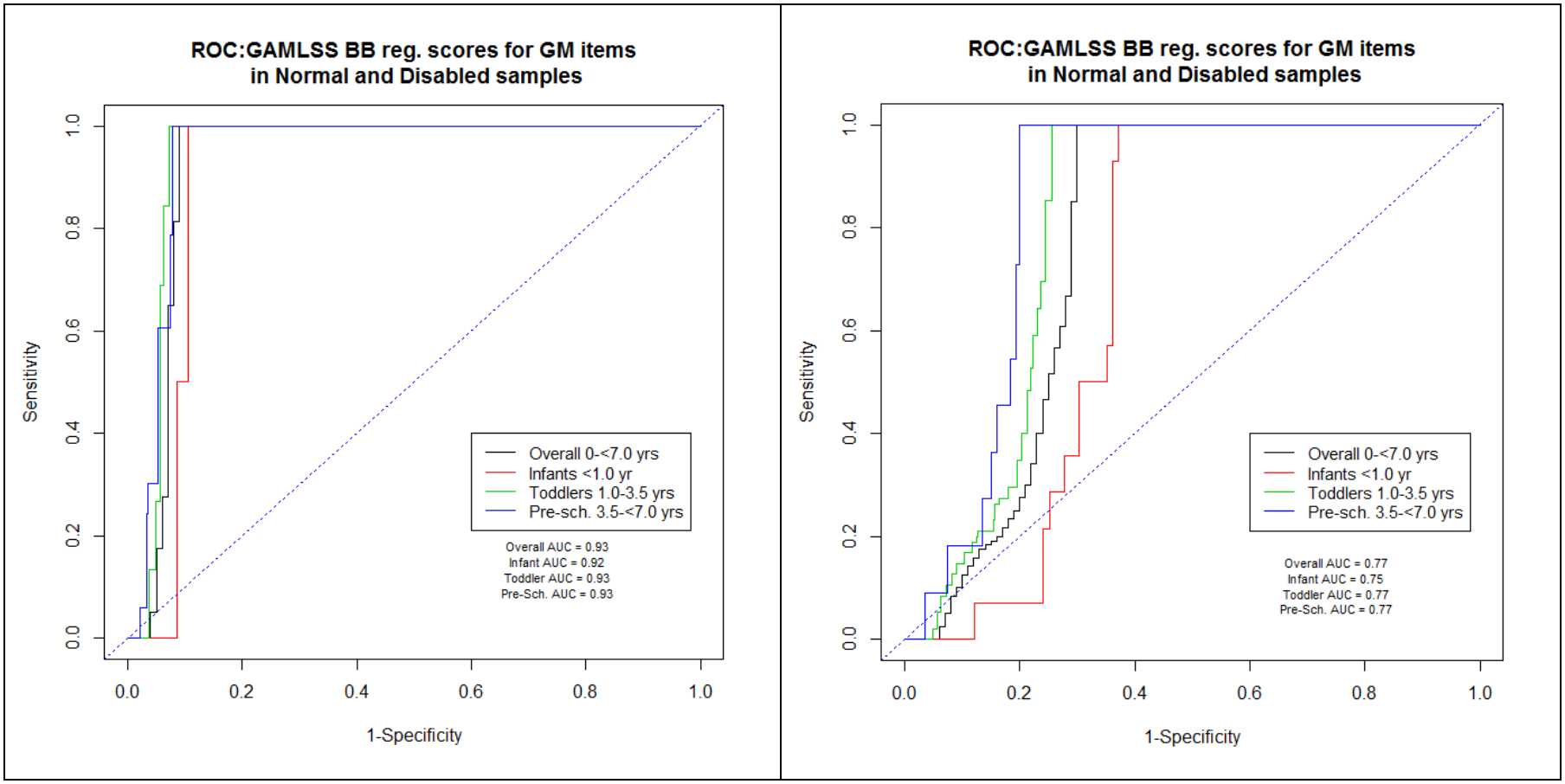


Figure 6.7: ROC curves: GAMLSS regression score method in GM domain.

6.3.1.3. Summary of simple scoring approach methods

Under the simple scoring framework, we have suggested weighting to ensure that there is a fair comparison of scores from two children of a similar age who have a different number of items administered. We then used a GAMLSS regression model to facilitate the correction of age on the naïve or the weighted scores. In their paper titled 'Discussion: A comparison of GAMLSS with quantile regression' Rigby, et al., (2013) reiterate the advantages and disadvantages that are showcased by both the GAMLSS modelling approach. In this child development context research we use these model based methods for 2 purposes; a) firstly, at the naïve simple raw score level to combine item responses and provide a framework to correct for the effect of age that has been shown to be strongly associated with the simple score counts. This correction of age allows the use of one cut-off threshold to classify disability or development delay but also facilitates comparison of different aged children; b) secondly, as a mechanism to create suitable confidence bands around scores that takes into account the effect of age and the observed increase in sparsity of the simple or weighted scores as age increases. The second purpose of using the model based methods also doubles up to provide a suitable way to create more appropriate age adjusted threshold values that can be used to classify development status if the researcher still insists on using the naïve simple scores.

6.3.2. Z-score methods

Section 5.2.2.2 explained how the Z-score approach that either internally or externally standardise the simple sum score of each child using the mean and standard deviation of the score of the age category that each case belongs to. This section now pits the classical Z-score approach against our suggested extension of using smoothed mean and standard deviation values to compute more reliable Z-scores that we call the smoothed Z-scores.

6.3.2.1. Comparison score characteristics of Z-Scoring Methods

Empirical Z-score

The summary distribution characteristics in column three in Table 6.4 suggest that the classical Z-scores in the normal sample are reasonably normally distributed. However, the density plots in Figures 6.8 below suggest that the disabled sample Z-scores are skewed to the left and the normal sample Z-scores have a bimodal distribution. The normality tests for the respective three samples confirm that the score distributions are not normal. The scatter plot of Z-scores against age with the loess curve overlaid in the third panel of Figure 6.8 shows that the Z-score approach is able to successfully correct for the previously observed effect of age in the simple score method. This is also confirmed by the weak correlation coefficient of 0.08 in the normal sample.

However, a closer look at the means of the age groups of interest from the empirical Z-score method summarised in Table 6.6 and illustrated in the first panel of Figure 6.8 reveals that they do not adhere to the assumption of monotonicity. The importance of the means used to standardise the simple score and compute the Z-score to be monotonically increasing was explained in Section 5.2.2 was not adhered to. See the red arrows in first panel of Figure 6.8. The lack of adherence of these age category means to monotonicity implies that it is possible that computing a Z-score using the unsmoothed means and standard deviations would be invalid as they would over or under estimate the Z-score. This weakness in the classical Z-score approach is our main motivation for recommending a smoothed

Z-score approach. For example, consider the child who is 0.54 years old and has a weighted simple score of 10.30 in the gross motor domain. The second and third columns of Table 6.6 contain means and standard deviations (SD) of the weighted simple scores for 'normal' children in the MDAT Gross Motor domain by age. As the child is 0.54 years, the mean and standard deviation values needed to standardise this child's weighted simple score are given in the 6 to 6.9 month row highlighted in the red dotted box of Table 6.6. As outlined in formula 5.35 the classical Z-score for this child for the gross motor domain can be calculated as follows;

$$\text{Empirical Classical Z-Score} = (10.30 - 9.93)/ 0.86$$

where the child's weighted score (10.30) is the number of items that this child passed (10) within the gross motor domain divided by the number of items administered within the domain (33), multiplied by the total number of items in the domain (34). The mean (9.93) and standard deviation (0.86) were calculated using the raw weighted simple counts of children who are 6 to 6.9 months of age as the actual age of the child is 0.54 years or 6.48 months. This child therefore has an unsmoothed empirical Z-score value of 0.43.

Smoothed Z-score using GAMLSS model

The distribution summary characteristics in the fourth column of Table 6.4 seem to suggest that the smoothed Z-scores in the normal sample are somewhat normally distributed. The density plots in the fifth panel of Figure 6.8 below show a slight bi-modal peak in the normal sample and that the previously observed skewness especially in the disabled sample Z-scores has been reduced. The formal normality tests for the respective three samples confirm that the score distributions are not normal. The scatter plot of the smoothed Z-scores against age with the loess curve overlaid in the sixth panel of Figure 6.8 shows that the smoothed Z-score approach is also able to successfully correct for the previously observed effect of age in the simple score method. This is also confirmed by the weak correlation coefficient of 0.05 and horizontal loess smoothed curve in the normal sample.

We have seen that the mean and standard deviation from the empirical score does not always adhere to the assumption of monotonicity which the smoothing approach attempts to rectify as shown in the third panel of Figure 6.8 below. Using a higher or lower mean or standard deviation to compute a Z-score has the potential of mis-classifying the development status of a child by either inflating or deflating the actual Z-score. Table 6.6 below shows the actual means and standard deviations (shaded in grey) that are used in the computation of the smoothed Z-scores are now monotonically increasing. However, beyond the benefits accorded by the Z-score approach the main purpose of smoothing the means and standard deviations used to compute Z-scores is to ensure that these summaries adhere to the monotonicity assumption and therefore avert the mis-classification potential highlighted earlier. The sensitivity analysis further confirms the potential danger of mis-classification resulting from using inflated or deflated means and standard deviations for Z-score computation.

Again, we consider the same child who is 0.54 years old and has a weighted simple score of 10.30 in the gross motor domain. The highlighted fifth and sixth columns in grey of Table 6.6 contain the corresponding smoothed means and standard deviations (SD) of the weighted simple scores for 'normal' children in the MDAT Gross Motor domain by age. These summary measures were obtained from fitting a GAMLSS model on the weighted simple scores as explained in Section 5.2.2.2 b). As the child is 0.54 years, the means and standard deviations values needed to standardise this child's weighted simple sum score are given in the 6 to 6.9 month row highlighted in the red dotted box of Table 6.6. As outlined in formula 5.36, the smoothed Z-score for this child for the gross motor domain can be calculated as follows;

$$\text{Smoothed Z-Score} = (10.30 - 10.52)/\ 1.89$$

where the child's weighted simple sum score (10.30) is the number of items that this child passed (10) within in the gross motor domain divided by the number of items administered within the domain (33), multiplied by the total number of items in the domain (34). The smoothed mean (10.52) and standard deviation (1.89) were calculated by smoothing the raw weighted simple counts of children

who are 6 to 6.9 months of age as the actual age of the child is 0.54 years or 6.48 months. This child therefore has a smoothed Z-score value of -0.11.

The following are the key points drawn from the classical Z-scoring approach;

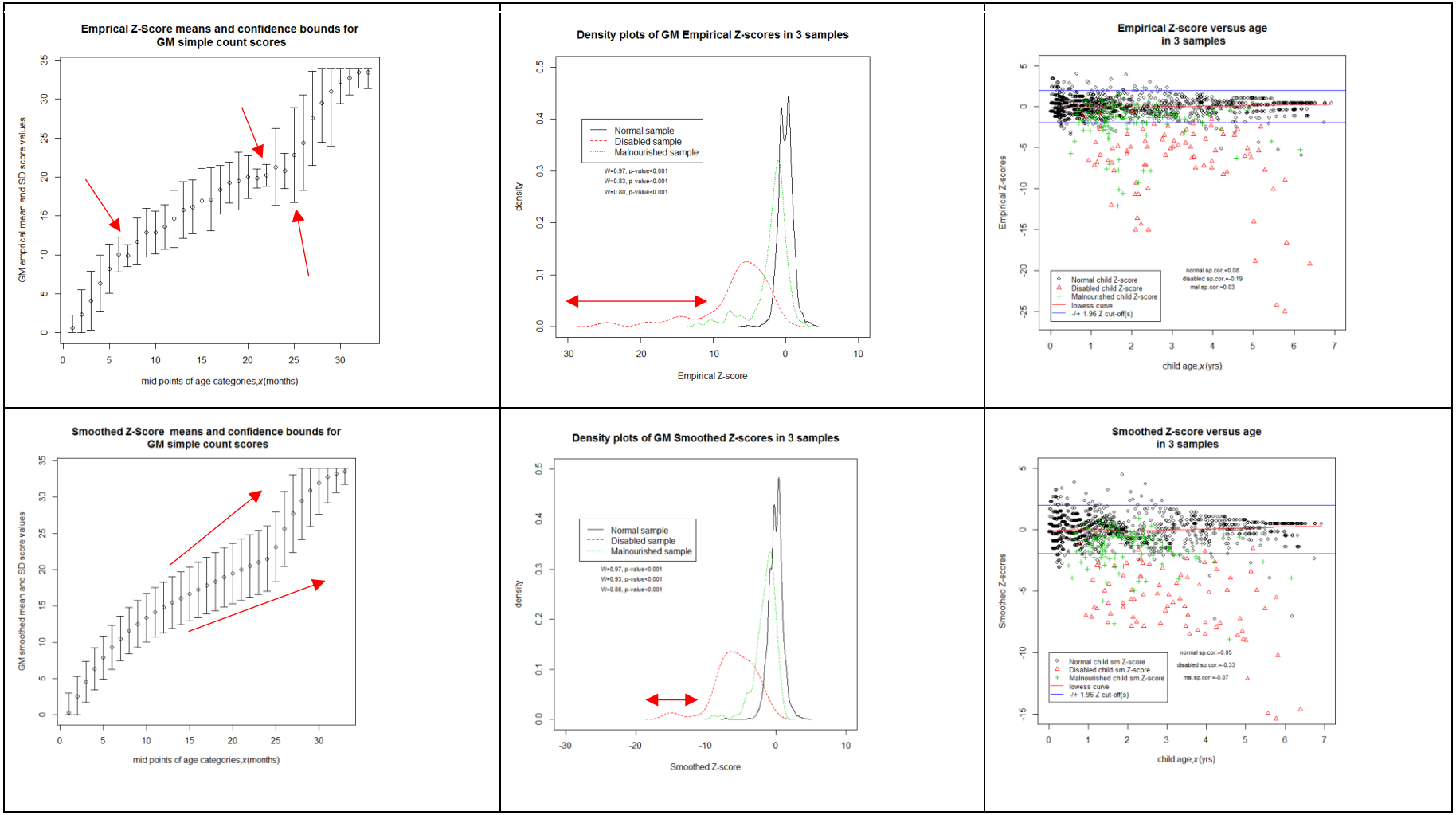
- As was seen in the previous model based approach using the GAMLSS BB regression on the weighted simple sum score, this method too does not directly take age of the child into account in the score computation. Instead while still using the total simple score it is able to successfully eliminate the previously strong association of naïve simple scores with age by standardising the overall raw scores by age. This allows us to be able to use one cut-off threshold for all ages. These are shown as blue lines in the scatter plots of Z-scores with age in Figure 6.8.
- Because the Z-score is a standardised form of the naïve sum of passed items;
    - The method still assumes that each item has similar difficulty.
    - Since the Z-score standardises the overall simple score, it does not allow the investigation of dependence between items.
- Table 6.8 shows total score summaries of mean and standard deviations for various age categories. Notice that while these can be used as references in typical research studies to evaluate development status of children, they are not very reliable if they are not monotonically increasing.
- The method also offers no direct framework for creating confidence bounds around the scores.

In addition to the points drawn from the classical Z-scoring approach, the smoothed Z-score approach;

- Ensures monotonicity avoiding the potential pitfall of mis-classifying the true developmental status of a child. Using an example, we have shown how the empirical classical Z-score could potentially classify a child as developing normally by reporting a slightly inflated Z-score of 0.43 yet their true Z-score is slightly lower, -0.11.
- The GAMLSS model offers a more realistic framework of creating confidence bands around Z-score estimates.

- As a result of using more realistic mean and variance values to standardise the simple score counts, the smoothed Z-score distribution is less skewed. This leads to a shorter smoothed Z-score range. Notice the range of unsmoothed Z-scores shown in Table 6.5 for the disabled and malnourished samples if far greater than those of the smoothed Z-score method.
- The GAMLSS model offers a suitable framework to account for other important variables beyond age that are strongly associated with the scores characterising the development of a child.

The means and standard deviations shown in the fifth and sixth columns of Table 6.8 that are highlighted in grey are instead used to compute smoothed Z-scores to classify development status of children. As the Z-score method does manage to adequately correct for the effect of age, the 2.5^{th} and 97.5^{th} percentile values of the smoothed Z-score can be used as cut-off thresholds for screening purposes to classify child development status. These thresholds are shown as blue lines in the Z-score scatter plots against age in Figure 6.8. The performance of these cut-off thresholds will be discussed in the sensitivity section of the Z-score methods.



W is the Shapiro-Wilk test for Normal, Disabled and Malnourished scores

Figure 6.8: Mean and SD summaries of classical and Smoothed Z-scores, Z-score density plots and Z-score scatter plots with age in GM domain in normal, disabled and malnourished data. Red arrows in 1st and 4th panel show change in variance of the mean due to smoothing. Red arrows in 2nd and 5th panel show skewness in score distributions (reduced skewness in smoothed Z-scores).

Table 6.6: Means, standard deviations and the 2.5th and 97.5th percentile values of the Z-score (confidence intervals) for use in creating both unsmoothed (classical) and smoothed Z-scores as well as ability classification for the GM domain of the MDAT tool.

Age category (months)	Unsmoothed			Smoothed		
	Mean (μ)	SD (σ)	95 % ŧC.I	Mean (μ)	SD (σ)	95 % ŧC.I
Less than 1	0.60	0.99	[0.00,2.23]	0.31	1.64	[0.00,3.01]
1 to 1.9	2.34	1.93	[0.00,5.51]	2.56	1.68	[0.00,5.32]
2 to 2.9	4.11	2.31	[0.31,7.91]	4.57	1.72	[1.74,7.39]
3 to 3.9	6.35	2.2	[2.73,9.97]	6.35	1.76	[3.46,9.24]
4 to 4.9	8.21	1.93	[5.04,11.38]	7.92	1.80	[4.96,10.88]
5 to 5.9	10.07	1.36	[7.83,12.31]	9.30	1.84	[6.27,12.33]
6 to 6.9	9.93	0.86	[8.52,11.34]	10.52	1.89	[7.41,13.62]
7 to 7.9	11.70	1.82	[8.71,14.69]	11.59	1.93	[8.41,14.77]
8 to 8.9	12.87	1.88	[9.78,15.96]	12.53	1.98	[9.28,15.79]
9 to 9.9	12.87	1.67	[10.12,15.62]	13.37	2.03	[10.04,16.70]
10 to 10.9	13.61	1.74	[10.75,16.47]	14.12	2.07	[10.71,17.53]
11 to 11.9	14.62	2.23	[10.95,18.29]	14.81	2.12	[11.31,18.30]
12 to 12.9	15.79	2.22	[12.14,19.44]	15.44	2.17	[11.87,19.02]
13 to 13.9	16.16	2.09	[12.72,19.60]	16.06	2.22	[12.40,19.72]
14 to 14.9	16.95	2.53	[12.79,21.11]	16.66	2.28	[12.92,20.40]
15 to 15.9	17.12	2.46	[13.07,21.17]	17.25	2.33	[13.42,21.08]
16 to 16.9	18.38	1.89	[15.27,21.49]	17.82	2.38	[13.91,21.74]
17 to 17.9	19.29	1.61	[16.64,21.94]	18.39	2.43	[14.39,22.39]
18 to 18.9	19.48	2.24	[15.80,23.16]	18.94	2.48	[14.85,23.03]
19 to 19.9	20.00	1.65	[17.29,22.71]	19.48	2.54	[15.31,23.65]
20 to 20.9	19.84	0.75	[18.61,21.07]	20.01	2.59	[15.75,24.26]
21 to 21.9	20.23	0.87	[18.80,21.66]	20.52	2.64	[16.18,24.87]
22 to 22.9	21.28	3.00	[16.35,26.21]	21.03	2.69	[16.60,25.45]
23 to 23.9	20.80	1.38	[18.53,23.07]	21.52	2.74	[17.01,26.03]
24 to 29.9	22.82	3.70	[16.73,28.91]	23.15	2.91	[18.36,27.94]
30 to 35.9	24.41	3.73	[18.27,30.55]	25.64	3.14	[20.47,30.81]
36 to 41.9	27.55	3.69	[21.48,33.62]	27.74	3.27	[22.37,33.11]
42 to 47.9	29.49	3.07	[24.44,34.00]	29.48	3.23	[24.17,34.00]
48 to 53.9	30.98	4.30	[23.91,34.00]	30.88	3.02	[25.92,34.00]
54 to 59.9	32.22	1.71	[29.41,34.00]	31.97	2.64	[27.63,34.00]
60 to 65.9	32.73	1.31	[30.58,34.00]	32.75	2.13	[29.24,34.00]
66 to 71.9	33.45	1.22	[31.44,34.00]	33.26	1.59	[30.65,34.00]
72 or greater	33.42	1.25	[31.36,34.00]	33.51	1.07	[31.74,34.00]

ŧC.I.-Confidence Interval or 2.5th and 97.5th percentile values of the Z-score. Red dotted box shows unsmoothed and smoothed summary standardisation values for a child who is 6 to 6.9 months old. Shaded in grey are the smoothed summary statistics.

6.3.2.2. Sensitivity of Z-scoring methods

Given the degree of overlap seen in the density and scatter plots of both Z-score methods shown in second and third panels of Figure 6.8, we expect that sensitivity will be higher while attempting to classify the normal versus disabled samples as compared to classifying the normal versus

malnourished samples. This expectation is confirmed by the ROCAUC values in Table 6.5 that are consistently higher for the normal verses disabled samples in comparison to the normal versus malnourished samples. Also we see that the sensitivity of the suggested smoothed Z-score extension is fairly comparable to the classical Z-score approach. Finer detail with regard to sensitivity performance is shown by the ROC curves for each of the two methods shown in Figures 6.9a) and Figures 6.9b) below. It is clear that both scoring methods are fairly comparable in terms of their sensitivity performance.

Further, although we can now use one cut-off to distinguish the normal versus disabled or malnourished cases as the effect of age has been eliminated, due to the difference in degree of skewness of the classical Z-scores shown in the density plots in the second panel of Figure 6.8, the choice of threshold cut-off needs to be adjusted slightly. Notice that the optimal (to achieve high sensitivity) overall score cut-off threshold for the classical Z-score method is 0.04 that translates to a Z-score value of ~-1.56 on the classical Z-score scale as opposed to an optimal Z-score threshold cut-off at the 0.05 percentile that translates to a slightly higher Z-score value of ~-1.52 on the smoothed Z-score scale for the normal and disabled sample data. However, to err of the side of caution, the clinically optimal Z-score is usually taken to be a Z-score value of -2.00.

Effect of age on Z-score sensitivity performance

The respective ROCAUC summary values that are used to make comparisons across the three age groups for infants who are less than 1 year old, toddlers who are 1 to less than 3.5 years old and pre-school age children aged 3.5 to less than 7 years old are summarised in Table 6.5. It is clear from the respective ROCAUC values that the Z-scoring approach appears to be in general more sensitive for the 3.5 to 7 year age group of children.

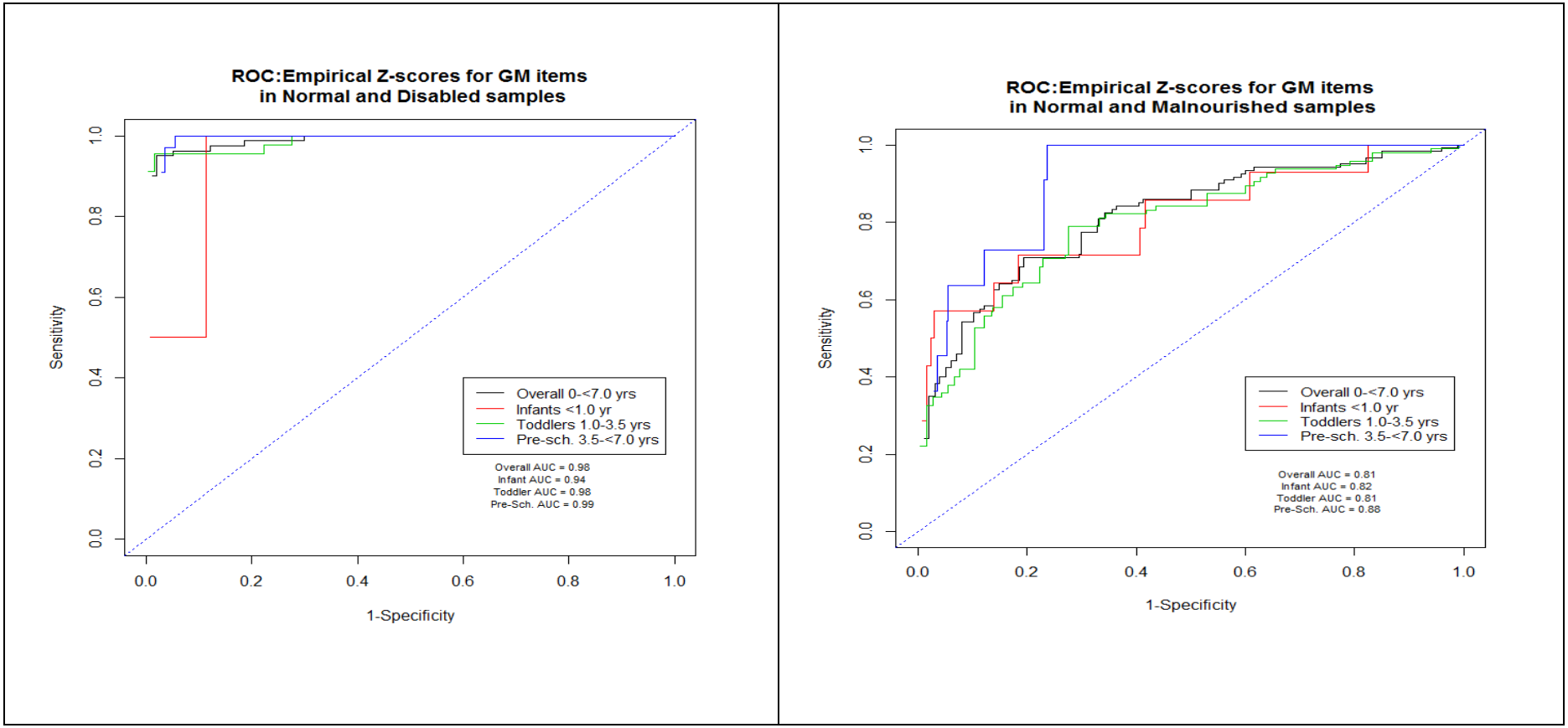


Figure 6.9a): ROC curves: Empirical Z- score method in GM domain

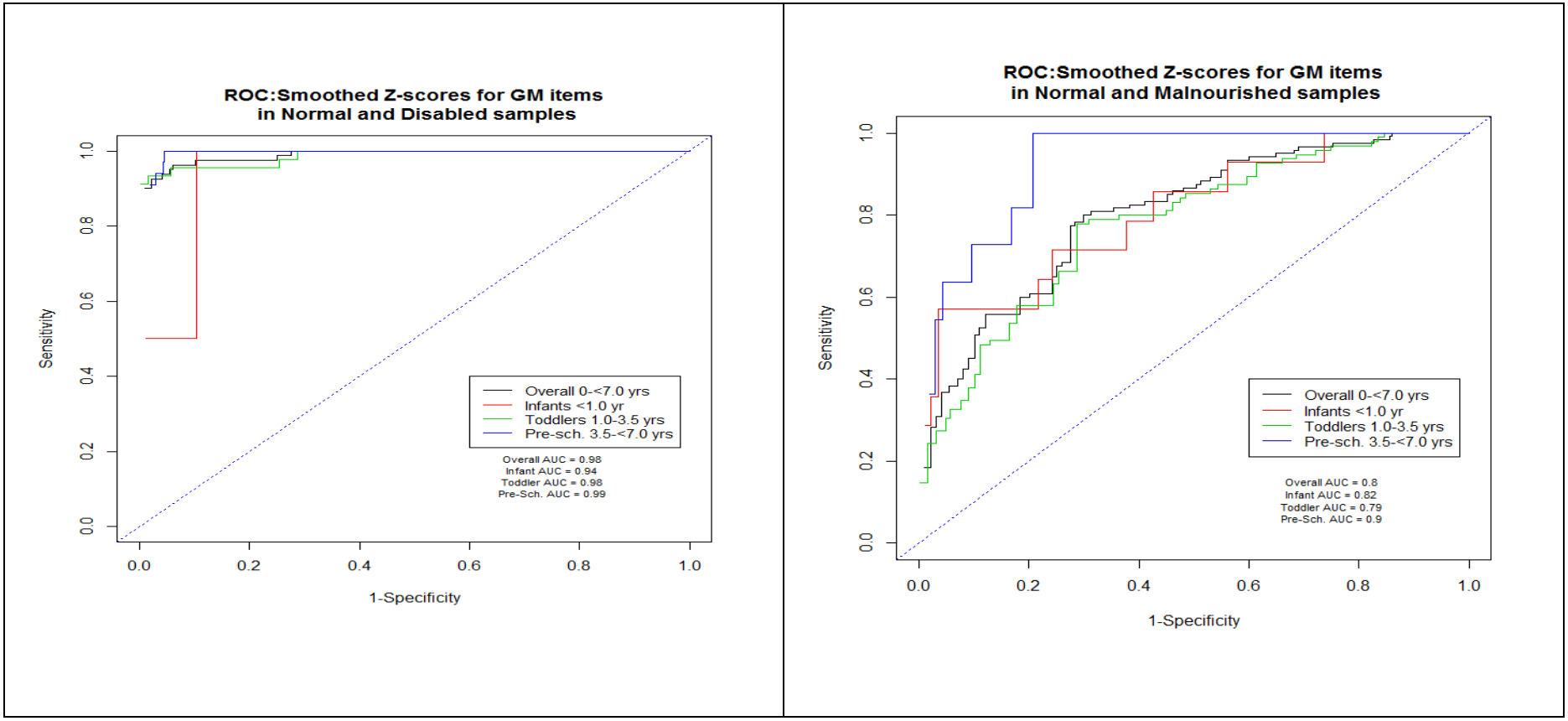


Figure 6.9b): ROC curves: Smoothed Z-score method in GM domain.

6.3.2.3. Summary of the Z-score methods

Within the Z-scoring approach, we have pitted the classical Z-score method against the smoothed Z-score method in terms of score characteristics, ability to correct for age and sensitivity performance. Apart from the benefits accorded by both approaches of successfully correcting for the effect of age on the simple sum score, through our sensitivity analysis we have demonstrated the superiority of the later method especially in averting the potential of mis-classification of the development status of a child as a consequence of either using unsmoothed empirical means that are not monotonic or standard deviations that vary considerably across respective age categories.

6.3.3. Item Response Theory Overall Scoring Methods

This section presents the findings of the third overall score computation method that uses the IRT framework. In this child development context, the important statistical issues of concern that have to be addressed in overall scoring against the backdrop of the remit of this thesis, the IRT frameworks perhaps offers the most ideal overall scoring approach. This is not only based on its ability to simultaneously address all of the highlighted issues but as we will see our suggested extensions were advised and motivated by lessons learnt from the item by item analysis and the other overall scoring methods considered previously.

Under the IRT approach we have considered two classical IRT scoring approaches; a 1 PL IRT model and its generalisation, the 2 PL IRT model. Given our objective to develop a methodology that directly adjusts for age we suggested and developed 1 extension to the 1 PL IRT model. Section 5.2.2.3b) explained how the 2 PL IRT model is a generalisation of the 1 PL IRT model as it allows the discrimination parameter α to vary across items. However, given our objective of devising a scoring method that corrects for age and owing to the connection of ability and age, as opposed to relaxing the discrimination assumption in line with the 2 PL model, we instead allow the discrimination parameter of each item to vary with age.

6.3.3.1. Comparison of score characteristics of IRT Scoring Methods

One Parameter Logistic IRT model (1 PL)

The summary distribution characteristics in the fifth column of Table 6.4 suggest that the classical 1 PL IRT standardisation normal sample scores are not normally distributed with a mean of approximately 0 and standard deviation of 1.20. The distribution plots in Figure 6.10 reveal a multimodal distribution especially in the normal sample hence the formal normality tests clearly underscores the fact that scores produced under this framework are not normally distributed. The scatter plot of 1 PL IRT scores against age in the second panel of Figure 6.10 shows a strong association of scores with age with a correlation coefficient value of 0.66. Hence it is clear that the 1 PL IRT model does not correct for age.

The third panel of Figure 6.10 shows plots of the item characteristic curves (ICCs) for the 1 PL IRT model for an item that is ideal for infants (<1 year), toddlers (1 to <3.5 years) and pre-school age children (3.5 to <7 years) in the normal (black lines), disabled (red lines) and malnourished samples (green lines) in the gross motor domain. A closer look at these 1 PL IRT model ICC curves reveals that;

- Firstly, there are some general expected features of the ICCs of this model; (a) the probabilities gradually increase with the ability trait level for each item. (b) The slopes of the curves are equal meaning that the items differ only in difficulty hence they do not cross (are parallel). The estimated constant item discrimination(α) parameter for the 1 PL model was constant, 8.41. (c) The point of inflection of the ICC occurs when the probability of passing an item is 0.50 is shown by the reference lines from at different ability trait levels to each of the types of items from the three data samples. The reference lines are drawn from the ability trait level that equals the item's difficulty when they cross the ICC at a 0.50 probability. This is the trait level at this probability and is interpreted as the threshold level for this items' difficulty i.e. the trait level at this point indicates the difficulty level at which the child is both as likely to pass or to fail an item e.g. as shown in the Table 6.7 below, the first item that is ideal for infants in the normal sample has a difficulty of -

2.17, the probability that a child with a trait level of -2.17 passes this item is 0.50. The table also presents the parameters for item discrimination(α) for the 2 PL model only as the item discrimination parameter for the 1 PL model is a constant, item difficulties (β_j), standard errors ($\sigma_\alpha \; and \; \sigma_\beta$) for both the 1PL and 2PL models whose specification was described in Section 5.2.2.3a) and 5.2.2.3b) respectively. Notice also that the increase in item difficulty is not always monotonically increasing. Therefore, given our previous findings in the item by item analysis discussed in Section 6.2.1 with regard to item ordering, this finding also supports the argument that this IRT model suggests that the MDAT items should be reordered to adhere to the assumption of monotonicity.

- Secondly, in the third panel of Figure 6.10, we intuitively would expect that the difficulty values for the three types of items should at least be equal or higher in both the non-normal samples of disabled and malnourished children i.e. the trait level of an item in the normal sample should be lower than the trait level of the same item in either the disabled or malnourished samples as a consequence of delayed development or disability in a specific trait being tested by the given item. This seems to be the case with the exception of the low difficulty item that is ideal for infants in the malnourished sample. We attribute this to the fact that children in the malnourished sample were older, and this would therefore result in them needing a lower ability level to be able to pass administered items. Recall that malnourishment does not necessarily directly imply delayed development.

Table 6.7: 1PL and 2PL IRT model parameter estimates for the MDAT tool in the normal sample

Type of item	Item	1 PL IRT Model		2 PL IRT Model			
		Difficulty parameters		Discrimination parameters		Difficulty parameters	
		β_j	σ_β	α	σ_α	β_j	σ_β
Ideal items for infants: <1 year old	1	-2.17	0.03	27.55	0.10	-2.12	0.15
	2	-1.86	0.03	9.29	1.50	-1.84	0.03
	3	-1.42	0.02	17.77	56.76	-1.43	0.03
	4	-1.66	0.04	8.02	0.81	-1.66	0.04
	5	-1.39	0.02	19.49	59.67	-1.37	0.02
	6	-1.23	0.02	19.81	58.18	-1.30	0.09
	7	-1.27	0.02	22.48	54.46	-1.32	0.10
	8	-1.19	0.03	9.17	1.47	-1.20	0.03
	9	-0.87	0.02	21.81	33.50	-0.85	0.03
	10	-0.77	0.02	30.16	0.10	-0.75	0.09
	11	-0.55	0.02	8.03	1.04	-0.70	0.02
	12	-0.42	0.03	24.17	39.52	-0.66	0.03
Ideal items for toddlers: 1 to <3.5 years old	13	-0.31	0.04	26.77	47.18	-0.65	0.04
	14	-0.10	0.02	11.36	1.52	-0.51	0.03
	15	-0.01	0.02	40.92	0.02	-0.14	0.02
	16	0.15	0.02	44.17	0.02	-0.03	0.02
	17	0.26	0.03	45.14	0.02	-0.02	0.02
	18	0.33	0.03	46.47	0.02	-0.02	0.01
	19	0.50	0.02	5.32	0.45	-0.03	0.02
	20	0.71	0.01	6.51	0.62	0.19	0.03
	21	0.63	0.01	6.53	0.67	0.09	0.02
	22	1.02	0.03	6.29	0.56	0.60	0.03
Ideal items for pre- School aged children: 3.5 to <7 years old	23	1.04	0.03	8.17	1.01	0.61	0.03
	24	0.93	0.02	6.63	0.55	0.52	0.03
	25	1.06	0.03	7.74	0.86	0.64	0.03
	26	1.20	0.02	28.63	0.10	0.67	0.03
	27	1.36	0.02	7.73	0.85	0.84	0.03
	28	1.41	0.02	6.77	0.58	0.93	0.03
	29	1.23	0.02	7.64	0.98	0.73	0.03
	30	1.54	0.03	6.72	0.61	1.10	0.03
	31	1.38	0.02	6.19	0.51	0.89	0.03
	32	1.48	0.02	5.87	0.47	1.08	0.03
	33	1.99	0.02	7.04	1.44	1.42	0.03
	34	2.01	0.02	7.10	2.21	1.45	0.04

Items response model estimates for babies (< 1 year old), toddlers (1 to < 3.5 years old) and pre-school aged children (3.5 to < 7 years old) in the blue, green and red dotted boxes respectively.

Estimate values are reported to nearest two decimal places.

α-alpha i.e. discrimination parameter for 1 PL model is constant=8.41.

Highlighted in grey are the questionable high variance parameter estimates due to very high item response pass rate patterns of these items.

Two Parameter Logistic IRT model (2 PL)

The summary distribution characteristics in column six of Table 6.4 also suggest that the classical 2 PL IRT scores are skewed and not normally distributed with a mean of approximately 0 and standard deviation of 1.21. The score distribution plots in the fourth panel of Figure 6.10 reveal a multimodal distribution especially in the normal sample hence the formal normality test rejects the null hypothesis and we can conclude that the 2 PL IRT score distribution is not normal. The scatter plot of the 2 PL IRT scores against age in the fifth panel of Figure 6.10 shows a strong association of these scores with age with a correlation coefficient value of 0.66. We can also conclude that the classical 2 PL IRT scores do not correct for age.

To facilitate comparison, we also plotted the item characteristic curves (ICCs) for the 2 PL IRT model for the same items that is ideal for infants (<1 year), toddlers (1 to <3.5 years) and pre-school aged children (3.5 to <7 years) in the normal (black lines), disabled (red lines) and malnourished samples (green lines) in the gross motor domain. These are shown in the sixth panel of Figure 6.10. Recall that the 2PL IRT model described in Section 5.2.2.3b) has a discrimination parameterα_j, which is allowed to vary for each item. Thus the 2 PL IRT model is relevant in this child development context where items are not equally related to the latent trait or items are not equally difficult meaning that they are not equally indicative of a child's standing on the ability trait. Based on the 2 PL IRT model ICC curves, we note that;

- Again, there are some general expected features of the ICC curves of this model; (a) the item discriminations of the items that are ideal for infants (<1 year), toddlers (1 to <3.5 years) and pre-school aged children (3.5 to <7 years) differ. Hence, the ICC curves are not parallel and may at times cross i.e. the slopes of the curves are not equal meaning that the items differ both in difficulty and discrimination. (b) In a similar fashion to the 1PL IRT model the point of inflection of the ICC occurs when the probability of passing an item is 0.50 as shown by the reference lines can be interpreted the same way e.g. in contrast to the 1 PL model as is shown in the Table 6.7, the

first item that is ideal for infants in the normal sample according to the 2PL IRT model now has a difficulty of -2.12. That means the probability that a child with a trait level of -2.12 passes this item is 0.50.

- Further, we also intuitively would expect that the difficulty values for the three types of items should increase in the non-normal samples of disabled and malnourished children i.e. the trait level of an item in the normal sample should be lower than the trait level of the same item in either the disabled or malnourished samples. What was seen earlier in the 1PL IRT model seems to be replicated by the 2 PL IRT model even when item discrimination is allowed to differ. We also attribute this observation to the fact that children in the malnourished sample were older, and this would therefore result in them needing a lower ability level to be able to pass the administered items in addition to the argument that malnourishment, unless severe does not necessarily imply a lack of ability.

In contrast to the 1PL IRT model, the 2 PL IRT item difficulty values (β_j) in the normal sample ranged from -2.12 to 1.45 while estimated item slope or discrimination variance values (σ_α) ranged from 0. 10 to 59.67. The 2 PL model allows items to differ in discrimination to cater for the fact that the MDAT items are designed to increase in difficulty, hence we argue that a model with constant item discrimination such as the 1 PL model is not suitable. However, although the 2 PL is a more flexible model that allows item discrimination to vary, it at times struggles to give reliable item discrimination and item difficulty variance estimates that were rather high especially for the gross motor items 3, 5, 6, 7, 9,12 and 13 (shaded in grey in Table 6.7). This could be as a result of the extra complexity of allowing discrimination of each item to vary, but also it could be related to an identifiability problem as these same seven items with questionable variance estimates for their discrimination and difficulty parameters did not have very good item by item model fits. It is for these reasons that the following section explores an alternative way of correcting for the effect of age. As we will realise in Section **Error! Reference source not found.**, this will have the added advantage of also finding one suitable

cut-off threshold (regardless of age) that can be used to correctly classify child development status with higher sensitivity.

One Parameter IRT Monotonic Spline Model

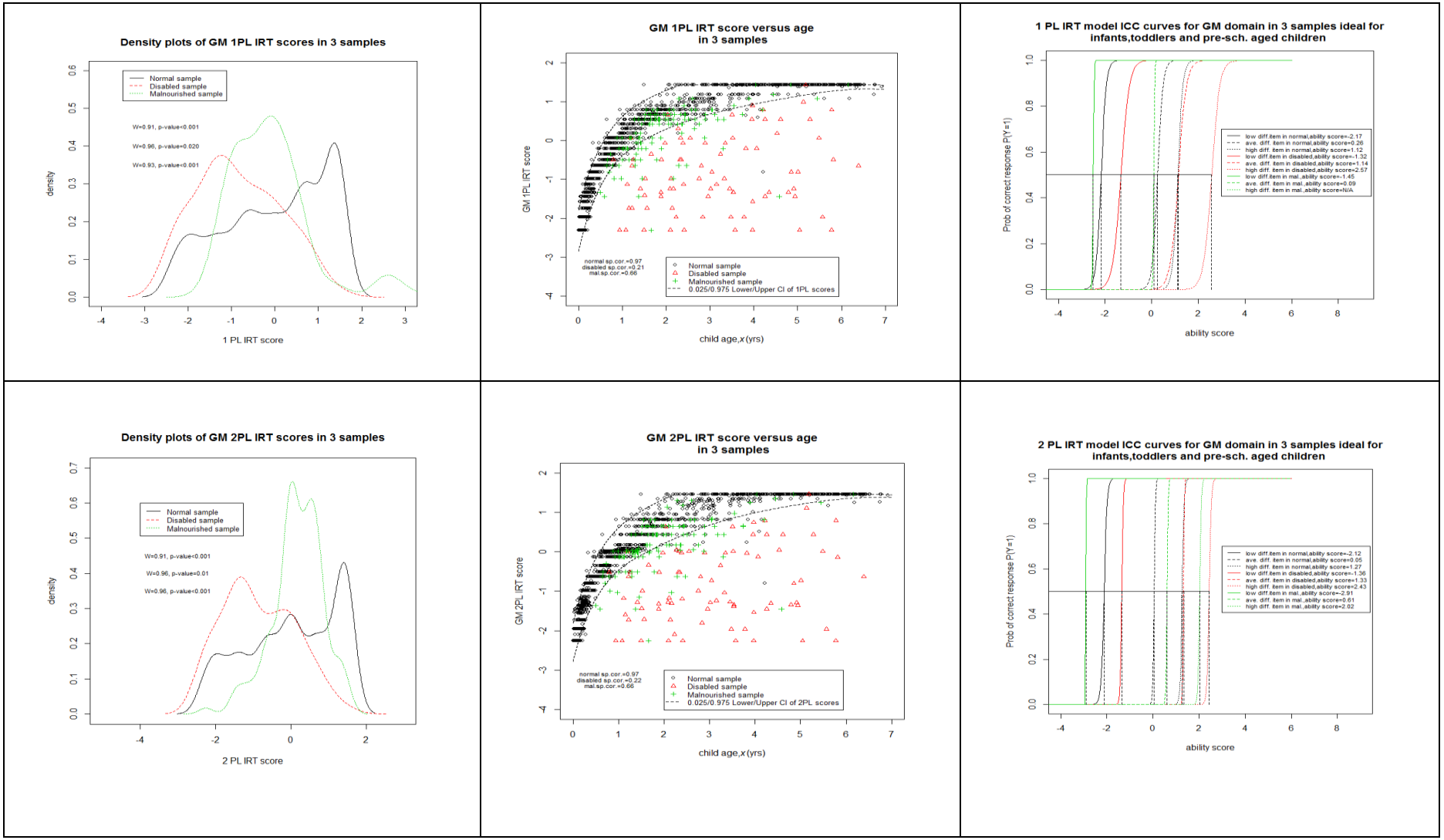
Section 5.2.2.3c) explained that the classical 1 PL IRT model assumes uniform item discrimination. We also further explained the fact that it is plausible to assume that the probability of passing an item increases with age. Therefore, as opposed to fitting a 2 PL IRT that allows item discrimination to vary, we can instead allow the discrimination parameter of each item to vary with age. Hence we can fit a 1 PL IRT model in the form of a mixed effects logistic regression model defined in equation 5.74 in Section 5.2.2.3.

We would like to first note that fitting the 1 PL Monotonic Spline IRT model using a mixed effects logistic regression approach can be frustrated by computational issues e.g. the failure of the model to appropriately converge due to subject specific item having response patterns with many failed or passed items. The seventh column in Table 6.4 above shows the summary statistics of this overall scoring method having a skewed distributions in the normal sample. The 1 PL IRT Spline Model has a score range of between -6.51 to 2.96 and a variance of 0.82. The skewness is also observed from the density plots in the first panel for the normal shown in Figure 6.11. The test of adherence to normality for each of the samples confirms that the score distributions of scores from fitting the 1 PL IRT Spline model are not normal. The second panel of Figures 6.11 is a scatter plots of the 1PL Spline model score's and age showing that the previously observed strong non-linear association in both the 1PL and 2PL classical IRT scores with age no longer exists i.e. is no systematic pattern between scores and age. These findings are confirmed by the reduced value of correlation coefficients of 0.02 between the respective scores and age as well as the almost horizontal loess smooth curve (red line) overlaid in the scatter plots of the 1 PL Spline model's scores and age. Therefore, this extension of the classical 1PL IRT scoring approach can be used to compute overall scores that are adequately age adjusted.

The classical IRT model framework outputs respective parameters for item discrimination(α) and item difficulties (β_j) for both the 1PL and 2PL models enabling the plotting of ICC curves over an ability spectrum (grid) of interest. Extracting these parameters and plotting the ICC curves for the extended version of the 1 PL IRT model is not very clear, as the mixed model approach computes a subject specific value (score) i.e. the random effect that represents a child's probability of passing a number of administered items given the age and ability of each child. For the 1 PL IRT Spline model we extracted the subject specific overall probabilities using the 'ranef' function in the lme4 R package that substitutes the mean of the subject specific residuals into the formulae to the response probability given in equation 5.74.

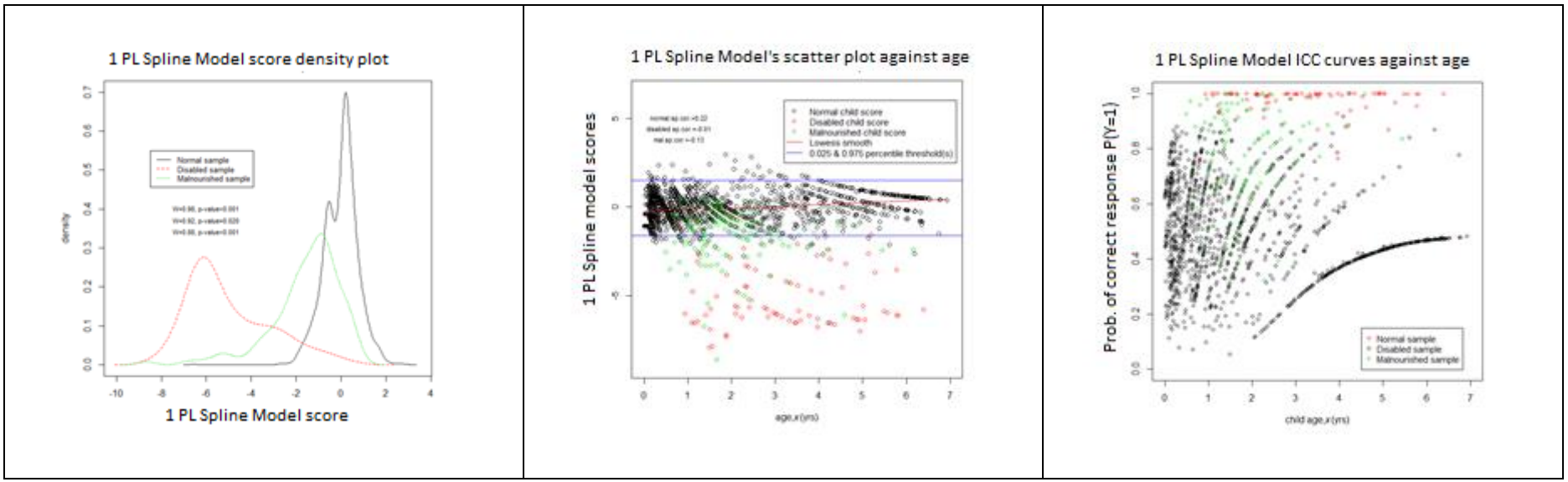
The third panel of Figure 6.11 shows the ICC plots against age for the 1 PL IRT Spline Mixed Model. From the plot we notice also that the ICC curves are not parallel i.e. the slopes of the curves are not equal meaning that the items differ in discrimination with respect to age. This finding is expected given the empirical ICC curves plotted in Section 4.4.5 and confirms the increase in item difficulty design of the MDAT tool. The following points are drawn from the IRT scoring approaches;

- Both the 1 PL and 2 PL scoring methods do not account for the effect of the age.
- The 1 PL IRT Monotonic Spline Model is a suitable scoring method that appropriately computes scores by simultaneously allowing a different number of items to be administered to each child, items to differ in difficulty as well as allowing discrimination of items to depend on the age of the child. As we see in the following section, this has now the added advantage of requiring only one cut-off thresholds to determine the development status of a child regardless of age that proves to be more sensitive. Further, any item correlation within a child's responses or between children can be addressed within this mixed model framework.



W is the Shapiro-Wilk test for Normal, Disabled and Malnourished scores

Figure 6.10: Density plots, scatter plots and ICC plots of 1PL IRT and 2PL IRT scores in GM domain for Normal, Disabled and Malnourished data



W is the Shapiro-Wilk test for Normal, Disabled and Malnourished scores

Figure 6.11: Density plots, score scatter plots, ICC plots of 1PL spline IRT scores in GM domain for Normal, Disabled and Malnourished data.

6.3.3.2. Sensitivity of IRT scoring methods

As expected, higher sensitivity was observed while classifying development status between the normal and disabled samples in all the models considered under this IRT framework. This was to a large extent facilitated by the use of more appropriate score cut-off thresholds like was done for the simple sum score but instead used both the classical 1 PL and 2 PL IRT models' scores in a GAMLSS model that took age into account. From the summary given in Table 6.5 we see that both the classical 1PL and 2PL models reported comparable overall ROCAUC values that improved in the 1 PL IRT Spline model between the normal and malnourished samples especially for the infant and toddler age groups.

Effects of age on sensitivity

Even though there was not a considerable increase or decrease in sensitivity with the different age categories as far as sensitivity is concerned in each type of IRT model, the benefit of the 1 PL IRT Spline model was evidenced by higher sensitivity in detecting delayed development or disability between the normal and malnourished samples. Recall that we have noted that detecting disability or delayed development is harder in malnourished children because malnourishment, unless severe does not necessarily imply a lack of ability. For example, the classical 1 PL IRT model reported ROCAUC values of 0.84 and 0.81 in the normal versus malnourished samples for the infant and toddler age categories respectively. However, 1 PL IRT Spline model highlighted in the red dotted box in Table 6.5 had consistently slightly higher sensitivity values of 0.85 and 0.84 in the same normal versus malnourished samples for the infant and toddler age categories respectively.

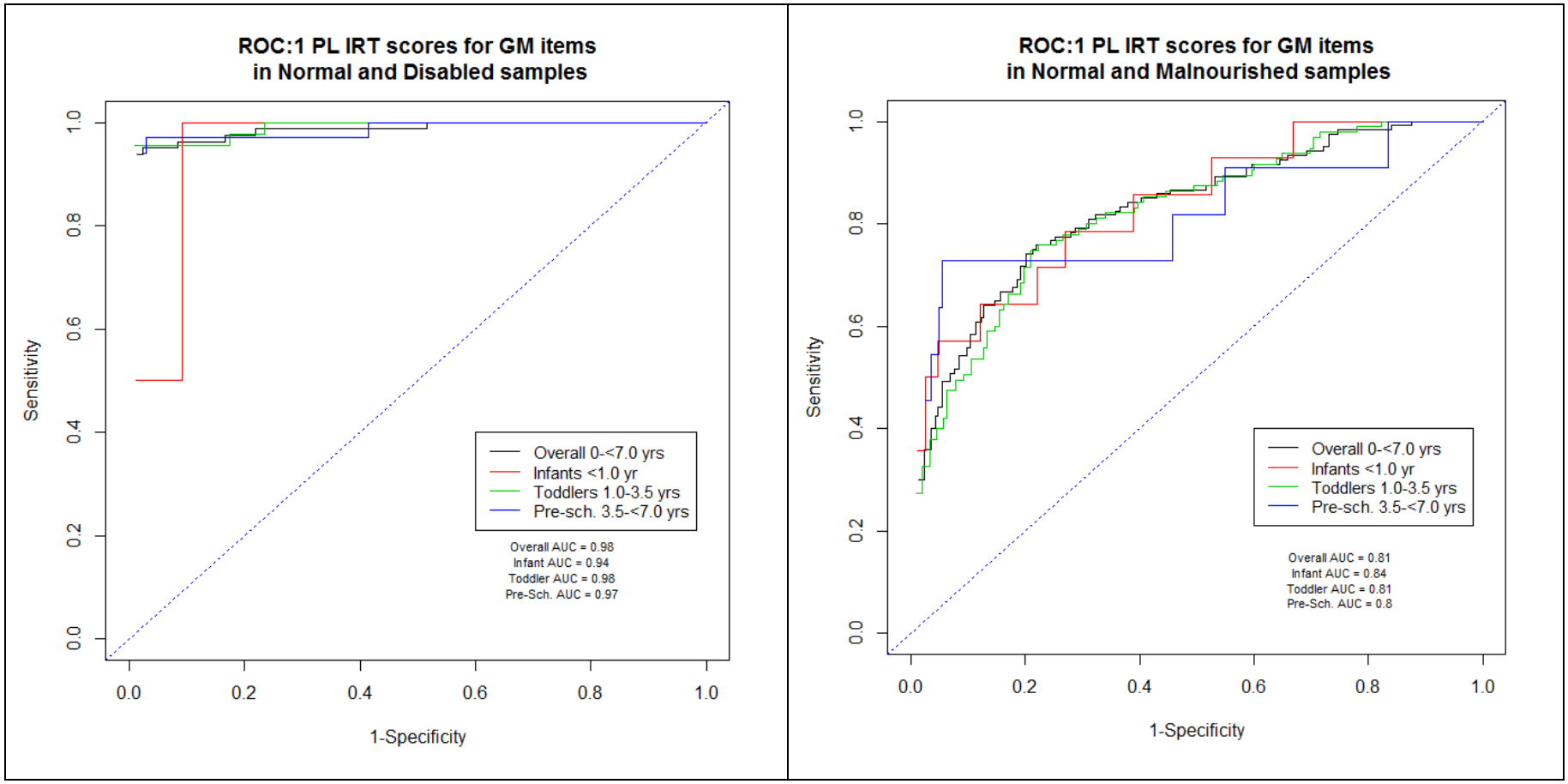


Figure 6.12a): ROC curves: 1PL score method in GM domain.

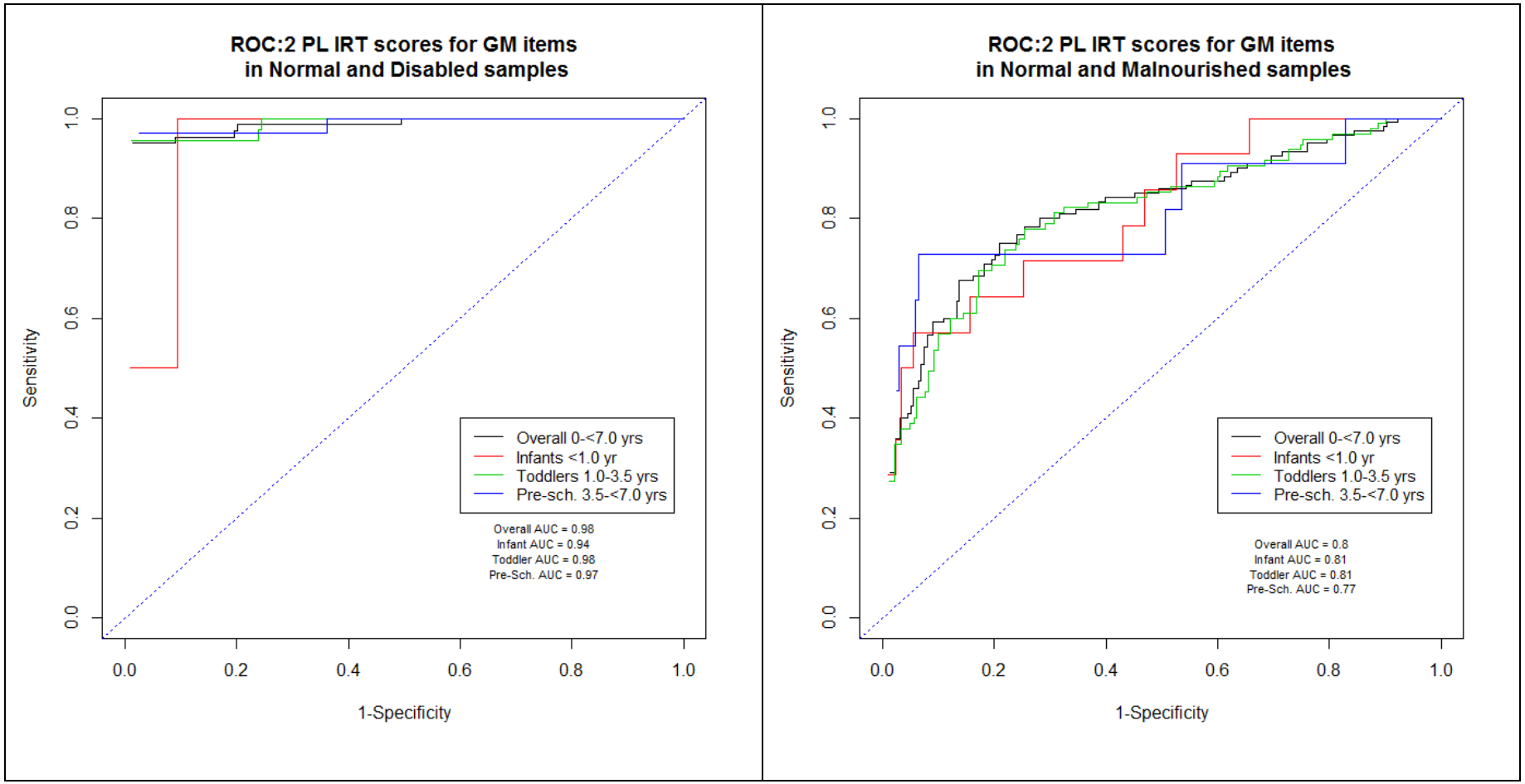


Figure 6.12b): ROC curves: 2PL score method in GM domain.

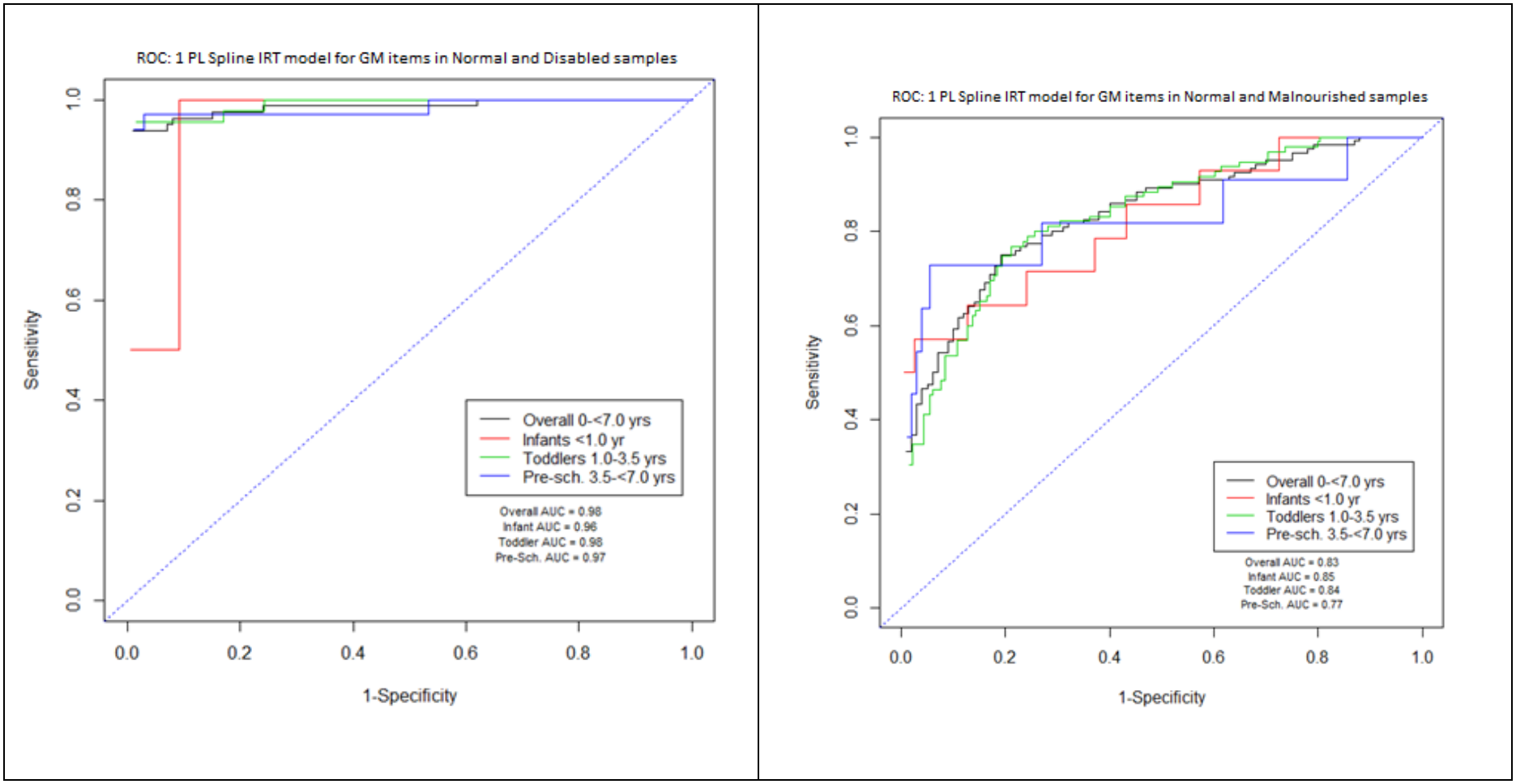


Figure 6.12c): ROC curves: 1PL spline score method in GM domain.

6.3.3.3. Summary of Item Response Theory methods

The IRT approaches' mettle was to be able to cater for the fact that children responded to a different number of items, adjust for the difference in item difficulty, the difference in item discrimination, cater for item correlation and allow the adjustment of the effect of age within one modelling framework. Starting with the classical 1 PL and 2 PL IRT models that are applicable within this child development context, we suggested and developed an extension that allows the item discrimination to change with respect to age. In the Section **Error! Reference source not found.** we have outlined the two classical IRT model score characteristics in terms of score distribution, ability to adjust for age (score correlation with age) and sensitivity. On the later issue, we made use of the flexibility of the GAMLSS model framework to define cut-off score thresholds to classify developmental status with respect to age as both classical IRT models produce scores that are strongly correlated with age.

Instead of relaxing the discrimination assumption in line with the 2 PL model, we allowed the discrimination parameter of each item to vary with age given that the probability of a child passing an item is primarily driven or dependant on age. The non-linear relation between score and age was captured using a monotonic spline function and we fitted an extension of the 1 PL IRT model under the GLMM framework. This was mainly to take advantage of the fact that the IRT model can be fitted within the backbone of a generalised mixed model and while avoiding the drawbacks of the other scoring methods also allow us to simultaneously address the four important issues of interest in this child development context; the fact that items in a child development context differ in difficulty, item pass probability is dependent on age, different children will respond to a different number of items and the underlying correlation or association between and within item responses. The 1 PL IRT Spline model was successful in adjusting for the effect of age on scores as it is accounted for directly in the score computation.

While giving a brief recap of the findings given the pros and cons of various age estimation and overall scoring approaches, the next chapter discusses the overall conclusions implications of our results.

PART IV – DISCUSSION

Chapter 7. Discussion and Conclusions

7. Discussion and Conclusions

7.1. Introduction

It is hoped that the sixth chapter that compared the characteristics of the different age estimate and overall scoring methods has at least managed to convince the reader of the relevance of addressing the statistical issues highlighted in the fifth Chapter in the form of the pros and cons of; a) each item by item analysis method whose output are the age estimates of defined milestones measured by the assessed item or b) each overall scoring method whose output is a score that characterises a child's ability using all items assessed in a domain at a given age. To define the importance and relevance of our research, and therefore justify why this research is important, we carried out both a literature and systematic review of statistical issues that underpin age estimation, overall scoring and the benefit of age adjustment. Our goal was to consolidate the extant literature on current assessment tool translation and adaptation methods and the corresponding statistical methodology used to compute scores. Beyond identifying specific gaps in our target research area, the reviews making up the first part of this thesis also had the added benefit of vividly outlining current methods both in terms of practice and reporting standards.

Following this evidence from the literature, we were best placed to highlight some common pitfalls in current practice in our methods chapter and therefore decide how to extend current methods and determine whether our suggested extensions are more suitable. These three aspects outlining current practice, statistical concerns and extensions formed the backbone of the broader outlook of this thesis each having dedicated chapters that define, justify and address our specific research themes. The evidence of our extensions' suitability or superiority formed the results chapter that compared and contrasted the properties of both classical age estimate and overall scoring methods against the suggested extensions. In this discussion chapter we revisit the highlights of each of the methods and explain the implications of our findings.

Just to reiterate, the primary main objective of this thesis was to create both robust age estimation methods to create item by item reference charts and overall scoring methods to create norms used to assess child development. Beyond defining the methods extensions, we were pragmatic and also developed workable processes and R code to easily use or calculate the scores. This is motivated by the fact that even if appropriate scoring methods are developed, they may not be implemented owing to software limitations or data quality. Section 7.2 revisits the highlights of each previous chapter by giving a brief recap of findings but more importantly outlining the take home messages and conclusions. Therefore, while enabling the appreciation of our suggested item by item age estimation and overall scoring methods given the posed research questions, we also point to feasible areas of future work in Section 7.3. Our final concluding remarks are given in Section 7.4.

7.2. General Discussion

The discussion is set in the form of questions that each of the chapters attempted to answer. This not only serves to identify each chapter's research starting and finishing points; it will also aid in outlining our perspective and assumptions in the build up to our conclusions. Therefore, as we outline how each step in tool development underpins the current child ability age estimation and overall scoring methods for ability status classification, it will link the result outputs thereby showing how this work adds to current methods by addressing their identified weaknesses.

7.2.1. What are the current tool development, age estimation and scoring practices?

The literature review chapter outlined the complex tool development process using a translation and adaptation strategy. The edifice of cultures definitely is an important factor that should be considered when developing tools from already established assessment items from western countries. Apart from influencing the tool content, we highlighted how culture and other important issues such as country specific laws may also influence its administration. In as much as every translation and adaptation

study is unique, still, the statistical methods used to address bias and establish equivalence between source and new tool should be appropriate and well thought out given available resources and study design.

Having a well translated and adapted tool achieving high levels of reliability and validity is only 'half' of the task. The computation of either item by item age estimates or developmental norms for defined mile stones is the next step using binary data collected using high quality tools and is the stage at which this thesis makes its direct contribution. Because this analysis stage also doubles up by offering further insight into item quality and item adaptation flaws, suitable statistical methods should be used. The current statistical analyses methods used were introduced in Section 2.3.2 that explained the necessity of robust methods to deal with the often skewed item outcome distribution and the importance of the monotonicity assumption especially in this child development context. With respect to the overall scoring methods, we outlined the importance of transforming raw scores to account for the often ignored fact that items in a child development context differ in difficulty, different children will be administered and respond positively to a different number of items depending on their age, and that there is an underlying correlation or association between and within items. Accounting for all these issues enables better interpretation of scores and enables valid ability development status classification.

Chapter one therefore adequately set the stage for the main objective of this project to extend both the current item by item age estimation and scoring methods. We realised that child development assessment encompasses a detailed and cyclic process. Our hope was to ensure the realisation that production of sensitive and high quality scores really begins at the inception of any research geared towards ability assessment. Without being put off by the complexity of the tool development process we highlighted issues that directly and indirectly affect the quality of scores and manifest themselves in different ways. However, we realise that even the indirect issues at the design and implementation stages of tool development are just as important hence it is thus clear that even the most robust or

sophisticated scoring methods may do little to alleviate the repercussions of poor study designs that eventually compromise computed age estimates or overall scores.

7.2.2. What is the current age estimation and scoring reporting practice?

Chapter two was a stock take of the current translation, adaptation strategies, the statistical methods used to establish the performance equivalence of translated items with the original source items as well as the scoring methods used in the developing world. The output of this chapter aside from a list of recommendations that were used to formulate an appraisal checklist that outlines a list of properties of a well-adapted child assessment tool was an in depth discussion of the current statistical methods used to compute item by item age estimates and overall scores. The checklist was created in an effort to encourage a standardised reporting architecture when carrying out a translation, adaptation, item by item age estimation or norm creation study. However, we wish to reiterate that the systematic review confirmed that there is indeed a consensus in acknowledgement of the importance of accounting for cultural aspects in child assessment tools. While some authors e.g. Sireci, (2005), Gladstone, et al., (2008), Borsa, et al., (2012) and Kammerer, et al., (2007) attested to the growing interest in the application of appropriate methods to culturally adapt assessment tools, others especially Sabanathan, et al., (2015) and Hambleton & Kanjee, (1995) also declared the importance of or need for developing new culturally adapted tests. This is done by the adequate application of translation, adaptation and scoring methodology in tool development.

Unfortunately, routine reporting especially in the developing world still falls short and lacks a standardised reporting format. Even if this is a consequence of either a difference in research priorities due to different child health burden rates or the fact that many studies in the developing world usually have many ambitions, at least a dedicated effort to adequately report findings given available resources should be evident. There is a clear weakness in that details used in adaptations come as an

afterthought and are not primary considerations in the design of these studies. However, while we hope that this problem declines and there is more investment into dedicated research, this strategy may work for the time being. While the most ideal method to create a culturally relevant assessment tool is to design a completely new tool, so as to avoid any bias due to cultural differences, the adaptation of existing tools saves in resources especially because the time frame needed to develop a test is substantially shortened. Further, by the fact that certain constructs are shared between cultures, the adaptation strategy also facilitates or makes it easier to compare populations across multiple cultures. This is vital for various organisations e.g. WHO and World Bank or governments concerned with international policy making or distribution of resources regionally.

7.2.3. What are the ideal properties of a typical assessment tool?

The third chapter focused on the MDAT tool as the illustrative example in this thesis. It was developed in Malawi through a series of studies as was described in Section 3.2 that addressed the item translation and adaptation using items of various established western tools. We went on to describe how the tool was used to assess children in this study but also especially recommended important properties an assessment tool should aspire to have. Most importantly we highlighted the potential source of bias resulting from the item response recording mechanism that has implications on the validity of both the age estimate and scoring analyses due to data quality. We recommended that the item response recording should be done very diligently invoking very reasonable assumptions even if any missing data can be explained against the backdrop of typical child development assessment particularity.

7.2.4. Why is item data quality important?

The fourth chapter discussed the initial or preliminary steps to undertake once the data have been collected and the exploratory data analysis process. The EDA is an important process as it gives insight into any underlying characteristics in the item response data to help decide on the ‘best’ statistical

modelling approaches to apply. We also saw that EDA flagged potentially problematic items that were not particularly useful in the development of the score and can thus be removed or modified. Within the child development evaluation context, test items are designed to increase in difficulty. Therefore, it is important for the item test data to reflect this fundamental property of increasing in difficulty. In Section 4.4 we outlined various empirical methods including pass probability histograms against age, discrimination index computation and plotting empirical item curves that can be used in investigating and quantifying the increase in item difficulty and discrimination properties when the test outcome is binary.

A second important data quality to investigate and quantify is the extent of correlation between test item responses and item to total correlation. We recommended the use of the polychoric correlation coefficient as the test item data was binary. The existence of a strong correlation within test item data means that methods that assume independence of test items may not be very suitable. The purpose of the item to total correlation serves to ensure that every item in the assessment domain is relevant and measures a related aspect as the overall target domain construct.

As noted earlier, the type of data dictates the appropriate statistical approach to be used for analysis. Our work has focused on the use of binary data for item by item age estimation and norm creation. This we believe separates our work from the age estimation and norm creation methods that mainly use continuous data measured on various physical growth attributes of children e.g. weight, height, head circumference and MUAC. The exploration of item response data characteristic mentioned above serve to; firstly, give insight to item quality and secondly, they advise on the correct statistical approaches to use and the important statistical assumptions that should be adhered to.

7.2.5. What are the important statistical implications?

This section discusses the important implications of the item by item analysis and the overall scoring analysis methods.

7.2.5.1. Item by item methods

The main purpose of the item by item analysis is to come up with age estimates and confidence intervals at which a normally developing child is expected to pass a certain item at a stipulated probability. While this approach greatly informs us about various important item properties that point to their specific suitability to assess the target development construct, we have demonstrated how our suggested extensions are better suited to deal with the often skewed item pass probability distributions. In place of the classical logistic regression framework that may be frustrated by limited flexibility to fit skewed item data while still conforming to model assumptions, we suggested use of the SCAM model in Section 5.2.1.2. The SCAM model did not only prove to be more flexible and robust at adequately fitting skewed data, it had the benefit of being independent of the subjective choice of knots used in the spline models. Further, apart from its flexibility that rivals traditional generalised additive models, the fact that it is monotonically constrained ensured adherence of the monotonicity assumption to which the computed estimates should adhere. We argue that this assumption heralds all other item model assumptions within this child development assessment context.

7.2.5.2. Overall scoring methods

The overall scoring method is tasked with utilising all the items in a particular tool domain simultaneously to compute one index, usually called a score, which summarises the performance of a child on all administered items in that domain. Using binary data, the overall scoring method should be able to address the fact that items differ in difficulty, different children will respond to a different number of items and there is an underlying correlation or association caused by the fact that multiple items test the same construct and multiple responses are given by one respondent. Further, the method should provide a suitable framework to adjust for the effect of age that is known to be strongly associated with the probability of passing an item. At this juncture, we trust that it is clear that the IRT overall scoring approach described in Section 5.2.2.3 is a generalisation of the item by item analysis.

The IRT approach is also only viable if certain model assumptions that are unique to this child development context are suitably adhered to.

The results chapter gave evidence of the pitfalls of the classical overall scoring approaches and displayed the superiority of our suggested extensions in achieving important score qualities both in terms of score distribution characteristics and the sensitivity of the development status classification performance. We showcased the suitability of the 1PL IRT Monotonic Spline model within the IRT framework over all the other considered overall scoring methods by its ability to simultaneously allow the change item difficulty, item discrimination, number of item responses and adjusting for age while addressing the underlying correlation between and within item responses. Further, despite the IRT framework being well supported by a sound statistical theory, its application or implementation may be frustrated by available statistical software due to model complexity or number of item responses. Therefore, we also developed the required R software code using various optimisation and maximisation functions to fit this 1PL IRT Monotonic Spline model in a GLMM framework.

7.2.6. What are the original contributions of this thesis and how are our suggested extensions more superior?

It would be remiss of us to simply highlight the suitability of our suggested extension methods solely based on better fitting models within the item by item analysis or score distributions that are less skewed and closer to the normal distribution. Therefore, the most important output of this thesis is to show that our extensions result in more meaningful age estimates and overall scores as far as accurate development classification is concerned. Beyond their superiority in quality that is reflected in higher quality development assessment charts and norm summaries used in ability status classifications, the extensions should be easily implementable in available statistical software. Therefore, our work has also explained how to implement the more complex extensions that address important issues in child development assessment in R statistical software.

7.2.6.1. Item by item methods

As discussed in Section 2.3.2.1, our EDA in Section 4.3.4 also found that the item pass probability distributions are often skewed. This is because the items that are ideal for the infants (<1 year old) will have very high pass rates as the toddlers (1 to <3.5 years) and pre-school aged children (3.5 to <7 years old) will easily pass these items. On the other hand items that are ideal for older children will have low pass rates as the younger children are likely not to pass them again leading to skewed pass probability distributions that often frustrate the fit of the logistic regression model. Further, the confidence band around the logistic regression that is based on asymptotic theory becomes invalid.

The SCAM model described in Section 5.2.1.2 showcased its flexibility to fit these items with skewed item response probability distributions and produce more valid age estimates used in the development charts. The bootstrapping process also enabled us to compute the necessary confidence band to quantify variability around age estimates at the pass probabilities of interest. If we had to choose between the logistic and SCAM model, we would say that both modelling frameworks are comparable when the item pass/fail probability distribution is an even S-shaped curve with the former being more preferable as it is easier to explain and implement in typical statistical software. The robust SCAM model is ideal in the skewed pass probability scenarios explained earlier.

7.2.6.2. Overall scoring methods

All the statistical analysis preceding the overall scoring stage from the exploratory analysis to the item by item analysis is aimed at advising the overall scoring methods. We reiterate again that an appropriate overall scoring approach in a child development assessment context has to especially account for the difference in item difficulty, different number of items passed per child, between item correlations and correct for the effect of age, which is known to be strongly associated with the probability of passing an item. Our work considered three (simple scoring, Z-scoring, and IRT) main approaches of overall scoring and within each, suggested and developed both direct and indirect methods to address the four important item response data characteristics that were explored in

Section 4.4. In Section 1.1.2 we explained that an indirect transformation of a score method meant making the transformation to already computed scores in an attempt to rectify certain score flaws or weaknesses while a direct transformation of a score implied making the required adjustment or transformation during score computation. Therefore, the scoring extensions suggested for the first two methods of simple sum scores and Z-scores are examples of indirect methods while the scoring extensions suggested for the later third method IRT frameworks are examples of direct score transformation methods. Our work makes the argument that even though the indirect methods may work well, issues that they ignore to address such as assuming uniform item difficulty are carried on to the new methods.

Inimical to the criticism by many authors in addition to ours outlined in Section 5.2.2.1, the simple naïve count is still widely used by many researchers. Understandably, it's very ease of use is perhaps the reason for its popularity despite its obvious pitfalls in this child development context. Therefore, our extensions took both an indirect and direct approach towards addressing the 4 issues listed in Section 7.2.5.2 above. We first suggested weighting the simple score with the number of administered items. This is especially important in instances where there is missing item response data i.e. when there is no response maybe due to the child being very ill or restless. The weighting of the simple score is suitable when assessment is done on a very sickly child cohort and there is a high propensity of missing data. In our case, the missing data rate was very minimal (<2 %) and in most cases we had complete data even for the disabled and malnourished samples. Therefore, the benefit of weighting was not really apparent and distribution properties and sensitivity results between the simple and weighted score methods were very similar. However, the weighting of the simple score should be done if the item missingness is very minimal and can safely be classified as MAR as it shows no systematic pattern to any item or covariate. Otherwise various imputation methods should be used to recover missing item responses.

By no means does either the simple score or the weighted score account for the difference in difficulty of items or the effect of age on scores. We then considered a model based approach; the GAMLSS regression models to mitigate the effect of age which in our view was the greater disadvantage of the simple score approach. The GAMLSS regression model assuming either a Beta Binomial or Normal score distribution were used to account for the effect of age on scores as explained in Sections 5.2.2.1 and 5.2.2.2. These had the benefit of; a) flexibly modelling and summarising the expected overall raw scores and producing high quality centiles that can be used as diagnosis cut-off thresholds for development status classification. b) ensuring that the captured score and age relation is monotonically increasing with age to give meaningful score estimates.

The Z-score is a popular overall scoring approach that standardises the simple score according to age specific reference data to compute the age-standardised Z-score described in Section 5.2.2.2. The Z-score method is still based on the simple score and therefore does not differentiate item difficulty. We discovered that both the mean and variance for certain age groups that are used to compute the empirical Z-scores may be sporadic to the extent that the mean summary measures may not be monotonic with respect to age. Our results presented in Section **Error! Reference source not found.** showed that the smoothing of the empirical Z-scores using a smoothed mean and standard deviation value from a GAMLSS model of the age categories was able to ameliorate many of the empirical Z-score weaknesses specifically by; a) firstly, controlling for the sporadic variability in the Z-scores produced from empirical mean and standard deviation summaries thus avoiding the potential to misclassify the developmental status of a child as a consequence of the use of inappropriate means and standard deviations, b) secondly, smoothing ensures that the age category summary means used in the Z-score computation are monotonically increasing, c) thirdly, the GAMLSS model framework also allows for the correction of other variables apart from age that may be considered important e.g. gender, underlying disease status and social economic status.

The third approach we considered to create an overall score is the IRT framework described in Section 5.2.2.3. The 2PL IRT model specifically distinguishes itself from the previously considered overall scoring alternatives for creating scores by not assuming that each item in the tool has equal discrimination. In this child development context, we have outlined the use of the 1 PL IRT model and highlighted its weakness of assuming uniform discrimination across all items which the 2 PL IRT model relaxes. However, given that both the classical 1PL and 2PL IRT models do not adjust for age, we allowed the discrimination parameter of each item to vary with age. This approach of correcting for age was seen as a more natural approach because the probability of a child passing an item is primarily driven or dependant on age. Therefore, to ensure that the model is able to adequately differentiate response patterns and hence underlying ability, the discrimination parameter was allowed to change with respect to age. Even at the risk of losing computation efficiency due to complexity, the results discussed in Section **Error! Reference source not found.** show that the suggested 1 PL IRT extension, the 1 PL IRT monotonic spline model managed to adequately eliminate the effect of age on scores as it is accounted for directly in the score computation.

Section **Error! Reference source not found.** described the methods that served to make an objective comparison of the current scoring methods versus the suggested extensions. This was in terms of score characteristics to provide evidence of the robustness of the methods, given the highlighted limitations or poor performance of the rudimentary or classical methods to produce reliable and valid scores. The results show that our suggested scoring approaches are at least in agreement with the classical methods in terms of distributional characteristics and sensitivity. More importantly our favoured IRT model variant, the suggested 1 PL IRT monotonic spline model extension, appropriately adjusts for the effect of age that was the primary objective of this thesis.

7.3. Limitations and future research

The foregoing results chapter has revealed both promising scoring methods that we hope will quickly be assimilated into common practise as well as methods that did not really increase benefit especially in terms of sensitivity of disability classification but offered a more novel framework to address important scoring issues. Apart from the successful delivery of the research objectives, any research work should admit to its limitations and walk the reader through where to take this scoring research next. This not only makes the consideration of the future work suggestions feasible as they are well motivated by the just concluded research but also ensures they are immediately achievable and impactful. This section will only discuss future work that this work directly points to while remaining relevant with regard to accurate classification of child ability development status.

While the systematic review confirmed that there is indeed a consensus in acknowledgement of the importance of accounting for cultural aspects in child assessment tools, there were some limitations identified. Therefore, the output of the systematic review was a list of recommendations that were used to formulate the checklist summarised in Section 2.4 that outlies**Error! Reference source not found.** the properties a well-adapted child assessment tool should possess. The so far developed reporting quality criteria can be further refined by; a) developing a reporting guideline development protocol as explained by Davidoff, et al., (2008), Hopewell, et al., (2008), Altman & Moher, (2005) and Phillips, et al., (2013) having three stages for the development process. The first stage, which is the systematic review that has identified the weakness in the reporting of translation, adaptation and scoring studies, is already complete. A second stage involving a Delphi process (Norman & Olaf, 1963; Brown, 1968; Sackman, 1972; Linstone & Turoff, 1975) to reach consensus on which and how best to report the most important study details of such studies. A thirst stage to carry out the guideline development process including the pilot testing of the refined reporting guidance statement.

With the devised robust statistical framework to compute item by item age estimates and scores to assess child development that is more sensitive cross-sectionally (i.e. at one time point), we can now

look to assessing both between-child and within-child differences longitudinally. As noted by Curran, et al., (2010) this interest is motivated by the fact that the foundational goal underlying the developmental sciences is the systematic construction of a reliable and valid understanding of the course, causes of normal human development as well as the consequences of his/her interaction with the environment on the same development. We too, in this child development context are interested in both the typical and atypical patterns of ability development in the general child population from the early years until the end of puberty or adolescence. There are two general frameworks that can be used to fit suitable longitudinal ability development models to observed data within this child development context. These are the multilevel framework or the structural equation modelling (SEM) framework which have been adequately explained by Bauer & Curran, (2003), Curran, (2003), Raudenbush, (2001), Willett & Sayer, (1994) and Willett, et al., (1998).

We had the expectation that the use of more appropriate modelling approaches especially the IRT framework would stave off many of the ignored issues pertaining to allowing items to differ in difficulty or adjusting for the effect of age. We hope the promise of the IRT framework will further illuminate the link of factor analysis (FA) and item response theory (IRT). These two methods are frequently used to determine whether a set of items reliably measures an underlying latent variable in a child development context. Several authors have described the theoretical relationship between FA and IRT. For example, the work of Ten Holt, et al., (2010), Takane & de Leeuw, (1987), Kamata & Bauer, (2008), Brown, (2006) and Mehta & Taylor, (2006) have demonstrated that certain variants of FA and IRT are similar. In line with the work of Wirth & Edwards, (2007), Moustaki, et al., (2004), Glöckner-Rist & Hoijtink, (2003), Jöreskog & Moustaki, (2001) and Knol & Berger, (1991) that compared FA and IRT using empirical data, we envisage that the pros and cons of each approach can be assessed depending on one's research objectives and more so investigate how the approaches can complement each other as far as ability measurement is concerned using the now realised improved overall scores. The comparison will be in terms of the assessment of various tool assumptions like

unidimentionality, local independence, differential item functioning and quality of scores in as far as detecting disability is concerned. Our hope is to find suitable ways to be able to 'correct' or adjust for certain important covariates known to influence ability in addition to age like gender, social economic status and presence of a disease or condition on the child or its guardians within each approach.

Further, the FA and IRT relation has also been used with respect to measurement equivalence and linking for example by Meade & Lautenschlager, (2004), Raju, et al., (2002), Raju, et al., (1995) and Reise, et al., (1993). Following the current interest in comparing child development across countries by various stake holders like the World Bank and World Health Organisation (WHO) thus necessitating the requirement to objectively identify assessment items that are not influenced by culture, we will be able to outline what is the best practice in terms of; a) objectively identifying items that can appropriately assess the ability of children from different countries irrespective of their different characteristics, b) why and when a certain approach is superior or inferior to another if one has binary or continuous item data or has a cohort or longitudinal survey design. The output will be a detailed strategy decision tree that practitioners can use to motivate their choice of statistical methodology. Further, following the development of an improved scoring framework from this work that will feed into a better item identification framework, a better multiregional comparison of child ability estimates will be facilitated.

Finally, we should not under estimate the importance of early disability or delayed development diagnosis hence our motivation to offer better scoring methods. In spite of the complexity of some of our proposed extensions, we believe software development to the point of allowing non-statistical researchers to easily and practically implement the methods will go a long way in easing the tool development process, with many of the methods also doubling up to ascertain item quality. For example, see the development of the WHO Anthro software for PC version 3.2.2 (2011) using different types of statistical software such as R, SPSS, SAS, S-Plus and STATA to easily compute physical growth Z-scores. Therefore, we envisage that with more rigorous testing facilitated by an easy to use suite of

score computation software programs applicable to various forms of empirical data in different disease contexts, the superiority of features of our scoring extensions can further be affirmed which will therefore hasten their uptake.

7.4. Concluding remarks

This research has developed a statistical framework to compute more accurate item by item age estimates and for overall scoring which allows correction or adjustment for age that is known to be strongly associated with child development. This work therefore serves as a primer for investigating the influence of other important clinical and social factors on a child's developmental trajectory. We have seen that this approach gives more accurate developmental scores to child populations known to have specific characteristics. While keeping the germane of disease (and disability) diagnosis, the more sensitive scoring methods will be of benefit to community health workers looking at the developmental outcomes of children. These can also be incorporated into national and international maternal and neonatal programs, and be used to monitor and evaluate not only child development, but also health-related quality of life in other contexts at a population level. They will further allow for evidence-based evaluation of identified factors influencing disability in respective age groups as well as the measurement of impact of preventive and treatment interventions in the community.

Bibliography

Agresti, A., 2002. *Categorical Data Analysis.* 2nd ed. Florida: John Wiley & Sons, Inc.

Aitkin, M., 1987. Modelling variance heterogeneity in normal regression using GLIM. *Applications Statistics,* Volume 36, pp. 332-339.

Aitkin, M., 1999. A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics,* Volume 55, pp. 117-128.

Aitkin, M., Anderson, D., Francis, B. & Hinde, J., 1989. Statistical Modelling in GLIM. *Statistics in Medicine,* 8(11), pp. 1418-1419.

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: B. Petrov & F. Csáki, eds. *2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, September 2-8, 1971.* Budapest: Akadémiai Kiadó, pp. 267-281.

Akaike, H., 1983. Information measures and model selection. Bull Inst Int Statist, 50(1), pp. 277-290.

Akantziliotou, K., Rigby, R. & Stasinopoulos, D. M., 2002. The R implementation of Generalized Additive Models for Location Scale and Shape, in Statistical modelling in Society. Chania, Greece, Proceedings of the 17th International Workshop on statistical modelling.

Albert, A. & Anderson, A., 1984. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika,* Volume 71, pp. 1-10.

Allison, P., 2001. *Missing data.* Newbury Park, CA: Sage Publications, Inc.

Altman, D. & Moher, D., 2005. Developing guidelines for reporting healthcare research: scientific rationale and procedures. *Medicina Clinica,* 125 (Supplementary 1), pp. 8-13.

Altman, D., 1991. *Practical Statistics for Medical Research.* 1st ed. Florida: CRC Press.

Altman, D. & Royston, P., 2006. The cost of dichotomising continuous variables. *British Medical Journal,* 332(7549), p. 1080.

Anastasi, A. & Urbina, S., 1997. *Psychological Testing.* 7 ed. Michigan: Prentice Hall.

Angoff, W., 1972. *A technique for the investigation of cultural differences.* Honolulu, Paper presented at the annual meeting of the American Psychological Association, Honolulu .

Angoff, W., 1982. Use of difficulty and discrimination indices for detecting item bias. In: R. Berk, ed. *Handbook of methods for detecting test bias.* Baltimore: Johns Hopkins University Press.

Angoff, W. & Ford, S., 1973. Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement,* Volume 10, pp. 95-105.

Aranda-Ordaz, F. J., 1981. On Two Families of Transformations to Additivity for Binary Response Data. Biometrika, 68(2), pp. 357-363.

Bangirana, P. et al., 2014. Severe malarial anemia is associated with long-term neurocognitive impairment. *Clinical Infectious Disease,* 4(59), pp. 336-344.

Bauer, D. & Curran, P., 2003. Distributional assumptions of growth mixture models: implications for overextraction of latent trajectory classes. *Psychological Methods,* 8(3), pp. 338-363.

Beaton, D. E. & Guillemin, F., 2000. Guidelines for the Process of Cross-Cultural Adaptation of Self-Report Measures. *Spine,* 25(24), p. 3186–3191.

Beck, J. & Shultz, E., 1986. The use of relative operating characteristic (ROC) curves in test performance evaluation. *Archives of pathology & laboratory medicine,* 110(1), pp. 13-20.

Beguin, A. & Glas, C., 2001. MCMC Estimation and Some Model-Fit Analysis of Multidimensional IRT Models. Psychometrika, 66(4), pp. 541-562.

Bickel, P. J. & Doksum, K. A., 2011. An Analysis of Transformations Revisited. *Journal of the American Statistical Association,* 76(374), pp. 296-311.

Blom, A. G., Jackle, A. & Lynn, P., 2010. The Use of Contact Data in Understanding Cross-National Differences in Unit Nonresponse. In: J. A. Harkness, et al. eds. *Survey Methods in Multinational, Multiregional, and Multicultural Contexts.* Hoboken, NJ, USA: John Wiley & Sons, Inc, pp. 335-354.

Bock, D. R. & Aitkin, M., 1981. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46(4), pp. 443-459.

Bonett, D. G. & Price, R. M., 2005. Inferential Methods for the Tetrachoric Correlation Coefficient. *Journal of Educational and Behavioral Statistics,* 30(2), pp. 213-225.

Borsa, J. C., Damásio, B. F. & Bandeira, D. R., 2012. Cross-cultural adaptation and validation of psychological instruments: some considerations. *Paideia,* 22(53), pp. 423-432.

Bossuyt, P. et al., 2015. *An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies,* Oxford: Equator Network.

Bradley, A. P., 1997. The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern recognition,* 30(7), pp. 1145-1159.

Briggs, A., Wolstenholme, J. & Clarke, P., 2003. Missing presumed at random: cost-analysis of incomplete data. *Health Economics,* Volume 12, pp. 377-392.

Brown, B. B., 1968. *Delphi Process: A Methodology Used for the Elicitation of Opinions of Experts.,* Santa Monica, California: The RAND Corporation.

Brown, T. A., 2006. *Confirmatory Factor Analysis for Applied Research.* 2nd ed. New York: Guilford.

Carey, V., Yong, F., Frenkel, L. & McKinney, R., 2004. Growth velocity assessment in paediatric AIDS:smoothing, penalized quantile regression and the definition of growth failure. *Statistics in Medicine,* 23(3), pp. 509-526.

Carmines, E. G. & Zeller, R. A., 1979. *Reliability and Validity Assessment.* London: Sage Publications.

Chang, C. Y., 2001. *Cross-Cultural Assessment: A Call for Test Adaptation.* [Online] Available at: http://aac.ncat.edu/newsnotes/y99sum1.html [Accessed 7 August 2012].

Cheung, Y.-B.et al., 2008. Comparison of four statistical approaches to score child development: a study of Malawian children. *Tropical Medicine and International Health,* 13(8), pp. 987-993.

Chin-Lun Hung, G. et al., 2015. Socioeconomic disadvantage and neural development from infancy through early childhood. *International Journal of Epidemiology,* 44(6), pp. 1889-1899.

Cleveland, W. S. & Devlin, S. J., 1988. Locally weighted regression: an approach to regression analysis by local fitting. American Statistical Association, 83(403), pp. 596-610.

Cole, T., 1988. Fitting smoothed centile curves to reference data (with discussion). *Journal of the Royal Statistical Society, Series A,* Volume 151, pp. 385-418.

Cole, T., 1990. The LMS method for constructing normalized growth standards. *European journal of Clinical Nutrition,* 44(1), pp. 45-60.

Cole, T., 1994a. Do growth chart centiles need a face lift?. *BMJ,* Volume 302, pp. 641-642.

Cole, T., 1994b. Growth charts for both cross-sectional and longitudinal data. *Statistics in Medicine,* Volume 13, pp. 2477-2492.

Cole, T., 1998. Presenting information on growth distance and conditional velocity in one chart: Practical issues of chart design. *Statistics in Medicine,* 17(23), pp. 2697-2707.

Cole, T., 2003. The secular trend in human physical growth: A biological view. *Economics & Human Biology,* pp. 161-168.

Cole, T., 2008. The WHO Child Growth Standards and current Western growth references. *Breastfeeding Review,* 16(3), pp. 13-16.

Cole, T. et al., 2008. Nonlinear growth generates age changes in the moments of the frequency distribution: The example of height in puberty. *Biostatistics,* 9(1), pp. 159-171.

Cole, T. & Green, P., 1992. Smoothing reference centile curves: the LMS method and penalized likelihood. *Statistics in Medicine,* 11(10), pp. 1305-1319.

Cole, T., Wright, C., Williams, A. & RCPCH Growth Chart Expert Group, 2012. Designing the new UK-WHO growth charts to enhance assessment of growth around birth. *Archives of disease in Childhood.Fetal and Neonatal Edition,* pp. F219-22.

Connelly, R., 2013. *Millennium Cohort Study Data Note: Interpreting Test Scores,* London: Centre for Longitudinal Studies.

Cook, D., 1959. A replication of Lord's study on skewness and kurtosis of observed test-score distributions. *Educational and Psychological Measurement,* Volume 19, pp. 81-87.

Cox, D. & Snell, E., 1969. *Analysis of Binary Data.* 1st ed. London: CRC Press.

Craven, P. & Wahba, G., 1979. Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation. *Numerische Mathematik,* Volume 31, pp. 377-403.

Cronbach, L. & Furby, L., 1970. How should we measure 'change' - or should we?. *Psychological Bulletin,* Volume 74, pp. 68-80.

Curran, P., 2003. Have multilevel models been structural equation models all along?. *Multivariate Behavioral Research,* Volume 38, pp. 529-569.

Curran, P. J., Obeidat, K. & Losardo, D., 2010. Twelve Frequently Asked Questions About Growth Curve Modeling. *Journal of Cognitive Development,* 11(2), pp. 121-136.

Davidoff, F. et al., 2008. Publication guidelines for improvement studies in health care: evolution of the SQUIRE Project. *Annals of Internal Medicine,* 149(9), pp. 670-676.

De Boeck, P. & Wilson, M., 2004. *Explanatory Item Response Models; A Generalized Linear and Nonlinear Approach.* 1st ed. New York: Springer.

DeCarlo, L. T., 1997. On the meaning and use of kurtosis. American Psychological Association, 2(3), pp. 292-307.

Dobson, A. J. & Barnett, A., 2008. *An Introduction to Generalized Linear Models.* 3rd ed. London: Chapman & Hall/CRC Texts in Statistical Science.

Dong, Y. & Peng, C.-Y. J., 2013. Principled missing data methods for researchers. *SpringerPlus,* 2(222).

Dorans, N. J. & Kulick, E., 1986. Demonstrating the Utility of the Standardization Approach to Assessing Unexpected Differential Item Performance on the Scholastic Aptitude Test. Journal of Educational Measurement, 23(4), pp. 355-368.

Downing, S. M. & Haladyna, T. M., 2006. Handbook of Test Development. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Drachler, M., Marshall, T. & de Carvalho, J. L., 2007. A continuous-scale measure of child development for population-based epidemiological surveys: A preliminary study using item response theory for Denver Test. *Paediatric and Perinatal Epidemiology,* 21(2), pp. 138-153.

Drasgow, F., 1988. Polychoric and polyserial correlations. In: L. Kotz & N. L. Johnson, eds. Encyclopedia of Statistical Sciences. 1st ed. New York: Wiley, pp. 69-74.

Dunn, G., 1992. Design and Analysis of Reliability Studies: The Statistical Evaluation of Measurement Errors. 2nd ed. New York: Oxford University Press.

Efron, B., 1979. Bootstrap methods: Another look at the Jackknife. *The Annals of Statistics,* 7(1), pp. 1-26.

Efron, B. & Tibshirani, R. J., 1993. *An Introduction to the Bootstrap.* 1st ed. Boca Raton, Florida: Chapman & Hall/CRC.

Ekstrom, J., 2011. A Generalized Definition of the Polychoric Correlation Coefficient. *Department of Statistics Papers,* pp. 1-24.

Elliott, C., Smith, P. & McCulloch, K., 1996. *British Ability Scales Second Edition (BAS II). Administration and Scoring Manual,* London: Nelson.

Embreston, S. E. & Reise, S. P., 2000. *Item response theory for psychologists.* 1st ed. Mahwah, NJ: Erlbaum Associates.

Engle, P., Bentley, M. & Pelto, G., 2000. The role of care in nutrition programmes: current research and a research agenda. *Proceedings of the Nutrition Society,* 59(1), pp. 25-35.

Engs, R. C., 1997. Construct Validity and Re-assessment of the Reliability of the Health Concern Questionnare in Advances in Health Education/Current Research. AMS Press Inc, Volume 4, pp. 303-311.

Fawcett, T., 2006. An introduction to ROC analysis. *Pattern Recognition Letters,* Volume 27, pp. 861-874.

Feigelman, S., 2011. *Middle childhood.* 19 ed. Philadelphia: Nelson Textbook of Pediatrics.

Fenske, N., Burns, J., Hothorn, T. & Rehfuess, E. A., 2013. Understanding Child Stunting in India: A Comprehensive Analysis of Socio-Economic, Nutritional and Environmental Determinants Using Additive Quantile Regression. *PLOS ONE,* 8(11).

Flegal, K. M. & Cole, T. J., 2013. Construction of LMS parameters for the Centers for Disease Control and Prevention 2000 growth charts, U.S.A: Natl Health Stat Report.

Forster, J. J., McDonald, J. W. & Smith, P. W., 1996. Monte Carlo Exact Conditional Tests for Log-Linear and Logistic Models. Journal of the Royal Statistical Society. Series B (Methodological), 58(2), pp. 445-453.

Frankenburg, W. et al., 1992. The Denver II: a major revision and restandardization of the Denver Developmental Screening Test. *Pediatrics,* 89(1), pp. 91-97.

Galea, S. et al., 2011. Estimated Deaths Attributable to Social Factors in the United States. *American Journal of Public Health,* 101(8), p. 1456–1465.

Galler, J., Ramsey, F., Salt, P. & Archer, E., 1987. Long-term effects of early kwashiorkor compared with marasmus. III. Fine motor skills. *Journal of pediatric gastroenterology and nutrition,* 6(6), pp. 855-859.

Galler, J. R. et al., 2013. Malnutrition in the First Year of Life and Personality at Age 40. *The Journal of Child Psychology and Psychiatry,* 54(8), pp. 911-919.

Geert, M. & Geert, V., 2005. Models for Discrete Longitudinal Data. 1st ed. New York: Springer Series in Statistics.

Geisinger, K., 1994. Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment,* Volume 6, pp. 304-312.

Ghosh, D., 2007. Incorporating monotonicity into the evaluation of a biomarker. Biostatistics, 8(2), pp. 402-413.

Giampiero, M. & Rosalba, R., 2010. Penalised regression splines: theory and application. Statistical Methods in Medical Research, 19(2), pp. 107-125.

Gladstone, M. et al., 2008. Can Western developmental screening tools be modified for use in a rural Malawian setting?. *Disease in Childhood,* 93(1), pp. 23-29.

Gladstone, M. et al., 2010a. The Malawi Development Assessment Tool (MDAT): The Creation, Validation, and Reliability of a Tool to Assess Child Development in Rural African Settings. *PLoS Medicine,* 7(5), pp. 1-14.

Gladstone, M. et al., 2010b. Perspectives of normal child development in rural Malawi – a qualitative analysis to create a more culturally appropriate developmental assessment tool. *Childcare, health and development,* 36(3), pp. 346-353.

Glaser, R., 1963. Instructional Technology and the Measurement of Learning Outcomes. *American Psychologist,* Issue 18, pp. 519-521.

Glöckner-Rist, A. & Hoijtink, H., 2003. The Best of Both Worlds: Factor Analysis of Dichotomous Data Using Item Response Theory and Structural Equation Modeling. Structural Equation Modelling, 10(4), pp. 544-565.

Goldstein, H. & Lewis, T., 1996. *Assessment: Problems, Developments and Statistical Issues, a volume of expert contributions.* 1st ed.:John Wiley & Sons Inc.

Graham, J., 2009. Missing data analysis: making it work in the real world. *Annual Review of Psychology,* Volume 60, pp. 549-576.

Grantham-McGregor, S. et al., 2007. Developmental potential in the first 5 years for children in developing countries. *The Lancet,* Issue 369, pp. 60-70.

Greenfield, P. M., 1997. You can't take it with you: Why ability assessments don't cross cultures. *American Psychologist,* Volume 52, p. 1115–1124.

Greenland, S., 1995a. Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. (Commentary). *Epidemiology,* 6(4), pp. 450-454.

Greenland, S., 1995b. Problems in the average-risk interpretation of categorical dose-response analyses. *Epidemiology,* 6(5), pp. 563-565.

Greenland, S., 1995c. Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. *Epidemiology,* 6(4), pp. 356-365.

Grieve, K. W., 1992. Play based assessment of the cognitive abilities of young children, Pretoria: Thesis/dissertation.

Hambleton, R., 1994. Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment,* Volume 10, pp. 229-244.

Hambleton, R. & Kanjee, A., 1995. Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptation. *European Journal of Psychological Assessment,* Volume 11, pp. 147-157.

Hambleton, R. K. & Russell, J. W., 1993. Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development. *Educational Measurement: Issues and Practice,* pp. 38-47.

Hand, D. J., 2009. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning,* Volume 77, pp. 103-123.

Hanley, J. & McNeil, B., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology,* 143(1), pp. 29-36.

Harkness, J. et al., 2010. *Survey methods in multinational, multicultural and multiregional contexts.* New Jersey: John Wiley & Sons.

Harvey, A., 1976. Estimating regression models with multiplicative heteroscedasticity. *Econometrica: Journal of the Econometric Society,* Volume 41, pp. 461-465.

Harzing, A.-W., 2006. Response styles in cross-national survey research: a 26-country study. *International Journal of Crosscultural Management,* 6(2).

Hastie, T. & Tibshirani, R., 1986. Generalized Additive Models. *Statistical Science,* 1(3), pp. 297-318.

Hastie, T. & Tibshirani, R., 1990. *Generalized Additive Models.* 1st ed. London: Chapman and Hall/CRC.

Hastie, T. J., Tibshirani, R. J. & Friedman, J., 2001. The Elements of Statistical Learning: Data Mining, Inference and Prediction. 2nd ed. New York: Springer Series in Statistics.

Hedeker, D. & Gibbons, R. D., 2006. *Longitudinal data analysis.* New York: Wiley Inc.

Hedeker, D. & Mermelstein, R. J., 1998. A Multilevel Thresholds of Change Model for Analysis of Stages of Change Data. *Multivariate Behavioural Research,* 33(4), pp. 427-455.

Hedeker, D., Mermelstein, R. J. & Flay, B. R., 2006. Application of Item Response Theory Models for Intensive Longitudinal Data. In: T. A. Walls & J. L. Schafer, eds. *Models for Intensive Longitudinal Data.* Oxford: Oxford University Press, p. Chapter 4.

Heinze, G. & Schemper, M., 2002. A solution to the problem of separation in logistic regression. *Statistics in Medicine,* 21(16), pp. 2409-2419.

Henning, G., 1989. Meanings and implications of the principle of local independence. *Language Testing,* 6(1), pp. 95-108.

Hens, N., 2005. *Non- and Semi-parametric Techniques for Handling Missing Data,* Hasselt: Faculteit Wetenschappen.

Hopewell, S. et al., 2008. CONSORT for reporting randomized controlled trials in journal and conference abstracts: explanation and elaboration. *PLoS Med,* 5(e20).

Hubley, A. M. & Zumbo, B. D., 2013. Psychometric Characteristics of Assessment Procedures: An Overview. In: K. F. Geisinger, ed. *APA Handbook of Testing and Assessment in Psychology.* Washington, DC: American Psychological Association, pp. 3-19.

Jacobusse, G. & van Buuren, S., 2007. Computerized adaptive testing for measuring development of young children. *Statistics in Medicine,* 26(13), pp. 2629-2638.

Jacobusse, G., van Buuren, S. & Verkerk, P., 2006. An interval scale for development of children aged 0-2 years. *Statistics in Medicine,* Volume 25, pp. 2272-2283.

Janes, H., Longton, G. & Pepe, M., 2009. Accommodating Covariates in ROC Analysis. *Stata Journal,* 9(1), pp. 17-39.

Jayasinghe, S., 2011. Conceptualising population health: from mechanistic thinking to complexity science. *Emerging Themes in Epidemiology,* 8(2).

Jiang, D. R. & Powell, W. B., 2015. An Approximate Dynamic Programming Algorithm for Monotone Value Functions. Operations Research, 63(6), pp. 1489-1511.

Joffe, H. et al., 2012. Lifetime history of depression and anxiety disorders as a predictor of quality of life in midlife women in the absence of current illness episodes. *Archives of general psychiatry,* pp. 484-492.

Jöreskog, K. G. & Moustaki, I., 2001. Factor analysis of ordinal variables: a comparison of three approaches. *Multivariate Behavioural Research,* 36(3), pp. 347-387.

Kamata, A. & Bauer, D. J., 2008. A Note on the Relation between Factor Analytic and Item Response Theory Models. *Structural Equation Modeling,* Volume 15, pp. 136-153.

Kammerer, J. et al., 2007. Adherence in patients in dialysis strategies for success. *Nephrology Nursing Journal,* 34(5), pp. 479-486.

Karras, D., 1997. Statistical methodology: II. Reliability and validity assessment in study design, Part B. *Academic Emergency Medicine,* 4(2), pp. 144-147.

Kim, J.-S. & Bolt, D. M., 2007. Estimating Item Response Theory Models Using Markov Chain Monte Carlo Methods. *Educational Measurement: Issues and Practice,* pp. 38-51.

Kline, R., 2011. *Principles and practice of structural equation modelling.* 3rd ed. New York: Guildford Press.

Kline, R. B., 2005. *Principles and Practice of Structural Equation Modeling.* 2nd ed. New York: Guilford Press.

Kline, T. J., 2005. Psychological Testing: A Practical Approach to Design and Evaluation. 2nd ed. New Delhi: Sage Publications.

Klinnert, M. D. et al., 2001. Onset and persistence of childhood asthma: predictors from infancy. *Pediatrics,* 108(4).

Kneib, T., 2013. Beyond mean regression. *Statistical Modelling,* 13(4), pp. 275-303.

Knol, D. L. & Berger, M. P., 1991. Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioural Research,* 26(3), pp. 457-477.

Kolen, M. J. & Brennan, R. L., 2014. *Test Equating, Scaling, and Linking.* 1st ed. New York: Springer.

Krieger, N., 1994. Epidemiology and the web of causation: has anyone seen the spider?. *Social Science and Medicine,* 39(7), pp. 887-903.

Lancaster, G. A., 2009. Statistical issues in the assessment of health outcomes in children: a methodological review. *Journal of the Royal Statistical Society: Series A,* 172(4), pp. 707-727.

Lancaster, G. A., Dodd, S. & Williamson, P. R., 2004. Design and analysis of pilot studies: recommendations for good practice. *Journal of Evaluation in Clinical Practice,* 10(2), pp. 307-312.

Langer, M. M. et al., 2007. Item response theory detects differential item functioning between healthy and ill children in QoL measures. *Clinical Epidemiology,* 61(3), pp. 268-276.

Lazarsfeld, P. F. & Henry, N. W., 1968. *Latent structure analysis.* New York: Houghton, Mifflin.

Lee, S., Poon, W. & Bentler, P., 1995. A two-stage estimation of structural equation models with continuous and polytomous variables. *British Journal of Mathematical and Statistical Psychology,* Volume 48, pp. 339-358.

Lemeshow, S. & Hosmer, D. J., 1982. A review of goodness of fit statistics for use in the development of logistic regression models. *American Journal of Epidemiology,* 115(1), pp. 92-106.

Lesaffre, E. & Albert, A., 1989. Partial Separation in Logistic Discrimination. *Journal of the Royal Statistical Society. Series B (Methodological),* 51(1), pp. 109-116.

Leys, C. et al., 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. Journal of Experimental Social Psychology, 49(4), pp. 764-766.

Liao, T. F., 1994. *Interpreting Probability Models; Logit, Probit, and other Generalized Linear Models.* Urbana-Champaign: Sage.

Li, J. & Karakowsky, L., 2001. Do we see eye-to-eye? Implications of cultural differences for cross-cultural management research and practice. *Journal of Psychology,* 135(5), pp. 501-517.

Linstone, H. A. & Turoff, M., 1975. *The Delphi Method: Techniques and Applications.* [Online] Available at: http://is.njit.edu/pubs/delphibook/ [Accessed 16 May 2016].

Little, R. J. & Rubin, R. B., 1991. Statistical Analysis with Missing Data. *Journal of Educational Statistics,* 16(2), pp. 150-155.

Little, R. & Rubin, D., 1987. *Statistical analysis with missing data.* New York: Wiley Inc.

Liu, D. & Zhou, X.-H., 2013. Covariate adjustment in estimating the area under ROC curve with partially missing gold standard. *Biometrics,* 69(1), pp. 91-100.

Lord, F., 1955. A survey of observed test-score distributions with respect to skewness and kurtosis. *Educational and Psychological Measurement,* Volume 15, pp. 383-389.

Lord, F. M. & Novick, M. R., 1968. *Statistical Theories of Mental Test Scores.* Reading: Addison-Wesley.

Madow, W., Nisselson, H. & Olkin, I., 1983. *Incomplete Data in sample Surveys, Volume 1: Report and Case Studies,* New York: America Press.

Magis, D., Beland, S., Tuerlinckx, F. & De Boeck, P., 2010. A general framework and an R package for the detection of dichotomous differential item functioning. *Behaviour Research Methods,* 42(3), pp. 847-862.

Mantel, N. & Haenszel, W., 1959. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of National Cancer Institute,* Volume 22, pp. 719-748.

Maulik, P. & Darmstadt, G. L., 2007. Childhood Disability in Low- and Middle-Income Countries: Overview of Screening, Prevention,Services, Legislation, and Epidemiology. *Pediatrics,* 120 (Suppliment).

May, S. & Bigelow, C., 2005. Modeling Nonlinear Dose-Response Relationships in Epidemiologic Studies: Statistical Approaches and Practical Challenges. *Dose Response,* 3(4), pp. 474-490.

Maydeu-Olivares, A., 2015. Evaluating the Fit of IRT Models. In: S. P. Reise & D. A. Revicki, eds. Handbook of Item Response Theory Modeling; Applications to Typical Performance Assessment. New York and London: Routledge, pp. 111-127.

McCulloch, C. E., Searle, S. R. & Neuhaus, J. M., 2008. *Generalized Linear and Mixed Models.* 2nd ed. Alexandria VA: Wiley.

Meade, A. W. & Lautenschlager, G. J., 2004. *Same Question, Different Answers: CFA and Two IRT Approaches to Measurement Invariance.* Chicago, IL, Symposium presented at the 19th Annual Conference of the Society for Industrial and Organizational Psychology.

Mehta, P. & Taylor, W., 2006. *On the relationship between item response theory and factor analysis of ordinal variables: Multiple group case.* HEC Montreal, Canada, Paper presented at the 71st annual meeting of the Psychometric Society.

Micerri, T., 1989. The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin,* Volume 105, pp. 156-166.

Moustaki, I., Joreskog, K. & Mavridis, D., 2004. Factor models for ordinal variables with covariate effects on the manifest and latent variables: a comparison of LISREL and IRT approaches. *Structural Equation Modelling,* 11(4), pp. 487-513.

Nakai, M. & Weiming, K., 2011. Review of the Methods for Handling Missing Data in Longitudinal Data Analysis. *International Journal of Mathematical Analysis,* 5(1), pp. 1-13.

Nelder, J., 1992. Joint modelling of the mean and dispersion. In: *Multivariate Statistical Modelling Based on Generalized Linear Models.* Amsterdam: Springer Series in Statistics, pp. 263-272.

Nelder, J. & Pregibon, D., 1987. An extended quasi-likelihood function. *Biometrika,* Volume 74, pp. 221-232.

Neter, J., Kutner, M. H., Nachtsheim, C. J. & Wasserman, W., 1996. *Applied Linear Statistical Models.* 4th ed. Chicago: The McGraw-Hill Companies, Inc.

Newton, P. E. & Shaw, S. D., 2014. Validity in Educational and Psychological Assessment. 1st ed. Cambridge: American Educational Research Association.

Norman, D. & Olaf, H., 1963. An Experimental Application of the Delphi Method to the use of experts. *Management Science,* 9(3), pp. 458-467.

Olive, D. J., 2013. Plots for Generalized Additive Models. Communications in Statistics - Theory and Methods, 42(18), pp. 3310-3328.

Ozer, D. & Reise, S., 1994. Personality assessment. *Annual Review of Psychology,* Volume 45, pp. 357-388.

Pastor, R. & Guallar, E., 1998. Use of Two-segmented Logistic Regression to Estimate Change-points in Epidemiologic Studies. *American Journal of Epidemiology,* 148(7), pp. 631-642.

Pepe, M., Longton, G. & Janes, H., 2009. Estimation and Comparison of Receiver Operating Characteristic Curves. *Stata Journal,* 9(1), p. 1.

Pérez-García, G., Guzmán-Quevedo, O., Da Silva Aragão, R. & Bolaños-Jiménez, F., 2016. Early malnutrition results in long-lasting impairments in pattern-separation for overlapping novel object and novel location memories and reduced hippocampal neurogenesis. *Scientific Reports.*

Petscher, Y. & Logan, J. A., 2014. Quantile Regression in the Study of Developmental Sciences. *Child Development,* 85(3), pp. 861-881.

Pfeifer, L., Queiroz, M. A., Santos, J. L. & Stagnitti, K. E., 2011. Cross-cultural adaptation and reliability of Child-Initiated Pretend Play Assessment (ChIPPA). *The Canadian Journal of Occupational Therapy,* 78(3), pp. 187-195.

Phillips, A. C. et al., 2013. Protocol for development of the guideline for reporting evidence based practice educational interventions and teaching (GREET) statement. *BMC Medical Education,* 13(9).

Pope, C., Ziebland, S. & Mays, . N., 2000. Qualitative research in health care; Analysing Qualitative Data. BMJ, Volume 320, pp. 114-116.

Pya, N., 2012. *SCAM:Shape Constrained Additive Models: R package version 1.1-5.* [Online]
Available at: http://cran.r-project.org/web/packages/scam/
[Accessed 5 January 2012].

Pya, N. & Wood, S., 2015. Shape Constrained Additive Models. *Statistics and Computing,* Volume 25, pp. 543-559.

Raju, N., Laffitte, L. & Byrne, B., 2002. Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology,* 87(3), pp. 517-529.

Raju, N., van der Linden, W. & Fleer, P., 1995. IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement,* 19(4), pp. 353-368.

Rasch, G., 1960. *Probabilistic models for some intelligence and attainment tests.* 1st ed. Copenhagen: Nielson and Lydiche.

Raudenbush, S., 2001. Toward a coherent framework for comparing trajectories of individual change. In: L. Collins & A. Sayer, eds. *Best methods for studying change.* Washington, DC: The American Psychological Association, pp. 33-64.

Reise, S., Widaman, K. & Pugh, R., 1993. Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin,* 114(3), pp. 552-566.

Rigby, R. & Stasinopoulos, D., 1996b. Mean and dispersion additive models. In: W. Härdle & M. G. Schimek, eds. *Statistical Theory and Computational Aspects of Smoothing.* Heidelberg: Physica, pp. 215-230.

Rigby, R. & Stasinopoulos, D., 2005. Generalized additive models for location, scale and shape (with discussion). Journal of the Royal Statistical Society: Applied Statistics-Series C, Volume 54, pp. 507-554.

Rigby, R. & Stasinopoulos, D., 2006. Using the Box-Cox t distribution in GAMLSS to model skewness and kurtosis. Statistical Modelling, Volume 6, pp. 209-229.

Rigby, R., Stasinopoulos, D. & Voudouris, V., 2013. Discussion: A comparison of GAMLSS with quantile regression. *Statistical Modelling,* 13(4), pp. 335-348.

Rigby, R. & Stasinopoulos, R., 1996a. A semi-parametric additive model for variance heterogeneity. *Statistics and Computing,* Volume 6, pp. 57-65.

Rijmen, F., Francis, T., De Boeck, P. & Kuppens, P., 2003. A nonlinear mixed model framework for item response theory. American Psychological Association, 8(2), pp. 185-205.

Ripley, B. D., 1996. Pattern Recognition and Neural Networks. 1st ed. Cambridge: Cambridge University Press.

Ross, S., 1998. *A First Course in Probability.* 5th ed. New Jersey: Prentice Hall.

Royston, P. & Altman, D., 1994. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). *Applied Statistics,* Volume 43, pp. 429-467.

Royston, P., Altman, D. & Sauerbrei, W., 2006. Dichotomizing continuous predictors in multiple regression: a bad idea. Statistics in Medicine, 25(1), pp. 127-141.

Rubin, D., 1987. *Multiple imputation for nonresponse in surveys.* New York: Wiley Inc.

Rubin, D. B., 1988. *An overview of multiple imputation,* One Oxford Street, Cambridge: Harvard University.

Rue, H. & Held, L., 2005. *Gaussian Markov random fields: theory and applications, vol.104.* London: Chapman & Hall/CRC.

Sabanathan, S., Wills, B. & Gladstone, M., 2015. Child development assessment tools in low-income and middle-income countries: how can we use them more appropriately?. *Disease in Childhood,* 100(5), pp. 482-488.

Sackman, H., 1972. *Delphi Assessment: Expert Opinion, Forecasting and Group Process,* California: RAND Corporation.

Schafer, J. L. & Graham, J. W., 2002. Missing Data: Our View of the State of the Art. *Psychological Methods,* 7(2), pp. 147-177.

Schmeiser, C. & Welch, C., 2006. Test Development. In: L. Brennan, ed. *Educational Measurement.* Washington DC: American Council on education (ACE)/Praeger, pp. 307-353.

Schwarz, G. E., 1978. Estimating the dimension of a model. Annals of Statistics, 6(2), pp. 461-464.

Seaman, S. & White, I., 2013. Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research,* 22(3), pp. 278-295.

Shapiro, S. S. & Wilk, M. B., 1965. An Analysis of Variance Test for Normality (complete samples). Biometrika, 52(3-4), pp. 591-611.

Sheng, Y. & Wikle, C., 2009. Bayesian IRT Models Incorporating General and Specific Abilities. *Behavirometrika,* Volume 36, pp. 27-48.

Sheskin, D. J., 2011. Handbook of parametric and nonparametric statistical procedures. 5th ed. Boca Raton, FL: Chapman & Hall/CRC.

Siegal, S. & Castellan, N. J., 1988. *Nonparametric statistics for behavioural science.* 2nd ed. New York: McGraw-Hill.

Sireci, S. G., 2005. Unlabeling the Disabled: A Perspective on Flagging Scores from Accommodated Test Administrations. *Educational Researcher,* 34(1), pp. 3-12.

Smith, P., 1979. Splines as a useful and convenient statistical tool. *The American Statistician,* 33(2), pp. 57-62.

Smith, W. et al., 2001. Risk factors for age-related macular degeneration: Pooled findings from three continents. *Ophthalmology,* 104(4), pp. 697-704.

Smit, J. et al., 2006. Translation and cross-cultural adaptation of a mental health battery in an African setting. *African Health Sciences,* 6(4), pp. 215-222.

Stigler, S. M., 1982. A Modest Proposal: A New Standard for the Normal. *The American Statistician,* 36(2), pp. 137-138.

Streiner, D. L. & Norman, G. R., 2008. *Health Measurement Scales: A practical guide to their development and use.* 4th ed. New York: Oxford University press.

Stroup, W. W., 2012. *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications.* 1st ed. London: CRC Press.

Syed, A. S., 2013. A brief review of risk-factors for growth and developmental delay among preschool children in developing countries. *Advanced Biomedical Research,* 2(91).

Tabachnick, B. G. & Fidell, L. S., 2012. *Using Multivariate Statistics.* 6th ed.:Pearson.

Tacq, J., 1997. *Multivariate Analysis Techniques in Social Science Research.* London: Sage.

Takane, Y. & de Leeuw, J., 1987. On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika,* 52(3), pp. 393-408.

ten Holt, J. C., van Dujin, M. A. & Boomsma, A., 2010. Scale construction and evaluation in practice: A review of factor analysis versus item response theory applications. *Psychological Test and Assessment Modeling,* 52(3), pp. 272-297.

Titman, A. C., Lancaster, G. A. & Clover, A. F., 2013. Item response theory and structural equation modelling for ordinal data: Describing the relationship between KIDSCREEN and Life-H. *Statistical Methods in Medical Research,* 0(0), pp. 1-33.

Tsai, A. H.-L., McClelland, M. M., Pratt, C. & Squires, J., 2006. Adaptation of the 36-Month Ages and Stages Questionnaire in Taiwan: Results From a Preliminary Study. *Journal of Early Intervention,* 28(3), pp. 213-225.

Tzavidis, N., Salvati, N., Schmid, T. & Flouri, E., 2016. Longitudinal analysis of the strengths and difficulties questionnaire scores of the Millennium Cohort Study children in England using M-quantile random-effects regression. *Journal of the Royal Statistical Society: Series A,* 179(2), pp. 427-452.

UNICEF, 2009. *The State of the World's Children Special Edition: Celebrating 20 Years of the Convention on the Rights of the Child,* Geneva: UNICEF.

United Nations, 1989. Convention on the Rights of the Child. *Treaty Series*, 20 November, Volume 1577, p. 3.

United Nations, 2013. *World Economic Situation and Prospects 2013,* New York: United Nations.

van der Linden, W. J. & Hambleton, R. K., 1997. *Handbook of Modern Item Response Theory.* 1st ed. New York: Springer.

von Davier, A. A., 2011. *Statistical Models for Test Equating, Scaling, and Linking.* 1 st ed. NJ: Springer.

Wahba, G., 1990. *Spline Models for Observational Data.* 1st ed. Philadlphia, Pennsylvania: Society for Industrial and Applied Mathematics.

Walker, S. P. et al., 2007. Child development: risk factors for adverse outcomes in developing countries. *The Lancet,* 369(9556), pp. 145-157.

Wang, Y. & Chen, H.-J., 2012. Use of Percentiles and Z -Scores in Anthropometry. In: *Handbook of Anthropometry: Physical Measures of Human Form in Health and Disease.* Baltimore: Springer, pp. 29-48.

Watson, N. & Wooden, M., 2012. The HILDA Survey: a case study in the design and development of a successful household panel study. *Longitudinal and Life Course Studies,* 3(3), pp. 369-381.

Wei, Y., Pere, A., Koenker, R. & He, X., 2006. Quantile regression methods for reference growth charts. *Statistics in Medicine,* 25(8), pp. 1369-1382.

WHO, 2006. WHO Child Growth Standards: Length/Height-for-Age, Weight-for-Age, Weight-for-Length, Weight-for-Height and Body Mass Index-for-Age, Geneva: WHO.

WHO, 2007. *The world health report,* Geneva: WHO.

WHO, 2011. *Monitoring maternal, newborn and child health: Understanding key progress indicators,* Geneva: World Health Organization.

WHO, 2012. *Process of translation and adaptation of instruments.* [Online]
Available at: http://www.who.int/substance_abuse/research_tools/translation/en/
[Accessed 5 January 2012].

WHO, W. H. O., 2011. *World report on disability,* Geneva: World Health Organisation and World Bank.

Wijedasa, D., 2012. Developmental screening in context: adaptation and standardization of the Denver Developmental Screening Test-II (DDST-II) for Sri Lankan children. *Child Care Health Development,* 38(6), pp. 889-899.

Willett, J. & Sayer, A., 1994. Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin,* Volume 116, pp. 363-381.

Willett, J., Singer, J. & Martin, N., 1998. The design and analysis of longitudinal studies of development and psychopathology in context: Statistical models and methodological recommendations. *Development and Psychopathology,* 10(2), pp. 395-426.

Wirth, R. & Edwards, M., 2007. Item factor analysis: current approaches and future directions. *Psychological Methods,* 12(1), pp. 58-79.

Wittchen, H., 1994. Reliability and validity studies of the WHO--Composite International Diagnostic Interview (CIDI): a critical review. *Journal of Psychiatric Research,* 28(1), pp. 57-84.

Wood, S. N., 2006. *Generalized Additive Models: An Introduction with R.* 1st ed. New York: Chapman & Hall/CRC.

Wood, S. N., 2008. Fast stable direct fitting and smoothness selection for generalized additive models. *Journal of Royal Statistical Society: Statistical Methodology-Series B,* 70(3), pp. 495-518.

Wood, S. N., Pya, N. & Safken, B., 2016. Smoothing Parameter and Model Selection for General Smooth Models. Journal of the American Statistical Association, 111(516), pp. 1548-1563.

Yu, K., Lu, Z. & Stander, J., 2003. Quantile regression: applications and current research areas. *Journal of the Royal Statitical Society: Series D (The Statistician),* 52(3), pp. 331-350.

Zhao, L. & Lipsitz, S., 1992. Designs and analysis of two-stage studies. *Statistics in Medicine,* Volume 11, pp. 769-782.

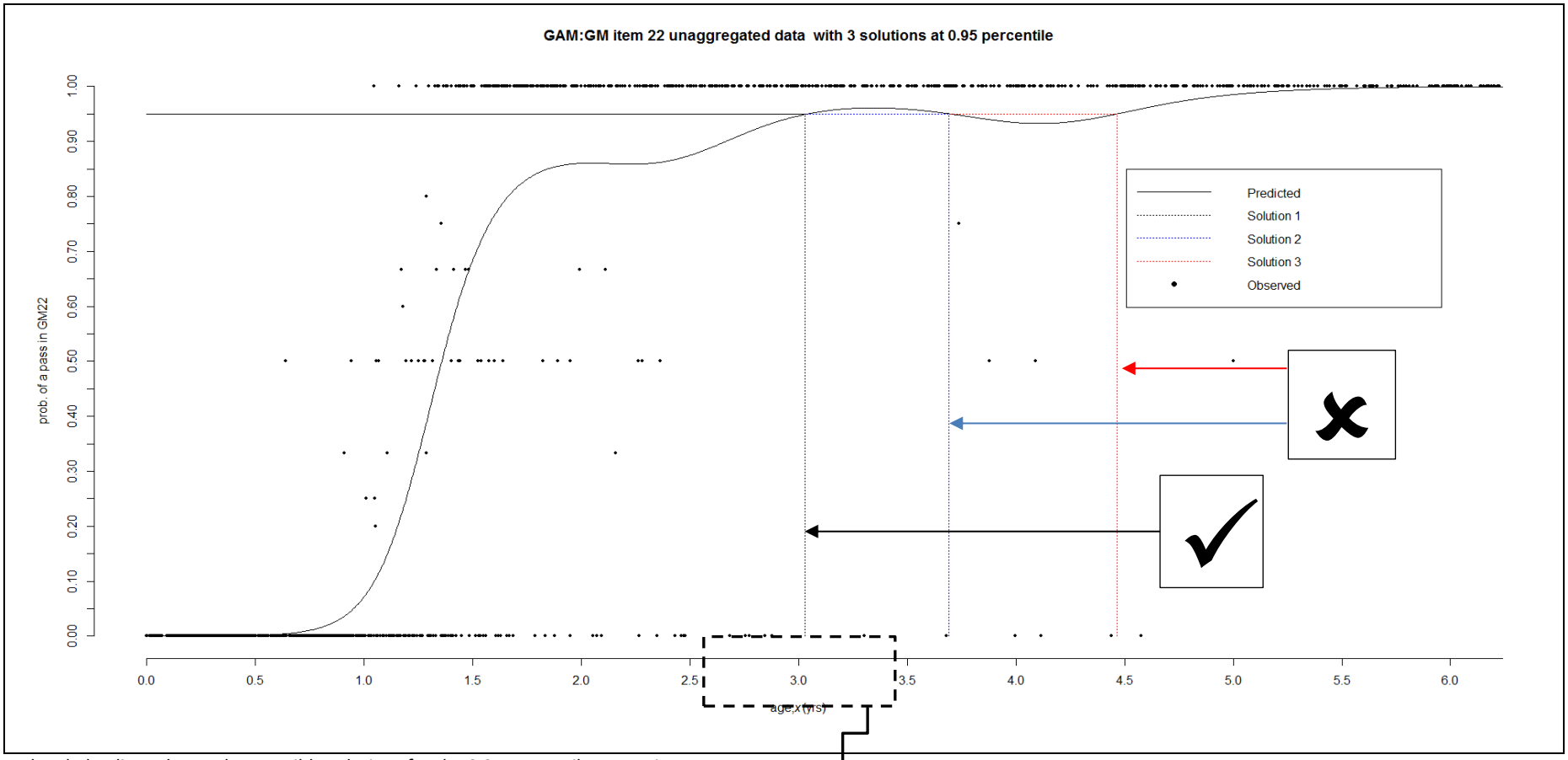
Zumbo, B. D., 2007. Three Generations of DIF Analyses: Considering Where It Has Been, Where It Is Now, and Where It Is Going. Language Assessment Quarterly, 4(2), pp. 223-233.

Zweig, M. H. & Campbell, G., 1993. Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine. *Clinical Chemistry,* 39(4), pp. 561-577.

8. Appendix

8.1. Appendix A – Supplementary Material for Chapter 2

Figure A.1: Creating normal reference ranges for each item under the GAM framework.



Red and Blue lines show other possible solutions for the 0.95 percentile age estimate.

One of the relevant age estimate(s) at the 0.95 pass percentile of interest used to draw scoring chart shown in Figure 3.5.

8.2. Appendix B – Supplementary Material for Chapter 3

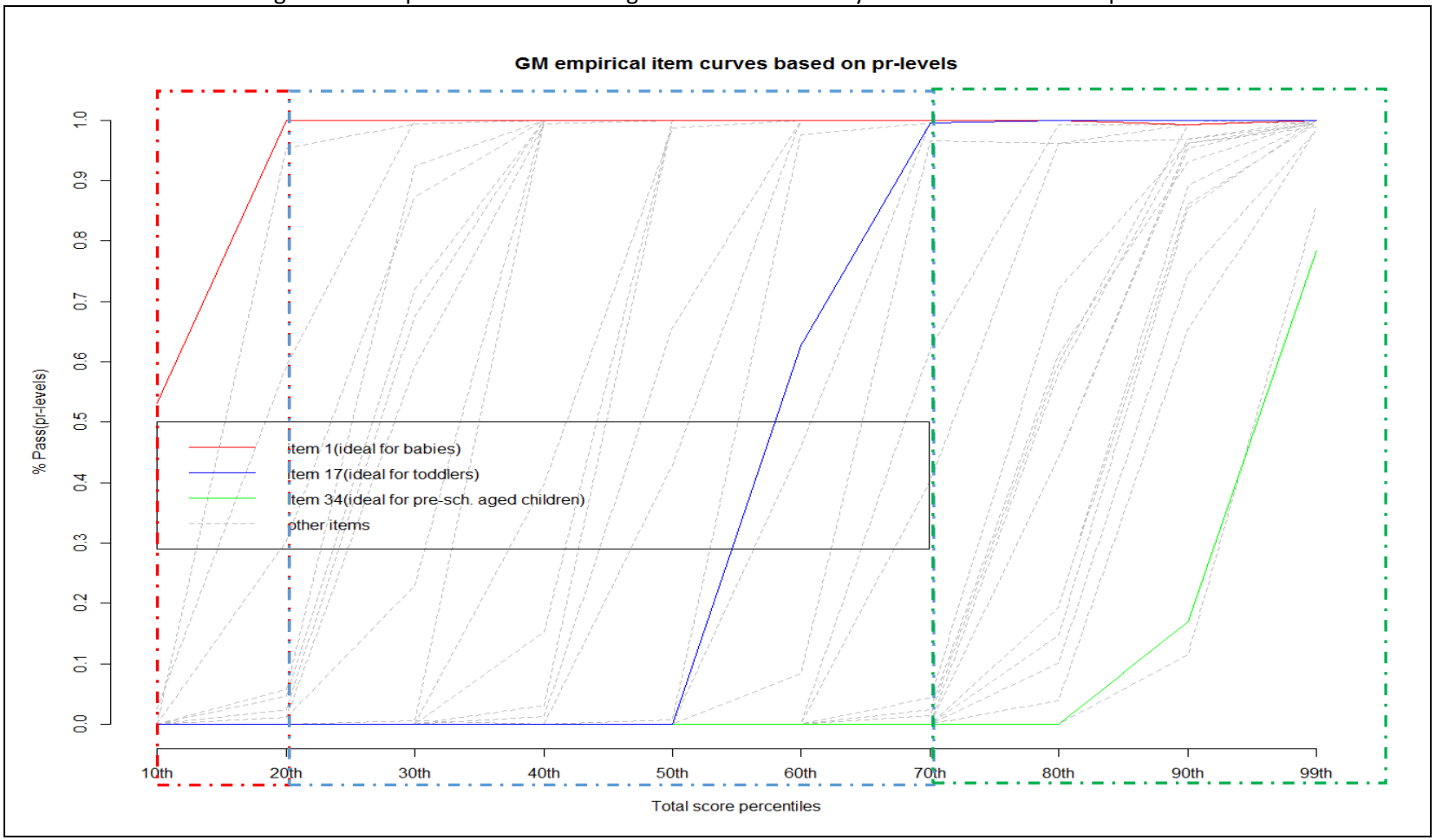
Figure B.1: Special expert and play equipment.

Weight measurement equipment	
	
Supplimentary play material	
	

Picture(s) were sourced from the MDAT tool manual.

8.3. Appendix C – Supplementary Material for Chapter 4

Figure C.1: Empirical item curves of gross motor domain by overall total raw score percentile.



pr-levels are the pass rates for each item

Items that are ideal for babies (< 1 year old), toddlers (1 to < 3.5 years old) and pre-school aged children (3.5 to < 7 years old) in the red, blue and green dotted boxes respectively.

8.4. Appendix D – Supplementary Material for Chapter 5

Figure D.1: Model fits under GLM and GAM extension (SCAM) model frameworks.