Transforming outcome measures in plastic surgery



Conrad Harrison

Balliol College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2022

For my father,
John Harrison,
1935-2017.

Sine qua non.

Acknowledgements

First and foremost I must thank my supervisors, Jeremy Rodrigues, Dominic Furniss, Chris Sidey-Gibbons, and Marc Swan, who have not only played critical roles in delivering every aspect of this work, but have directly supported my personal and professional development beyond what any DPhil student could expect.

I first approached Jeremy and Dom in 2015 with the idea of developing a PROM for wound closure, having read a paper which inappropriately applied a burns scar scale to compare Caesarian section skin closure techniques. They told me that this was an unrealistic idea, and it was. But since then, they have encouraged and enabled my journey into academic plastic surgery and measurement science, and coached me through every career milestone, including this DPhil.

Watching Marc repair a cleft lip was an inspiring and career-defining moment for me. He and Tim Goodacre showed me first hand the magic that is possible with modern surgical techniques and described the area of unmet need that I have gone on to address in this thesis. Marc has given up a great deal to support the clinical translation of this work, and to help me develop my own professional network. He sent me to Cambridge, where I met Chris for a five-day crash course in psychometrics and machine learning. Within five days, Chris convinced me that quantitative psychology is one of the most interesting and exciting areas of research today. I have been following in his footsteps ever since.

I have been fortunate to work with very supportive collaborators, without whom this work would not have happened. I am grateful to Anne Klassen and the Q-portfolio team, including Andrea Pusic, Karen Wong-Riff, and Charlene Rae, for their hard work in developing the full-length CLEFT-Q, and for granting me access to their field test data. I thank contributors to the validation dataset, and all members of the Spires cleft team, who went to great lengths to welcome me and support the use of the CLEFT-Q CAT in clinical practice. I thank the University of Cambridge Psychometrics Centre team and Rhian Davies at Jumping Rivers for going above and beyond to train me in various facets of data science. I am grateful for the support with patient and public involvement I received from Kenny Ardouin and the Cleft Lip and Palate Association. I thank all of my research participants - these studies would not have been possible without you.

I must also thank my funders, the Clarendon fund, Balliol College, the University of Oxford Medical Sciences Division, and the National Institute for Health and Care Research. I am thankful for the training and administrative support I have received from the Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences.

When my father died, he left me his copy of Encyclopedia Britannica, which was our go-to learning resource when I was young enough to sit on his knee. Learning has changed a lot since then, and in particular the open science movement has completely transformed the way we transfer knowledge. Open science is about more than making research data and publications freely available. The R community, and its collaborative and altruistic approach to open-source licensing, open educational resources, and open networking is a moving demonstration of what teamwork can achieve, and has dramatically enhanced this work. To that end, I must thank the package developers, particularly Phil Chalmers and the other contributors to his mirt package Google group. I thank contributors to CRAN, GitHub, and Stack Overflow, and all other researchers who challenge restrictively conventional approaches to knowledge dissemination.

Finally, I thank my family. My parents did not receive a university education, but worked tirelessly, making many personal sacrifices, to ensure that I did. You have given me roots and wings. My wife Nina has also made sacrifices so that I might pursue my academic and clinical interests. She has lightened the load of this thesis considerably, while excelling in her own career with humbling modesty. Thank you for showing me what I aspire to be.

Declaration

This thesis is submitted for the degree of Doctor of Philosophy. It is the result of my own work, and except where otherwise stated, describes my own research. Any figures that have been taken from other resources are clearly indicated and the permission reference is provided when necessary. Any assistance I have received is declared below.

- The University of Cambridge Psychometrics Centre team assisted me with the development of the Concerto-based CLEFT-Q CAT interface. Aiden Loe and I worked on the R functions together, while Vesselin Popov and Przemisław Lis provided supported with front-end development, databasing, and hosting.
- Rhian Davies, at Jumping Rivers, coached me in R package development and kindly fixed a bug in my CLEFT-Q CAT Score Checker code which made the application run considerably faster.

I further state that no substantial part of my work has already been submitted, or is being concurrently submitted, for any such degree, diploma, or other qualification at the University of Oxford or any other University or similar institution.

Finally, I can confirm that this thesis adheres to the University of Oxford Medical Sciences Division regulations on word limits on DPhil theses. This dissertation does not exceed the prescribed limit of 50,000 words.

This document presents independent research funded by the NIHR and the University of Oxford. The views expressed are my own and not necessarily those of the University of Oxford, the NHS, the NIHR, or the Department of Health and Social Care.

Conrad Harrison
Balliol College, University of Oxford

Abstract

The CLEFT-Q is a patient reported outcome measure (PROM) designed to measure the elements of health most important to people born with a cleft lip or palate (CL/P). Response burden is considered a major barrier to its uptake in routine clinical practice in the UK and abroad. In this thesis, I aimed to develop, validate, and deploy a computerised adaptive testing (CAT) version of the questionnaire, that can provide accurate scores for a respondent, based on an individualised selection of items. This is anticipated to reduce response burden and facilitate the questionnaire's uptake. I also aimed to explore future research avenues for clinical CAT research, including the quantification of measurement precision, and the application of multidimensional item response theory and regression tree learning.

In a series of Monte Carlo simulations, I compared popular score estimators and item selection criteria for use in CLEFT-Q CAT algorithms, based on Rasch models calibrated from the item responses of 2434 participants in the CLEFT-Q field test. I validated these algorithms in an external sample of 536 patients undergoing routine CL/P care, and set stopping rules for the assessments during a workshop that involved patients, parents, researchers, and clinicians. I developed a user interface for the software and a web application to assist with score interpretation. I deployed the platform during routine clinical practice at a single centre and interviewed 6 patients and 4 clinicians to understand the perceived impact of the system on clinical care.

Through simulation studies, I compared the measurement precision of Rasch-based CAT algorithms to CAT algorithms built with graded response models (GRMs). I developed a multidimensional GRM to describe the relationship between items and factors in the CLEFT-Q appearance scales, built this into a CAT algorithm, and demonstrated how it might work in practice. Finally, I built a range of CAT algorithms that use regression trees to incorporate clinician-reported data into the assessment, and compared the efficiency of these algorithms to Rasch-based counterparts.

There were no meaningful differences between the performance of popular score estimators and item selection criteria, and I chose to proceed with standard Bayesian techniques (expected *a posteriori* scoring and minimum expected posterior variance item

selection). In the validation study, algorithms were able to provide accurate measurements from fewer items, but accuracy decreased with decreasing assessment length. Standard error of measurement (SE) was higher for these Rasch-based algorithms than what has been recommended for GRM-based CAT algorithms. In the stopping rule workshop, stakeholders voted to reduce the CLEFT-Q scale lengths by 1-4 items each, and at these assessment lengths the root mean squared error between full-length scale scores and CAT scores was 2-5 points out of 100. In clinical practice, the platform was perceived to enhance communication between patients and clinicians, improve consultation focus, facilitate multidisciplinary team working, and improve the early detection of health related quality of life issues.

I demonstrated that while GRM-based algorithms achieve a lower SE, they are unlikely to provide more accurate point-estimates than Rasch-based algorithms. The multidimensional GRM CAT algorithm was able to measure the appearance of the face based on item responses to scales relating to facial components (for example, appearance of the nose, lips, and teeth). However, it did not provide obvious gains in efficiency. Regression trees preferentially used item responses over clinical variables for latent construct measurement and demonstrated inferior performance to Rasch-based algorithms, likely due to a loss in measurement granularity from binary data splitting.

The CLEFT-Q CAT system reduces assessment length and its use in clinical care is perceived to enhance communication and patient experience. This may drive routine, standardised, patient-centred outcome measurement in CL/P, and can be rapidly translated into other clinical fields. The high SE values are likely an artefact of Rasch model parsimony and SE heuristics used in GRM-based CAT systems should not necessarily be generalised to these Rasch-based algorithms. Future work should continue to cautiously explore the potential for multidimensional item response theory to improve assessment efficiency, and other methods for incorporating non-PROM data into latent variable models should be evaluated.

Contents

List of Abbreviations **xii**

Preface **1**

1 Introduction **2**
1.1 The origins of psychometrics 3
1.2 Patient-reported outcome measures 4
1.3 Classical test theory 5
1.4 Modern test theory 8
1.4.1 Rasch measurement theory 8
1.4.2 Item response theory 13
1.4.3 Likelihood, Fisher information, and standard error of measurement 17
1.4.4 Bayesian and frequentist estimators 20
1.4.5 Item parameter estimation and model fit 21
1.4.6 The prerequisites of modern test theory 22
1.5 Computerised adaptive testing 25
1.6 Cleft lip and/or palate 30
1.6.1 Embryology, epidemiology, and classification 30
1.6.2 Treatment 32
1.6.3 Outcome measurement 34
1.6.4 The CLEFT-Q 35
1.7 Summary of background 38
1.8 Aims, objectives, and thesis outline 39

2 A comparison of estimators and item selection criteria for the CLEFT-Q CAT **41**
2.1 Introduction 42
2.1.1 Latent construct estimators 42
2.1.2 Item selection criteria 43
2.2 Methods 44
2.2.1 Software 44
2.2.2 Data 45

2.2.3 Rasch model parameterisation . . . 45
2.2.4 Latent construct estimators . . . 46
2.2.5 Item selection criteria . . . 47
2.3 Results . . . 47
2.3.1 Demographics, clinical characteristics, and missing data . . . 47
2.3.2 Latent construct estimators . . . 48
2.3.3 Item selection criteria . . . 52
2.4 Discussion . . . 54
2.5 Conclusion . . . 56

3 External validation and stopping rules for the CLEFT-Q CAT 57
3.1 Introduction . . . 58
3.2 Methods . . . 59
3.2.1 Validation data . . . 59
3.2.2 Monte Carlo simulation . . . 60
3.2.3 cleftqCATsim . . . 61
3.2.4 Stopping rule workshop . . . 61
3.3 Results . . . 62
3.3.1 Demographics, missing data, and outliers . . . 62
3.3.2 Simulation results . . . 63
3.3.3 Workshop results . . . 66
3.4 Discussion . . . 67
3.5 Conclusion . . . 70

4 The CLEFT-Q CAT in clinical practice 72
4.1 Introduction . . . 73
4.2 Methods . . . 74
4.2.1 A user interface for CAT administration . . . 74
4.2.2 The CLEFT-Q CAT score checker . . . 76
4.2.3 Qualitative evaluation of the CLEFT-Q CAT in clinical practice 79
4.3 Results . . . 81
4.4 Discussion . . . 90
4.5 Conclusion . . . 92

5 The CLEFT-Q CAT's standard error of measurement 93
5.1 Introduction . . . 94
5.2 Methods . . . 97
5.2.1 A comparison of Rasch and GRM scores . . . 97
5.2.2 A comparison of CAT assessments built with Rasch models and GRMs . . . 98

5.3 Results . . . 98
5.3.1 A comparison of Rasch and GRM scores . . . 98
5.3.2 A comparison of CAT assessments built with Rasch models and GRMs . . . 100
5.4 Discussion . . . 101
5.5 Conclusion . . . 104

6 Multidimensional computerised adaptive testing 105
6.1 Introduction . . . 106
6.2 Methods . . . 108
6.2.1 Conceptual model . . . 108
6.2.2 Sample size . . . 109
6.2.3 Missing data in the structural equation modelling sample . . . 110
6.2.4 Structural equation modelling . . . 110
6.2.5 Multidimensional item response theory modelling and computerised adaptive testing . . . 111
6.3 Results . . . 113
6.3.1 Structural equation model fit . . . 113
6.3.2 Multidimensional computerised adaptive testing simulation . . . 113
6.4 Discussion . . . 116
6.5 Conclusion . . . 119

7 Recursive partitioning as an alternative approach to computerised adaptive testing 120
7.1 Introduction . . . 121
7.2 Methods . . . 123
7.2.1 Software . . . 123
7.2.2 Participants . . . 123
7.2.3 Ground truth . . . 123
7.2.4 Data splitting . . . 124
7.2.5 Patient-reported data . . . 124
7.2.6 Clinical variables . . . 124
7.2.7 Decision tree training . . . 125
7.2.8 Rasch-based CAT assessments . . . 126
7.2.9 Assessment length . . . 126
7.2.10 Comparison . . . 127
7.3 Results . . . 127
7.4 Discussion . . . 129
7.5 Conclusion . . . 131

8 Discussion **133**
8.1 A summary of completed work 134
8.2 The CLEFT-Q CAT 135
8.2.1 Measurement philosophy 135
8.2.2 Strengths 139
8.2.3 Limitations 140
8.2.4 Impact 144
8.3 The future of clinical psychometrics 145
8.3.1 Measurement error 145
8.3.2 Multidimensional measurement 147
8.3.3 Incorporating non-PROM data into latent variable models . . . 148
8.4 Conclusion 150

A Appendix A: Chapter 2 results **152**
A.1 Demographics and clinical characteristics 153
A.2 Missing item response data 166
A.3 Item selection criteria 178

B Appendix B: Chapter 3 results **185**
B.1 Simulation results including outliers 186
B.2 Simulation results excluding outliers 191
B.3 Score frequency distribution and standard error of measurement 196
B.4 Bland-Altman plots 201
B.5 Stopping rule workshop vote results 205

C Appendix C: Chapter 4 interview schedules **207**
C.1 Interview schedules 208

D Appendix D: Chapter 5 results **213**
D.1 A comparison of Rasch and GRM scores 214
D.2 A comparison of CAT assessments built with Rasch models and GRMs . 219

E Appendix E: Chapter 7 results **228**
E.1 Comparative simulation study results 229

References **236**

List of Abbreviations

ABG . . . . . Alveolar bone graft

AISP . . . . . Automated item selection procedure

BCLP . . . . . Bilateral cleft lip and palate

CAT . . . . . Computerised adaptive testing

CFI . . . . . Comparative fit index

CHAID . . . . . Chi-squared automatic interaction detection

CL . . . . . Cleft lip

CLA . . . . . Cleft lip and alveolus

CLAP . . . . . Cleft lip, alveolus, and palate

CP . . . . . Cleft palate

CSAG . . . . . Clinical Standards Advisory Group

CTT . . . . . Classical test theory

DIF . . . . . Differential item functioning

DT . . . . . Decision tree

EAP . . . . . Expected *a posteriori*

EHR . . . . . Electronic health record

EM . . . . . Expectation maximisation

EORTC . . . . . European Organisation for Research and Treatment of Cancer

GRM . . . . . Graded response model

ICC . . . . . Item characteristic curve

ICHOM . . . . International Consortium for Health Outcomes Measurement

IRT . . . . . Item response theory

IQR . . . . . Inter-quartile range

LA . . . . . Limit of agreement

MAE . . . . . Mean absolute error

MAP . . . . . Maximum *a posteriori*

MCAT . . . . . Multidimensional computerised adaptive testing

MDT . . . . . Multidisciplinary team

MEI . . . . . Maximum expected information

MEPV . . . . . Minimum expected posterior variance

MI . . . . . Maximum information

MIC . . . . . Minimal important change

MIRT . . . . . Multidimensional item response theory

MLE . . . . . Maximum likelihood estimation

MLWI . . . . . Maximum likelihood weighted information

MPWI . . . . . Maximum posterior weighted information

NIHR . . . . . National Institute for Health and Care Research

PROM . . . . . Patient-reported outcome measure

PROMIS . . . . Patient-Reported Outcomes Measurement Information System

RCT . . . . . Randomised controlled trial

RMSE . . . . . Root mean square error

RMSEA . . . . Root mean square error of approximation

RMT . . . . . Rasch measurement theory

SE . . . . . Standard error of measurement

SEM . . . . . Structural equation modelling

SRMR . . . . . Standardised root mean square residual

TLI . . . . . Tucker-Lewis index

TOPS . . . . . Timing of palate surgery

UCL . . . . . Unilateral cleft lip

UCLP . . . . . Unilateral cleft lip and palate

WHO . . . . . World Health Organisation

WLE . . . . . Weighted likelihood estimation

WLSMV . . . . Weighted least squares (mean and variance adjusted)

Preface

Applied and theoretical measurement science has progressed more slowly in surgery than in other fields. These concepts do not form part of clinical training, they can be arduous to learn, and often they do not immediately appear clinically relevant.

The first chapter of this thesis contains a significant amount of measurement theory that has grown from other scientific disciplines. I would urge unfamiliar readers to persevere with this, as these concepts will make it considerably easier to follow the subsequent experimental chapters, which go on to demonstrate the clinical relevance of this theory.

"Measurement achieved in the physical sciences as some kind of mapping on a continuum partitioned into equal units from an origin is understood readily. However, measurement is also an advanced concept, the achievement of which has been integral to the remarkable progress of physical science."

— David Andrich

1 Introduction

Contents

1.1 The origins of psychometrics 3
1.2 Patient-reported outcome measures 4
1.3 Classical test theory 5
1.4 Modern test theory 8
1.4.1 Rasch measurement theory 8
1.4.2 Item response theory 13
1.4.3 Likelihood, Fisher information, and standard error of measurement 17
1.4.4 Bayesian and frequentist estimators 20
1.4.5 Item parameter estimation and model fit 21
1.4.6 The prerequisites of modern test theory 22
1.5 Computerised adaptive testing 25
1.6 Cleft lip and/or palate 30
1.6.1 Embryology, epidemiology, and classification 30
1.6.2 Treatment 32
1.6.3 Outcome measurement 34
1.6.4 The CLEFT-Q 35
1.7 Summary of background 38
1.8 Aims, objectives, and thesis outline 39

1.1 The origins of psychometrics

Washington Irving Bishop was arguably the most famous mind-reader of the 19th Century. During his act, a member of the audience would hide an object in a room which Bishop had vacated. Bishop would return to quickly find the object by walking around the room while holding the hand of the audience member. Bishop's performance caught the attention of notable scientists at the time, including the editors of *Nature* and the *British Medical Journal*, and Sir Francis Galton, the polymath half-cousin of Charles Darwin and father of the eugenics movement. After a series of experiments involving Bishop, Galton and his colleagues concluded that the performer was not a genuine telepath but did possess a remarkable ability to measure his proximity to hidden objects by observing audience members' subconscious and involuntary 'ideomotor' body movements [1, 2].

The ability to interpret mental processes through observable human behaviour had been recognised as an art in business, love, and war for millennia, but it was only around the time of Galton and Bishop that it gained recognition as a *science*. Galton became fascinated by the idea that latent mental constructs could be formally modelled through observable behaviours and started to collect psychological assessment data at the International Health Exhibition, London. Together with mathematicians including James Cattell, Karl Pearson, and Charles Spearman, Galton developed tools and theory to evaluate unobservable, mental characteristics from observable behaviour and the science of psychometrics was born [3].

Psychometric theory was rapidly adopted into the field of educational psychology, where Galton believed it could be used to identify and measure desirable human characteristics, such as intelligence, as part of his eugenics movement. Meanwhile, Alfred Binet and Théodore Simon applied these techniques to revolutionise educational assessments and schooling policy, through landmark advances such as the Stanford-Binet intelligence test [3]. The early rooting of psychometrics in educational psychology contextualises some of the technical terminology that will be introduced later in this chapter, such as the terms

item difficulty and *person ability*, which are most intuitive in this context.

Today, psychometrics as a science has grown both in sophistication and in the breadth of its application. Besides educational assessment, advanced psychometric techniques are now used in studies of human personality [4], consumer marketing exercises [5], finance [6], and have even been implicated in high-profile allegations of electoral fraud [7]. The use of psychometric theory in healthcare is also growing. As clinical medicine and surgery move away from the paternalistic treatment of disease, and towards the holistic treatment of people, a greater focus is being placed on patient-reported outcome measures (PROMs) [8].

1.2 Patient-reported outcome measures

Patient-reported outcome measures are questionnaires that can measure the severity of symptoms such as pain, nausea, and physical impairment, from a patient's perspective, and highlight specific health concerns to clinicians. These can be generic or specific to a particular condition or population [9]. Patient-reported outcome measures are now frequently used as primary outcome measures in clinical trials [10, 11], national registries [12], and for policy setting and commissioning decisions by the Food and Drug Administration, the European Medicines Agency [13], and the National Institute for Health and Care Excellence [14].

In addition to their role as outcome measures, evidence is building to support the use of PROMs as direct adjuncts to clinical care. Across a range of settings, PROMs can enhance communication between clinicians and patients, improve health-related quality of life outcomes, and drive-up patient experience [15]. In cancer, patients randomised to remote symptom monitoring through PROMs have demonstrated improved survival compared to those followed up with routine imaging, and this is attributed to an earlier detection of disease recurrence [16–19]. In paediatric care, there is evidence to suggest

that PROMs lead to higher referral rates, faster consultations, and earlier detection of health-related quality of life problems [20–24].

But PROMs are yet to reach their full potential, particularly in the field of plastic and reconstructive surgery, where several barriers exist to their uptake in both research and clinical practice [25]. Patient-reported outcome measures can be burdensome to complete and difficult to interpret, and most available PROMs were not developed to the contemporary measurement standards outlined by the consensus-based standards for the selection of health measurement instruments (COSMIN) [26]. For example, many plastic surgery PROMs were not developed with patient input and lack evidence to support their content validity (their ability to comprehensively capture all relevant elements of the health construct the instrument sets out to measure) [27–31]. This detracts from their perceived relevance and risks measurement that fails to capture the elements of health that are important to patients. Evidence to support the construct validity of many plastic surgery PROMs (the degree to which scores derived from these instruments accurately measure the latent construct) is also lacking [27–31]. Studies that examine the psychometric performance of plastic surgery PROMs have typically relied on crude statistical techniques or classical test theory (CTT), which have both largely been superseded in other fields by modern psychometric techniques.

1.3 Classical test theory

The early work of Galton and colleagues led to the development of CTT. Under this paradigm, a respondent's observed test score, X, is a combination of their true score, T, and some stochastic error, e [32].

$$X = T + e. \tag{1.1}$$

In Equation 1.1, the respondent's observed test score, X, is an imperfect reflection of their true score, T. It may be an overestimate, or an underestimate, depending on the value of e. Figure 1.1 illustrates this concept as applied to a population of test takers. In this graph, each respondent's observed score is explained by their true score and the residual term, e.

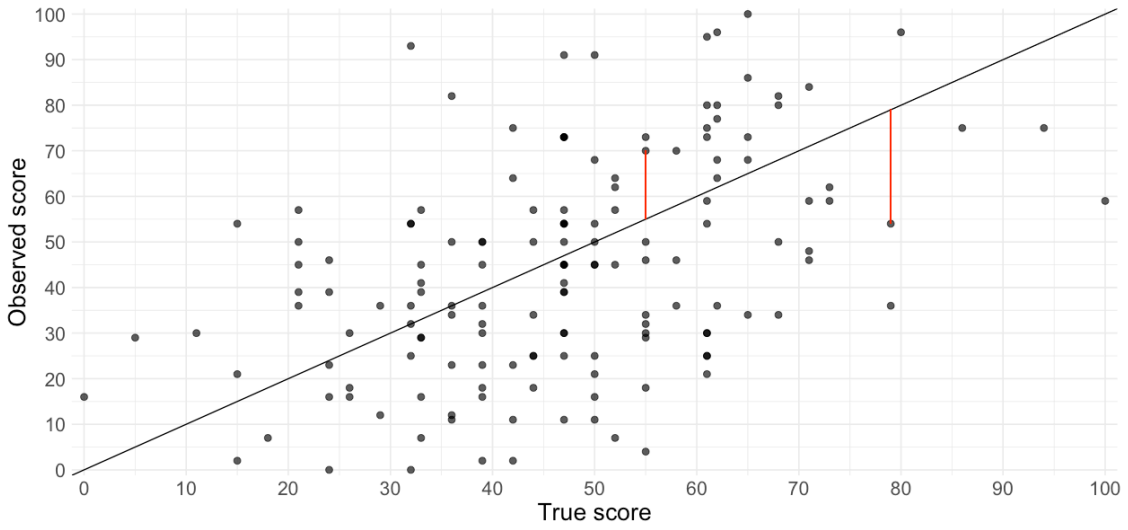


Figure 1.1: An illustration of classical test theory. For each test taker, the observed test score is explained by their true score and residual variance (test error). For two test takers in this illustration, the residual variance is shown as vertical red lines. The diagonal line, $y = x$, shows where points would fall in an errorless test (where $e = 0$).

Classical test theory assumes a linear relationship between observed scores and true scores, where error terms are normally distributed and homoscedastic (independent to the underlying level of the latent construct). Therefore, the positive error terms are assumed to approximately cancel the negative error terms across a population of test takers (Figure 1.1) [32]. But for an individual respondent with a large error term, the observed score may differ considerably from the true score. This is relevant for PROM use in clinical practice, where high-stakes decisions are made at the individual level. Furthermore, Equation 1.1 contains two unobservable terms (T and e), and there is no way to quantify both from the single observable X term. In other words, CTT does not provide a means to understand how much error may have contributed to an individual observed score. This is a problem for clinicians who wish to use psychometric assessments with individual patients, as we need to know how accurately a latent construct is reflected in a person's score before we

can use that score for clinical decision making.

The assumptions of classical test theory are conceptually challengeable. Firstly, many questionnaire items, especially in PROMs, provide categorical response options. For example: “severe pain”, “moderate pain”, “mild pain”, and “no pain”. In terms of true score, T, these response options are more likely to represent ordinal categories than equidistant graduations of a linear continuum, and this makes a linear relationship between observed score and true score improbable.

Secondly, it is difficult to accept that error terms are unrelated to the underlying latent construct level. Questionnaires are only able to provide meaningful measurement within certain ranges of the measured construct, a property known as *targeting* [33]. That is to say, the minimum score on a depression questionnaire does not represent the minimum level of depression that is humanly possible, nor does the maximum score represent the maximum level of depression that can exist in nature. Consider a test that produces observed scores that range from 0 to 100. A person with a true score of 100 might obtain an observed score of 100, in which case the assessment has been accurate. But a person with a true score of 200 can only obtain a score as high as the test’s maximum, 100. In this second case, the test is inaccurate because the respondent’s latent construct level is outside of the assessment’s target range. In this example, the assumption of homoscedasticity is violated.

Perhaps the most relevant limitation of CTT to this thesis is the “circular dependency” of item and person statistics [33, 34]. In CTT, the *difficulty* of a dichotomous item relates to the proportion of respondents that endorse it positively (in a true or false educational test question, the proportion of people who answer correctly – the more people who answer correctly, the less difficult the item). The item *discrimination* relates to the ability of an item to differentiate between high overall test scorers and low overall test scorers. This is typically derived by correlating item and test scores within a sample of test takers. Under this framework, the item statistics (difficulty and discrimination) are dependent on the sample of test takers. A sample of respondents with a high *ability* (a

propensity for positive endorsement, for example mathematical aptitude in the context of a mathematical assessment) will cause items to appear less difficult than a sample of respondents with a low ability. Similarly, respondents who answer less difficult items will appear more able than those who answer more difficult items. This means that for fair comparison under CTT, respondents must complete identical sets of items. This is a problem in clinical research and practice, as under CTT, a single missed item can invalidate the comparison of PROM scores.

1.4 Modern test theory

The 1960s saw a revolution in psychometric theory as academics such as Georg Rasch, Frederic Lord, and Paul Lazarsfeld sought to overcome the limitations of CTT in high-stakes educational assessments [35]. They aimed to employ probabilistic modelling to derive truly continuous (as opposed to ordinal) measurement, with quantifiable certainty at the person level, irrespective of the items used.

Two distinct philosophies emerged from these efforts: Rasch measurement theory (RMT) and item response theory (IRT). These approaches, collectively known as modern test theory, are considered superior to CTT by most psychometricians, but the choice of RMT or IRT remains divisive.

1.4.1 Rasch measurement theory

Rasch measurement theory is a prescriptive philosophy, in which items are selected for inclusion in a scale where their responses fit the Rasch model, a strict model where items differ only by *difficulty* and respondents differ only by *ability* [36, 37]. As in CTT, difficulty relates to the level of latent construct required for a positive endorsement and ability relates to the level of latent construct present in the respondent, but these parameters are derived differently. Critically, they are measured on the same continuous

scale. Thus, they are sometimes referred to as item or person *location*.

The Rasch model is based on the standard logistic function

$$f(x) = \frac{e^x}{1 + e^x} \tag{1.2}$$

which produces a sigmoid curve with asymptotes at $y = 0$ as x tends to -∞, and $y = 1$ as x tends to +∞.

For a given person, and given dichotomous item (one with two possible responses), the Rasch model takes the form

$$P(X = 1) = \frac{e^{(\theta - b)}}{1 + e^{(\theta - b)}}, \tag{1.3}$$

where $P(X = 1)$ is the probability of a positive endorsement, θ is the ability of the person and b is the item difficulty parameter [38]. Equation 1.3 is simply the standard logistic function (Equation 1.2) translated along the x axis by b. As item difficulty increases, the function shifts to the right (Figure 1.2). The curves derived by this model are known as item characteristic curves (ICCs).

Variations of this model exist for polytomous (multiple choice) items, the most relevant of which is the partial credit Rasch model. This name relates to the model's ability to assign partial credit to an exam response that is partially correct. In polytomous PROMs, responses are neither correct nor incorrect, but ordered response categories may reflect an increasing level of latent construct in the same way that answers ordered by correctness might reflect increasing intelligence in an educational assessment. The partial credit model is similar to the dichotomous Rasch model but relies on successive dichotomisation of consecutive response options [36].

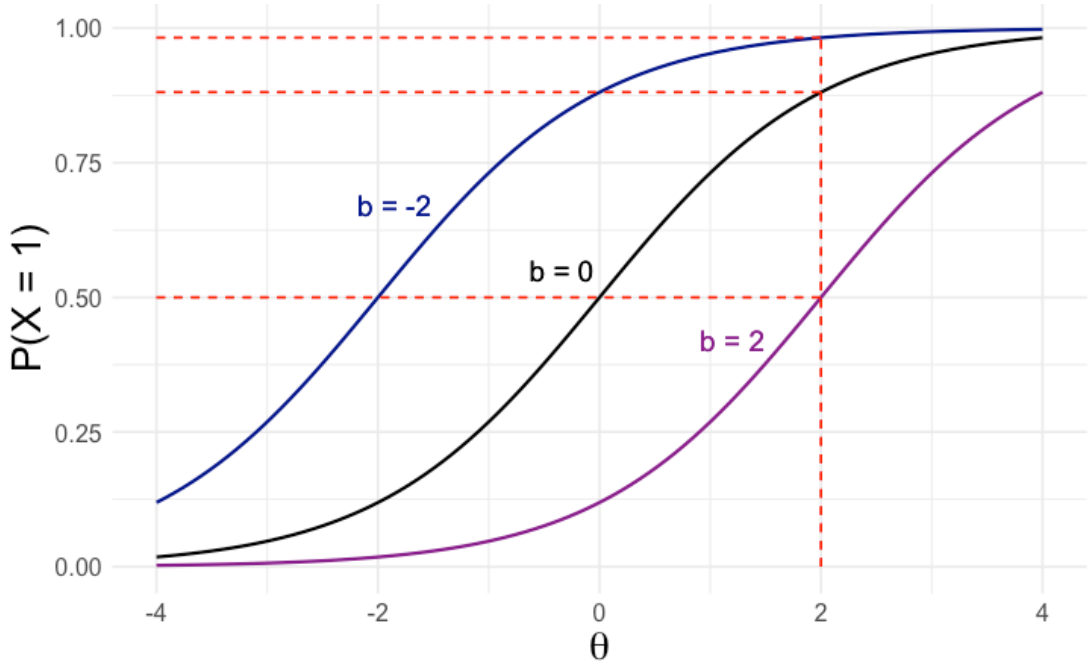


Figure 1.2: Item characteristic curves in the dichotomous Rasch model. The blue, black, and magenta curves represent items with varying levels of difficulty, b. For a given respondent with ability, θ, of 2 logits, the probability of a positive endorsement varies with item difficulty. The blue curve represents the item with the lowest difficulty, and this has the highest probability of positive endorsement at any given point. The black curve, where $b = 0$, is equivalent to the standard logistic function (Equation 1.2).

To illustrate this, Equation 1.3 can be rewritten

$$P(X = 0) = \frac{1}{1 + e^{(\theta - b)}} \tag{1.4}$$

$$P(X = 1) = \frac{e^{(\theta - b)}}{1 + e^{(\theta - b)}}, \tag{1.5}$$

where b represents the level of θ where $P(X = 0)$ and $P(X = 1)$ are equally probable and $P(X = 0) + P(X = 1) = 1$ [39].

In other words, when the person ability and item difficulty parameters are equal, the test-taker is as likely to endorse response option $X = 0$ as response option $X = 1$ (the equations both produce a probability of 0.5, see Figure 1.2).

If we add a third response category to this item, $X = 2$, we can similarly define the level of θ where $P(X = 1)$ and $P(X = 2)$ are equal using a second b parameter, and we can continue to add response categories in this manner [39].

The partial credit model is a divide-by-total model, meaning that b parameters for each category boundary (threshold) are defined such that the probability of responding to a given response category is modelled by the parameter combination for that category (see Equations 1.4 and 1.5), divided by the sum of parameter combinations across all categories [39–41].

Figure 1.3 illustrates an ICC that might be obtained from a partial credit model with four response categories. In this figure, the probability of endorsing successive response options increases and decreases as we move across the latent continuum.

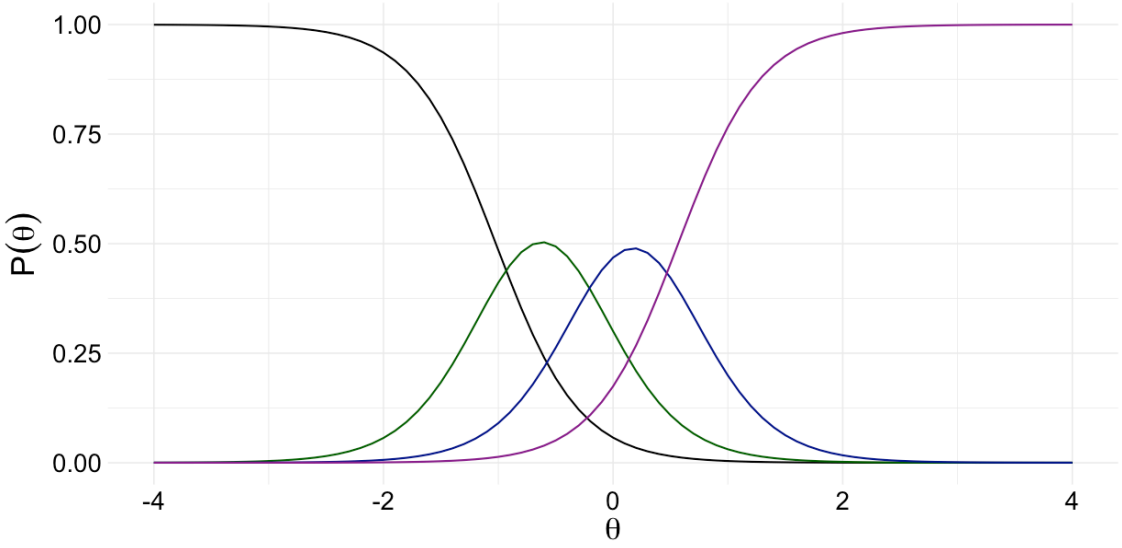


Figure 1.3: Item characteristic curves in the partial credit Rasch model. This graph represents a single item with four ordered response options ranging from least difficult (the black curve on the left) to most difficult (the magenta curve on the right). In a PROM, these four options could represent "no pain", "mild pain", "moderate pain", and "severe pain", for example.

Because divide-by-total models define item difficulty parameters as the level of θ where successive response options are equally probable, they can be visualised as intersections in the ICC (Figure 1.4). For example, in Figure 1.4, a respondent with a latent construct

level higher than b_3 is most likely to endorse the fourth response option, represented by the magenta curve.

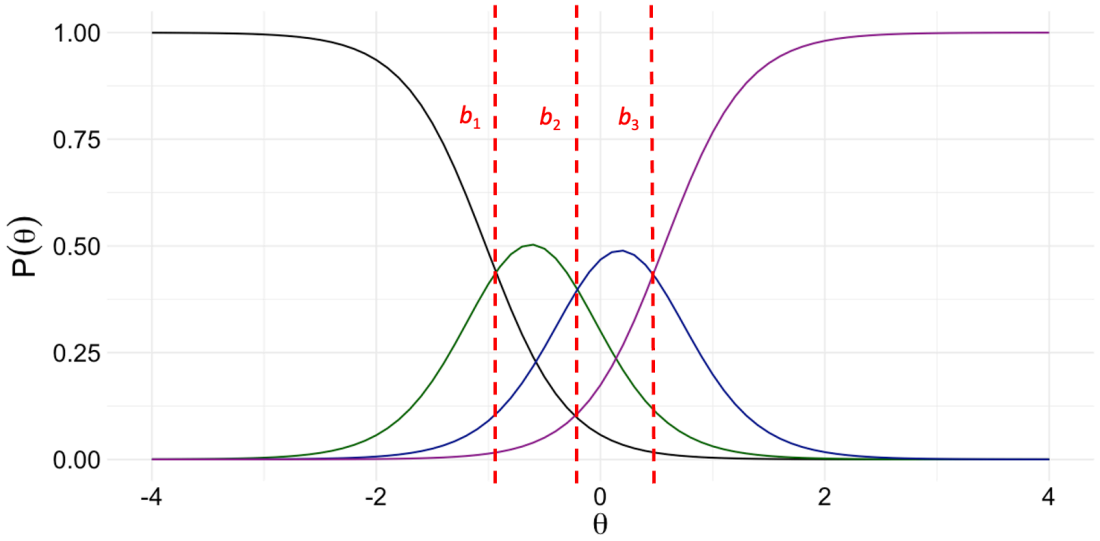


Figure 1.4: Difficulty parameters in divide-by-total models. This graph illustrates the difficulty parameters for the item presented in Figure 1.3 as the intersections of sequential response category functions. A respondent with an ability lower than b_1 is most likely to endorse the response option represented by the black curve, a respondent with an ability between b_1 and b_2 is most likely to endorse the response option represented by the green curve, and so on.

Under the partial credit model, difficulty parameters are not necessarily spaced equally. The distance between each difficulty parameter is free to vary within an item, and between items within a scale [42]. This contrasts with the rating scale version of the Rasch model, which requires equally spaced difficulty parameters between items [36].

A unique property of Rasch models is the *separability* of item and person parameters [37]. Operationally, the marginals of the item-by-person response matrix are *sufficient* statistics to calculate item parameters independently of person parameters and vice versa [36]. This is because item parameters can be calculated through conditional likelihood equations in which person parameters are eliminated [43, 44]. In contrast to the cyclic dependency of CTT, the difficulty parameters in Rasch models are not related to the ability of test takers from which they were derived. It is purported that this allows Rasch

model parameters to transcend person samples in a way that even the other models described by IRT cannot [44].

1.4.2 Item response theory

Item response theory is a descriptive (rather than prescriptive) philosophy. Instead of items being selected or discarded based on their fit to the Rasch model, models are selected and parameterised based on their fit to the data at hand [37]. Under IRT, a range of models might be chosen to describe the relationship of item response and latent construct, including Rasch models. Many of these models contain additional parameters that are freely estimated with the aim of describing item properties other than difficulty.

For example, Samejima's graded response model (GRM), which will be revisited later in this thesis, is expressed

$$P_m^*(\theta) = \frac{e^{a(\theta - b_m)}}{1 + e^{a(\theta - b_m)}}, \tag{1.6}$$

where $P_m^*(\theta)$ is the probability of a respondent endorsing response category m or higher, a is the discrimination parameter (i.e. the gradient of the inflection in a dichotomous ICC), and b_m is the level of latent construct at which point the probability of endorsing category m or higher is 0.5 [45, 46].

For dichotomous items, the GRM's discrimination parameter alters the shape of the ICC (while under the Rasch model, only its position along the x axis can be changed). This is illustrated in Figure 1.5, where the blue and black ICCs share a difficulty of 0 but differ in discrimination. For respondents with a latent construct level of $-1 < \theta < 1$ responding to the item with the blue ICC, the probability of a positive endorsement rises sharply as the level of the latent construct increases. For the same respondents responding to the item with the black ICC, the probability of a positive endorsement rises less sharply. After

respecifying both the difficulty and discrimination parameters, the magenta ICC takes on a different shape and position to the black ICC. While items that behave in this way might be ruled out by RMT, they are accommodated by the GRM [42].

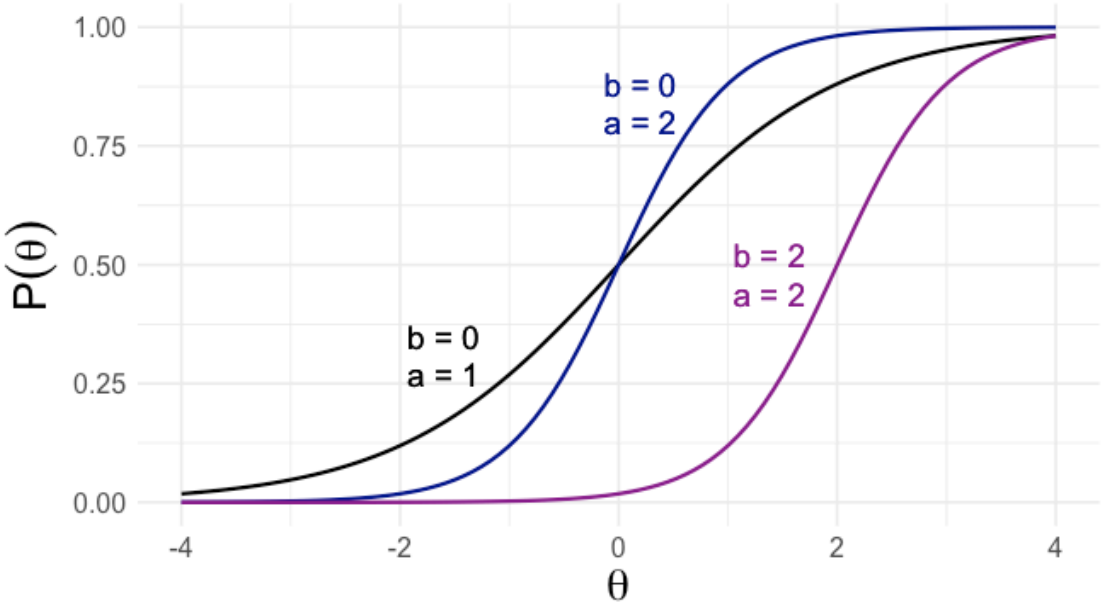


Figure 1.5: Item characteristic curves in the dichotomous graded response model. The blue, black, and magenta curves represent items with varying levels of difficulty, b and discrimination, a (see Equation 1.6). The blue and black curves share an inflection point (b) of $\theta = 0$, although the gradient of the blue curve is steeper as a result of its larger discrimination parameter. The item generating the blue ICC therefore discriminates more effectively between respondents with an ability of $-1 < \theta < 1$ than the item generating the black ICC. The item generating the magenta ICC, which has a difficulty of $b = 2$ and a discrimination of $a = 2$, is better targeted to discriminate between respondents with a higher level of ability. While the blue ICC is appropriate for discriminating between respondents with an ability of $-1 < \theta < 1$, all respondents at this range are unlikely to respond positively to the item with the magenta ICC.

Another important contrast between the polytomous GRM and the partial credit model is that the GRM is a difference model, as opposed to a divide-by-total model [46]. When the GRM is applied to an item with categories $m \in \{0, 1, 2 \ldots M + 1\}$, the difficulty parameters b_m are defined as the level of θ where the probability of choosing category m or higher is 0.5.

The probability of endorsing response m is then given as

$$P_m(\theta) = P^*_m(\theta) - P^*_{m+1}(\theta). \tag{1.7}$$

For this reason, the difficulty parameters in the GRM do not necessarily correspond to the level of θ at which one response option becomes more probable than another (i.e. the intersections of the ICC, Figure 1.6) [41].

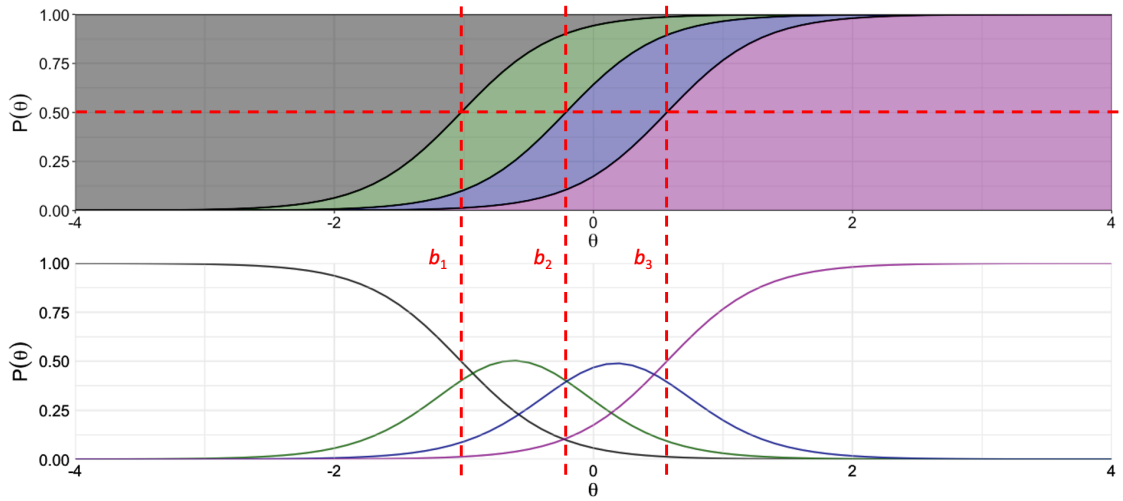
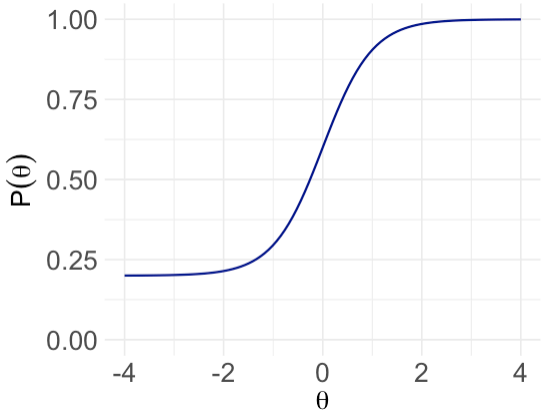


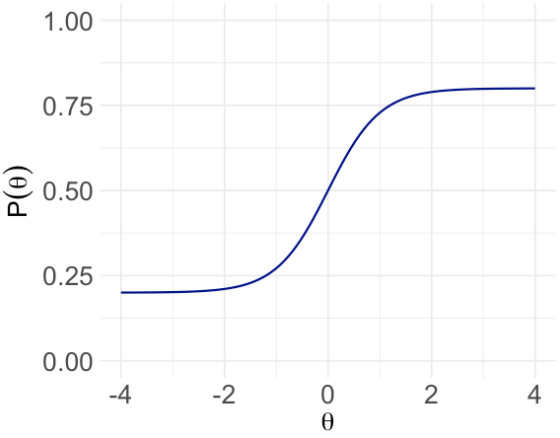
Figure 1.6: Difficulty parameters in difference models. This graph illustrates a four-category item parameterised under the GRM with a = 2.7, $b1$ = -1.0, $b2$ = -0.21, and $b3$ = 0.57. The categories, in ascending position along the latent continuum, are coloured grey, green, blue, and magenta. The upper panel demonstrates the cumulative probability of endorsing each response option. The contribution of each response option to this cumulative probability is given by the area shaded in that category's colour. The discrete probability of endorsing each response option is illustrated as an ICC in the lower panel. As a difference model, difficulty parameters in the GRM relate to the level of θ where the probability of endorsing a response at least as high as the next category is 0.5. These values do not necessarily correspond to the ICC crossings in the lower panel, as demonstrated by the red dashed lines.

Other IRT models include the generalised partial credit model (a 2-parameter logistic, divide-by-total model) [40], and 3- and 4- parameter models that attribute some probability of positive or negative endorsement to *guessing* and *inattention* parameters respectively (see Figure 1.7, Equation 1.8, Figure 1.8, and Equation 1.9) [37, 47] .



$$P(\theta) = (1 - c)\frac{e^{a(\theta-b)}}{1 + e^{a(\theta-b)}} + c \tag{1.8}$$

Figure 1.7: Item characteristic curve of a dichotomous, 3-parameter logistic model. In the 3-parameter model, the addition of a guessing term, c, raises the lower asymptote of the ICC. In this case (c = 0.2), no matter how low the respondent's level of θ, there is always a > 0.2 probability of positive endorsement. In an intelligence test, this may represent the probability of guessing a correct answer.



$$P(\theta) = (d - c)\frac{e^{a(\theta-b)}}{1 + e^{a(\theta-b)}} + c \tag{1.9}$$

Figure 1.8: Item characteristic curve of a dichotomous, 4-parameter logistic model. In the 4-parameter model, the addition of an inattention parameter, d, lowers the upper asymptote of the ICC.

These models allow far greater flexibility than the RMT approach, which some argue is an unrealistic representation of real-world item performance and may lead to the rejection of valuable items during scale development [48]. On the other hand, as the number of freely estimated model parameters increases so does the propensity to overfit, particularly when models are calibrated with small samples, and the separability of item and person

parameters is unique to Rasch models [44].

Both RMT and IRT allow items to function independently to one another, such that two respondents' scores can be compared regardless of which items they have been administered from a full-length assessment. For each respondent, scores are measured on a continuous logit scale with a quantifiable level of certainty relating to a given set of item responses.

1.4.3 Likelihood, Fisher information, and standard error of measurement

Figures 1.2 and 1.5 illustrate how different dichotomous items may be better suited for measurement at different levels of the latent construct. Where the ICC is steep, small changes in ability correspond to large changes in the probability of a positive endorsement – the item is targeted for measurement at the level of θ where the ICC inflects. Similarly, polytomous items are targeted to levels of the latent construct where small changes in θ result in large changes in the probability of endorsing different response options. For example, in Figures 1.3 and 1.4, the item provides poor measurement at a latent construct level of $\theta < -3$. At this level, even a large change in θ is unlikely to change a person's response, which will almost certainly be the lowest response option. However, at a latent construct level of $-1 < \theta < 1$, small changes in θ result in rapid ascents and descents of the response curves, making it easier to discriminate between people with differing latent construct levels, based on the responses they have endorsed.

When a person endorses the lowest response option in Figures 1.3 and 1.4, we might estimate that they have a $\hat{\theta}$ of -3, as at that level of ability we are most likely to observe that response (hereafter, $\hat{\theta}$ denotes the estimated value of θ). But our measurement is very imprecise. In the absence of other data or assumptions (for example Bayesian priors, which will be discussed later in this chapter), our respondent may be almost equally likely to have a θ of -4, or -5. Our confidence that $\hat{\theta} = -3$ is low. On the other hand, if

our respondent selects the second response option (represented by the green curve) we might estimate a $\hat{\theta}$ of -0.6; at this level, the second response option is more likely than any of the others. If the respondent actually had a θ of < -1, we would have been more likely to observe the lowest response, and if the respondent had a θ of > -0.2, we would have been more likely to observe the third response. This means that we can be more confident in estimating $\hat{\theta}$ = -0.6, given the second response option has been endorsed, than in estimating $\hat{\theta}$ = -3, given the lowest response option has been endorsed. One of the principal advantages of modern test theory over CTT is the ability to quantify this confidence (or precision), associated with our measurement.

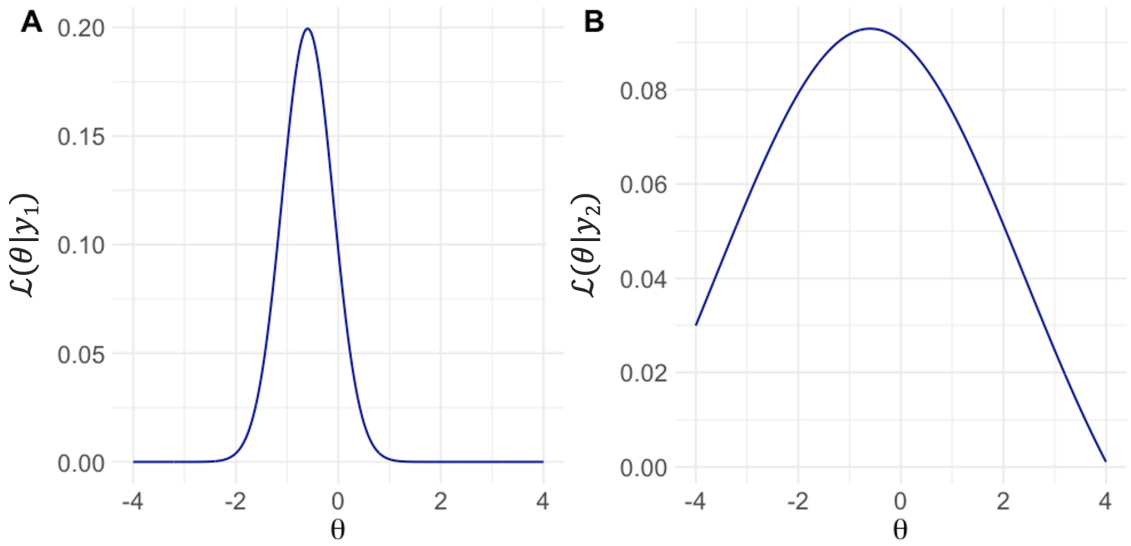


Figure 1.9: Two illustrative likelihood functions. Panel A demonstrates a relatively sharp likelihood function for θ given a set of item responses, y_1. It is at its maximum where θ = -0.6. Small deviations from this estimate result in a rapid decrease in the likelihood of observing y_1. Panel B shows a relatively broad likelihood function for θ given a different set of item responses, y_2. It is also at its maximum where θ = -0.6, but deviations from this estimate do not result in such a rapid decrease in the likelihood of observing y_2. The responses y_2 contain less information than y_1 about the true value of θ, and y_1 provides more precise measurement than y_2.

For a given model, the probability of observing a series of item responses, y, varies with the level of θ. This relationship is known as the likelihood function, $\mathcal{L}(\theta|y)$, (Figure 1.9), and can be computed from a set of item responses and model parameters, essentially by multiplying the probabilities of each observed item response across a range of possible θ values (see Chalmers [49] for an elegant derivation). For computational ease in

subsequent calculations, the log-likelihood function, $\ell(\theta|y)$, which shares a vertex with the likelihood function, is often used to describe this relationship [49–51].

The shape of the likelihood function, or log-likelihood function, relates to the precision of the θ estimate. In panel A of Figure 1.9, the most likely value of θ is -0.6, given the observed responses y_1. As θ deviates from -0.6, the likelihood of observing y_1 decreases rapidly. If θ was in fact -1, the likelihood of observing y_1 would be significantly lower (0.15 as opposed to 0.20, in this illustration). This sharp curve reflects a high confidence in our estimate of θ (a high measurement precision). In panel B of Figure 1.9, the most likely value of θ is also -0.6, given the observed responses y_2, but the likelihood of observing y_2 falls less rapidly as we deviate from θ = -0.6. In fact, the likelihood of observing y_2 is very similar when θ is -0.6 or -1. This broad curve reflects a low confidence in our estimate of θ (a low measurement precision). The responses y_1 provide more information about the true value of θ than responses y_2. This information is known as Fisher information [52], and is formally defined as

$$I(\theta) = Var\left(\frac{\partial}{\partial\theta}\ell(\theta|y)\right). \tag{1.10}$$

Assuming model parameters are fixed, the more rapidly the gradient of the log-likelihood function increases with respect to θ, the more information our observation contains about the true value of θ. Information is therefore inversely related to the variance of the ability estimate (Equation 1.11) [37],

$$Var(\theta) = \frac{1}{I(\theta)}. \tag{1.11}$$

In RMT and IRT, the confidence around a measurement is often provided as the standard error of measurement (which hereafter I abbreviate to SE, not to be confused with standard error of the mean or structural equation modelling). It is given

$$SE(\theta) = \sqrt{\frac{1}{I(\theta)}} \tag{1.12}$$

and can also be expressed as a confidence interval using z scores, $\hat{\theta} \pm z(SE(\hat{\theta}))$ [37, 53].

1.4.4 Bayesian and frequentist estimators

Figure 1.9 demonstrates how the likelihood function can be used to estimate a person's ability, given a set of responses to parameterised items. The level of θ where the curve reaches its vertex can be estimated through iterative computation of $\mathcal{L}(\theta|y)$ or $\ell(\theta|y)$ across increments of θ, and this represents the level of θ that maximises the likelihood of observing the given response set, y, which is at θ = -0.6 in Figure 1.9. This is called the maximum likelihood estimation (MLE) [37, 50, 51].

Consider a respondent who picks the lowest possible response option to each item in an assessment. The level of θ that maximises the likelihood of this observation is -∞ [37]. This statement is true but unhelpful, and illustrates a potential drawback of MLE. In the real world, it is impossible for a person to have an ability of -∞. And if a patient were to achieve a score of -∞ on a PROM, it would be difficult for a clinician or trial statistician to interpret. We might instead report the score as "too low to measure". This approach, which only uses observable data and makes no assumptions about the underlying distribution of θ amongst the population (for example, that it is unlikely to ever be -∞), is consistent with the frequentist approach to statistical inference [54].

An alternative philosophy is Bayesian inference, which incorporates prior assumptions about the distribution of an unknown parameter into its estimate [55]. Bayes' theorem describes a posterior distribution, defined

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1.13}$$

where $P(A)$ is the assumed distribution of A in a population (known as the prior) and $P(B)$ is the marginal probability of B (which, in the case of dependent probabilities, is the probability of B summed across all values of A). This can be applied to our ability estimate as

$$P(\theta|y) = \frac{\mathcal{L}(\theta|y)P(\theta)}{\int_{-\infty}^{+\infty} \mathcal{L}(\theta|y)P(\theta)d\theta} \tag{1.14}$$

to generate a posterior probability distribution $P(\theta|y)$ that accounts for both our likelihood function and prior understanding of the distribution of θ amongst the population [51, 56]. In a similar fashion to MLE, one might take the mode of this posterior distribution to be the ability estimate, known as the maximum *a-posteriori* (MAP, also known as the Bayes modal estimate). Alternatively, one might take the posterior distribution's expected value (i.e. the weighted mean), known as the expected *a-posteriori* (EAP) [49, 56].

Frequentists will be quick to point out that erroneous assumptions about the distribution of θ will lead to biased ability estimates, and priors are often selected based on an arbitrary or convenience basis. On the other hand, disciples of Thomas Bayes will advocate the use of posterior-based estimators to generate far more meaningful measurements at the extreme ends of a scale's range [57].

1.4.5 Item parameter estimation and model fit

It is relatively simple to derive estimates of ability given a set of item responses and model parameters, as described in the previous subsection. But difficulty arises when we need to estimate person ability and item parameters at the same time. For example, the shape of the likelihood functions in Figure 1.9 may differ considerably for a different set of item parameters. This creates a seemingly impossible chicken-and-egg situation: how

can ability be estimated if the item parameters are not known in the first place?

This problem is commonly addressed through an expectation maximisation (EM) algorithm. Put simply, EM algorithms iteratively adjust item and person parameters in turn until the likelihood of $\hat{\theta}$, given the observed response set, is maximised [49, 58]. The first set of unknown parameters are set arbitrarily and used to estimate the second set. These are then used to re-estimate the first set with increasing likelihood, and so on, until a set of item and person parameters that best explain the observed data, are identified. In practice, the EM algorithm and similar alternatives are run through specialist software platforms such as Phil Chalmers' *mirt* package for the R statistical computing environment. The precise mechanisms of these algorithms and their potential limitations are discussed in detail elsewhere [49, 59].

Once a model has been calibrated (fitted) in this way, a range of statistical tests can be used to assess how closely the model represents the observed data (model fit). These can include assessments of global model fit across a scale, local fit of specific items or item pairs, examination of the ICCs, and a comparison of different models through likelihood ratio tests [60]. Under an RMT paradigm this might guide a PROM developer as to which items should be included in a scale, or under an IRT paradigm it might guide a PROM developer towards the most appropriate model to describe the nature of their items.

1.4.6 The prerequisites of modern test theory

To fit a modern test theory model, data should meet four important prerequisites. These are commonly known as the assumptions of modern test theory, although that term is slightly misleading as violations of these prerequisites tend to result in model misfit, rather than in a fitting model that makes incorrect assumptions. These prerequisites are monotonicity, unidimensionality, invariance, and local independence [61]. I introduce them here, *after* the concept of model fit, as in practice these can be considered *possible*

explanations of poor model fit.

Monotonicity, describes a non-decreasing relationship between the probability of endorsing a positive item response (or successive item responses) and the level of the latent construct [62]. For example, consider a PROM item that measures pain with the response options "no pain", "moderate pain", and "severe pain". As pain increases, a patient with "moderate pain" should not become more likely to endorse "no pain". A violation of this pattern (for example, when a second patient with a higher level of pain endorses a lower response option) is known as a Guttman error [63]. Monotonicity can be examined through Mokken analysis, which is a nonparametric form of IRT that examines the ability of individual items to rank respondents by total score (respondents who endorse "severe pain" in this item should rank higher than those who endorse "no pain" in this item, when all item responses are combined, assuming no Guttman errors). Scalability coefficients, known as Loevinger's H statistics, can be used to compare the number of observed Guttman errors to the number that would be expected in a group of completely unrelated items [63].

Unidimensionality describes a set of items that all measure a single latent construct [61]. When item scores are combined into a single index, that score should reflect a single construct. Consider a PROM that contains items about pain and function but provides a single score that combines these constructs. If a patient's function improves but pain worsens, these changes might cancel each other out when combined to a single score. It would appear that the patient has not changed at all when in fact two important changes have occurred. For that reason, items should measure one latent construct only, and all items in a scale should measure the same latent construct. There are potential exceptions to this rule when working with more complex multidimensional item response theory (MIRT) [64], and these will be explored later in this thesis.

There are a number of data-driven approaches to exploring the dimensionality of an item response dataset. For example, factor analysis, which describes a dimension reduction

technique that identifies covariance between item responses [65]. Correlation between responses in a group of items suggests that these items measure the same latent construct. If the responses to a second group of items correlate with each other, but not with the first, it suggests that this second group of items measure a second latent construct. The strength of the correlation between responses to each item and the latent construct (or factor) can then be quantified (the factor loading). Factor analysis can be exploratory, in which the factor loadings of every item onto every factor are freely estimated, or confirmatory, where certain parameters are constrained to check the fit of a prespecified model that represents the relationship between items and factors as hypothesised by the researcher. An extension of factor analysis is structural equation modelling (SEM). In addition to the relationships between items and factors, SEM explores the relationship between one factor and another through regression modelling [66].

A second approach to assessing dimensionality is to use another branch of Mokken analysis, specifically a tool known as the automated item selection procedure (AISP), which divides a set of items into unidimensional scales, using Loevinger's H statistics, to create scales where each item's responses can monotonically order respondents by their total score within that scale [62, 67].

The third prerequisite, measurement invariance, describes the ability of an item to provide consistent measurement across heterogeneous groups [61]. Items that cannot are said to exhibit differential item functioning (DIF). An example of this is an item that aims to measure physical functioning by asking "do you need assistance to use the toilet?" Respondents who use squat toilets will require assistance at much higher levels of physical functioning than respondents who use western toilets. A positive endorsement to this item implies different levels of physical functioning when used in these two groups [68]. Typically, DIF is checked for by demographics such as age, sex, and country of origin. This can be done by fitting logistic regression models that predict the response to an item using the level of the latent construct [69]. The addition of other covariates to this model (such as age, sex, or country of origin) should not significantly improve the

model's ability to predict the response to that item.

Finally, modern test theory requires items to function independently [61]. The response to one item should not be conditional on the response to another and the covariance between two items should be explainable by the level of the latent construct. In other words, the responses to two items should only be related because they both measure a single, intended, latent construct. Items can be checked for residual covariance, after the level of the latent construct has been accounted for. A common heuristic is that Yen's Q3 statistic of residual covariance between any two items should be < 0.2 [70]. A higher level may suggest that two items are related for another reason, for example one item's response influences the other, or both items are measuring a second, unintended, latent construct.

1.5 Computerised adaptive testing

As we have seen in subsection 1.4.3, items provide differing levels of information, $I(\theta)$, when posed to respondents with differing levels of ability. Well-targeted items will provide large amounts of information and poorly targeted items will provide little information. When constructing a pen-and-paper or linear questionnaire (one in which all items are posed to all respondents), it is important to include items that together provide information across the range of ability levels that are to be measured. Unfortunately, this creates redundancy at the person-level, particularly when a scale aims to provide measurement across a broad range of ability. Only a small proportion of items will be precisely targeted to the individual completing them, the other items are primarily included to provide measurement in respondents with different ability levels.

Because items function independently under modern test theory, the scores of two people can be compared even when they have answered a different set of items from the assessment. This creates an opportunity to selectively administer only the items that are productive for measurement in any given individual. This can be achieved through

computerised adaptive testing (CAT) [71, 72]. In CAT, items are selected one-by-one for efficient measurement, based on the respondent's level of ability, estimated in the assessment so far. There are three key components to CAT algorithms: an item selection criterion, a score estimator, and a stopping rule.

At the start of the CAT assessment, before any items have been administered, an ability estimate is chosen for the respondent. This is often $\hat{\theta} = 0$, although under a Bayesian framework, it can be adjusted based on prior assumptions about the distribution of θ in the population, or previous $\hat{\theta}$ estimates for that person [73].

Next, the first item is selected for the assessment. This is often the item that provides the most information at the chosen level of $\hat{\theta}$. Based on the response to this item, the person's ability (score) estimate is updated. Bayesian and frequentist estimators that are commonly used for this, including EAP, MAP, and MLE, have been discussed in subsection 1.4.4.

The next item is selected for administration based on this updated ability estimate. A number of item selection criteria have been described, and as with score estimators, these can align with either the principles of Bayesian or frequentist inference. Possibly the simplest item selection criterion is maximum Fisher information (MI). For each remaining item, $I(\hat{\theta})$ is calculated and the item with the highest value is selected. The disadvantage of this technique is that the estimate, $\hat{\theta}$, may be inaccurate and this could lead to an inefficient item selection [73]. This potential bias can be addressed by weighting the Fisher information by either the likelihood function of the items that have been administered so far (maximum likelihood weighted information, MLWI), or by the posterior distribution (maximum posterior weighted information, MPWI). Another approach to item selection is to consider how the expected response to each remaining item would affect the current measurement precision, and select the next item based on this. For example, maximum expected information (MEI) aims to maximise $I(\hat{\theta})$ given the expected response to each remaining item. A Bayesian alternative to MEI is

minimum expected posterior variance (MEPV), which aims to minimise the variance of the posterior, $P(\theta|\mathrm{y})$, instead (see Equation 1.14) [73].

This cyclic process of score estimation and item selection continues until a stopping rule has been met. This might be a certain number of items, a time limit, a threshold of $SE(\hat{\theta})$, a minimum change in $\hat{\theta}$ or $SE(\hat{\theta})$ between items, or a combination thereof. Within the limited literature on CAT for PROMs, a stopping rule of $SE(\hat{\theta}) < 0.3$, which roughly equates to a marginal reliability of 0.9, is emerging as a threshold level of precision that is deemed desirable [74–76], although I challenge this heuristic later in this thesis.

Other algorithmic components include exposure controls and content balancing, although these are not commonly used within healthcare assessments. With CAT, there is a danger that selecting only the items that lead to highest measurement precision reduces the content validity coverage of the questionnaire. It may also be undesirable to always start an assessment with the same item, particularly in educational assessments. Exposure control mitigates excessive selection of any given item, for example by randomly selecting one of the n most informative items. Content balancing is another way to ensure adequate variety in item selection. For example, items can be grouped by content, then selected from desired content groups to ensure appropriate group representation across the assessment [77].

Computerised adaptive testing is potentially most effective when combined with item banking. In linear assessments, we must make a parsimonious selection of items to cover a broad range of ability while minimising the burden on the assessment taker (the number of items). Breadth of measurement is favoured over fine precision at any level of the latent construct. However, when an assessment is built with CAT in mind, a very large number of items can be generated to finely target a broad range of ability. Only the most relevant items from this bank need be administered, with burden limited by the stopping rule. Computerised adaptive testing can result in both shorter *and* more precise

assessments than short-form versions of the item bank [75].

Since the 1980s, CAT has become increasingly popular for educational assessments and is now widely used in this sphere, including for school assessments, medical licensing exams, and military aptitude testing [78]. More recently, CAT has been applied to PROMs that have been calibrated with modern test theory. The most notable CAT system for PROMs is the Patient Reported Outcomes Measurement Information System (PROMIS), which contains CAT assessments for a range of latent constructs, including anxiety, depression, cognitive function, fatigue, and physical function [79]. The PROMIS platform is available through the EPIC electronic health records (EHR) platform (which is widely used in the USA), and this has undoubtedly contributed to its impressive uptake. Within the platform, scores are displayed as standardised T scores with reference to general population norms [80]. The PROMIS CAT assessments use a stopping rule of 12 items or $SE(\hat{\theta}) < 0.3$ for adults, and 12 items or $SE(\hat{\theta}) < 0.4$ for children [76].

Another imminent advancement is likely to come from the European Organisation for Research and Treatment of Cancer (EORTC) working group, who have recently developed a bank of 262 items for assessing the impact of cancer and its treatment. It is intended that these items will be administered through CAT [81], and it is hoped that this could revolutionise quality of life measurement in cancer, much like the EORTC-QLQ-C30 questionnaire did in the 1990s [82, 83].

Currently, most of the literature on CAT in PROMs has focused on the development and evaluation of CAT systems, with some clinicians reporting treatment outcomes that have been captured by the PROMIS system. Very few CAT studies have followed a randomised controlled trial (RCT) design, and these have tended to examine CAT as an intervention rather than leverage the technology as a primary outcome measure [84–86]. But a recent review of ongoing clinical trials suggests that the uptake of CAT in healthcare research is growing and the literature is likely to reflect this shortly [87].

There are limitations to currently available CAT platforms that may have hindered their uptake. Firstly, many CAT assessments, such as those available through PROMIS, are generic measures, as opposed to condition-specific PROMs. While generic PROMs are broadly applicable across a range of settings, they may fail to comprehensively capture the elements of health that are important to patients with specific health issues or less common symptoms. For example, there is no PROMIS CAT that specifically addresses cystic fibrosis chest symptoms, the impact of eating disorders on social function, or the appearance outcomes of breast reconstruction surgery.

Secondly, CAT assessment scores are often interpreted through comparison with normative values from generic populations [80]. This is potentially less useful for clinicians than comparison to populations matched by demographic and clinical characteristics. For example, a 10-year-old boy with Becker muscular dystrophy may have very poor *Strength Impact* and *Physical Activity* PROMIS CAT scores, as these are compared to a general population. But when compared to other 10-year-old boys with Becker muscular dystrophy, his scores may be clinically excellent and not warrant further investigation.

Thirdly, as already alluded to, current CAT systems impose stopping rules based on psychometric heuristics such as an $SE(\hat{\theta})$ under 0.3 or 0.4 [74–76]. These are not necessarily patient-centred, low-burden, or generalisable across assessments or populations. I would suggest that a more relevant approach would be to include patients and other stakeholders in the development of CAT algorithms and the setting of stopping rules.

Finally, many CAT platforms must relay item responses from an EHR to an external assessment centre where the next item is selected [88]. A more secure and efficient solution would be a locally implementable system that could perform score estimation and item selection off-line.

1.6 Cleft lip and/or palate

This work forms part of a broader movement towards applying and developing modern psychometrics across and beyond plastic surgery. As a subspecialty, the care of people born with a cleft lip and/or palate (CLP) is a natural starting point for this research.

Cleft lip and/or palate describes a spectrum of congenital anomalies resulting in partial or total discontinuity of the upper lip and oral cavity. These anomalies can have profound implications for a child's feeding, appearance, speech, and psychosocial development [89]. The impact on the child's family can also be significant [90].

1.6.1 Embryology, epidemiology, and classification

The lip, alveolus, and primary palate (the portion of the hard palate anterior to the incisive foramen) share an embryological journey that is distinct to the secondary palate (the hard palate posterior to the incisive foramen, and the soft palate). By the 6th week of gestation, the lip, alveolus, and primary palate are formed by the fusion of the medial nasal processes with one another, and with the maxillary processes. By the 10th week, the palatal shelves, arising from the expanding maxillary processes, fuse to form the secondary palate. Failure of these mechanisms can occur in isolation or together, resulting in isolated cleft lips (which can be associated with clefts of the alveolus and primary palate), isolated clefts affecting the secondary palate, or clefts affecting the lip, alveolus, primary palate, and secondary palate [89]. Between 5% and 7% of orofacial clefts occur as part of a genetic syndrome (there are currently over 600 known syndromes associated with CLP) [89], and a number of environmental causes, such as smoking, maternal malnutrition, and certain medicines, have also been implicated in this process [91–93].

Globally, CLP occurs in approximately 1 in 700 births, and while this has remained relatively stable over time, the incidence and prevalence does vary by country (for example, in Venezuela, the rate is thought to be almost 1 in 260) [94]. Cleft lip with or

without cleft palate is approximately twice as common in males than in females, whereas isolated cleft palates are approximately 50% more common in females [95]. In most countries, cleft lip with or without cleft palate is more common overall than isolated cleft palate [94].

Many classification systems have been proposed for categorising orofacial clefts into aetiological, anatomical, or clinically relevant subgroups [96]. In this thesis, I categorise orofacial clefts morphologically as follow: isolated cleft lip (CL); cleft lip and alveolus (CLA); cleft lip, alveolus, and palate (CLAP); and isolated cleft palate (CP).

Cleft lips are described as unilateral (where the cleft affects either the left or right side) or bilateral (where the cleft affects both sides). Cleft lips are then further subdivided into clefts which are complete, incomplete, or in the case of bilateral cleft lips, asymmetric. A complete cleft lip refers to one in which the cleft extends vertically from the lip into the soft tissue of the alar base, often accompanied by clefting of the alveolar arch and premaxilla. An incomplete cleft lip refers to a less severe form which spares the alar base and deeper structures [97]. An asymmetric bilateral cleft lip comprises a complete cleft lip on one side and an incomplete cleft lip on the other [98]. Bilateral cleft lips tend to have poorer appearance outcomes than unilateral cleft lips, and complete clefts lips tend to have poorer appearance outcomes than incomplete cleft lips [99].

In categorising palatal clefts, I apply Victor Veau's 1931 classification [100], which defines four cleft palate morphologies:

I Clefts isolated to the soft palate

II Clefts of the whole secondary palate

III Clefts of the whole secondary palate, extending into the primary palate and unilaterally through the alveolus

IV Clefts of the whole secondary palate, extending into the primary palate and bilaterally through the alveolus

While the Veau classification is primarily an anatomical classification, ascending classes are expected to indicate increasing severity and greater challenges to manage [101]. Notably, patients with ascending Veau class report increasingly poor speech function and more significant speech distress [99].

1.6.2 Treatment

Generally, people born with a CLP benefit from a series of treatments from infancy to early adulthood [102]. A treatment plan might include reconstructive surgery (of the lip, alveolus, palate, oropharynx, nose, jaw, or maxilla), orthodontics, reconstructive dentistry, psychological therapy, speech and language therapy (including prostheses), audiological treatments, genetic counselling, and social care. To illustrate, a treatment pathway for a patient with a unilateral CLAP, might take a similar form to the following:

1. Prenatal screening detects the cleft by 20 weeks of gestation and the parents are put in contact with the regional cleft centre.

2. The child undergoes a feeding assessment within 48 hours of birth.

3. An initial appointment with the cleft multidisciplinary team (MDT) takes place within 3 months of birth. This may include an audiology assessment, genetic screening, and counselling for the parents.

4. Between 3 months and 6 months of age, the child undergoes primary lip repair surgery to correct the lip and nasal alar. This may be accompanied by closure of the hard palate with vomerine flap surgery.

5. At 12 months, the child undergoes primary repair of the soft palate.

6. At 18 months, the child starts speech and language therapy.

7. At 5 years, a thorough assessment is undertaken by the MDT, which may include psychological screening, video fluoroscopy, and a dental examination.

8. The child then undergoes orthodontic treatment, prior to alveolar bone grafting, which takes place around the age of 8 to 10 years, prior to the eruption of the secondary canines.

9. At 10 and 15 years, further assessments are undertaken by the MDT. Ongoing dental, orthodontic, speech and language, social, and psychological support is provided as needed.

10. From the age of 16, the child may be offered a rhinoplasty to address any problems with the form or function of their nose. Scar revision surgery may be offered to improve the appearance of the cleft lip scar, and orthognathic surgery may be offered to alter the profile of the jaw. Further treatments, such as pharyngoplasty, may be required to correct velopharyngeal dysfunction.

The choice and timing of treatment is contentious, highly variable, and lacks a robust evidence base. In 2001, the Eurocleft study surveyed all 201 European cleft units and found 194 different treatment protocols for a unilateral cleft lip and palate. The only shared treatment protocols were those mandated by ongoing clinical trials [103]. Within the UK, undesirable treatment variation was linked to poor patient outcomes by the Clinical Standards Advisory Group (CSAG) cleft lip and palate study in the late 1990s [104]. These findings led the UK and Republic of Ireland to centralise cleft care to 12 specialist teams across 18 hospital sites. Since this change, outcomes have improved nationally [105], but there remains a paucity of high-level evidence to support most treatment decisions, and wide variation internationally. For example, proponents of early palate repair argue that this benefits speech development and later speech outcomes, whereas proponents of delayed or two-stage palate repair argue that their strategy is less detrimental to maxillary growth [106].

Prospective clinical trials are particularly difficult to conduct in CLP. The final result of an intervention may not be apparent for many years, and trials with lengthy follow-ups can be very expensive. Patients with CLP will often need many treatments in succession, which can complicate the comparison of a single discrete intervention and control. Until recently,

there has also been very little consensus on which outcome measures to use in research and clinical practice, and this hinders studies of comparative treatment effectiveness.

1.6.3 Outcome measurement

The metrological concepts described in this chapter have been introduced in the context of PROMs and other psychometric assessments, but many are directly applicable to any health outcome measure. Outcome measures should capture meaningful elements of health. In general, they should provide continuous data, and they should be monotonic, unidimensional, and demonstrate measurement invariance. This is not the case for most outcome measures used in CLP. Most popular assessments, including the Asher-McDade rating scale of facial photographs [107], the Great Ormond Street Speech Assessment tool [108], and the Goslon yardstick for dental arch relationships [109], provide ordinal, clinician-derived, operator-dependent measurement. One of the most anticipated ongoing clinical trials in CLP is the Timing of Palate Surgery (TOPS) study, which will compare 5-year speech outcomes in children who have undergone palate repair at either 6- or 12-months [110]. The primary outcome measure of the TOPS trial is the Velopharyngeal Composite Score, which is based on clinical assessments of hypernasality, non-oral errors, and velopharyngeal insufficiency. A score of 0, 1, or 2 is achievable with increasing severity in each domain, and these scores are summed to an overall index ranging from 0 to 6. A threshold of ≥ 4 has been chosen to indicate "insufficient" velopharyngeal function, and the proportion of patients with this classification will be compared between trial arms for the primary analysis [110]. The TOPS trial is at the very forefront of evidence generation in this field, but with a lack of alternatives, the trialists have had to use a primary outcome measure which is subjective, ordinal, multidimensional, arbitrarily dichotomised, and does not directly address the patients' own impression of their health.

The World Health Organisation (WHO) has all called for improvement and consensus in the measurement of CLP outcomes in the hope that this will support the evidence base behind treatment options in CLP, and in turn drive up standards in care internationally.

Specifically, the WHO has advocated outcome measures that "matter to ordinary people rather than sophisticated surrogate measures that may have little relevance in everyday life" [111].

In 2017, the International Consortium for Health Outcomes Measurement (ICHOM) released a standard set of outcome measures for the "comprehensive appraisal of cleft care", based on an international, multidisciplinary Delphi consensus study. The standard set makes recommendations about the timing and method of assessment of health outcomes across a broad range of domains: eating and drinking; dental and oral health; speech and communication; otologic health; breathing; appearance; psychosocial development; and burden of care [112].

Of the 20 outcome measures recommended by the ICHOM set, 11 are PROMs, reflecting the culture shift towards holistic, person-centred assessment, research, and care. Of these, 9 are components of the CLEFT-Q, the first PROM developed specifically to assess health outcomes in CLP [113].

1.6.4 The CLEFT-Q

A series of systematic reviews published in 2008, 2011, and 2012, failed to identify any PROM that had been developed specifically to address the health needs of people born with a CLP [114–116]. At the time, 29 different PROMs had been applied to CLP populations [117] and some form of validity had been demonstrated in 5 of these instruments: the Youth Quality of Life – Facial Differences questionnaire, the Pediatric Voice-Related Quality of Life questionnaire, the Cleft Audit Protocol for Speech – Augmented questionnaire, the Child Oral Health Impact Profile, and the Child Oral Health Quality of Life questionnaire. Only the development of the Youth Quality of Life – Facial Differences questionnaire had included qualitative patient input, and overall, the content validity of these instruments was suboptimal [113].

Following expert input, a systematic review of health constructs which had been measured previously in CLP [116], and 136 qualitative interviews with patients across 6 countries, the CLEFT-Q developers generated a conceptual framework and item bank to comprehensively assess the health constructs most important to people born with a CLP [113]. Items and scales were refined in 3 rounds of cognitive interviews and expert feedback. The psychometric properties of these scales were then evaluated in a large international field test [118].

Between October 2014 and November 2016, the CLEFT-Q field test recruited 2434 participants aged 8 to 29 years, at 30 cleft centres across 12 countries. The principal aim of the study was to establish construct validity through RMT analysis (specifically using the partial credit Rasch model). Participants were also asked about the impact of completing the questionnaire to explore acceptability. Following Rasch analysis, 110 items were retained in the final CLEFT-Q questionnaire. These were grouped into 12 scales across 3 domains (appearance, health-related quality of life, and facial function) [118].

The appearance domain contains 7 scales, which measure the appearance of: the face, the nose, the nostrils, the lips, the cleft lip scar, the jaws, and the teeth. Within the health-related quality of life domain, scales measure: psychological function, school function, social function, and speech distress. The facial function domain contains one scale: the speech function scale, and also a 9-item eating and drinking checklist which did not fit the partial credit Rasch model [118].

Each scale contains 6-12 items, and each item has either 3 or 4 response options. The appearance scales require participants to reflect on their appearance momentarily, whereas the other scales have a recall period of a week. Some of the items contain illustrations based on cognitive interview feedback, and the mean Flesch-Kincaid grade level (a readability statistic roughly corresponding to the US school grade system) for the items is 1.4 (range 0 to 5.2) [113]. The CLEFT-Q is currently available in 26 different

languages [119], with one scale per page and a conversion table for each scale that translates the sum of item scores into a 0-100 Rasch score, with a higher score indicating a better clinical status. When asked, the majority of field test respondents stated that they enjoyed completing the questionnaire [120].

Besides partial credit Rasch model fit, construct validity has been examined through hypothesis testing, where the scores of groups that are expected to differ clinically are compared statistically. For relevant scales, the cross-sectional scores of people awaiting rhinoplasty, lip revision, speech surgery, and orthognathic surgery were found to be significantly poorer than those who had completed their surgery [121]. Bilateral cleft lips are associated with poorer CLEFT-Q appearance scale scores than unilateral cleft lips, and CLEFT-Q speech function and speech distress scores become worse with ascending Veau class [99]. Convergent validity (correlation) has been demonstrated between CLEFT-Q scales and instruments that are intended to measure similar constructs, including scales from the Child Oral Health Impact Profile, and the Cleft Hearing Appearance and Speech Questionnaire, a modified version of the Satisfaction With Appearance questionnaire for facial differences, which was developed by the Cleft Psychology Special Interest Group of the Craniofacial Society of Great Britain and Northern Ireland to assess facial appearance, hearing, and speaking in CLP [122, 123]. A recent prospective study has confirmed the responsiveness of relevant CLEFT-Q scales to changes following rhinoplasty, scar revision and orthognathic surgery, and has proposed provisional target differences for clinical trials based on distributional statistics [124].

While the CLEFT-Q has demonstrated excellent content validity and construct validity, it is perceived as burdensome and this may have hindered its uptake [125]. While CLEFT-Q scales are independently functioning, and only contain up to 12 items each, there are points in the patient journey where the ICHOM recommend administering up to 75 CLEFT-Q items at once [112]. Qualitative interviews of clinicians from the Netherlands, Sweden, United States, and United Kingdom have identified burden as a key barrier to

implementing the CLEFT-Q and ICHOM standard set in clinical practice [126, 127].

In a survey published in 2021, clinicians at 15 of the 18 cleft units in the UK and Republic of Ireland described burden as a barrier to implementing the CLEFT-Q, and clinicians at 16 of these units felt that an online CAT platform that could reduce the response burden of the CLEFT-Q would directly facilitate its implementation in their service [128].

1.7 Summary of background

To summarise, modern test theory can improve the quality of measurement derived from PROMs, and this may be particularly relevant for the care of people born with a CLP. The CLEFT-Q is a novel PROM designed to measure the health constructs that matter most to people born with a CLP. It has been rigorously developed, with superior content validity and construct validity to other available measures. The CLEFT-Q could directly improve clinical practice by enhancing communication between patients and clinicians, and expediting the detection and treatment of health-related quality of life problems. It could also enhance the evidence base in cleft care internationally, as a high-quality outcome measure for clinical research and audit. But the perceived response burden of completing the CLEFT-Q is limiting its potential.

Because the CLEFT-Q has been built with RMT, each item functions independently. This means that it would be possible to apply CAT to each scale, either through RMT or IRT. This might make the scales shorter, reduce response burden, and facilitate the CLEFT-Q's implementation in clinical practice and research.

Furthermore, there is an innovation gap in the field of healthcare CAT research that could be exploited by a system that:

1. Provides condition-specific measurement

2. Sets patient-centred stopping rules
3. Executes CAT algorithms locally
4. Contextualises results through comparison to population reference values, stratified by relevant clinical and demographic characteristics

1.8 Aims, objectives, and thesis outline

The research aims addressed in this thesis are to:

1. Develop a CAT version of the CLEFT-Q (the CLEFT-Q CAT) that could facilitate the questionnaire's uptake in routine clinical practice, address some of the limitations of current CAT systems, and easily translate into other clinical fields.
2. Explore potential future research directions for CAT, in the context of the CLEFT-Q CAT's limitations.

Chapters 2, 3, and 4 address the first aim through the following objectives:

Chapter 2. Optimise CAT algorithms for the CLEFT-Q scales, by evaluating the performance of different score estimators and item selection criteria in Monte Carlo simulation studies, using simulated data.

Chapter 3. Evaluate the performance of optimised CAT algorithms in a real patient dataset, and determine stopping rules for these assessments by multistakeholder consensus.

Chapter 4. Develop user interfaces for CLEFT-Q CAT administration and score interpretation, deploy these in clinical practice, and qualify the impact of this system as perceived by patients and clinicians.

Chapters 5, 6, and 7 address the second aim through the following objectives:

Chapter 5. Compare the SE of CAT algorithms built through RMT and IRT, and explore whether relatively higher SE values achieved by RMT algorithms translate into poorer agreement between CAT scores and full-length assessment scores.

Chapter 6. Develop and evaluate a conceptually-driven structural equation model that describes the relationships between CLEFT-Q appearance scales, and explore the potential to use these relationships to enhance CAT efficiency through multidimensional assessment.

Chapter 7. Compare the accuracy and efficiency of conventional CAT algorithms to novel techniques that might incorporate EHR-derived data as part of the assessment.

In **Chapter 8**, I discuss the principal findings of this work, its strengths and limitations, and future research priorities for clinical psychometrics and outcome measurement in plastic surgery.

2

A comparison of estimators and item selection criteria for the CLEFT-Q CAT

Contents

2.1 Introduction 42
2.1.1 Latent construct estimators 42
2.1.2 Item selection criteria 43
2.2 Methods 44
2.2.1 Software 44
2.2.2 Data 45
2.2.3 Rasch model parameterisation 45
2.2.4 Latent construct estimators 46
2.2.5 Item selection criteria 47
2.3 Results 47
2.3.1 Demographics, clinical characteristics, and missing data . . . 47
2.3.2 Latent construct estimators 48
2.3.3 Item selection criteria 52
2.4 Discussion 54
2.5 Conclusion 56

2.1 Introduction

In the previous chapter, I introduced the three major components of a modern test theory CAT algorithm (a latent construct estimator, an item selection criterion, and a stopping rule), and described different techniques for both latent construct estimation and item selection. The purpose of this chapter is to decide which of these techniques should be used in the development of the CLEFT-Q CAT, by evaluating the performance of different score estimators and item selection criteria in Monte Carlo simulations.

2.1.1 Latent construct estimators

To reiterate, latent construct estimators can be based on the likelihood function (or log-likelihood function), following a frequentist approach to statistical inference, or they can be based on the posterior probability distribution, following a Bayesian approach to statistical inference [50, 51, 73].

I introduced two Bayesian estimators in the previous chapter:

- **EAP**: the mean of the posterior distribution
- **MAP**: the mode of the posterior distribution

and a single frequentist estimator:

- **MLE**: the mode of the likelihood (or log-likelihood) function

As I alluded to in the previous chapter, a limitation of MLE is that it produces positively biased θ estimates for high scoring respondents and negatively biased θ estimates for low scoring respondents [129]. This bias is inversely proportional to the number of

---

This chapter is based on Harrison, C. *et al* Optimising the computerised adaptive test to reliably reduce the burden of administering the CLEFT-Q: A Monte Carlo simulation study. *Journal of Plastic Reconstructive and Aesthetic Surgery* 74(6), 1355-1401 (2021).

items administered [130]. In 1989, Thomas Warm demonstrated that this bias could be mitigated by applying a weighting function to the MLE. Warm's weighted likelihood estimation (**WLE**) has been shown to provide less bias θ estimates than MLE [56, 130, 131]. Some simulation studies have suggested that WLE provides less bias estimates than Bayesian methods [130, 131], while other simulations have found similar performance between WLE, EAP, and MAP [50, 56].

It is unclear which latent construct estimator would provide the most accurate CLEFT-Q scale $\hat{\theta}$ estimates from truncated CAT assessments. While MLE results in "outward" bias (overestimating high scores and underestimating low scores), Bayesian techniques result in "inward" bias, drawing the score towards the mean of the prior [50]. Well-specified priors will reduce bias, while inappropriate priors might increase bias, particularly when few items are used [50].

2.1.2 Item selection criteria

In the previous chapter, I also introduced Bayesian and frequentest approaches to item selection. These were:

Bayesian

- **MEPV**: the remaining item which would minimise the variance of the posterior distribution, given the expected response to each remaining item
- **MPWI**: the remaining item with the highest Fisher information, given the current $\hat{\theta}$ estimate weighted by the posterior distribution

Frequentist

- **MI**: the remaining item with the highest Fisher information, given the current θ estimate
- **MEI**: the remaining item which would maximise the Fisher information associated with the $\hat{\theta}$ estimate, given the expected response to each remaining item
- **MLWI**: the remaining item with the highest Fisher information, given the current $\hat{\theta}$ estimate weighted by the likelihood function

As with latent construct estimators, simulation studies comparing item selection criteria have produced contrasting results, although there is a tendency for MI to produce slightly poorer outcomes than other techniques [132–136].

The variability in the findings of these studies might be explained by variation in the number of items used, the number of response options per item, the IRT or Rasch model, the targeting of items, the appropriateness of the prior, or any other factor that might affect the likelihood function or posterior probability distribution. For this reason, it is difficult to generalise existing literature to address this chapter's goal. The best way to decide which estimator and item selection criterion to use for the CLEFT-Q CAT is through *in silico* experimentation.

2.2 Methods

2.2.1 Software

I performed all analyses using R (version 3.6.1) with the following packages installed: *mirtCAT* (version 1.9.3) [77], *mirt* (version 1.31) [49], and *foreign* (version 0.8-71).

2.2.2 Data

I generated Partial credit Rasch models for each of the 12 CLEFT-Q scales using data from the CLEFT-Q field test [118]. This dataset included CLEFT-Q scale responses from 2434 participants, aged 8 to 29 years, and undergoing treatment for CL, CP, CLA, or CLAP at one of 30 hospital sites across 12 countries. The jaw scale was only completed by participants aged 12 years or older, the cleft lip scar scale was only completed by participants born with a cleft lip, and the school scale was only completed by participants aged 8 to 18 years and attending school. Participants completed the CLEFT-Q at a single time point, which was unrelated to the course of their treatment. Demographics and descriptive statistics for this dataset, which have been published previously [118], are representative of intended CLEFT-Q end users.

I used this dataset for consistency with existing CLEFT-Q Rasch models. It is the same sample used to perform the initial Rasch modelling of the CLEFT-Q and derive the questionnaire's Rasch scoring table.

The sample size is considered large for Rasch analyses. For example, one heuristic suggests that 250 participants is enough to provide model parameter estimates with over 99% confidence to within 0.5 logits [137].

For each scale, I handled missing item responses through listwise exclusion. I chose this over imputation to avoid potentially inflating statistical relationships between item responses which Rasch analysis goes on to identify.

2.2.3 Rasch model parameterisation

Before fitting partial credit Rasch models, I collapsed the middle response options in the Speech Distress and Speech Function scales, and reversed their scoring for consistency

with the final version of the CLEFT-Q [118]. I used the EM algorithm available in the *mirt* package for item parameterisation [49].

2.2.4 Latent construct estimators

I programmed a Monte Carlo simulation to compare EAP, MAP, MLE, and WLE estimators, using the *mirtCAT* package.

First, I simulated 1000 θ values for each CLEFT-Q scale. To ensure these scores had a realistic distribution, I created the sample by bootstrapping $\hat{\theta}$ values from the CLEFT-Q field test, which I computed with the partial credit Rasch models described above, and EAP θ estimation. In this simulation, these numbers represented *true* θ values from 1000 simulated respondents, for each CLEFT-Q scale.

Next, I computed plausible response sets to each CLEFT-Q scale, for each simulated respondent, based on their simulated true θ value and the Rasch model parameters. In other words, I created a simulated dataset of CLEFT-Q item responses, and corresponding θ values. The premise behind this experiment is that an accurate estimator should be able to closely reproduce each simulated respondent's θ value from their response set.

Next, for each CLEFT-Q scale, I estimated $\hat{\theta}$ for each respondent, using the respondent's complete response set. I did this with each estimator in turn. For each scale, and each estimator, I compared the true θ value to the estimate, $\hat{\theta}$. To do this, I used Pearson's correlation coefficient and root mean square error (RMSE, the quadratic mean of the differences between the pre-specified θ values and their estimates). A higher correlation, and lower RMSE, reflect a more accurate estimator (one which more closely reproduced the true θ value from its corresponding response set).

2.2.5 Item selection criteria

I programmed a second Monte Carlo simulation to compare MEPV, MPWI, MI, MEI, and MLWI item selection criteria.

In this experiment, I developed CAT algorithms for each CLEFT-Q scale using the *mirtCAT* package. In this experiment, I applied CAT to reproduce θ values from the simulated response sets, to a pre-specified level of precision, and assessed algorithmic efficiency. The most efficient CAT algorithm would reproduce θ from the fewest item responses.

The algorithms mimicked real-world CAT assessments, by selecting item responses one at a time for each respondent, and estimating $\hat{\theta}$ after each response, until a stopping rule was met. For consistency, MLE was used to estimate $\hat{\theta}$ each time. I repeated the experiment with stopping rules of SE < 0.65, SE < 0.55, and SE < 0.45, for each item selection criterion.

For each respondent and each scale, I recorded the mean number of items used to reach the stopping rule, as well as the Pearson correlation coefficient and RMSE between pre-specified θ values and their estimates. In this case, the mean number of items is preferable to the median, which can only be an integer or number ending in .5, and is therefore less appropriate for comparing subtle differences in the number of items administered.

2.3 Results

2.3.1 Demographics, clinical characteristics, and missing data

The demographics and clinical characteristics of CLEFT-Q field test respondents are presented in Table 2.1. Generally, there was a slight preponderance of males in the sample, which had a median age of 14 years. Cleft lip, alveolus, and palate was more

common than CL, CP or CLA. Unilateral and complete cleft lips were more common than bilateral or incomplete cleft lips, and Veau III was the most common class of cleft palate. A more detailed description of demographics and clinical characteristics is presented in Appendix A (Section A.1), for each of the 12 CLEFT-Q scales.

Missing item response data are presented in Appendix A (Section A.2). Missingness was mostly explainable by other variables and deemed unlikely to limit the generalisability of Rasch model parameters. For example, 84% (837/991) of participants missing one or more jaw scale item(s) were under the age of 12 years. This is because only CLEFT-Q field test participants aged 12-29 years were asked to complete jaw scale items [118]. Similarly, 50% (307/614) of participants missing one or more school scale items were not attending school (and therefore not administered these items in the CLEFT-Q field test). In this sample, 43% (263/615) of participants missing speech distress items, and 39% (263/670) of those missing speech function items were born with a cleft lip only, and therefore unlikely to use these scales in a real-world setting. There was no obvious pattern in the specific items missed in each scale.

2.3.2 Latent construct estimators

The correlation and RMSE between true and estimated θ values varied between scales. Scales with more items (for example the nose and speech function scales, each with 12 items), produced more accurate estimates (lower RMSE) than scales with fewer items (for example the jaw scale, with 7 items or the nostrils scale, with 6 items).

For each scale, there was very little difference in the correlation or RMSE between estimators. Weighted likelihood estimation achieved a marginally higher RMSE (poorer accuracy) than other methods in every scale, but there were no other consistent trends. These results are presented in Table 2.2 and Table 2.3.

Table 2.1: Demographics and clinical characteristics. IQR: inter-quartile range; CL: cleft lip; CP: cleft palate; CLA: cleft lip and alveolus; CLAP: cleft lip, alveolus and palate.

Demographics and clinical characteristics	
Age	
Median (IQR)	14 (7)
Missing	1
Gender	
Male	1351
Female	1081
Missing	2
Country	
Australia	25
Canada	624
England	339
Ireland	100
USA	362
The Netherlands	206
India	232
Sweden	100
Turkey	54
Columbia	210
Chile	89
Spain	93
Missing	0
Cleft type	
CL	263
CP	568
CLA	204
CLAP	1399
Missing	0
Lip type	
Unilateral	1366
Bilateral	488
Missing	12
Palate type	
Veau I	258
Veau II	287
Veau III	943
Veau IV	394
Missing	85

Table 2.2: Comparison of latent construct estimators (lip, scar, teeth, jaw, nose, and nostril scales). RMSE: root mean square error; EAP: expected a-posteriori; MAP: Maximum a-posteriori; MLE: Maximum likelihood; WLE: Weighted likelihood estimation.

Scale	Estimator	Correlation	RMSE
Lip	EAP	0.97	0.65
	MAP	0.97	0.67
	MLE	0.97	0.65
	WLE	0.97	0.68
Scar	EAP	0.97	0.67
	MAP	0.97	0.69
	MLE	0.97	0.67
	WLE	0.97	0.70
Teeth	EAP	0.96	0.57
	MAP	0.96	0.57
	MLE	0.96	0.57
	WLE	0.96	0.60
Jaw	EAP	0.97	0.70
	MAP	0.97	0.73
	MLE	0.97	0.72
	WLE	0.97	0.76
Nose	EAP	0.97	0.52
	MAP	0.97	0.53
	MLE	0.97	0.53
	WLE	0.97	0.55
Nostrils	EAP	0.96	0.74
	MAP	0.96	0.76
	MLE	0.96	0.74
	WLE	0.96	0.78

Table 2.3: Comparison of latent construct estimators (face, psychological function, social function, school function, speech distress, and speech function scales). RMSE: root mean square error; EAP: expected a-posteriori; MAP: Maximum a-posteriori; MLE: Maximum likelihood; WLE: Weighted likelihood estimation.

Scale	Estimator	Correlation	RMSE
Face	EAP	0.94	0.58
	MAP	0.94	0.58
	MLE	0.94	0.60
	WLE	0.94	0.61
Psychological function	EAP	0.95	0.65
	MAP	0.95	0.66
	MLE	0.95	0.64
	WLE	0.95	0.67
Social function	EAP	0.93	0.58
	MAP	0.93	0.59
	MLE	0.93	0.57
	WLE	0.92	0.62
School function	EAP	0.92	0.63
	MAP	0.92	0.63
	MLE	0.93	0.62
	WLE	0.92	0.66
Speech distress	EAP	0.94	0.56
	MAP	0.94	0.56
	MLE	0.94	0.56
	WLE	0.94	0.59
Speech function	EAP	0.96	0.58
	MAP	0.96	0.59
	MLE	0.96	0.57
	WLE	0.96	0.60

2.3.3 Item selection criteria

The full results from the simulations comparing item selection criteria are presented in Appendix A (Section A.3). For illustration, the results from the simulations involving the face and psychological function scales are presented in Table 2.4 and Table 2.5.

All item selection criteria performed similarly in terms of accuracy and item reduction. At higher SE thresholds (lower precision requirements), the CAT algorithms posed fewer items. In general, item reduction was greatest for longer scales. For example, the nose scale, which is usually 12 items long, was reduced to a mean of approximately 9 items at the SE < 0.55 stopping rule, whereas the Nostrils scale, which is 6 items long, achieved no item reduction at this stopping rule. As fewer items were posed, RMSE increased. At a threshold of SE < 0.45, all algorithms posed all items.

Table 2.4: Comparison of item selection criteria for the face scale. SE: standard error of measurement; RMSE: root mean square error; MEPV: minimum expected posterior variance; MPWI: maximum posterior weighted information; MI: maximum information; MEI: maximum expected information; MLWI: maximum likelihood weighted information

Face scale (9 items)

	SE < 0.65			SE < 0.55			SE < 0.45		
Criterion	**Mean items administered**	**Correlation**	**RMSE**	**Mean items administered**	**Correlation**	**RMSE**	**Mean items administered**	**Correlation**	**RMSE**
MEPV	6.44	0.93	0.66	8.02	0.94	0.61	9.00	0.94	0.60
MPWI	6.43	0.93	0.66	8.01	0.94	0.61	9.00	0.94	0.60
MI	6.44	0.93	0.66	8.02	0.94	0.61	9.00	0.94	0.60
MEI	6.68	0.93	0.66	8.13	0.94	0.61	9.00	0.94	0.60
MLWI	6.44	0.93	0.66	8.02	0.94	0.61	9.00	0.94	0.60

Table 2.5: Comparison of item selection criteria for the psychological function scale. SE: standard error of measurement; RMSE: root mean square error; MEPV: minimum expected posterior variance; MPWI: maximum posterior weighted information; MI: maximum information; MEI: maximum expected information; MLWI: maximum likelihood weighted information

Psychological function scale (10 items)

	SE < 0.65			SE < 0.55			SE < 0.45		
Criterion	**Mean items administered**	**Correlation**	**RMSE**	**Mean items administered**	**Correlation**	**RMSE**	**Mean items administered**	**Correlation**	**RMSE**
MEPV	8.02	0.94	0.69	9.61	0.95	0.65	10.00	0.95	0.64
MPWI	8.03	0.95	0.69	9.61	0.95	0.65	10.00	0.95	0.64
MI	8.02	0.94	0.69	9.60	0.95	0.65	10.00	0.95	0.64
MEI	8.18	0.95	0.68	9.62	0.95	0.65	10.00	0.95	0.64
MLWI	8.02	0.94	0.69	9.61	0.95	0.65	10.00	0.95	0.64

2.4 Discussion

The principal finding of this study was that the choice of estimator or item selection criterion is likely to have a minimal impact on the accuracy or efficiency of the CLEFT-Q CAT. Each estimator achieved a similar accuracy, and at a given level of precision, CAT algorithms built with each item selection criterion posed a very similar mean number of items. Based on the findings of this study, the CLEFT-Q CAT has been built using EAP θ estimation and MEPV item selection. The rationale behind this decision is that EAP is less computationally intensive than the other estimators, which all aim to maximise a function through iterative calculation [50], and a Bayesian approach to θ estimation will never produce the unhelpful score estimates of -∞ or +∞ [37], as described in Chapter 1, Subsection 1.4.4. The natural partners to EAP under a Bayesian paradigm are MEPV or MPWI, and while these approaches produced very similar results, MEPV does not require the additional computation of a weighting function.

A secondary finding was that accuracy increased as more items were administered. This has been described previously [131] and is not surprising given the processes explained in Chapter 1, Subsections 1.4.3 and 1.4.3. As more items are administered (and responded to accurately), the likelihood function (or posterior distribution) becomes steeper, $I(\hat{\theta})$ increases as $\hat{\theta}$ approaches its true value, and $SE(\hat{\theta})$ decreases. This also explains why greater item reduction was observed with higher SE thresholds. Across all scales, the mean number of items required to reach the SE < 0.65 threshold was approximately 6-8, regardless of the number of items in the original scale, and CAT algorithms may therefore be more clinically useful for the longer scales.

Another notable finding was that CAT algorithms achieved no item reduction with a SE < 0.45 stopping rule. This SE threshold is relatively high, compared to other CAT platforms [74–76] and at first, this finding might be interpreted as poor performance of the CAT

algorithms built during this study. However, a key difference between these algorithms and most comparators, is that these algorithms were based on partial credit Rasch models, as opposed to IRT models with a freely estimated discrimination parameter, employed by other systems, such as PROMIS [138]. In Chapter 5, I explore this finding in more detail and demonstrate that the relatively high SE values generated by Rasch-based CAT algorithms do not necessarily translate into poorer accuracy.

The use of simulated data is a limitation of this study. I aimed to simulate data with a realistic distribution by bootstrapping $\hat{\theta}$ values from the CLEFT-Q field test, but this approach assumes that the CLEFT-Q field test dataset has a representative distribution of future CLEFT-Q CAT assessment takers. Furthermore, by simulating plausible response sets based on these values, I created a dataset where no items were responded to in error. In real life, participants may misread an item, or accidentally select an incorrect response, potentially increasing $SE(\hat{\theta})$. For these reasons, the use of simulated data in this study may have over-estimated the accuracy and efficiency of CLEFT-Q CAT algorithms. To address this, I perform external validation of these algorithms using real patient data, in the next chapter.

A second limitation of this study is that the accuracy statistics are not easy to interpret in real-world terms. Pearson's correlation coefficients were generally high (above 0.93), but this is a very crude measure of accuracy, and high correlation can exist in the presence of poor agreement [139]. The RMSE does measure agreement, but it averages out the agreement between all true θ values and their estimates. It is possible that large errors at uncommon θ levels bias this statistic. In the next chapter, I expand the analysis of CAT performance to evaluate respondent-level agreement between full-length assessment and CAT assessment, across latent construct levels, using real patient data.

2.5 Conclusion

In this chapter, I have demonstrated that the choice of score estimator and item selection criterion is unlikely to have a clinically meaningful impact on the CLEFT-Q CAT's accuracy or efficiency. Therefore, I have chosen to proceed with less computationally intensive techniques, under a Bayesian framework (EAP and MEPV). I have shown that reducing the number of items administered comes at the cost of precision (SE) and accuracy (RMSE), and to achieve item reduction, the SE threshold in these Rasch-based CAT algorithms is higher than the heuristics proposed for CAT algorithms built through IRT models with more parameters (for example, SE < 0.3) [74–76].

The next step in the development of the CLEFT-Q CAT was to explore this trade-off between accuracy and burden (number of items) in more detail, using real patient data, before setting stopping rules that reflect a reasonable balance between accuracy and burden in the eyes of stakeholders. This work is described in Chapter 3.

3

External validation and stopping rules for the CLEFT-Q CAT

Contents

3.1 Introduction 58
3.2 Methods 59
3.2.1 Validation data 59
3.2.2 Monte Carlo simulation 60
3.2.3 cleftqCATsim 61
3.2.4 Stopping rule workshop 61
3.3 Results 62
3.3.1 Demographics, missing data, and outliers 62
3.3.2 Simulation results 63
3.3.3 Workshop results 66
3.4 Discussion 67
3.5 Conclusion 70

3.1 Introduction

The three basic components of a modern test theory CAT algorithm are an estimator, an item selection criterion, and a stopping rule. In Chapter 2, I outlined the rationale for choosing the EAP estimator and MEPV item selection criterion for use with the CLEFT-Q CAT. The purpose of the work described in this chapter was to set stopping rules for each of the PROM's scales, that balance the trade-off between error and burden.

Typically, stopping rules are chosen by researchers, and often based on heuristics relating to measurement precision, such as SE < 0.3 or SE < 0.4 [74–76]. While this approach is straightforward, it does not necessarily capture the optimal balance of accuracy and burden as perceived by the algorithms' end users. These thresholds are unhelpful if they result in an assessment that is still unacceptably burdensome. In the case of the CLEFT-Q CAT, the results of Chapter 2 suggest that these stopping rules (SE < 0.3 or SE < 0.4) would result in very little item reduction, if any at all.

When SE-based stopping rules are used for CAT, different respondents may be posed a different number of items, depending on their response pattern. This might lead to a situation where some patients are posed an acceptable number of items, whereas others are not. For example, those participants with very high or low latent construct levels, where Fisher information is lowest and SE is highest, might be posed more items than respondents with latent construct levels that match the targeting of the questionnaire (see Chapter 1, subsection 1.4.3). This could have implications for care equity, and in a research setting, if assessment burden was related to both engagement with the assessment and the latent construct level, it could result in bias.

Involving end users in the setting of fixed-length stopping rules is potentially more equitable, and potentially more patient-centred, than striking the balance of error and

burden on their behalf. The research described in this chapter determines the error associated with CLEFT-Q CAT algorithms at different fixed-length stopping rules through Monte Carlo simulation with real patient data. Based on these results, I set stopping rules for each scale through a multi-stakeholder consensus workshop involving the algorithms' end users.

3.2 Methods

3.2.1 Validation data

I evaluated the performance of CLEFT-Q CAT algorithms at different fixed-length stopping rules through Monte Carlo simulation with real patient data. These data were obtained as part of routine care, from 536 patients aged 7 to 24 years and undergoing treatment for either CL, CP, CLA, or CLAP at Duke Children's Hospital (USA), Boston Children's hospital (USA), or Erasmums University Medical Center (the Netherlands). CLEFT-Q response sets were collected from these patients at 561 clinic appointments between November 2015 and April 2019.

During these appointments, patients completed the CLEFT-Q scales included in the ICHOM standard set for CLP (these are the face, jaw, teeth, psychological function, social function, school function, speech distress, and speech function scales). Scales were administered in line with the timings (patient ages) recommended in the ICHOM standard set [112]. This means that not every patient completed every scale at every appointment. No responses were collected for the lip, cleft lip scar, nose, or nostrils scale.

I tabulated missing data frequency and determined outliers through Mahalanobis distance (χ^2 $p < 0.05$). In the primary analysis, I excluded incomplete response sets listwise

and included outliers. In subsequent sensitivity-type analyses, I repeated simulations without outliers, and after handling missing data through multiple imputation with a proportional odds model.

3.2.2 Monte Carlo simulation

In this experiment, I programmed the Rasch-based CAT algorithms for each CLEFT-Q scale to follow fixed-length stopping rules. In other words, they terminated after asking a prespecified number of questions, but were free to pick which items to administer via EAP scoring and MEPV item selection. I tested all possible stopping rules. For example, I first applied the CAT algorithm for the face scale (which normally contains 9 items) to the validation data with a stopping rule of 8 items, then again with a stopping rule of 7 items, then again with a stopping rule of 6 items, and so on. I compared the scores generated by each CAT algorithm, at each stopping rule, to scores generated from a full-length linear assessment (where all items were administered sequentially to each participant). Specifically, I compared CAT scores and linear assessment scores through Pearson's correlation coefficient, RMSE, and Bland-Altman analysis, including 95% limits of agreement (the maximum error expected in 95% of cases) [139].

For ease of interpretation, I scaled logit scores so that they ranged from 0 to 100. I used the following formula to do this

$$\hat{\theta}_{scaled} = \frac{100(\hat{\theta} - \theta_{Min})}{(\theta_{Max} - \theta_{Min})} \tag{3.1}$$

where $\hat{\theta}$ represents the latent construct estimate in logits, θ_{Min} represents the latent construct estimate for the lowest possible response pattern, $\{1, 1, 1\ldots.\}$, and θ_{Max}

represents the latent construct estimate for the highest possible response pattern, {4, 4, 4...} or {3, 3, 3...} depending on the number of response categories in the scale.

3.2.3 cleftqCATsim

To perform these simulations, I developed and released a bespoke R package called *cleftqCATsim*. The package contains 15 functions that allow even novice R users to reproduce these experiments with their own validation data from only a few lines of code. The package mostly runs as a convenience wrapper for the mirtCAT package [77], with additional features for logit score scaling, Bland-Altman analysis, and data visualisation. I have made all source code for cleftqCATsim available for open appraisal under the 3-clause Berkeley Source Distribution license. The package is free to download from GitHub with an instructional vignette [140].

3.2.4 Stopping rule workshop

I discussed the findings of the Monte Carlo simulation at a multi-stakeholder consensus workshop, which was attended by 5 patients undergoing treatment for CLP, aged 11-16 years and each accompanied by a parent, 3 adults who had undergone treatment for CLP previously, 2 cleft surgeons, 2 psychologists, 2 speech and language therapists, 2 specialist cleft nurses, a cleft orthodontist, and a reconstructive dentist. Before the workshop, each participant read through the CLEFT-Q questionnaire and considered how many items they felt would be appropriate to pose as part of a routine cleft clinic appointment.

The workshop was held virtually, using the Zoom videoconferencing platform [141], on the 25th March 2021, and lasted for 2 hours. During the workshop, members of

my supervisory team moderated 3 virtual breakout rooms where participants discussed the results of the Monte Carlo simulation within the context of each scale's length, the wording of the items, and participants' experiences of completing or administering the questionnaire. Following the breakout session, all participants voted on the number of items each CAT algorithm should be limited to.

Participants aged 18 years or older were reimbursed by £20 for their involvement at the workshop and participants aged under 18 years were provided with a £10 online retail voucher. I obtained ethical approval for this work from the University of Oxford Medical Sciences Interdivisional Research Ethics Committee (R74005/RE001).

3.3 Results

3.3.1 Demographics, missing data, and outliers

The demographics and clinical characteristics of the validation dataset are presented in Table 3.1, along with the number of outliers per scale.

In the validation dataset, one respondent missed 9 out of the 10 items in the speech distress scale and was excluded. There were 141 missing responses to item 9 of the social function scale. Item 9 was added to the scale during data collection (in other words, responses were *missing at random* [142] and missingness was not expected to affect the generalisability of response patterns). Additionally, there were 9 missing responses to item 7 of the social function scale "I feel like I fit in". These all occurred at the same treatment centre.

Table 3.1: Demographics and clinical characteristics of the validation dataset. CL: cleft lip; CP: cleft palate; CLA: cleft lip and alveolus; CLAP: cleft lip, alveolus, and palate. All values represent sample sizes except age, which is expressed in years. Outliers were determined by Mahalanobis distance (χ^2 p < 0.05).

	Face	**Jaw**	**Teeth**	**Psychological function**	**Social function**	**School function**	**Speech distress**	**Speech function**
Sample size	530	314	529	247	345	247	258	274
Outliers	22	15	13	9	6	8	10	1
Gender								
Male	292	164	290	134	189	134	134	144
Female	238	150	239	113	156	113	124	130
Median age	11	12	11	12	9	12	12	12
Cleft type								
CL	39	22	40	13	27	13	4	4
CP	151	94	151	71	99	70	86	93
CLA	51	29	50	24	29	24	7	7
CLAP	289	169	288	139	190	140	161	170
Treatment centre								
Erasmus	354	214	358	130	226	130	157	174
Boston	95	47	95	47	48	47	47	47
Duke	81	53	76	70	71	70	54	53

3.3.2 Simulation results

Correlation and agreement between CAT scores and full-length scale scores decreased as fewer items were posed. A decrease in 2 items per scale had a minimal impact on accuracy. At this level of item reduction all correlations were 0.97 or above and RMSE ranged from 2-5 out of 100 points. These results were not significantly altered by imputation of missing data or exclusion of outliers. For illustration, the results from the CAT simulation involving the face and jaw scales are presented in tables 3.2 and 3.3. Full results for this simulation are provided in Appendix B, Section B.1 and Section B.2.

The CAT algorithms generally achieved lower $SE(\hat{\theta})$ towards the middle of each scale, where respondents were more frequently distributed, suggesting appropriate targeting.

Table 3.2: Face scale CAT algorithm performance at different fixed-length stopping rules Scores are scaled from 0 to 100. Outliers have been included and missing data removed through listwise exclusion. RMSE: root mean square error; lower LA: lower 95% limit of agreement; upper LA: upper 95% limit of agreement.

Face scale (9 items)

Items administered	Correlation	RMSE	Lower LA	Upper LA
9	1.000	0.00	0.00	0.00
8	0.997	1.67	-3.52	2.92
7	0.989	3.19	-6.80	5.46
6	0.983	4.01	-8.48	7.00
5	0.972	5.07	-10.41	9.35
4	0.964	5.71	-11.56	10.79
3	0.949	6.76	-13.49	13.04
2	0.912	8.80	-17.42	17.09
1	0.812	12.64	-23.70	25.74

Table 3.3: Jaw scale CAT algorithm performance at different fixed-length stopping rules Scores are scaled from 0 to 100. Outliers have been included and missing data removed through listwise exclusion. RMSE: root mean square error; lower LA: lower 95% limit of agreement; upper LA: upper 95% limit of agreement.

Jaw scale (7 items)

Items administered	Correlation	RMSE	Lower LA	Upper LA
7	1.000	0.00	0.00	0.00
6	0.997	2.24	-4.09	4.64
5	0.992	3.68	-6.82	7.57
4	0.985	5.23	-9.70	10.72
3	0.980	6.28	-11.68	12.86
2	0.967	8.24	-15.04	17.07
1	0.915	12.73	-23.20	26.39

This is illustrated in Figure 3.1 for the face scale. The bimodal score distribution observed in the face scale was typical for the other scales, which all demonstrated a modest ceiling effect. Similar plots for the other scales can be found in Appendix B, section B.3.

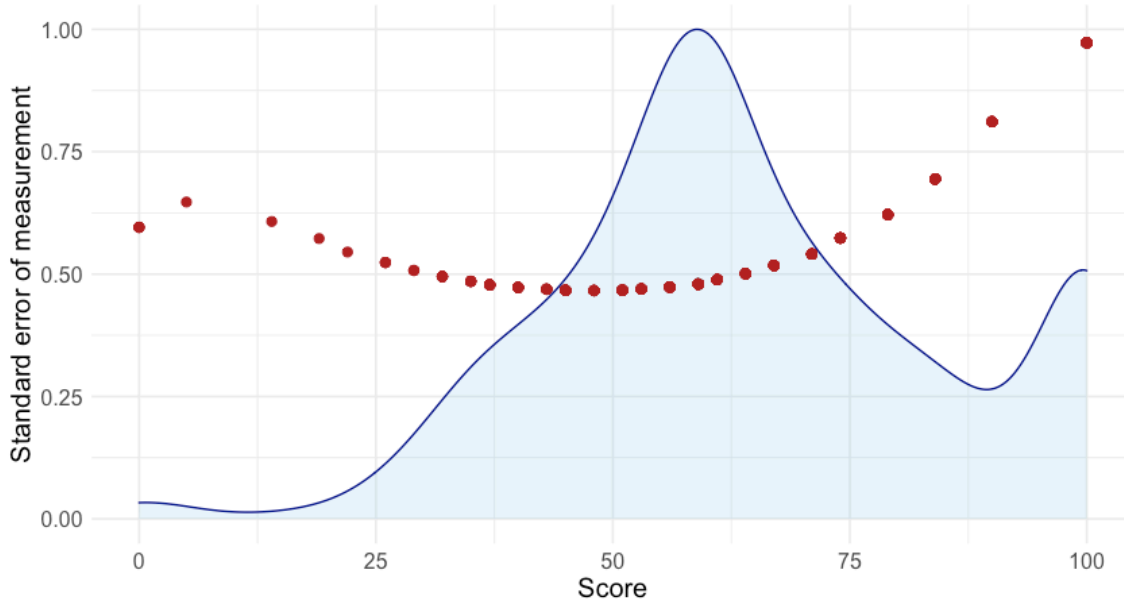


Figure 3.1: Score frequency distribution and standard error of measurement for the face scale validation data. This kernel density plot demonstrates the relative frequency distribution of respondents based on their full-length assessment score, which has been scaled to range from 0 to 100. The red points demonstrate the standard error of measurement (SE) associated with full-length assessment scores across the range of latent construct levels. The SE is lowest, and the measurement is most precise, around the middle of the scale, where respondents are more frequently distributed.

At the individual level, the greatest error between CAT score and linear assessment score was observed at extreme ends of the scale, (where $I(\hat{\theta})$ was lowest and $SE(\hat{\theta})$ was highest). This error distribution is illustrated for the face scale CAT assessment with a 7-item limit in Figure 3.2.

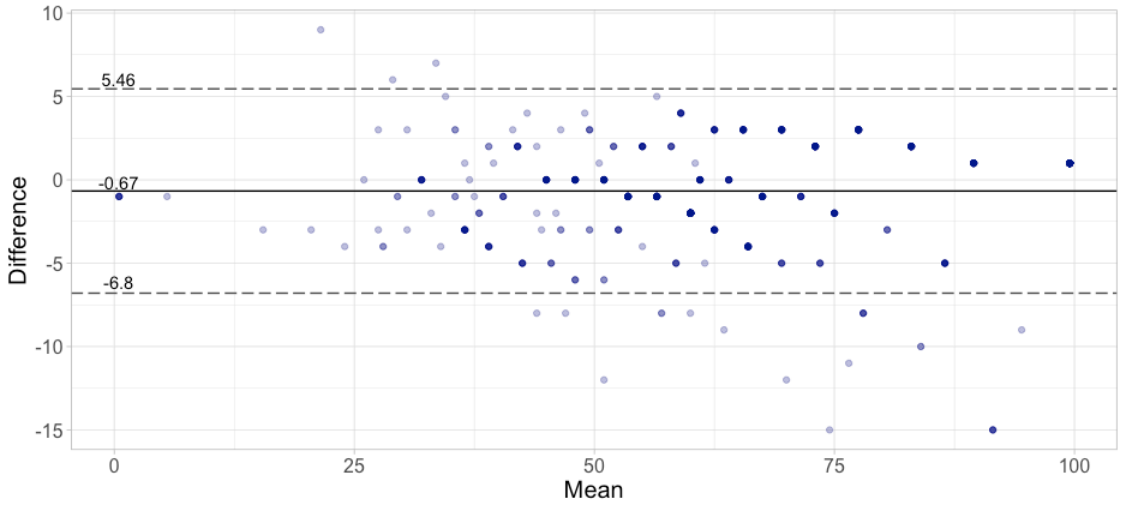


Figure 3.2: Bland-Altman plot for the face scale CAT at 7 items. The x axis represents the mean of each respondent's paired CAT and linear assessment scores. These have been scaled to range from 0 to 100. The y axis shows the difference between CAT and linear assessment scores for each respondent (the CAT score has been subtracted from the linear assessment score). The closer a point is to 0 along the y axis, the better the agreement between that respondent's CAT and linear assessment score. Points that overlie each other appear darker. The solid horizontal line is the mean absolute difference between CAT and linear assessment scores (this can be considered the mean level of error across a population of assessment takers). The dashed horizontal lines represent the 95% limits of agreement (95% of respondents fall between these lines). The scores are more narrowly spread along the y axis between a mean face scale score of 25 to 50. This suggests that CAT accuracy is best when a respondent's face score is approximately 25 to 50.

3.3.3 Workshop results

Workshop participants voted on stopping rules for each scale. These stopping rules are presented in Table 3.4 together with accuracy statistics and median SE. At the chosen stopping rules, the CLEFT-Q CAT presents 1 to 4 fewer items per scale than the linear assessment. These stopping rules were believed to represent the optimal balance between error and assessment burden. Bland-Altman plots demonstrating CAT algorithm performance in the valdiation dataset, at each of these stopping rules, are available in Appendix B, Section B.4. The number of votes per stopping rule is presented in Appendix B, Section B.5, for each scale.

Table 3.4: Chosen stopping rules for the CLEFT-Q CAT and associated algorithmic performance with the validation dataset. RMSE: root mean square error; Lower LA: lower 95% limit of agreement; Upper LA: upper 95% limit of agreement; SE: standard error of measurement.

Scale	Stopping rule	Correlation	RMSE	Lower LA	Upper LA	Median SE
Face (9 items)	7 items	0.989	3.19	-6.80	5.46	0.546
Jaw (7 items)	6 items	0.997	2.24	-4.09	4.64	0.743
Teeth (8 items)	6 items	0.989	3.17	-6.33	6.12	0.554
School function (10 items)	7 items	0.987	3.53	-7.28	6.52	0.552
Psychological function (10 items)	8 items	0.994	2.72	-5.27	5.41	0.587
Speech distress (10 items)	9 items	0.995	2.15	-4.48	3.85	0.716
Speech function (12 items)	8 items	0.981	4.98	-10.75	8.12	0.697
Social function (10 items)	8 items	0.995	2.16	-4.14	4.32	0.521

3.4 Discussion

In this chapter, I have demonstrated the accuracy (correlation and agreement with full-length assessment) achieved by CLEFT-Q CAT algorithms, at different fixed-length stopping rules, in an independent sample of real patient data. Similarly to the results of Chapter 2, error increased as CAT assessment length decreased.

At a population level, where positive and negative errors are averaged out, for example in a clinical trial, CAT error is very small. This is demonstrated by the low RMSE values in Tables 3.2, 3.3, 3.4 and Appendix B sections B.1 and B.2, and supported by the mean absolute error in the Bland-Altman plots (Figure 3.2 and Appendix B, section B.4). For example, after asking only 2 questions, the RMSE between CAT and full-length

assessment was less than 9 out of 100 points for both the face scale (Table 3.2) and jaw scale (Table 3.3).

On the other hand, at the individual level CAT error can be severe, as demonstrated by the 95% limits of agreement in the aforementioned tables and Bland-Altman plots. This is especially relevant for clinicians who must make high stakes decisions at the individual level. While positive and negative errors might mitigate each other in a clinical trial, they would still render the CLEFT-Q CAT unhelpful to the individual patient in clinic. For example, after asking 2 questions, 5% of jaw scale CAT respondents will have a score that is more than 15 points lower or 17 points higher than their full-length assessment score (Table 3.3). This error is similar to the difference between patients who have had, or are awaiting, orthognathic surgery [121].

A strength of this work is that intended end users of the CAT algorithms determined stopping rules themselves. These stopping rules represent a balance between error and assessment burden that was acceptable to the majority of stakeholders attending the consensus workshop. Stakeholders voted to reduce the length of each of the 8 scales included in the ICHOM standard set by 1-4 items. For context, this means that a 12 year old born with a CP, assessed in accordance with the ICHOM standard set [112], would be posed 53 questions instead of 66. At this level of item reduction, RMSE ranged from approximately 2 to 5 points out of 100 and the 95% limits of agreement between CAT and full-length assessment scores ranged from approximately 4 to 11 (Table 3.4). Provisional estimates of the minimal important change (MIC) values for CLEFT-Q scales (the minimum change in a scale score that is considered to represent clinically meaningful change) have been published recently, based on standardised effect sizes, and are larger than the 95% limits of agreement between CAT and full-length assessment scores at the stopping rules chosen in this study [124]. This implies that for the vast majority of respondents, the latent construct level measured by the CAT algorithm will

not be meaningfully different to that measured by the full-length assessment.

In this study, I have demonstrated at least a modest ceiling effect for all 8 scales included in the validation dataset (see Figure 3.1 and Appendix B, Section B.3). In other words, the distribution of scores is bimodal, with a relatively large proportion of respondents scoring between 90 and 100. At this latent construct level, the SE of each scale is relatively high (and the measurement precision is therefore relatively low). This means that very high CLEFT-Q scores (which are relatively common) should be interpreted more cautiously than mid-range scores. One advantage of electronic assessment (including CAT) is that SE can be presented alongside a respondent's score (for example, as a confidence interval). This means that the clinician not only has access to the patient's measurement, but also the level of precision associated with that measurement. It is worth noting that the validation dataset was exclusively obtained from 3 hospitals in the USA and the Netherlands, and the ceiling effect observed in this dataset may not exist in other populations.

While the CLEFT-Q CAT algorithms demonstrated good accuracy in this study, the median SE of measurements at the chosen stopping rules was relatively high, compared to IRT-based CAT algorithms with SE-based stopping rules [74–76]. The median SE ranged from 0.521 to 0.743 at the chosen stopping rules, while PROMIS measures will typically achieve a SE of less than 0.3 or 0.4 [76]. This is similar to the findings of Chapter 2, and raises questions about whether heuristics for desirable precision, that have been popularised by IRT methodologists, are generalisable to CAT algorithms built with Rasch models. This finding warrants further exploration and is addressed in Chapter 5.

The major limitation of this study is that there were no validation data for lip, cleft lip scar, nose, and nostrils scales, which are not part of the ICHOM standard set. Once these become available, the cleftqCATsim software should be used to externally validate the

CAT algorithms for these scales. In the meantime, I have taken a pragmatic approach to setting stopping rules for these scales, based on those chosen for scales with similar lengths at the consensus workshop. For the lip scale, the CAT algorithm will terminate after 7 of the 9 items (similarly to the face scale). For the cleft lip scar scale, the CAT algorithm will terminate after 6 of the 7 items (similarly to the jaw scale). The nose scale CAT algorithm will terminate after 8 of the 12 items (similarly to the speech function scale), and the nostrils scale will remain 6 items long (6 items was the lowest stopping rule picked at the workshop). These recommendations should be reevaluated following external validation.

A limitation of CAT simulations in general is that they assume the order of items administered has no bearing on the item responses. The correlation, RMSE, and limits of agreement reported in this study compare full-length linear assessment scores to CAT scores that would be obtained from respondents, provided that they would answer each question in exactly the same way, regardless of whether the items were administered sequentially or adaptively. The impact of item order on item response has been studied in other fields, such as assessments of personality, attitude, and *cyberloafing* (a form of procrastination), and in some circumstances it can be significant [143]. The impact of CAT assessment length and item ordering on PROM responses should be investigated in future prospective research.

3.5 Conclusion

The studies described in Chapter 2 and Chapter 3 have resulted in CAT algorithms for the CLEFT-Q which are ready to operationalise. The estimator, item selection criteria, and stopping rules have been selected through experimentation and pragmatism, and these algorithms theoretically work to balance error with assessment burden in the eyes of

stakeholders. The next steps in the development of a real-world CAT assessment platform were to build user interfaces for CAT administration and score interpretation, and then to pilot the system in clinical practice. These steps are described in Chapter 4.

4

The CLEFT-Q CAT in clinical practice

Contents

4.1 Introduction 73
4.2 Methods 74
4.2.1 A user interface for CAT administration 74
4.2.2 The CLEFT-Q CAT score checker 76
4.2.3 Qualitative evaluation of the CLEFT-Q CAT in clinical practice 79
4.3 Results 81
4.4 Discussion 90
4.5 Conclusion 92

4.1 Introduction

The research described in the previous chapters has resulted in CAT algorithms for the CLEFT-Q that have been developed *with* stakeholders, through experimentation and pragmatism. But to deliver any real-world impact, these algorithms must be operationalised as a software platform, with a user interface that allows patients, clinicians, and researchers to interact meaningfully with the underlying model parameters and source code.

The most popular system for administering clinical CAT algorithms is the PROMIS platform [79], which is widely available through the EPIC EHR. Platforms such as PROMIS are undoubtedly enhancing the quality of clinical measurement in many settings, but they are not perfect, and as I described in Chapter 1, sections 1.5 and 1.7, the limitations of platforms such as PROMIS represent innovation opportunities for the CLEFT-Q CAT. These opportunities are to:

1. Provide condition-specific measurement
2. Use patient-centred stopping rules
3. Execute CAT algorithms locally
4. Contextualise results through comparison to population reference values, stratified by relevant clinical and demographic characteristics

The previous chapters have addressed opportunities 1 and 2. The work described in this chapter addresses opportunities 3 and 4. Here, I aim to develop and pilot a system for CLEFT-Q CAT administration that can run locally, without transferring patient data to and from an external assessment centre, and to create a platform for score interpretation that compares results to relevant reference populations. Through this work, I also aim to provide open-source software infrastructure that can be used to rapidly translate this

system into other settings (for example, to use with PROMs in other conditions).

To develop a locally executable CAT system, I used Concerto, an open-source CAT administration platform maintained by the University of Cambridge Psychometrics Centre. Concerto delivers CAT functionality through a highly adaptable R back-end, which can be accessed easily using a point-and-click interface, or modified directly for more specialist assessments. Concerto-based assessments can be hosted locally (for example on an NHS hospital server) or through a cloud-hosting service, making it an ideal tool for this project. Concerto has been used previously to develop CAT assessments of personality and verbal reasoning [144], and the versatile R environment has even accommodated open-text assessments of doctors' performance through natural language processing and supervised machine learning [145]. I have previously published a detailed user guide for Concerto which provides a practical introduction to the platform [72].

4.2 Methods

4.2.1 A user interface for CAT administration

I wrote back-end R code that exactly mimicked the Monte Carlo simulation process described in Chapters 2 and 3, but at the individual respondent-level, by deconstructing and modifying functions from the mirt [49] and mirtCAT [77] packages. I used an HTML and CSS front-end for the patient-facing and clinician-facing side of the platform (see Figures 4.1 and 4.2), a MySQL database management system for storing data, and a UK Amazon cloud server for deployment. The assessment collected item responses but no identifiable or personal data and met all data security requirements for use in Oxford University Hospitals (see subsection 4.2.3).

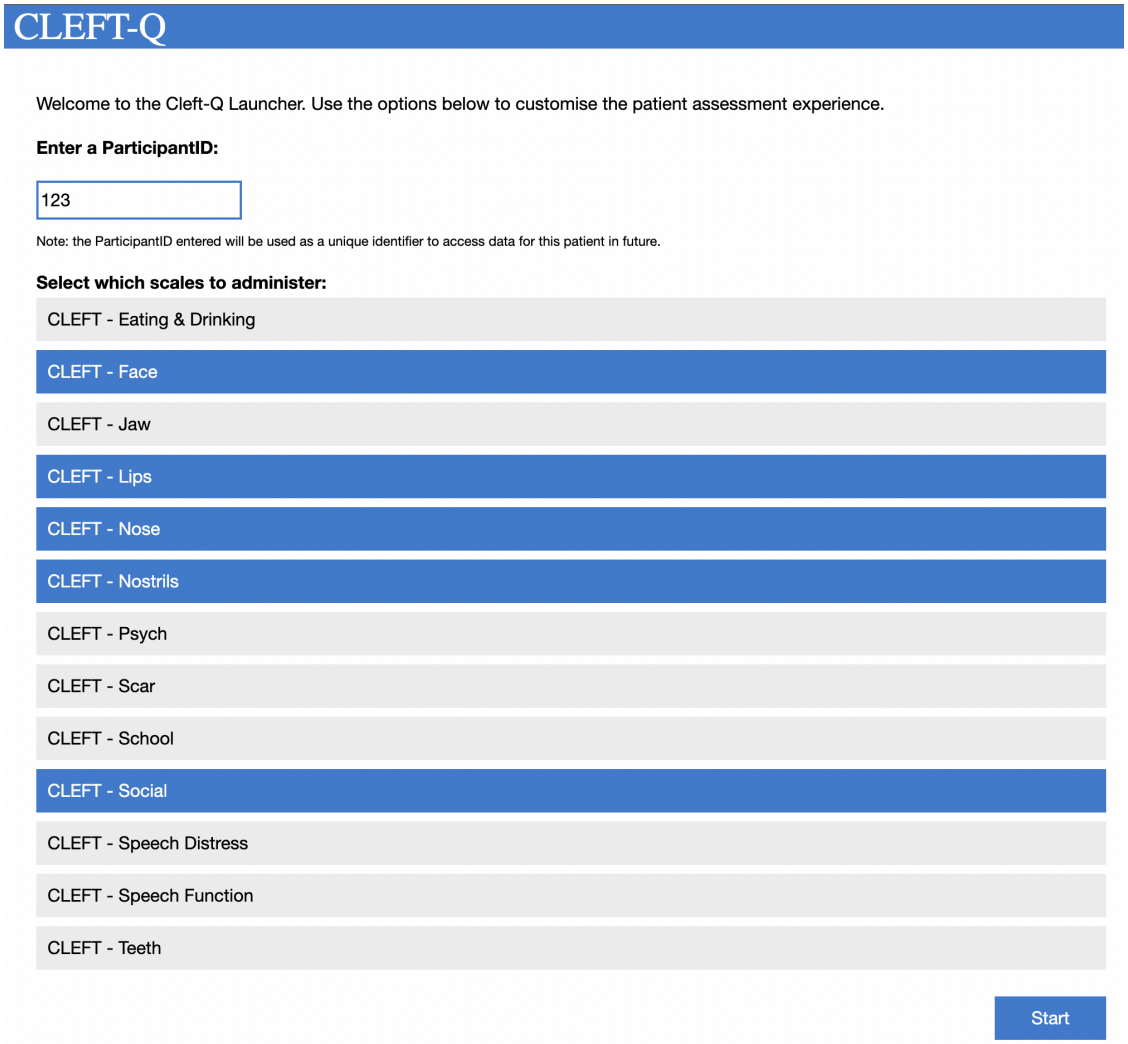


Figure 4.1: The CLEFT-Q CAT administration panel. Prior to the assessment, the clinician or researcher can select which scales they wish to administer.

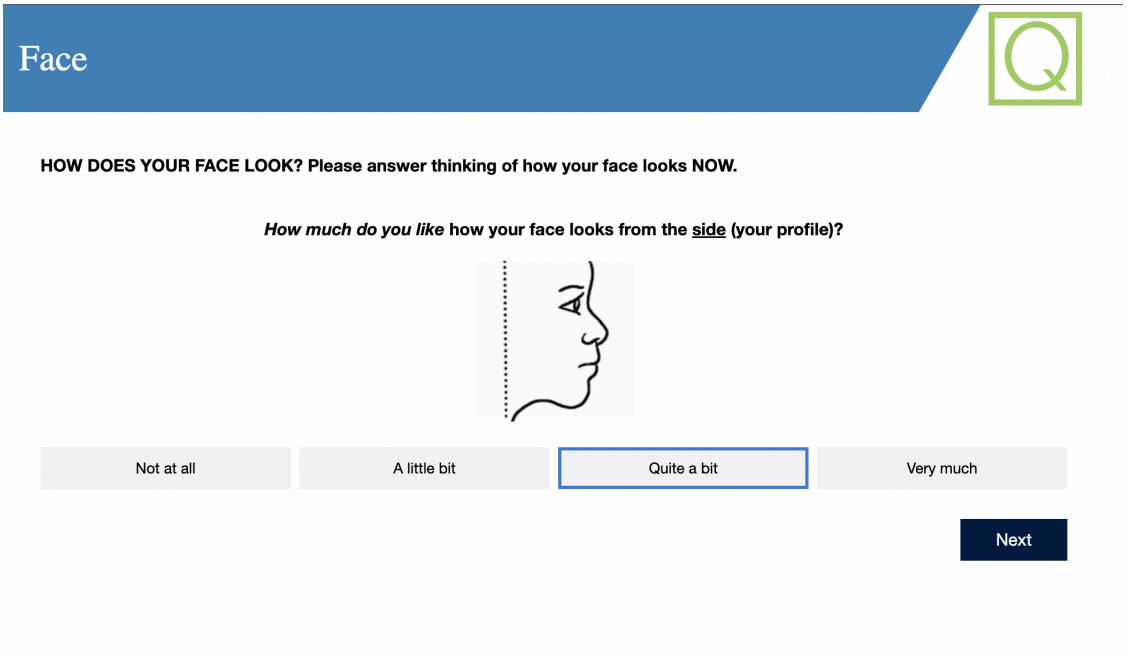


Figure 4.2: The CLEFT-Q CAT assessment. Respondents are posed items from the CLEFT-Q, together with illustrations included in the original questionnaire.

4.2.2 The CLEFT-Q CAT score checker

I built a web application using R Shiny to help CLEFT-Q CAT users interpret a respondent's score. This is called the CLEFT-Q CAT Score Checker [146]. The application compares a score to those that were obtained by CLEFT-Q field test participants [118] that are matched by age, gender, cleft type and cleft laterality (unilateral or bilateral). It then presents this comparison as either a radar chart or a kernel density plot. First, the user enters details about the respondent (age, gender, cleft type, and laterality) through drop down boxes in the user interface. The user interface then reacts with a server function, which filters a dataset of CLEFT-Q field test respondents based on the information provided. The filtered dataset is then used to produce a plot, which is then passed back to the user interface.

In the radar chart (Figure 4.3), a respondent's score on each scale is presented as a magenta point, connected by solid magenta lines. The median CLEFT-Q field test scores achieved by respondents with similar characteristics are displayed in the same plot as red points connected by dashed red lines. In the kernel density plot (Figure 4.4), the distribution of scores obtained by CLEFT-Q field test respondents with similar characteristics, for a given scale, is displayed in magenta, while the individual respondent's score is superimposed as a blue, vertical, dashed line, together with the cumulative percentile of that score within the distribution (the percentage of matched field test scores at or below the respondent's score). This information is also provided as text underneath the plot with the sample size of the filtered dataset. For example, when no filters are applied, "a score of 50 on the CLEFT-Q face scale is higher than 31% of scores reported by people with similar phenotypes, in the CLEFT-Q field test, based on a sample size of 2301".

The CLEFT-Q CAT Score Checker is hosted on a Shiny server, is free to use, and I have made all source code and data available through GitHub under the 3-clause Berkeley Source Distribution license [147].



Figure 4.3: Radar plot tab of the CLEFT-Q CAT Score Checker. The magenta points represent a respondent's score in each of the CLEFT-Q CAT scales. These are input through open text boxes below the plot. The red points represent the median scores in each scale obtained by CLEFT-Q field test respondents with similar demographics and clinical characteristics. These are filtered using the drop-down boxes to the bottom left of the page.

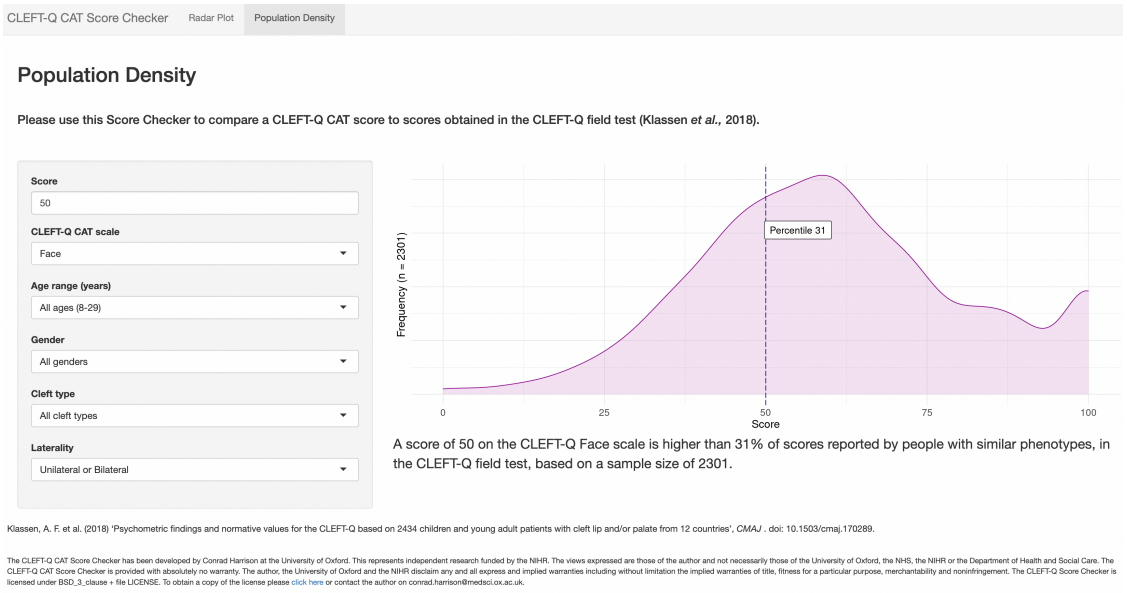


Figure 4.4: Population density tab of the CLEFT-Q CAT Score Checker. The magenta kernel density plot represents the score distribution of face scale respondents in the CLEFT-Q field test. The respondent's score is presented as a vertical dashed line, together with their cumulative percentile. Users may change the distribution by selecting a different scale or entering information about the respondent with the drop-down menus to the left of the plot. Written information is then updated under the plot, including the sample size of the filtered dataset.

4.2.3 Qualitative evaluation of the CLEFT-Q CAT in clinical practice

I piloted the CLEFT-Q CAT system at the Spires Cleft Centre, Oxford, during April and May of 2021. During this time, 24 patients completed the CLEFT-Q CAT as part of routine care. I asked patients to complete the assessment on an iPad while waiting for their clinic appointment. Patients were administered all 12 scales, unless they were born with a cleft palate only, in which case I omitted the cleft lips scar scale. Before their appointment, I reviewed patients' scores with the clinical team, using the Score Checker application. Clinicians were able to discuss and action these results as they wished.

I then recruited a deliberately diverse sample of patients and clinicians who had experience of using the CLEFT-Q CAT, for semi-structured interviews that explored

their experience of the platform. I conducted interviews either in person or remotely via the Zoom videoconferencing platform, recorded the audio content, and transcribed this verbatim. Participants under the age of 18 years were interviewed in the presence of an adult with parental responsibility. All patients were interviewed within 7 days of using the CLEFT-Q CAT. Interview transcripts were not returned to participants. The interview schedules are presented in Appendix C, Section C.1.

I single coded the content of these interview transcripts, using the NVivo platform (version 1.4), under the following categories: experience of the software; experience of the CLEFT-Q CAT's content; barriers to implementing the software in clinical practice; and facilitators to implementing the software in clinical practice. I synthesised emergent themes within and outside of these categories through an inductive approach [148] following grounded theory [149].

The project's protocol was reviewed by the University of Oxford Joint Research Office study classification group, who classified it as service evaluation. Ethics committee review was not required. Nonetheless, all participants provided written, informed consent, including consent to use de-identified quotations in any publication. The project was registered with clinical governance and information governance teams at Oxford University Hospitals (QIP 6703). This included a full data protection impact assessment.

The qualitative aspect of this chapter is reported in keeping with the Consolidated Criteria for Reporting Qualitative Research checklist [150].

4.3 Results

I recruited 6 patients and 4 clinicians for semi-structured interviews. The characteristics of these participants are presented in Tables 4.1 and 4.2. Interviews ranged in duration from 6 minutes and 49 seconds to 20 minutes and 25 seconds. While I did not deliberately recruit patients until thematic saturation had been reached, no new emergent themes were elicited from the final interview.

Table 4.1: Characteristics of patient participants. M: male; F: female; UCLP: unilateral cleft lip and palate; UCL: unilateral cleft lip; BCLP: bilateral cleft lip and palate. Ethnicity was self reported.

Patient participants

Interview number	Gender	Cleft type	Age (years)	Ethnicity
1	M	UCLP	28	White British
2	F	UCL	13	Kurdish
3	M	UCLP	16	Asian (not specified)
4	M	BCLP	8	Indian
5	F	BCLP	24	White British
6	F	BCLP	18	British Pakistani

Table 4.2: Characteristics of clinician participants. M: male; F: female.

Clinician participants

Interview number	Gender	Occupation
7	M	Surgeon
8	F	Specialist nurse
9	F	Speech and language therapist
10	M	Dentist

Emergent themes relating to the CLEFT-Q CAT user interface were its ease of use, and a preference for electronic assessment over pen-and paper. Even when asked directly, no

participant felt as though the CAT caused excessive burden.

I synthesised both positive and negative themes relating to the content of the CLEFT-Q CAT. Positive themes included: its person-centredness; its ability to cause patients to think about previously unconsidered health aspects; and its ability to normalise health concerns. Negative themes included its potential to cause upset to patients who would rather not answer sensitive questions, and its repetitiveness.

Both patient and clinician interviewees identified barriers and facilitators to implementing the CLEFT-Q CAT in clinical practice. Barriers included: difficulties integrating the system across multiple different EHR platforms used in the NHS and further afield; maintaining equity of access, particularly between hub and spoke services; resource requirements (for example, iPads, staffing, and space); opportunity cost; change resistance; and reluctance or difficulty with using technology. The facilitators identified were: the opportunistic use of waiting room time; training clinicians and support staff to administer the CLEFT-Q CAT; and educating patients and clinicians about the benefits of using the system. The option to complete the assessment at home was seen as a facilitator by some, but not by others.

The use of CAT as a clinical communication aid emerged as a theme in both patient interviews and clinician interviews. Within this theme, I synthesised the following sub-themes: improving consultation focus; improving patient-to-clinician information flow; facilitating a multi-disciplinary approach to care; improving patient readiness; and facilitating shared decision making.

The coding tree from this thematic analysis is presented in Figure 4.5 and illustrative quotes are presented for each theme, below.

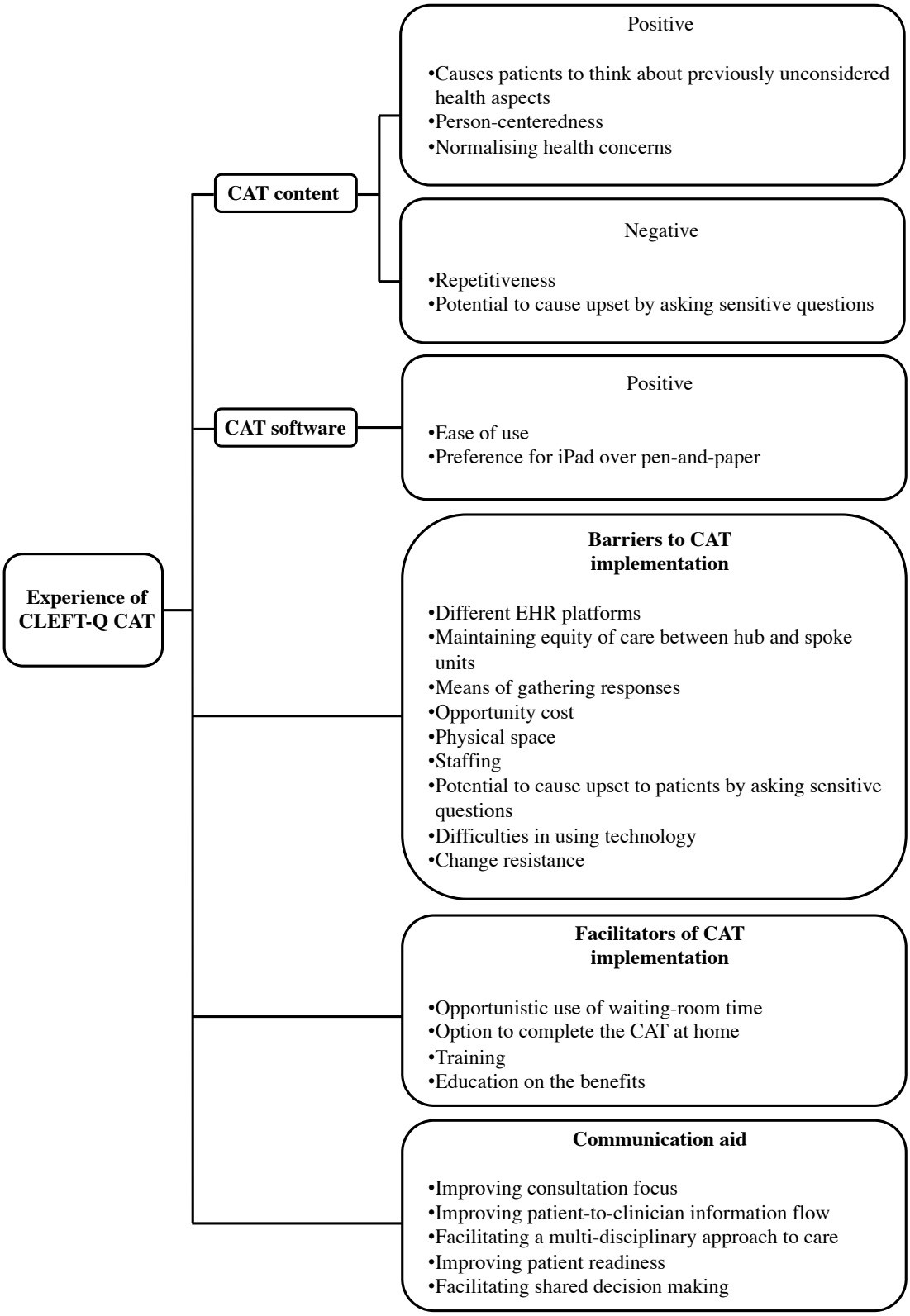


Figure 4.5: Coding tree derived from interview transcripts. CAT: computerised adaptive test. EHR: electronic health records.

Ease of use:

"It was really straightforward" – patient.

"For me it was pretty quick, so anyone could fill in this form" – patient.

"I think the visual appearance [of scores presented as a radar chart] is helpful. Rather than just having scores… you can absorb more information quite quickly" – clinician.

A preference for using electronic tablets over pen-and-paper:

"The iPad is much more convenient [than pen and paper]… you get the results much quicker" – patient.

Person-centredness:

"The questions are I think on topic and are what are relevant to me personally" – patient.

"I found that that questionnaire, compared to the paper one I've done in the past, they were similar, but it was more about how I felt, rather than other people" – patient.

Causing patients to think about previously unconsidered health aspects:

"I think it's really detailed, it makes you consider aspects that you don't really consider thinking about too often, so that can help the doctors as well as you" – patient.

Normalising health concerns:

"It made it feel way better, like I'm more like other people" – patient.

Potential to cause upset to patients by asking sensitive questions:

"It's good, because I guess you get to express yourself and say what you feel, I think. But also, I feel like it makes you more conscious about how you look… I didn't find it

upsetting personally, but I feel that for other people it probably might be" – patient.

Repetitiveness:

"*Some of them could have been a bit repetitive. It was almost like the same question but not"* – patient.

Integration across different electronic health record platforms:

"*They [other clinicians] will all tend to have different electronic record needs, so it's trying to find something generic, or a platform that can be used across systems, which may be a challenge.*" – clinician.

Maintaining equality between hub and spoke services:

"*You need to have that equality of care, because the reason for a lot of outlying clinics in cleft is for more deprived patient groups. So, in a way it's even more important that you can roll it out to those groups than it is for the hub patients"* – clinician.

Resource requirements:

"*Aside from the availability of iPads and then you know the logistics of them missing and signing them in and signing them out, I don't really see any barriers going forwards"* – clinician.

"*You need a space for them [the patients] to use [the CLEFT-Q CAT]"* – clinician.

"*You need a person who can explain to the patient or the family, what to do [to use the CLEFT-Q CAT]"* – clinician'

Opportunity cost:

"*If you're an adult, you're probably thinking about how you've taken a couple of hours*

out of your day and you've got your phone on you so you can do some work" – patient.

Interviewer: *"Anything that would make you want to do it less?"*

Patient: *"Yes"*

Interviewer: *"Like what?"*

Patient: *"Playing games"*

Interviewer: *"So if you could play games instead, you would rather do that? Is that what you're saying?"*

Patient: *Nods and grins cheekily.*

Change resistance:

"When we had similar projects, the barriers tend to be partly personality driven, so some centres will not adopt anything new, or certainly won't adopt anything they didn't develop, almost out of principle. So, there will be some late adopters, where hopefully they'll come on board later" – clinician.

Difficulties in using technology:

"I don't know [what might make the CLEFT-Q CAT difficult to implement], it is straightforward to be fair. But it might put them [other patients] off when they see the iPad and they think 'oh no', or unless you actually do it, because it was really straightforward, but it could put them off if they see the iPad. Maybe that's older people because a lot of young people know what an iPad is. I wouldn't be put off by doing it" – patient.

Opportunistic use of waiting room time:

"If you're a child you're bored and you've got the option of a tablet in front of you, you'd probably do it, personally. It's either that or watching some terrible TV you get in a waiting room... if it's like the 3-4 years I had of brace treatment, I was stuck in the

waiting room for 40 minutes, so 10 minutes doing that is probably a welcome distraction" – patient.

Training:

"It would probably be good to have somebody who understands it [the CLEFT-Q CAT] well to come in and have a short training session [with new users]" – clinician.

Education surrounding benefits:

"A lot of the time you're asked to fill out a questionnaire, and they're [the questionnaire administrators] like 'win a £10 Amazon voucher' – you're not bothered about that. But actually, if they say this will actually really benefit you from a health perspective, and our consultants will understand you as a person more, I think they'll go 'actually, yeah, let's not make this a dreaded experience, let's make this an experience we can benefit from'" – patient.

Option to complete the CAT at home:

"So, personally, I think [that if] we had that [the CLEFT-Q CAT] sent out with their appointment, I don't know, a few weeks ahead, that would be really useful because, then, to have that information going in, and we could have had it, if there was anything really significant that was coming up when we reviewed them, then we would be able to discuss that prior to their appointment, so I think, yeah, kind of getting ahead with time would be really good" – clinician

"I'm of the opinion that it should be done in the clinical setting with the clinician present at the time, or just before the appointment. Ideally patients would aim to arrive at their appointment a few minutes early anyway, and I think it's a good time to reflect" – clinician.

Improving consultation focus:

"We had a patient where they have cleft as part of a complex craniofacial condition, and there had been planning and suggestion of doing really quite major surgery, reconstructing around the nose area, but when the patient did the CAT before they came into the clinic, actually that was one of their lowest priorities, which meant that we could then refocus the consultation to focus on their priorities rather than what had been perceived as what should be discussed" – clinician.

"It's not slowing things down, if anything it's helping to speed up our focus. Because otherwise a lot of the questions will be things that we'd be having to explore during the clinic from scratch and this allows us to home in much more efficiently" – clinician.

"And I suppose in one way you could look at it to say that, because it sort of hones what you're actually focusing in on, it makes your, like, prioritisation of what you're doing, would therefore be much more, kind of, aligned with what your patients want. So you might be, kind of, putting a lot of work in where it's not really warranted [whereas with the CLEFT-Q CAT] you'd be focusing in on the right, the right thing for the right patient" – clinician.

Improving patient-to-clinician information flow:

"I was forearmed and so forewarned so I could broach things differently with her [my patient], have a slightly different dialogue, and then facilitate some supportive therapy with clinical psychologists" – clinician.

"I found the questionnaire really helped me be more honest about what was bothering me" – patient.

"So previously they [the patients] may not have mentioned it [their concern], and they

might have come into the clinic, because sometimes it's a little bit of a scary environment for them because there's so many clinicians in there. But this time it's made sure that we've actually addressed all of the things that they're concerned about" – clinician.

Facilitating a multi-disciplinary approach to care:

"I think what it allows patients to do, is be seen as a whole patient rather than just an element of treatment" – clinician.

"All of the patient's focus was on his teeth and jaws and I perhaps wouldn't have been thinking very much about that in my own sort of uni-disciplinary way, and it meant really that, you know, it's [the CLEFT-Q CAT is] actually guiding the treatment pathway for him" – clinician.

Improving patient readiness:

"It made me more alert as to why I was there" – patient.

"I think they [patients] do enjoy it because I think it's an opportunity for them to reflect on often a fairly intensive care pathway, and I think it gives them time to sit and think about every element of their cleft" – clinician.

"It'll make you think more about yourself than how you would have beforehand" – patient.

"I think it's really good that the patients can come into the clinic just being a little bit more aware and having previously thought about what they might like to discuss with the cleft team" – clinician.

"The patient came in having already sort of had time to reflect and, if you like, sort of, not primed for the conversation, but sort of, like she's ready, she's ready to talk and so I think

we kind of really got a lot, we got in there a lot faster than you probably normally would, because of it [the CLEFT-Q CAT]" – clinician.

Facilitating shared decision making:

"What it [the CLEFT-Q CAT] does do is open doors for patients and clinicians to rethink the direction sometimes they were taking [in their care plan]" – clinician.

"It helps set the plan of what's more important and what we can do first [which heath interventions should be prioritised]" – patient.

4.4 Discussion

The CLEFT-Q CAT is perceived to positively impact patient care, by both clinicians and patients at Spires Cleft Centre, Oxford. Many of the positive themes I drew from my interviews are similar to those that have been reported previously in studies of PROM implementation. For example, enhancing communication between patients and clinicians, improving consultation focus, facilitating multidisciplinary team working, and the early detection of health related quality of life issues [15, 20–24]. I have also identified additional potential benefits to using PROMs in clinical practice, including their potential to make patients think about previously unconsidered elements of their health, and to normalise health concerns - *"It made it feel way better, like I'm more like other people"*.

These can be considered benefits to using PROMs in clinical practice, rather than CAT specifically. From this study, it is difficult to distill the benefits of CAT from the benefits of PROMs in general, but there was a clear preference for electronic administration over pen-and-paper, and it is possible that the CLEFT-Q CAT's ability to omit poorly targeted items led one patient to perceive it as more individualised than the full-length linear assessment - *"I found that that questionnaire, compared to the paper one I've done in the*

past, they were similar, but it was more about how I felt, rather than other people".

One of the patients described the possibility of some items causing distress - *"I didn't find it upsetting personally, but I feel that for other people it probably might be"*. This opinion is shared by some members of the UK cleft speech and language community, who have previously voiced concerns over the negative wording of items in the CLEFT-Q speech scales [128]. I identified repetitiveness as another negative theme, and while this may reflect the similar nature of items in the CLEFT-Q, it is also possible that CAT enhances this by selecting items that are targeted to similar latent construct levels (and therefore potentially similar in content).

The barriers and facilitators to CLEFT-Q CAT implementation identified in this study should be considered in future efforts to roll out the software nationally and internationally. This might include lowering resource requirements (for example by enabling patients to complete the assessment on their own devices), providing training and education, and accommodating different options for integration of the assessment into the clinical workflow (for example, through the opportunistic use of waiting room time, or by administering the assessment remotely, prior to the clinic appointment).

There are limitations to this study that should be noted when interpreting its findings. All patients and clinicians came from a single treatment centre, transcripts were single coded, patients did not have the opportunity to review their transcripts or comment on the thematic analysis, and my own personal biases as a developer of the CLEFT-Q CAT may have affected both the quality of the interviews and the thematic synthesis. But even with these limitations in mind, the positive impact of the CLEFT-Q CAT on patient care seems plausible.

4.5 Conclusion

This chapter concludes the first half of this thesis, in which I have described the development, validation, and implementation of an open-source, locally executable system for CLEFT-Q CAT administration, which uses stopping rules that have been set with end-users, and can contextualise a respondent's score through comparison to relevant reference populations. The system is perceived by stakeholders to improve clinical care, but it has potential limitations which require further exploration. In the second half of this thesis, I focus on enhancing our understanding of CAT for PROMs, and identifying future research directions for CAT, in the context of the CLEFT-Q CAT's potential limitations.

5

The CLEFT-Q CAT's standard error of measurement

Contents

5.1 Introduction 94
5.2 Methods 97
5.2.1 A comparison of Rasch and GRM scores 97
5.2.2 A comparison of CAT assessments built with Rasch models and GRMs 98
5.3 Results 98
5.3.1 A comparison of Rasch and GRM scores 98
5.3.2 A comparison of CAT assessments built with Rasch models and GRMs 100
5.4 Discussion 101
5.5 Conclusion 104

5.1 Introduction

Chapters 2-4 have described the development, external validation, and initial deployment of the CLEFT-Q CAT, which has been perceived by patients and clinicians to have a positive impact on clinical communication and care. This work has largely involved the application of existing knowledge and techniques to a new clinical area with the aim of directly benefiting patients. Chapters 5-7 are less translational and more exploratory. These chapters aim to develop new ideas and future research targets for CAT, in the context of the CLEFT-Q CAT's limitations. The first limitation, which I address in this chapter, is the apparently large SE values achieved by the CLEFT-Q CAT.

As I have described, SE-based stopping rules are popular within the field of CAT for PROMs, and heuristics are emerging for SE thresholds that indicate a desirable level of precision. For example, thresholds of SE < 0.3 or < 0.4 have been proposed [74–76]. In Chapter 3 (Table 3.4) one can see that the median SE achieved by the CLEFT-Q CAT scales is higher than these thresholds. Even the lowest SE values obtained by CLEFT-Q CAT assessments are high, compared to these thresholds (see Appendix B section B.3). At first glance, this might be interpreted as unacceptably poor precision, but that might not be the case.

The rationale for SE-based thresholds, proposed by other authors [75], is illustrated below:

1. Marginal reliability can be defined

$$r_{xx} = 1 - SE(\theta)^2 \tag{5.1}$$

2. Previous heuristics for Cronbach's alpha (an estimate of marginal reliability) have

stated that a value of > 0.7 is desirable for group-level measurement and > 0.9 is desirable for individual-level measurement [151]

3. This roughly equates to $SE(\theta) > 0.55$ and > 0.32, respectively

I would respectfully challenge this rationale. Under an IRT framework, marginal reliability is intended to represent the variance in true score as a proportion of observed score variance, with respect to the latent construct level [152]. In other words, if a respondent was to complete an assessment an infinite number of times, with an unchanging latent construct level, no recall, and perfect experimental conditions, the marginal reliability would be inversely proportional to the spread of scores obtained by the respondent around their true score. In this case, deviations from the true score are purely a result of stochastic error. This score distribution is known as the propensity distribution [153].

The idea of estimating an individual's propensity distribution from cross-sectional data is difficult to reconcile in real world terms, and some have questioned whether cross-sectional data can contain any information about a person's propensity distribution whatsoever [153]. The use of marginal reliability in the choice of precision thresholds is questionable, as are the limits of > 0.7 for group-level and > 0.9 for person-level measurement. One could argue that test-retest reproducibility is far more useful for quantifying within-person measurement error than SE or marginal reliability, despite potential biases that might arise from experimental design. Unfortunately, these data do not exist yet for the CLEFT-Q or CLEFT-Q CAT.

Based on the theory discussed in Chapter 1, the choice of psychometric model is likely to have a large impact on $SE(\theta)$. Recall Figure 1.5, which demonstrates the effect of increasing the GRM's discrimination parameter on the ICC for a dichotomous item. A higher discrimination parameter (steeper ICC gradient) means that small differences in θ

have a greater impact on the probability of endorsing an item response. With this in mind, it is quite easy to imagine that as the discrimination parameter increases, the likelihood function of θ, given a set of item responses, narrows. This would cause a drop in $SE(\theta)$ (see Chapter 1, subsection 1.4.3).

Put simply, IRT models with more freely estimated parameters can be *moulded* to more closely represent calibration data, and measurement can be estimated more precisely within tighter-fitting models. But the chance of model overfit increases as we add additional parameters and this point leads us back to the debate between RMT and IRT (see Chapter 1, subsection 1.4.2). It is likely that the reason the CLEFT-Q CAT achieves high SE values is because it is based on partial credit Rasch models, whereas GRMs have been used by authors who quote the SE thresholds above.

I have developed the CLEFT-Q CAT using RMT for consistency with the full-length CLEFT-Q, but the CLEFT-Q CAT could be recalibrated using GRMs. This is likely to drive down the SE. But it is unclear whether doing this would have any impact on the assessment's test-retest reproducibility, or its *criterion validity* (the ability of the CLEFT-Q CAT to accurately reproduce full-length assessment scores, which are calculated through Rasch models). These are arguably far more important than marginal reliability. At present, data are unavailable to evaluate test-retest reproducibility, but criterion validity can be evaluated through Monte Carlo simulation, using the datasets described in Chapter 3. It would be timely and pragmatic to demonstrate that SE is unrelated to the CLEFT-Q CAT's ability to reproduce CLEFT-Q scores, as clinicians are now faced with a choice between implementing the full-length questionnaire or the CAT version and may be deterred by the CLEFT-Q CAT's seemingly poor reliability statistics.

This chapter aims to test whether the high SE values achieved by the CLEFT-Q CAT are related to the choice of psychometric model, explore whether this affects the criterion

validity of the platform, and understand whether emerging SE heuristics are applicable in an RMT framework. Specifically, I hypothesise that:

1. CLEFT-Q scale scores estimated via Rasch models and GRMs will be similar

2. At fixed-length stopping rules, the criterion validity (ability of the CLEFT-Q CAT to accurately reproduce full-length, Rasch model, CLEFT-Q scale scores) will be similar between CAT algorithms that use Rasch models and GRMs

3. Despite similar criterion validity, the GRMs will achieve lower SE values than Rasch models

5.2 Methods

5.2.1 A comparison of Rasch and GRM scores

For both experiments in this chapter, I used the calibration dataset described in Chapter 2 (subsection 2.2.2) to fit models, and the validation dataset described in Chapter 3 (subsection 3.2.1) to evaluate them. For each of the 8 scales included in the validation dataset (face, jaw, teeth, psychological function, social function, school function, speech distress, and speech function), I used item responses in the calibration dataset to fit GRMs via the EM algorithm available in the *mirt* package [49]. Using these GRMs, I then calculated full-length CLEFT-Q scale scores for each respondent in the validation dataset through an EAP approach with a standard normal prior. These were compared to scores derived from the Rasch models described in Chapters 2 and 3, with EAP scoring. For this comparison, I used Pearson's correlation coefficient, RMSE, and 95% limits of agreement. I did not scale logit scores into a 0-100 format as θ_{Min} and θ_{Max} may have differed between these models (see equation 3.1). This experiment addressed hypothesis 1: that CLEFT-Q CAT scale scores estimated via Rasch models and GRMs would be similar.

5.2.2 A comparison of CAT assessments built with Rasch models and GRMs

Using the GRMs fitted in the first experiment, I developed new CAT algorithms for the 8 CLEFT-Q CAT scales included in the validation dataset. For consistency with the Rasch-based CAT algorithms, these used EAP scoring with a standard normal prior and MEPV item selection. In a Monte Carlo simulation, I applied these algorithms to item responses in the validation dataset at all possible fixed-length stopping rules (as described for Rasch-based CAT algorithms in Chapter 3). At each stopping rule, I compared the median SE achieved by Rasch- and GRM-based CAT algorithms.

I assessed the criterion validity of CAT algorithms through correlation and agreement with full-length CLEFT-Q scale Rasch scores. I chose this approach (rather than comparing Rasch-CAT scores with full-length Rasch scores and GRM-CAT scores with full-length GRM scores) because in this experiment I wanted to test the appropriateness of CAT algorithms as replacements for the published CLEFT-Q scales, which have been scored according to Rasch models. This addresses hypotheses 2 and 3: that GRMs would drive down CLEFT-Q CAT SE values without improving the system's criterion validity. The specific statistics I used to assess criterion validity were Pearson's correlation coefficient, RMSE, and 95% limits of agreement.

5.3 Results

5.3.1 A comparison of Rasch and GRM scores

There was a close linear relationship between Rasch and GRM scores (with a Pearson's correlation coefficient over 0.98 for each scale, see Table 5.1, Figure 5.1 and Appendix D, section D.1). However, there were differences between the absolute values of logit

scores derived by each method (demonstrated by the RMSE statistic and 95% limits of agreement). These differences were more pronounced towards the extreme ends of the scales, where the Rasch models estimated a broader range around the sample mean (0 logits, see Figure 5.1). This suggests that Rasch and GRM scores could be reliably mapped onto each other through a simple linear model and scores derived from each method are similar around the population mean. But, particularly at extreme ends of the scale, they should not be considered equivalent.

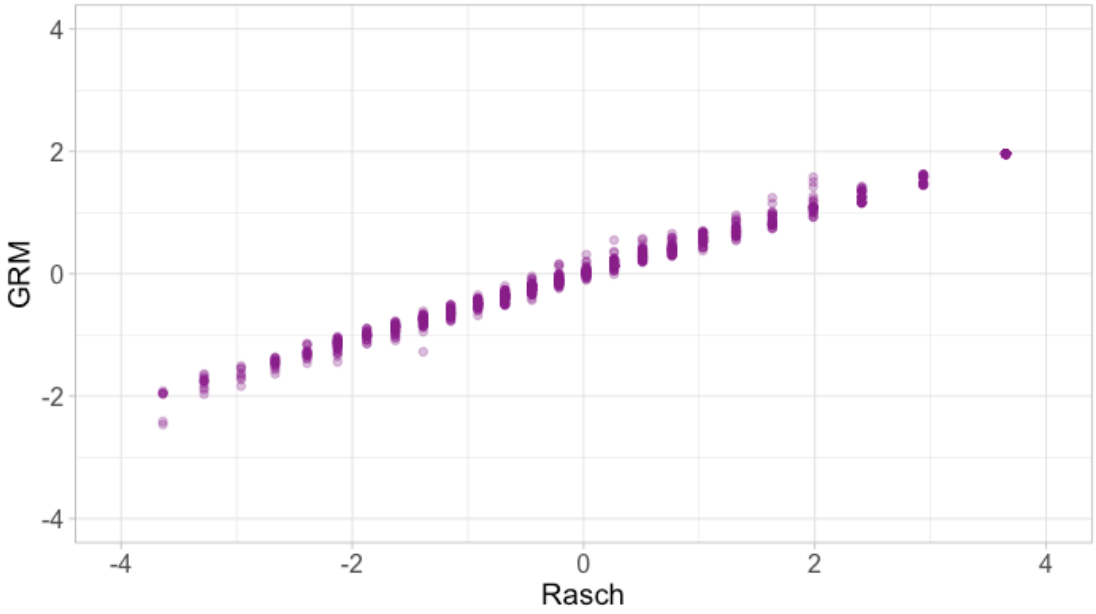


Figure 5.1: The relationship between Rasch and graded response model (GRM) scores for the full-length face scale. For each respondent in the validation dataset, Rasch scores are presented as logits on the x axis and GRM scores are presented as logits on the y axis. There is a strong linear relationship between Rasch- and GRM-based scores for the full length face scale, but relative to GRM scores, Rasch scores are more widely spread around the sample mean (0 logits). This is most apparent at extreme ends of the scale.

Table 5.1: Comparison of partial credit Rasch models and graded response models for scoring full-length CLEFT-Q scales. GRM: graded response model; SE: standard error of measurement; RMSE: root mean square error; MAE: mean absolute error; Lower LA: lower 95% limit of agreement; Upper LA: upper 95% limit of agreement.

Scale	Median SE (Rasch)	Median SE (GRM)	Correlation	RMSE	Lower LA	Upper LA
Face	0.524	0.270	0.996	0.910	-1.34	2.00
Jaw	0.691	0.186	0.992	1.601	-2.76	3.42
Teeth	0.551	0.223	0.995	1.215	-1.11	2.62
School function	0.512	0.263	0.993	0.962	-2.09	1.52
Psychological function	0.541	0.240	0.994	1.200	-2.47	2.22
Speech distress	0.411	0.244	0.992	1.480	0.38	2.39
Speech function	0.417	0.194	0.981	2.149	0.80	3.31
Social function	0.493	0.273	0.997	0.784	-1.65	1.39

5.3.2 A comparison of CAT assessments built with Rasch models and GRMs

The full results of the second experiment are presented in Appendix D, section D.2, and illustrated for the face scale in Table 5.2. For every scale, at any given fixed-length stopping rule, Rasch-based CAT algorithms reproduced full-length Rasch scores more accurately than their GRM-based counterparts. Despite this, median SE was substantially lower for GRM-based CAT algorithms in every comparison. This strongly supports hypotheses 2 and 3. At the fixed-length stopping rules chosen for the real world CLEFT-Q CAT scales (see Chapter 3, Table 3.4), GRM-based CAT algorithms for each scale achieved a SE < 0.3.

Table 5.2: Comparison of CAT algorithms that use partial credit Rasch models and graded response models for scoring the CLEFT-Q face scale. GRM: graded response model; SE: standard error of measurement; RMSE: root mean square error; Lower LA: lower 95% limit of agreement; Upper LA: upper 95% limit of agreement. Comparisons are made to full-length scale scores derived from partial credit Rasch models.

Face scale (9 items)

Number of items	Model	Median SE	RMSE	Correlation	Upper LA	Lower LA
9	Rasch	0.501	0.000	1.000	0	0
	GRM	0.270	0.910	0.996	2.00	-1.34
8	Rasch	0.516	0.146	0.997	0.24	-0.28
	GRM	0.277	0.917	0.992	2.01	-1.35
7	Rasch	0.546	0.244	0.989	0.46	-0.57
	GRM	0.289	0.931	0.986	2.04	-1.38
6	Rasch	0.586	0.335	0.983	0.60	-0.70
	GRM	0.308	0.951	0.978	2.09	-1.39
5	Rasch	0.634	0.417	0.972	0.78	-0.87
	GRM	0.331	0.980	0.963	2.15	-1.45
4	Rasch	0.695	0.506	0.963	0.91	-0.98
	GRM	0.359	1.008	0.950	2.20	-1.51
3	Rasch	0.778	0.614	0.950	1.08	-1.13
	GRM	0.400	1.066	0.921	2.33	-1.62
2	Rasch	0.922	0.790	0.912	1.43	-1.46
	GRM	0.473	1.127	0.898	2.45	-1.76
1	Rasch	1.164	1.098	0.812	2.14	-2.00
	GRM	0.564	1.262	0.835	2.76	-1.91

5.4 Discussion

The purpose of the CLEFT-Q CAT is to reproduce full-length CLEFT-Q scale scores from fewer items, and as I demonstrated in Chapter 3, it does this with a high level of accuracy, but also a high SE. When we consider the SE threshold heuristics that are emerging in the IRT literature, we might infer that the CLEFT-Q CAT produces measurement with

an unacceptably poor precision. The research described in this chapter challenges that inference.

I have demonstrated that one could drive down SE values by replacing the CLEFT-Q CAT's partial credit Rasch models with GRMs fitted to the calibration data. Doing this would bring the median SE for each of the 8 CLEFT-Q CAT scales tested here under the SE threshold of 0.3 advocated by other authors [74–76]. But doing this would make the CLEFT-Q CAT less fit for purpose (the agreement between CLEFT-Q CAT scores and full-length assessment scores would drop).

Instead of blindly following SE heuristics, we should recognise that SE is partly an artefact of model complexity. In IRT, we seek closely fitting models that might provide detailed representations of the relationship between item responses and latent construct levels. But in RMT, we strive for model parsimony. Proponents of IRT might argue that this increase in precision is a benefit to their approach, but Rasch advocates could also argue that these precision gains are merely an artificial manifestation of model overfit. I cannot comment on which philosophy provides *truer* measurement, but I can suggest that SE benchmarks should not be translated between these frameworks.

The first experiment showed a close correlation, but systematic differences, between Rasch and GRM scores. These differences become more pronounced as we measure further from the centre of the scale. Most participants are located towards the centre of each scale, where differences are small. The choice of CTT, IRT, or RMT for producing point estimates of latent construct levels has been investigated previously by the International Society for Quality of Life Research psychometrics special interest group, who have suggested that differences are minor in real world terms, and an undogmatic approach to measurement is favourable [154–157]. This chapter supports that conclusion by demonstrating reassuringly small differences between RMT and IRT measurement

towards the centre of each scale.

But this reassurance may be short-lived. While the EAP point estimates of latent constructs may be similar between IRT and RMT, the distribution of plausible scores that may have resulted in the observed response pattern, which can approximated by $N(\hat{\theta}, SE(\hat{\theta}))$, may be quite different. At present, in surgical research, benchmarking exercises, and clinical care, the measurement error of PROMs is typically either ignored or assumed to cancel out over a population (see Chapter 1, section 1.3). But in other fields with more developed psychometric practices there is a movement towards quantifying and accounting for measurement error. This can be done through techniques such as plausible value imputation, which involves randomly drawing latent construct estimates from their plausible distribution. When we account for measurement error in this way, the results and conclusions of observational studies and randomised controlled trials can change [158–160]. We are some time away from applying these techniques in surgical research, but when we do the question of measurement precision may resurface. I am sure trialists will enjoy the idea of 'more precise' IRT measurements, which generate smaller confidence intervals, and potentially reduce sample size requirements. But is this precision a true reflection of reality, or the artificial byproduct of an over-specified model? It may not be possible to answer this question, but future work could aim to gather real-world information on the test-retest reproducibility of Rasch and IRT measurements, and compare this to propensity distributions suggested by SE. Other work might focus on the identification and prevention of over- (or under-) specification. This might take the form of guidelines for model selection, which could potentially be based on parsimony-adjusted fit statistics or generalisability demonstrated through external validation.

5.5 Conclusion

The CLEFT-Q CAT achieves higher SE values than other CAT algorithms that have been built with IRT models, such as those available in the PROMIS platform. This is not necessarily a weakness of the CLEFT-Q CAT, but rather it is the result of a philosophical difference between IRT and RMT. We should not seek to 'improve precision' through the application of GRMs, as this would be detrimental to the platform's criterion validity. Instead, SE values should be interpreted cautiously, with consideration granted to the precision of the model's specification, and its generalisability to the data at hand.

6

Multidimensional computerised adaptive testing

Contents

6.1 Introduction 106
6.2 Methods 108
6.2.1 Conceptual model 108
6.2.2 Sample size 109
6.2.3 Missing data in the structural equation modelling sample . . . 110
6.2.4 Structural equation modelling 110
6.2.5 Multidimensional item response theory modelling and computerised adaptive testing 111
6.3 Results 113
6.3.1 Structural equation model fit 113
6.3.2 Multidimensional computerised adaptive testing simulation . . 113
6.4 Discussion 116
6.5 Conclusion 119

6.1 Introduction

The previous chapters have treated the CLEFT-Q as 12 independent scales. This is consistent with the approach taken by the developers of the PROM: under the unidimensional Rasch framework, the responses to items in a scale monotonically reflect the level of the latent construct measured by that scale. In this Chapter, I explore a different approach to measurement which might lead to more efficient CAT assessments. Under MIRT (multidimensional IRT, see Chapter 1, section 1.4.6), it is possible for a single item to reflect more than one latent construct, and experiential relationships between latent constructs can be accounted for within the psychometric model [64].

For example, consider a patient who feels very unhappy about the appearance of their nose, nostrils, lips, cleft lip scar, jaws, and teeth, and this is reflected in their CLEFT-Q scale scores for each of these constructs. Given how the patient feels about their facial components, we might reasonably assume that they do not feel happy about the appearance of their face in general, and this feeling would be reflected in their responses to the CLEFT-Q face scale items. If we wanted to assess their perceived facial appearance with the CLEFT-Q face scale CAT assessment, it might be reasonable to start the assessment with an item targeted towards the more negative end of the scale.

When this person responds to a nose scale item under the unidimensional Rasch framework, we are provided with information only about their perceived nasal appearance. But under a MIRT framework, it is possible that this response provides us with information about their perceived nasal appearance and also their perceived general facial appearance. This may be a more realistic model of how these mental constructs affect

---

This chapter is partly based on Harrison, C. *et al*. Deeper understanding of appearance in orofacial clefts: a structural equation model of the CLEFT-Q appearance scales. *Plastic and Reconstructive Surgery Global Open*, 9(9), e3806 (2021).

item responses, and by measuring two or more constructs at once, we can potentially realise efficiency gains in CAT.

Multidimensional IRT is more novel and more complex than unidimensional IRT or RMT. In essence, parameter values for person ability, θ, and item discrimination, a, are replaced by vectors with as many elements as measured constructs. This way, the probabilistic relationship between an item response and multiple latent construct levels can be described simultaneously [161]. While MIRT has been very well explored in the context of theoretical psychometrics [162, 163] and educational psychology [164, 165], and has been suggested for use with PROMs [166, 167], it is not commonly applied in mainstream and real world clinical measurement.

Within surgery, MIRT is particularly novel. On one occasion, I have proposed the application of MIRT to surgical PROMs with the aim of creating efficient multidimensional CAT (MCAT) assessments. Specifically, I have implied that an MCAT algorithm for the 2-factor, 30-item, Disabilities of the Arm, Shoulder, and Hand questionnaire provides more precise and more structurally valid measurement in Dupuytren's disease than its 11-item short-form, even while using fewer items [168]. But there seem to be few other authors exploring these techniques in the context of condition-specific latent variable measurement in surgery.

The first part of this chapter is hypothesis-driven. I propose a conceptually-based structural equation model (SEM) that might represent the relationship between item responses and latent constructs in the CLEFT-Q appearance scales more realistically than the current, unidimensional models. I then test the fit of this SEM to CLEFT-Q field test data to understand whether the data support the hypothesised model.

The second part of this chapter is exploratory. Based on my proposed SEM, I fit a

7-dimension MIRT model to CLEFT-Q appearance scale data, and apply this in an MCAT simulation to understand how it might work in practice.

6.2 Methods

6.2.1 Conceptual model

Figure 6.1 demonstrates the current conceptual model used for the appearance scales in both the full-length CLEFT-Q and the CLEFT-Q CAT. Under this model, each factor only affects the item responses of its respective scale. In SEM diagrams, factors are illustrated as circles and indicators (observable behaviours, such as item responses) are illustrated as squares. To simplify the diagrams in this chapter, I have grouped common items from each scale together as a single square and omitted error terms.

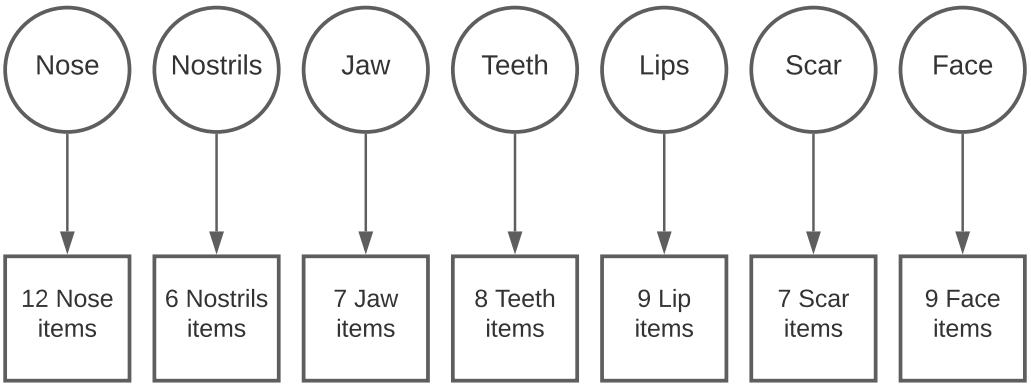


Figure 6.1: The factorial structure of the CLEFT-Q appearance scales as proposed by the instrument's developers. Reproduced from Harrison, C. *et al.* Deeper understanding of appearance in orofacial clefts: a structural equation model of the CLEFT-Q appearance scales. *Plastic and Reconstructive Surgery Global Open*, 9(9), e3806 (2021). Available under the creative commons license (CC BY 4.0): https://creativecommons.org/licenses/by/4.0/

My conceptual model is illustrated in Figure 6.2. This model contains six first-order

factors (the perceived appearance of the nose, nostrils, jaw, teeth, lips and scar). Each factor is measured by the respective scale items. The perceived appearance of the face in general is a second-order factor, measured by the first-order factors, or directly by the face scale items.

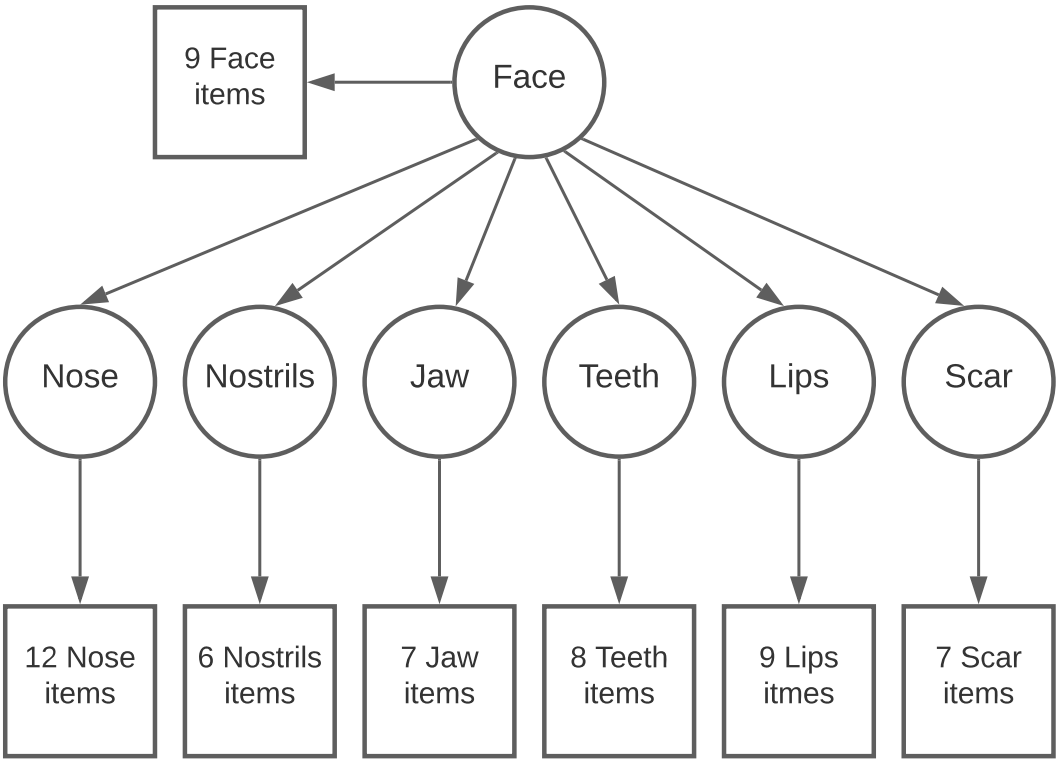


Figure 6.2: The hypothesised factorial structure of the CLEFT-Q appearance scales.
Reproduced from Harrison, C. *et al*. Deeper understanding of appearance in orofacial clefts: a structural equation model of the CLEFT-Q appearance scales. *Plastic and Reconstructive Surgery Global Open*, 9(9), e3806 (2021). Available under the creative commons license (CC BY 4.0): https://creativecommons.org/licenses/by/4.0/

6.2.2 Sample size

It is difficult to determine the optimal sample size for SEM analyses. A small sample size may yield unstable parameter estimates, while a large sample size will predispose SEM fit statistics to type I error (when trivial differences in the observed and expected

data structures are found to be statistically significant). A typical SEM study will involve 200-500 respondents, and complex models, non-normality, ordinality, incomplete datasets, and a high indicator to factor ratio are considered to drive up sample size requirements [169].

Sample size heuristics include 10 respondents per indicator, or 10 respondents per estimated parameter [169]. The proposed model (Figure 6.2) contains 58 indicators (items) and 64 freely estimated parameters (representing the relationships between indicators and factors, and the relationships between first-order factors and the second-order factor). In view of the model's high number of indicators and the ordinal response data, I planned for a sample size of 640 respondents from the CLEFT-Q field test (10 respondents per parameter) and accepted the chance of type I error.

6.2.3 Missing data in the structural equation modelling sample

I have described the CLEFT-Q field test's missing data in Chapter 2 and Appendix A section A.2. There were 1092 participants missing fewer than 2 (5%) of the 58 CLEFT-Q appearance items. From these, I randomly selected 640 respondents and replaced 146 missing item responses with imputation via a proportional odds model, as has been described for factorial analyses [170].

6.2.4 Structural equation modelling

Mardia's test demonstrated a violation of the multivariate normality assumption, and therefore I used mean and variance adjusted weighted least squares (WLSMV) to fit the SEM, as this is considered robust for large samples of non-multivariate normal, ordinal data [171, 172]. I specified and fit the model using the lavaan R package (version

0.6-5.7).

I applied five different tests of model fit to compare the fit of my hypothesised model to a null model of factorial independence. These were X^2, the comparative fit index (CFI), the Tucker-Lewis index (TLI), the root mean squared error of approximation (RMSEA), and the standardised root mean square residual (SRMR). I used the following thresholds to indicate good model fit: X^2 p > 0.050, TLI ≥ 0.950, CFI ≥ 0.950, RMSEA < 0.060, and SRMR ≤ 0.080 [66].

6.2.5 Multidimensional item response theory modelling and computerised adaptive testing

The specification and calibration of hierarchical MIRT models in R is not straightforward. According to Professor Phil Chalmers (developer of the mirt and mirtCAT packages), "second order models can be understood as a constrained version of the bifactor model" and can be fitted by imposing specific proportionality constraints to these models [173]. Under a regular bifactor model, each item directly loads onto a general factor (appearance of the face) and onto a specific factor (for example, appearance of the nose, see Figure 6.3 for an illustration) [174]. Typically, bifactor models assume an orthognal relationship between specific and general factors, which is why I had opted for the second-order SEM as my conceptual model instead (in real life, I believe that the appearance of the face and appearance of the facial components are related).

Following Professor Chalmers' example [173], I used all complete CLEFT-Q field test response sets to the 58 appearance items (n = 895) to fit a modified bifactor model where the general and specific factor loadings of each item were constrained to be equal (I did

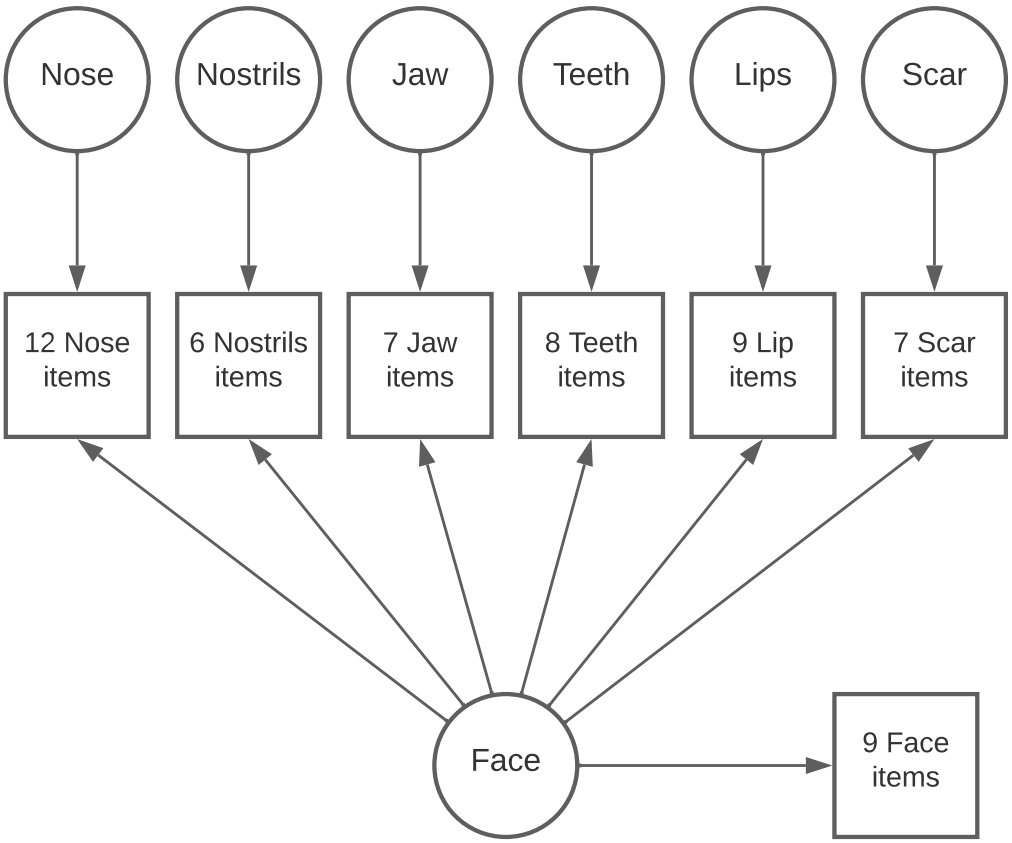


Figure 6.3: Representation of a bifactor model of CLEFT-Q appearance scales. All items load onto the general factor (appearance of the face), and the nose, nostrils, jaws, teeth, lip, and scar items load onto a respective specific factor.

not include a specific factor for the face scale items, in keeping with the hypothesised model). I then deployed the fitted model in a Monte Carlo MCAT simulation. Different CAT settings are recommended for low- and high-dimensional assessments [175]. I used MAP score estimation (which can provide more precise parameter estimates than EAP in a high dimensional setting [175]), and chose the D-rule, weighted by the posterior of the latent construct scores as my item selection criterion [77]. This aims to select the item that maximises information across all dimensions. I used a stopping rule of 46 items for the simulation. This is the number of items that would be used by the 7 unidimensional CLEFT-Q CAT appearance scales combined (see Chapter 3).

To simulate realistic response pattern data, I bootstrap resampled 1000 complete response

sets from the CLEFT-Q field test. I applied the MCAT to these response sets, recording the total number of items administered from each scale, and $\hat{\theta}$ and $SE(\hat{\theta})$ in each dimension, for each respondent, after each item.

6.3 Results

6.3.1 Structural equation model fit

The hypothesised model fit statistics are presented in Table 6.1. The model demonstrated excellent fit across all criteria, expect for the X^2 statistic, which is particularly prone to type I error when large sample sizes are used [176].

Table 6.1: Hypothesised model fit statistics compared to the thresholds implying good model fit. CFI: comparative fit index; TLI: Tucker-Lewis index; RMSEA: root mean squared error of approximation; SRMR: standardised root mean square residual.

	X^2 (p)	CFI	TLI	RMSEA	SRMR
Threshold	> 0.050	≥ 0.950	≥ 0.950	< 0.060	≤ 0.080
Model fit	< 0.001	0.998	0.998	0.043	0.039

The regression parameters for the structural component of the model are presented in Figure 6.4.

6.3.2 Multidimensional computerised adaptive testing simulation

Table 6.2 illustrates the MCAT process for the first 5 items administered to the first simulated respondent. The assessment starts with an item from the nostrils scale.

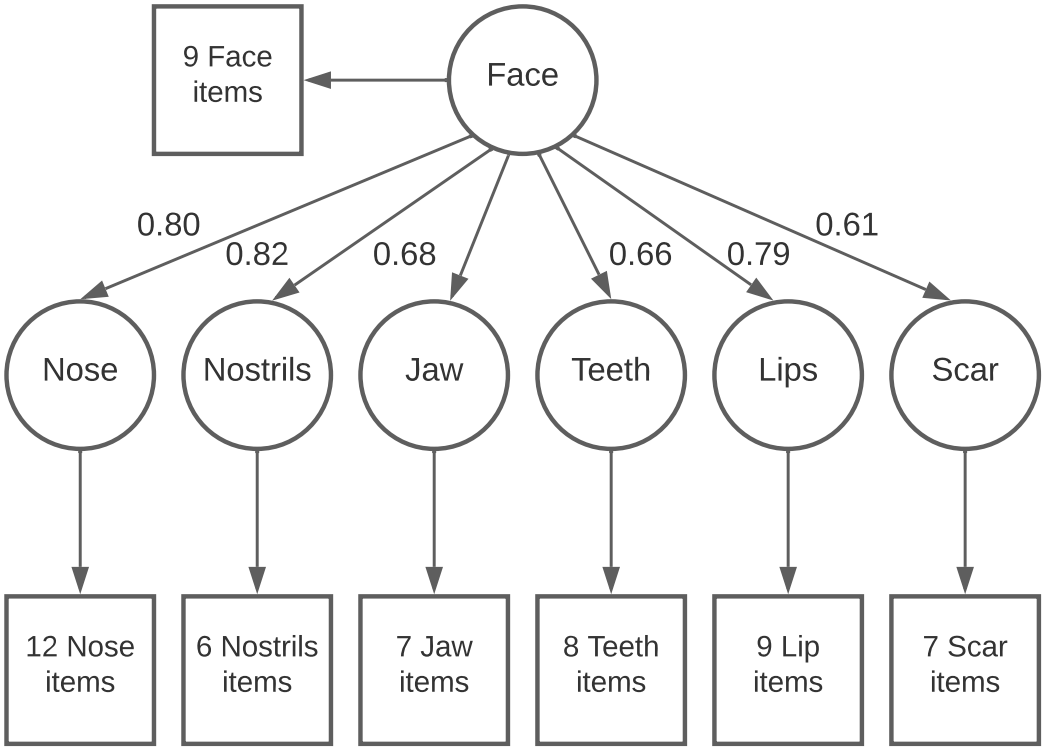


Figure 6.4: Regression parameters in the fitted model. Reproduced from Harrison, C. *et al.* Deeper understanding of appearance in orofacial clefts: a structural equation model of the CLEFT-Q appearance scales. *Plastic and Reconstructive Surgery Global Open*, 9(9), e3806 (2021). Available under the creative commons license (CC BY 4.0): https://creativecommons.org/licenses/by/4.0/. Item parameters differ in this diagram, to reflect the fitted model developed from this experiment.

Following this, we obtain measurement in both the nostrils dimension and face dimension. Based on these measurements, an item from the jaw scale was selected next. This provides information about the respondent's jaw score, but also adjusts the measurement in the nostrils and face scales with increasing precision (decreasing SE). The next two items come from the teeth and lips scales and add to the measurement picture in a similar way. After 6 items, we have a face scale measurement of 0.112 with an SE of 0.455 (brought down from an SE of 1.000), although no face scale items have been administered whatsoever. The first face scale item posed to this respondent was the 21st item in the assessment. Before this point, the SE for the face scale measurement had dropped to 0.355. As discussed in the last chapter, some might interpret this as a highly precise measurement, obtained purely from items that belong to other scales.

Table 6.2: An illustration of the multidimensional computerised adaptive testing simulation. This table demonstrates how the measurements in each dimension changed for the first simulated respondent as they respond to items in turn. For the first 5 items, I have highlighted the measurements that become more precise following the item response. The algorithm had produced a face scale measurement with an $SE(\hat{\theta})$ of 0.355 from the first 20 items, although none of these items belonged to the face scale.

Item	Item	Face		Nose		Nostrils		Jaw		Teeth		Lips		Scar	
number	name	$\hat{\theta}$	$SE(\hat{\theta})$	$\hat{\theta}$	$SE(\hat{\theta})$	$\hat{\theta}$	$SE(\hat{\theta})$	$\hat{\theta}$	$SE(\hat{\theta})$	$\hat{\theta}$	$SE(\hat{\theta})$	$\hat{\theta}$	$SE(\hat{\theta})$	$\hat{\theta}$	$SE(\hat{\theta})$
		0.000	1.000	0.000	0.639	0.000	0.727	0.000	1.147	0.000	1.154	0.000	0.769	0.000	1.087
1	Nostrils 2	0.304	0.688	0.000	0.639	0.161	0.618	0.000	1.147	0.000	1.154	0.000	0.769	0.000	1.087
2	Jaw 1	0.484	0.628	0.000	0.639	0.056	0.596	0.497	0.821	0.000	1.154	0.000	0.769	0.000	1.087
3	Scar 3	0.173	0.548	0.000	0.639	0.240	0.564	0.693	0.761	0.000	1.154	0.000	0.769	-0.952	0.650
4	Teeth 7	0.063	0.506	0.000	0.639	0.307	0.546	0.766	0.736	-0.491	0.718	0.000	0.769	-0.864	0.630
5	Lips 7	0.112	0.455	0.000	0.639	0.277	0.531	0.733	0.727	-0.526	0.700	0.113	0.572	-0.904	0.603
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
20	Nose 2	-0.240	0.355	-0.360	0.415	0.473	0.415	0.572	0.445	-0.084	0.501	-0.039	0.451	-0.659	0.460
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
46	Teeth 8	0.405	0.251	-0.602	0.310	0.041	0.331	0.045	0.339	-0.505	0.392	-0.352	0.341	-1.096	0.338

Table 6.3 displays the mean number of items used from each scale in the 46-item assessment, across all 1000 simulated respondents, together with the median $SE(\hat{\theta})$ for each dimension. Face scale items were used less frequently than items from other scales.

Table 6.3: Mean number of items administered from each scale, and median $SE(\hat{\theta})$ in each dimension.

	Mean items	**Median $SE(\hat{\theta})$**
Face	4.75	0.245
Nose	8.13	0.312
Nostrils	5.95	0.325
Lips	7.84	0.329
Scar	6.25	0.339
Teeth	6.29	0.410
Jaw	6.80	0.368

6.4 Discussion

Through SEM, I have demonstrated that there are statistical relationships between CLEFT-Q scales than could potentially be leveraged by MCAT algorithms. My conceptually-based structural model was supported by CLEFT-Q field test data, and demonstrated excellent performance against all fit measures except for the X^2 test. This is particularly sensitive to large sample sizes, and almost always produces a $p < 0.050$ with sample sizes larger than 400 [176]. Considering how well the model performed against the other fit thresholds, this is very likely to be a type I error.

Further, I have shown that by accounting for relationships between latent variables, MCAT can measure multiple latent constructs simultaneously. In this simulation, face scale scores were produced with a low $SE(\hat{\theta})$ from few items, but other latent constructs were measured with lower precision and a higher item demand. Chapter 5 and

Appendix D provide a sense of the precision achieved by unidimensional GRM-based CAT algorithms at different assessment lengths across different constructs. With these in mind, this MCAT system does not appear to offer dramatic efficiency gain over the multiple unidimensional system, but a direct and formal comparison is not advisable, as these simulations were performed with different datasets, under different experimental conditions (different estimators and item selection criteria), and as discussed in Chapter 5, it is not necessarily helpful to compare the SE achieved by models of different complexity.

I have previously published a more complex structural model to the one described in this chapter [177]. In that publication, I had aimed to explore whether detailed statistical patterns in the CLEFT-Q field test responses were consistent with a clinical understanding of how these appearance constructs may relate. To that end, I made data-driven modifications to the hypothesised model (Figure 6.2) that captured cross loadings of items onto multiple factors, and residual covariance between items. This resulted in a slightly better fitting model than the one described above. In this chapter, I aimed to take my structural theory forward as an MCAT algorithm. I have described the potential dangers of deriving IRT measurements from over-specified models in Chapter 5 (loss of generalisability and artificially inflated precision). For these reasons, I tried to respect model parsimony here by proceeding with the simpler structural model.

It is tempting to make post-hoc adjustments to psychometric models based on their ability to improve fit (this can be quantified through a statistic known as the modification index). A more complex model would have been likely to improve MCAT efficiency by allowing certain items to carry information in a higher number of dimensions. But in doing so, we may veer towards describing patterns that exist exclusively in the calibration dataset, rather than patterns that exist in reality. This practice has been likened to eating salted peanuts, since "one [modification] is never enough" [66]. In my view, freely estimated MIRT parameters should have a very firm conceptual grounding, as is the case in this

chapter.

I would consider this work as early and experimental, rather than suggest that these models and algorithms are suitable for real world deployment. There are a number of aspects to this work that could and should be challenged. For example, it is possible that the conceptual model could be improved. In the current model, the relationship between the appearance of the face and the facial components is reflective - the appearance of the face explains the common variance in facial component scores. But this relationship may actually be formative - the appearance of the facial components may group together to explain the overall appearance of the face (imagine the causal arrows between factors in Figure 6.2 were reversed). It may be possible to specify such reflective-formative models using partial-least-squares-based (as opposed to covariance-based) SEM [178], which is available in some newer R packages [179]. The model might also be extended to understand how appearance constructs impact on psychological, social, or school function. And some of the post-hoc modifications described in the 2021 paper [177] may be perfectly reasonable here. Large datasets for calibration and validation would be useful for this work.

One question this study raises is that of content validity. While it may be possible to generate a precise face scale measurement by only asking a respondent about their facial components, is it correct to do so? Do we risk missing face scale items that are very important to the respondent, or which might flag otherwise undetected quality of life issues to their clinician? Perhaps there is a role for content balancing in these types of assessment, to ensure that a certain number of items (or certain specific items) from each scale are asked. Another area to contemplate is the potential framing effect of grouping items into unidimensional scales. In a multidimensional assessment, sequential items may be drawn from different scales (see Table 6.2) and it is not clear whether this would have an impact on a person's responses, compared to multiple unidimensional assessments.

6.5 Conclusion

Given the current standards of psychometrics in surgery and CLP, a hierarchical MIRT solution for the CLEFT-Q appearance scales is not the most pressing addition to the literature. We should focus resources on driving a culture shift towards understanding the importance of high quality PROMs such as the CLEFT-Q, standardising the measures used in research and clinical practice, and streamlining their implementation. These actions will have a far greater, and more direct impact on patient care than the research described in this chapter. Nonetheless, MIRT may play a role in enhancing unidimensional CAT assessments once these are accepted as common practice, and this work narrows the gap between the study of theoretical psychometrics within surgery and other fields, potentially laying groundwork for the higher quality measurement systems of the future.

7

Recursive partitioning as an alternative approach to computerised adaptive testing

Contents

7.1 Introduction 121
7.2 Methods 123
7.2.1 Software 123
7.2.2 Participants 123
7.2.3 Ground truth 123
7.2.4 Data splitting 124
7.2.5 Patient-reported data 124
7.2.6 Clinical variables 124
7.2.7 Decision tree training 125
7.2.8 Rasch-based CAT assessments 126
7.2.9 Assessment length 126
7.2.10 Comparison 127
7.3 Results 127
7.4 Discussion 129
7.5 Conclusion 131

7.1 Introduction

In Chapters 5 and 6, I explored alternative approaches to deriving measurement information from item responses. In this chapter, I explore the idea that we might derive measurement information from other sources altogether. For example, previous work has shown that the distribution of CLEFT-Q lip scale scores achieved by people born with complete bilateral clefts is poorer than that achieved by people born with incomplete unilateral clefts. Similarly, ascending Veau classification is associated with decreasing speech function and speech distress scores [99]. This sort of clinical information could, in theory, be automatically obtained from a person's medical record, at no extra burden to the respondent. It might be possible to use this information to home in on a person's scale location, or select the most appropriate item to start a CAT assessment. I believe that this strategy has the potential to lead to more efficient, or more personalised, CAT assessments.

One approach to operationalising this idea is *recursive partitioning*. Recursive partitioning (also known as divisive hierarchical clustering) is a from of supervised machine learning that iteratively splits labelled data into subgroups, with the aim of minimising the intra-subgroup variance with respect to a given outcome (for example, a score on a PROM) [180]. Put simply, and in the context of CAT assessments, the algorithm creates the most efficient decision tree (flowchart) possible for dividing people into groups with similar scores. These divisions (branches of the flow chart) can be based on the responses to an item on a PROM, dictate which item should be posed next, and ultimately lead to a prediction of the score that would be obtained from the full-length assessment [181–183]. There are several different forms of recursive partitioning,

---

This chapter is based on Harrison, C. *et al*. Recursive partitioning vs computerized adaptive testing to reduce the burden of health assessments in cleft lip and/or palate: comparative simulation study. *Journal of Medical Internet Research*, 23(7), e26412 (2021).

but in this chapter the algorithms I refer to are what Breiman *et al* first described as "classification and regression trees" in 1984 [184]. Specifically, I consider only the regression form of these algorithms, which is designed to handle continuous outcome data.

Recursive partitioning might be an appealing alternative to modern test theory for the creation of CAT assessments. Firstly, decision trees are not restricted by modern test theory requirements (which many legacy PROMs fail to meet). Secondly, decision trees are constructed *a priori*, meaning that score estimation and item selection computations are not required between items. This might reduce lag time [185], particularly when data are transferred to and from an external assessment centre. It might also be possible to construct decision trees that split data based on both item responses and clinical variables. For example, in an assessment of physical function, item selection might be guided by step count, or in an assessment of asthma symptom burden, item selection might be guided by the frequency of inhaler prescriptions.

In this chapter, I explore whether CAT assessments built through recursive partitioning, including both clinician-reported and patient-reported information, perform more efficiently than CAT assessments built through a modern test theory approach. I test the hypothesis that at a given assessment length (number of patient-reported items), CAT assessments built through recursive partitioning would reproduce full-length scale scores more accurately than modern test theory CAT algorithms due to the extra clinical information at their disposal.

7.2 Methods

7.2.1 Software

I performed this analysis with R v 4.0.0 and the following packages: *foreign* v 0.8-80, *dplyr* v 1.0.0, *mirt* v 1.32.1, *mirtCAT* v 1.9.3, *stringr* v 1.4.0, *rpart* v 4.1-15, *partykit* v 1.2-8, *rpart.plot* v 3.0.8, and *ggplot2* v 3.3.2.

7.2.2 Participants

I used CLEFT-Q field test data for this study, as described in Chapter 2, section 2.2.2. I excluded participants with a cleft palate only from analyses relating to the lip and cleft lip scar scales. I excluded participants without a cleft palate from analyses relating to the speech function and speech distress scales. This is consistent with populations expected to use these assessments in real world settings. Missing response set data were handled through listwise exclusion and missing clinical data were handled through pairwise exclusion.

7.2.3 Ground truth

For each respondent in the dataset, I calculated EAP scores for each (full-length) CLEFT-Q scale, using the same Rasch models that I used to build the CLEFT-Q CAT (see Chapter 2, sections 2.2.2 and 2.2.3). I recorded these as logit scores. These represent the *ground truth* which CAT assessments attempted to reproduce in this study.

7.2.4 Data splitting

I randomly split data into training and test datasets, for each scale, in a ratio of 2:1. For each scale, I developed both Rasch-based CAT algorithms and tree-based CAT algorithms on the training datasets and evaluated their performance in the test datasets.

7.2.5 Patient-reported data

For Rasch model parameterisation, I used all complete response sets to each CLEFT-Q scale in the training dataset. To train the decision trees, I used these complete response sets and additional data. Additional data included clinical variables (described below) and another question posed to participants for each of the 7 appearance scales. These were anchoring questions used in the CLEFT-Q field test and did not form part of the Rasch-validated scales. For example, the question "how much do you like how your lips look overall?" was administered alongside the CLEFT-Q lip scale, with 4 response options: "not at all", "a little bit", "quite a bit", and "very much". Because these questions were not part of the Rasch scale, they cannot be used by Rasch-based CAT algorithms, but the decision trees were free to use these additional questions. The wording of these questions can be found in Appendix A, section A.1.

7.2.6 Clinical variables

Decision trees were permitted to make splits based on clinical variables as well as item responses. Clinical variables included:

1. Age
2. Gender

3. Cleft type (CL, CP, CLA, CLAP)
4. Laterality (unilateral or bilateral)
5. Type of unilateral CL (complete, incomplete, lesser form)
6. Type of bilateral CL (complete, incomplete, lesser form, asymmetric)
7. Whether future treatment was planned, including:
    (a) Primary lip repair
    (b) Lip revision surgery
    (c) Primary palate repair
    (d) Palatal fistula repair
    (e) Alveolar bone grafting
    (f) Orthodontic treatment
    (g) Orthognathic surgery
    (h) Rhinoplasty
    (i) Speech surgery
    (j) Speech therapy
    (k) Psychological therapy
8. Current speech problem perceived by the clinician

7.2.7 Decision tree training

I developed decision trees which made binary data splits based on either patient-reported or clinical variables. For example, the algorithm might calculate that the most efficient

way to partition training data into groups with similar CLEFT-Q lip scores is to start with a split based on cleft laterality. Respondents with unilateral clefts are then split again depending on their response to one item (or another clinical variable), and respondents with bilateral clefts are split by their response to a different item or clinical variable. Once the algorithm has computed the optimal set of splitting rules, these can be brought forward into a decision tree and applied to test dataset respondents as simulated assessments.

I limited the number of splits the algorithms could make to match the number of items posed by their Rasch-based comparators. Before evaluating the performance of the decision trees, I applied cost-complexity pruning, which involves retrospectively removing branches that do not improve model fit by a given amount (the complexity parameter). This complexity parameter was chosen to minimise prediction error during a ten-fold cross validation with the training dataset.

7.2.8 Rasch-based CAT assessments

I used the EM algorithm in mirt to parameterise Rasch models, and built CAT algorithms from these using MEPV item selection and EAP scoring. I chose fixed-length stopping rules for Rasch-based assessments, matched to the number of splits permitted during decision tree training. I trialled each assessment at two different stopping rules which represented approximately 50% and 75% of the scale's length.

7.2.9 Assessment length

I aimed to compare the accuracy of tree-based and Rasch-based CAT assessments that posed a similar number of items to respondents. I was able to predetermine the length of

Rasch-based CAT assessments via fixed-length stopping rules, but it was not possible to precisely predetermine the length of tree-based assessments for every respondent. This is because splits can be based on either item responses or clinical variables, and the removal of branches via cost-complexity pruning meant that respondents could be posed tree-based CAT assessments with a differing number of splits, depending on their item responses or clinical characteristics.

For tree-based assessments, the length of the assessment was defined as the number of items a respondent would have had to complete. Clinical variables were not counted as item responses, but responses to the anchoring questions (see subsection 7.2.5) were.

7.2.10 Comparison

I applied the tree-based and Rasch-based CAT algorithms to the test data with the aim of recreating ground truth scores (see subsection 7.2.3). For both assessment types at each assessment length, I recorded the mean number of items administered, which items or clinical variables were used, correlation between CAT score and ground truth score, and the RMSE and mean absolute error (MAE) with respect to ground truth scores.

7.3 Results

For each scale, I used the same CLEFT-Q field test samples that I described in Chapter 2. The demographics, clinical characteristics, and missing data distributions for these samples are available in Appendix A, section A.1.

Generally, the mean number of items administered was similar between tree-based and Rasch-based CAT algorithms, but tree-based algorithms achieved poorer accuracy

than Rasch-based algorithms (a lower correlation with ground truth, and higher RMSE and MAE values). Table 7.1 illustrates this comparison with the social function scale assessment, limited to 7 items. I have provided the full set of results tables in Appendix E, section E.1.

While the tree-based algorithms were free to use clinical variables to make splits, they tended to use scale items preferentially. Clinical variables were either not used at all, removed through cost-complexity pruning, or used at deeper levels of the tree. The one exception to this was the decision tree for the nose scale assessment, which made its first split based on the anchoring question "how much do you like how your nose looks overall?" This question is not included in the Rasch validated scale and cannot be used by the Rasch-based algorithms.

As an illustrative example, Figure 7.1 depicts the dendrogram for the tree-based nostrils scale algorithm, limited to a depth of 3 splits.

Table 7.1: Comparison of Rasch-based CAT algorithms and tree-based CAT algorithms for the social function scale assessment (7 item limit). DT: decision tree; RMSE: root mean square error; MAE: mean absolute error.

Social function scale assessment (7 item limit)

	Variables used	Mean items	Correlation	RMSE	MAE
DT	Age, cleft type, laterality, and all social function scale items	6.64	0.97	0.42	0.31
Rasch	All social function scale items	7.00	0.99	0.25	0.19

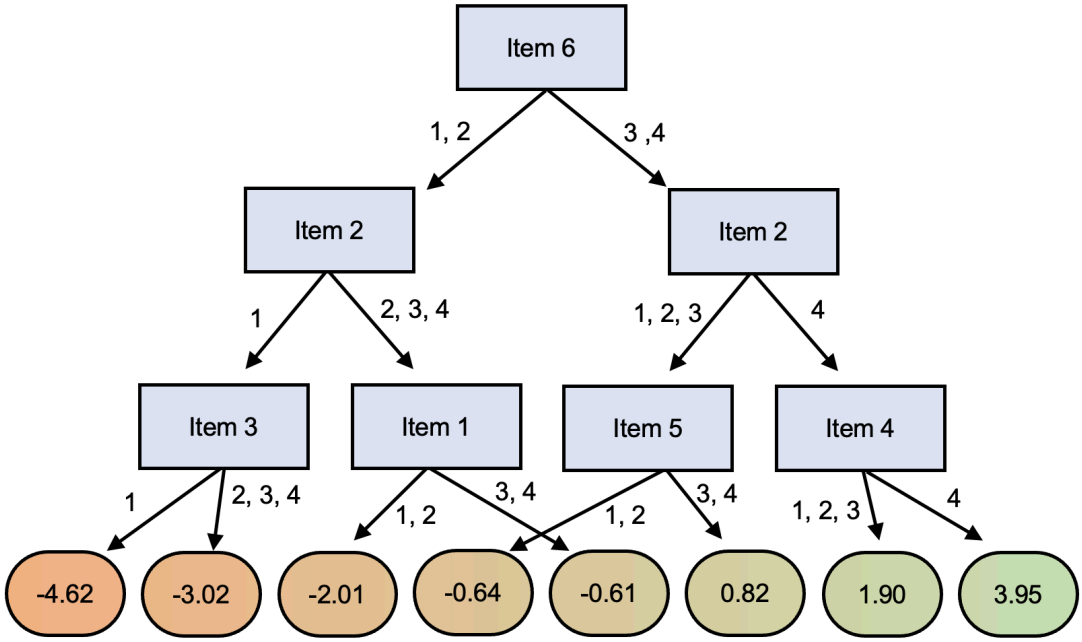


Figure 7.1: Dendrogram depicting the 3-item, tree-based, nostrils scale CAT algorithm. The rectangles represent the items posed and the arrows are labelled with the respondent's potential item responses. In this case, the assessment starts with item 6. If the respondent endorses response options 1 or 2, they are then presented with item 2. At that point, if they endorse response option 1 they are posed item 3. The ovals (leaves of the tree, also known as terminal nodes) are labelled with the respondent's predicted score, based on their set of item responses. These are the mean scores of respondents who had endorsed the same set of response options in the training dataset. No branches have been pruned in this example. Scores are presented as logits.

The figure is reproduced from recursive partitioning vs computerized adaptive testing to reduce the burden of health assessments in cleft lip and/or palate: comparative simulation study. *Journal of Medical Internet Research*, 23(7), e26412 (2021). Available at: https://www.jmir.org/2021/7/e26412. This figure is available under the creative commons license (CC BY 4.0): https://creativecommons.org/licenses/by/4.0/. Shading and labels have been adjusted.

7.4 Discussion

This study demonstrates that decision trees can combine both patient-reported data and healthcare record data to create truncated, personalised CAT assessments. While the approach I explored in this chapter is feasible, I would not recommend it for clinical use. At similar assessment lengths, most tree-based algorithms were considerably less

accurate than Rasch-based algorithms.

As it was not possible to predetermine the exact length of tree-based assessments, these often posed a smaller mean number of items than Rasch-based comparators (see Table 7.1 and Appendix E, section E.1). This does bias the accuracy comparisons in favour of Rasch-based algorithms, but even with this bias in mind, most Rasch-based assessments clearly outperformed their comparators.

The tree-based algorithms explored in this study can only make binary data splits and this is likely to be a major contributor to their inferior performance. In Figure 7.1, one can see that a 3-item tree-based assessment (which has not been pruned) is capable of producing 8 different measurements (2^3). The equivalent Rasch-based assessment, which treats each of the 4 item response options independently, would be capable of 64 (4^3) different EAP estimates.

An alternative approach might be to use a Chi squared automatic interaction detection (CHAID) algorithms [186]. These algorithms also create decision trees through recursive partitioning, but unlike classification and regression trees, they are able to make non-binary data splits. This might be more appropriate for polytomous PROM responses. Other researchers have applied CHAID algorithms to make purely patient-reported CAT assessments [187, 188], and I have previously suggested that they may be capable of combining PROM data and clinical variables with the aim of predicting surgical outcomes [189]. But their ability to use medical record data to improve the efficiency of CAT assessments remains unexplored.

An important limitation of tree-based approaches, including CHAID, is that they provide discrete measurements. Rasch-based CAT assessments will provide a distribution of plausible θ values, from which $\hat{\theta}$ is chosen (via EAP or MAP, for example). This allows

Rasch-based CAT algorithms to quantify measurement error, as described by their underlying Rasch models. It may be possible to provide some sense of measurement error for decision trees by examining the distributions of scores in the terminal nodes, or the distribution of error in an external validation sample. But this would only account for the error in the approximation to $\hat{\theta}$, and not the potential error of $\hat{\theta}$ itself. This argument will become more relevant once we move away from using PROMs to generate point estimates, and towards a culture of understanding and accounting for measurement error by treating latent construct measurements as plausible distributions.

In this study, the tree-based assessments preferentially used scale items for measurement, rather than clinical variables. This may be because the CLEFT-Q scale items have been specifically chosen to clearly reflect their respective latent constructs, and are mostly well-targeted to the population. It may also be that clinical variables, such as cleft type, have less of an impact on latent construct levels than I had expected. Tree-based algorithms might be most effective for use with legacy instruments that have not been developed under a modern test theory framework, or where clinical variables have a greater influence on latent construct levels (for example, a history of spinal cord injury might have a very large impact on physical function levels).

7.5 Conclusion

In some cases, clinical data can provide information about a respondent's likely latent construct level, and could potentially be automatically captured from an EHR platform to streamline CAT assessments. Classification and regression trees are not the optimal strategy for achieving this, and CLP may not be the best clinical substrate for this innovation. Recursive partitioning algorithms that can make non-binary splits, such as CHAID, may show a superior performance to those explored in this study when applied to polytomous PROM data, but all tree-based approaches will make the quantification

of measurement error more challenging. A more elegant solution might be to model the relationship between clinical variables and latent construct levels within a modern test theory framework, and further work in this area would be worthwhile.

8
Discussion

Contents

8.1 A summary of completed work 134
8.2 The CLEFT-Q CAT 135
8.2.1 Measurement philosophy 135
8.2.2 Strengths 139
8.2.3 Limitations 140
8.2.4 Impact 144
8.3 The future of clinical psychometrics 145
8.3.1 Measurement error 145
8.3.2 Multidimensional measurement 147
8.3.3 Incorporating non-PROM data into latent variable models . . 148
8.4 Conclusion 150

8.1 A summary of completed work

The research aims of this thesis were to:

1. Develop a CAT version of the CLEFT-Q that could facilitate the questionnaire's uptake in routine clinical practice, address some of the limitations of current CAT systems, and easily translate into other clinical fields.

2. Explore potential future research directions for CAT, in the context of the CLEFT-Q CAT's limitations.

Chapters 2 to 4 describe work which meets the first aim. I have calibrated and validated CAT algorithms for the CLEFT-Q scales, worked with key stakeholders to set patient-centred stopping rules for these assessments, operationalised them through the open-source and locally executable Concerto platform, released a free web application to compare CAT assessment scores to relevant reference populations, and demonstrated that both patients and clinicians at Oxford University Hospitals perceive the system to improve clinical communication and facilitate patient-centredness.[1]

Chapters 5 to 7 describe work which meets the second aim. In this second part of the thesis I have developed novel ideas that serve as targets for future clinical psychometrics research. In Chapter 5 I challenge emerging heuristics relating to measurement precision, in Chapter 6 I demonstrate how multidimensional frameworks could reduce assessment burden in complex conditions that otherwise require multiple unidimensional assessments, and in Chapter 7 I argue that we currently neglect rich clinical data that might contribute to latent variable measurement, at no added burden to the respondent.

[1] I am grateful to have received the International Society for Quality of Life Research 2021 student prize for the work described in Chapters 2 to 4. This work has also directly contributed to a successful application for a Royal College of Surgeons Hunterian Lectureship.

In this chapter, I first discuss the CLEFT-Q CAT, including the philosophical approaches to measurement which I have adopted during its development, the platform's strengths and limitations, its impact, and areas where further work is required for the system to fulfil its potential. I then provide suggestions for how we might expand on the broader ideas described in the second part of this thesis.

8.2 The CLEFT-Q CAT

8.2.1 Measurement philosophy

In developing the CLEFT-Q CAT, I have adopted certain measurement philosophies which can be considered separately to the platform's strengths and limitations, as the advantages and disadvantages of alternative approaches to measurement are still debated [42, 44, 48, 157, 158].

Firstly, I have applied Bayesian techniques (EAP scoring and MEPV item selection), assuming a standard normal prior. My rationale for this decision was to improve the quality of measurement at extreme ends of the scales (as discussed in Chapter 1 subsection 1.4.4, a frequentist approach might return a score of ∞ in a patient who endorses the most positive set of response options, unless quadrature limits are applied to the computations). I think this is a sensible strategy, considering the ceiling effect demonstrated in some of the CLEFT-Q scales (see Appendix B, section B.3). It is also reassuring that the PROMIS initiative share this philosophical approach, for exactly the same reason [190]. However, I do not have a compelling theoretical basis to assume that θ in the population follows a standard normal distribution, in any CLEFT-Q scale, and it is likely that CLEFT-Q CAT measurements have an 'inward' bias towards the mean of

that prior [50]. An alternative approach would be to apply quadrature limits to frequentist estimators (limiting the minimum and maximum possible values of $\hat{\theta}$). But I would not have a theoretical basis for setting these limits either. Fortunately, the results of Chapter 2 show a minimal difference in CLEFT-Q scale scores when comparing Bayesian and frequentist methods.

An exciting extension to EAP estimation for the CLEFT-Q, that could be explored in future, is full Bayesian modelling [158]. In this approach, the model parameters (as well as the person parameters) follow a posterior distribution. This would allow us to not only capture the uncertainty of $\hat{\theta}$ given the model ($SE(\hat{\theta})$), but also the uncertainty of the model itself. Again, bias can arise here if inappropriate priors are chosen.

Unlike the PROMIS team, I have followed an RMT (rather than an IRT) approach. This was primarily for consistency with the full-length CLEFT-Q. Chapter 5 explores an alternative IRT approach and Chapter 6 extends this into a multidimensional setting. I have touched upon some of the common points of discussion in the debate between rating scale models, partial credit Rasch models, IRT, and MIRT, including the separability of item and person parameters, and whether these models realistically represent the relationship between item responses and the underlying latent constructs. But to a non-specialist, this debate can be broadly considered one of model parsimony.

In January 1946, Pablo Picasso completed a series of 11 lithographs collectively known as Le Taureau, in which he progressively simplified a wash drawing of a bull into its most basic form (Figure 8.1) [191].

These prints are a perfect allegory for the parsimony spectrum that exists within modern test theory approaches to latent variable modelling. At one end of the spectrum we have a detailed representation of what one bull, or one dataset, might look like. The first

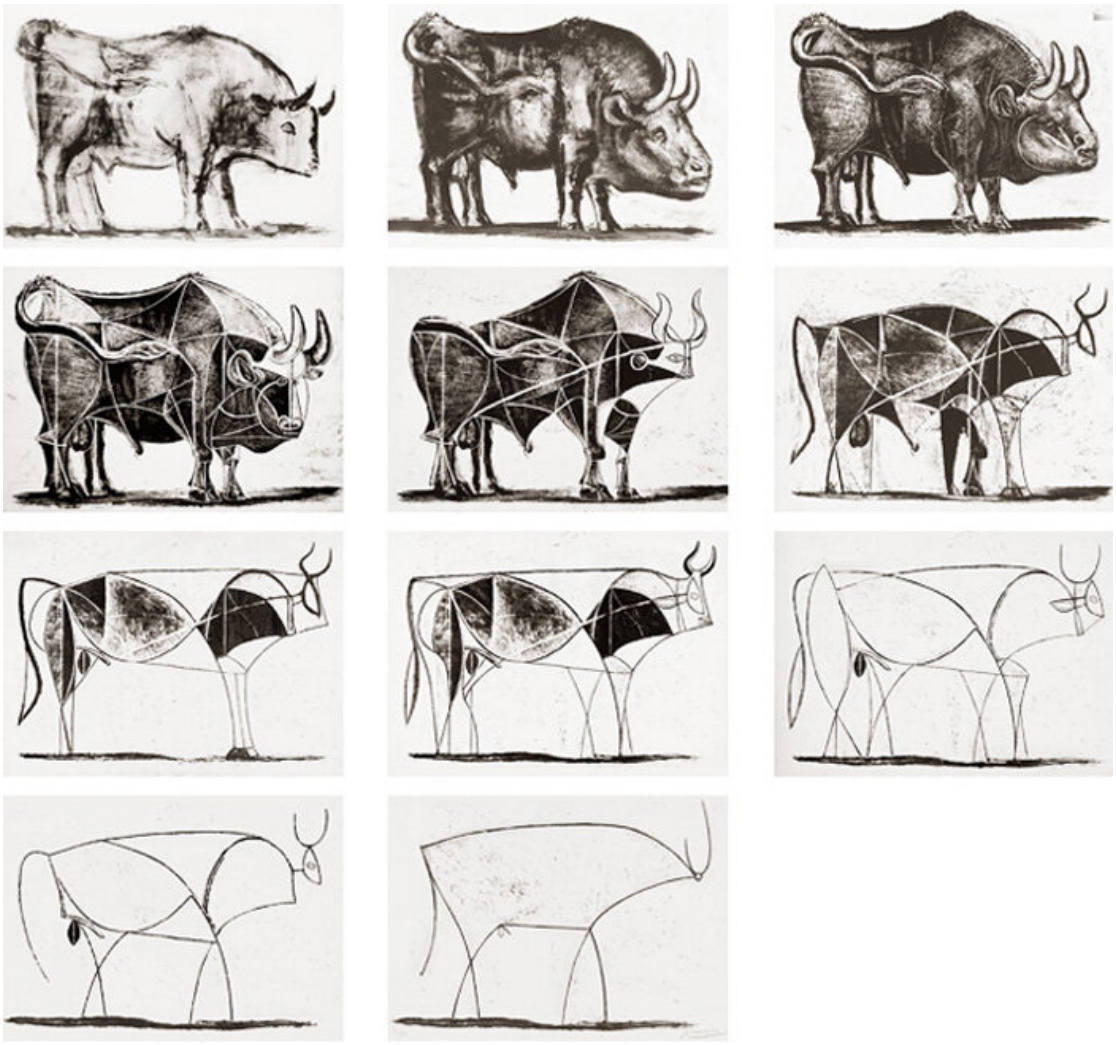


Figure 8.1: Pablo Picasso's Le Taureau. Photograph by Wikimedia Commons user Expiredpate. Available under the creative commons license (CC BY-SA 4.0): https://creativecommons.org/licenses/by-sa/4.0/.

lithograph includes subtle features such as intercostal recessions and the shading shows us exactly how the bull stands, facing away from the sun. This drawing clearly illustrates the appearance of the real-world subject on which it is modelled. This is possible because Picasso has been liberal with brush strokes, which can be adapted to take any form, as can the many freely-estimated parameters of a complex SEM or MIRT model. The final lithograph has a different aim. It captures the essence of the subject, through a general depiction of consistent and representative features that distinguish it from most other creatures. Here, Picasso has been very constrained with his quill strokes, resulting in an image which is less detailed, but more generalisable.

An understanding of this concept is important for interpreting model fit statistics and SE (as discussed in Chapter 5). But I do not believe that one approach necessarily provides *better* measurement than another. I have suggested that (with caution, large sample sizes, and external validation) more detailed models might allow more precise CLEFT-Q measurements and CAT efficiency gains, but as Picasso shows us, complexity is not always sophistication.

Another choice I made in the development of the CLEFT-Q CAT was to use fixed-length stopping rules. This ensures that all participants are posed the same number of items from each scale, and no participant is administered more items than were deemed acceptable by majority voting in the stopping rule workshop (see Chapter 4). This may be a more equitable approach than using precision-based stopping rules, and potentially mitigates any bias that might arise when respondents with certain latent construct levels (associated with a low Fisher information) are exposed to higher burden assessments. The disadvantage of this approach is that respondents with differing latent construct levels will be measured with differing levels of precision. This precision difference might be substantial when comparing respondents with scores near the middle or extreme ends of each scale (see Appendix B, section B.3). To mitigate this, PROMIS use conjugate stopping rules, for example $SE(\hat{\theta}) < 0.3$ or 12 items [76], but this approach results in both unequal assessment lengths and variation in precision (the variation in precision may be less than that seen with purely fixed-length stopping rules). To me, one approach is not clearly superior to the other. Currently, the CLEFT-Q CAT platform does not present SE alongside a respondent's measurement. Given the decision to proceed with fixed-length stopping rules, a valuable addition to the platform might be a graphical representation of a respondent's score, $\hat{\theta}$, together with confidence intervals based on that score's precision $SE(\hat{\theta})$.

8.2.2 Strengths

The CLEFT-Q CAT platform shows some convincing strengths over other CAT platforms. Firstly, it is based on a condition-specific measure that captures elements of health that have been shown to be highly important to its target user-base [118]. There are clear advantages to this for CLP care, compared with generic instruments that may fail to capture the specific needs of this population.

Rather than selecting stopping rules for patients, I selected stopping rules *with* patients. This is a more contemporary approach to assessment development and has potentially helped to align the platform with the wants and needs of end-users.

The Concerto platform is locally implementable. This means that patient data do not need to be sent to and from an external assessment centre for score estimation and item selection. The platform can be installed on an NHS server and run offline. This may be appealing with data security and computational efficiency in mind.

The CLEFT-Q CAT score checker web application can be easily used to contextualise a respondent's score through comparison to scores achieved by populations with matched clinical and demographic characteristics. It is straightforward to use and freely available. I believe that the use of condition-specific reference populations is a strength here, as it makes for a more clinically relevant comparison than the use of generic reference populations.

I have kept as much of this work open-source as possible. While the CLEFT-Q itself does require a license, Concerto is open-source, I have made Rasch model parameters available, and I have released the score checker, together with its source-code, under the 3-clause Berkeley Source Distribution license [146, 147]. I have also made the cleftqCATsim

R package freely available [140]. This is a fairly novel approach to open science. By condensing CLEFT-Q CAT simulation code into a few simple functions, even novice R users are now able to recreate the experiments described in Chapter 3. More advanced R users who wish to appraise these methods in detail are free to explore the functions. This open-source approach will make the translation of this work to other measures and clinical conditions considerably more rapid and cost-effective.

8.2.3 Limitations

There are limitations of CAT in general and limitation of the CLEFT-Q CAT specifically. While in this thesis I have primarily considered PROMs a tool for measuring latent constructs, they can also be considered as conversation starters or screening tools for important symptoms or health concerns [192]. By reducing the number of items in a PROM, we potentially reduce the content validity coverage of that questionnaire. This might result in important questions being missed, and important issues being undetected. The impact of CAT on content validity could be explored further through qualitative research and strategies to uphold content validity in CAT could be explored (for example, content balancing, where a certain number of items from a certain group of items are always administered).

The CLEFT-Q CAT scales will always start with the same item, and subsequent items are selected consistently, based on a given response pattern. It is possible that this might feel repetitive if the CAT is used by a respondent multiple times. In these circumstances, alternative item selection criteria (for example, the randomesque method, in which one of the n best items is randomly selected [193]), could be explored. This might also improve content validity coverage.

In all the simulation studies described in this thesis, I have assumed that the ordering of items does not affect the item response. I do not have empirical evidence to support this assumption. This is highly relevant, given that the Rasch models were calibrated on sequential item responses and that the CLEFT-Q scales are structured such that the questions at the start of each scale are targeted to more negative latent construct levels. This could be investigated further by fitting Rasch models to CLEFT-Q item responses that have been collected in different orders and comparing model parameters.

I have also assumed that by reducing the assessment length, I have reduced assessment burden. This might not be the case. For example, there may be additional burden associated with using an iPad, maintaining a network connection, and waiting for the algorithm to select the next question. While the patients interviewed in Chapter 4 generally perceived the system as low-burden, I made no formal comparison to the full-length questionnaire. Nor have I presented any quantitative data to support this claim (for example by comparing completion times between the full-length and CAT versions of the questionnaire).

The usability of the CLEFT-Q CAT system could be improved. Currently, scores for a respondent are displayed numerically on the clinician-facing interface of the Concerto platform. These then need to be manually entered into a patient's EHR or the CLEFT-Q CAT score checker. An integrated system, where scores are presented graphically and automatically uploaded to the patient's medical record, would be preferable. This is almost certainly possible with future upgrades. Graphical feedback can be provided within the Concerto environment, although this may require more funding and specialist technical input, and the platform's application programming interface can already be used to communicated with EHRs. Work is underway to permit more seamless integration between Concerto assessments and major EHR systems, and this could significantly enhance the CLEFT-Q CAT's ease of use.

Currently, it would be challenging to use the CLEFT-Q or CLEFT-Q CAT as a primary outcome measure in a clinical trial because it is difficult to understand what level of difference between groups, or change within a group, should be considered clinically meaningful. One study has reported distribution-based estimates for minimal important change in CLEFT-Q scale scores following rhinoplasty, orthognathic surgery, and scar revision [124]. But ideally target differences for clinical trials should be triangulated from multiple techniques, including anchor-based methods [194, 195]. Future test-retest work involving the CLEFT-Q and CLEFT-Q CAT will help us to understand the instruments' minimal detectable change values (the smallest change in score not attributable to measurement error) and target differences for clinical trials in various contexts.

In this thesis, I have developed CAT assessments for existing PROM scales which were relatively short (6 to 12 items) to begin with. But this is potentially not the most effective use CAT. Other CAT researchers, including the PROMIS team, have taken a different approach to CAT development, known as item banking [196]. In this strategy, a large number of items (for example, several hundred) are generated and incorporated into psychometric models. Together they cover a broad range of latent construct levels with a high precision. Both a wide breadth of targeting and high precision is possible because there is no limit to the number of items in the item bank. A short-form can be generated by selecting items from the item bank to cover a certain target θ range. Alternatively, items can be administered through CAT. The CLEFT-Q scales (which have undergone item reduction) are similar to the short-forms generated in an item banking approach. They have been developed and selected to cover a certain target θ range and precision is limited by the number of items they each contain. When item banks are used, CAT assessments can return higher-precision measurements than short-forms, in some cases even while using fewer items [75, 168]. It is likely that the CLEFT-Q CAT assessments would have had a higher precision had they been built from large item banks. This may

have also mitigated the ceiling effect demonstrated in Appendix B, section B.3. It would be technically possible to expand the set of items used by the CLEFT-Q CAT algorithms (either by calibrating items from other questionnaires onto a common psychometric model [158, 197–199], or by generating new CLEFT-Q items). This might leverage the advantages of item banking, but would also come with practical considerations surrounding instrument licensing and short-form version control.

Following this DPhil, I propose the following next steps:

1. A test-retest cohort study to establish minimal detectable change estimates for the CLEFT-Q and CLEFT-Q CAT, conducted in line with COSMIN guidance [26]. This would involve administering the assessments twice, with a two-week interval in which no change in latent construct level is expected. As an extension to this work, the test-retest variation in individuals' scores could be compared to their expected test-retest variation, based on baseline $SE(\hat{\theta})$ values. This would provide valuable data on the precision of the CLEFT-Q, how this varies with CAT, and the appropriateness of quantifying reliability through $SE(\hat{\theta})$.

2. A series of (potentially registry-based) cohort studies that measure preoperative and postoperative CLEFT-Q CAT scores, together with anchoring questions, from which target differences for clinical trials might be triangulated in line with DELTA2 guidance [195]. An anchor question, such as the global rating of change, can be used to identify patients who have had a small but meaningful improvement, and those who have not changed following surgery. Postoperative score differences between these groups can be used to guide target differences for clinical trials, together with distributional target difference estimates.

3. Cognitive debriefing to explore the content validity coverage of the CLEFT-Q CAT with the stopping rules I have recommended in this thesis. This should involve

unstructured interviews with a maximally diverse sample of patients, until thematic saturation is reached.

These work streams would support the use of the CLEFT-Q CAT as a primary outcome measure in high-stakes clinical research, which may then go on to improve evidence-based treatment protocols.

8.2.4 Impact

Assessment burden has been identified as one of the biggest barriers to cleft centres implementing the CLEFT-Q as part of routine care [125–128]. In a preparatory survey for this project, clinicians at 16 of the 18 UK hospitals providing cleft care felt that a CAT platform would directly facilitate the routine use of the CLEFT-Q as part of their practice [128]. This work has resulted in a CLEFT-Q CAT platform which is perceived to improve clinical communication and patient experience, and assist in shared decision making, without adding excessive burden. This has the potential to drive high quality, standardised PROM use for CLP care in the UK and beyond.

The routine use of the CLEFT-Q CAT requires a resource commitment from the clinical team. There is a small cost to setting up and maintaining either a local or cloud-based server and staff time is required to ensure assessments are administered to patients. So far, at least one bid for a National Institute for Health and Care Research (NIHR) programme grant has been submitted to fund research which would involve 16 UK cleft centres implementing the CLEFT-Q CAT platform. Senior leaders of a North American hospital network and a national cleft registry have also expressed an interest in using the platform routinely. Important potential collaborators to target in future include the Cleft Registry and Audit Network, who oversee an audit of clinical outcomes across all UK cleft centres each year, and ICHOM.

This research is also already translating into other areas of clinical care. Based on this work, CAT systems for PROMs in breast reconstruction and massive weight loss are under development, and I have recently launched a study which uses a similar Concerto-based platform to monitor symptom variation in common hand conditions [200].

8.3 The future of clinical psychometrics

In the second half of this thesis, I explore some future research avenues for clinical psychometrics: how to interpret measurement error, the role of multidimensional modelling, and leveraging non-PROM data in latent variable measurement.

8.3.1 Measurement error

Generally, when PROMs are used in research and clinical practice, respondents are assigned a single score following the assessment, for example a score of 36 on the CLEFT-Q nose scale. This might simply be the sum of scored item responses, or a modern test theory point estimate of the latent construct level (based on the likelihood function or posterior distribution), or a transformation of the former onto the later (commonly presented as a look-up table). This is accepted practice in both clinical care, and in high-stakes research, for example in multi-million pound NIHR-funded clinical trials [10, 11]. It is also the approach taken by the developers of the CLEFT-Q, and in the development of the CLEFT-Q CAT. But this approach ignores measurement error (or assumes that at the population-level, measurement error is stochastic, normally distributed, and does not alter the mean measurement). In reality, PROM scores are

imperfect reflections of latent construct levels. The PROM scores themselves are not important (patients do not come to see their doctor complaining of poor PROM scores), it is the latent constructs which PROM scores imperfectly reflect that are important. All measurement instruments carry some degree of error, and this might be quite significant for some PROMs.

There are modern test theory strategies to quantify the measurement error of PROMs. For the CLEFT-Q CAT, this might be as simple as presenting a respondent's score as a distribution, with a confidence interval based on $SE(\hat{\theta})$. Plausible value imputation (see Chapter 5, section 5.4) can incorporate measurement error into the statistical analysis of population-level PROM data, again by treating latent construct estimates as plausible distributions. This can potentially alter the results and conclusions of clinical trials and observational studies [158–160].

These approaches are probably more helpful than treating PROM scores as point estimates, but ultimately they rely on cross-sectional estimates of a person's propensity distribution, and as I demonstrate in Chapter 5, SE is influenced considerably by model flexibility. It is not clear how these techniques translate across different psychometric models, nor how correctly propensity distributions can be estimated from cross-sectional data. Further work is needed to understand how to best define measurement error in different settings (for example, whether there is merit in conducting test-retest experiments as part of a trial's pilot). Work is also needed to understand how measurement error might influence PROM interpretation (for example in the statistical analysis of trial data or in clinical decision making).

8.3.2 Multidimensional measurement

Structural equation modelling and MIRT are not often used in surgical PROM research, but they are potentially very powerful tools. We should aim to make these techniques more familiar and accessible to those involved in the analysis of surgical PROM data. In this thesis, I used SEM to guide the development of a MIRT model for the CLEFT-Q appearance scales. But SEM can also provide valuable psychological insights from cross-sectional data. For example, the model described in Chapter 6 could be extended to understand how much variance in social function is explained by variance in appearance or speech distress. These insights have direct clinical relevance and potential implications for treatment planning. In this example, the PROM might not only help us to understand *how* respondents feel, but *why* they feel that way.

The use of MIRT in PROMs is relatively new, and it is not clear yet where this will be most useful. I have shown that there may be circumstances where MIRT can be used to derive latent construct measurements from PROM data, and that this strategy may lead to efficiency gains in CAT. Typically, items that cross load onto multiple factors are discarded from scales during PROM development, but under a MIRT framework this property may be desirable, as multiple constructs might be measured simultaneously.

Another use of MIRT might be in assessing changes in latent construct levels from longitudinal PROM data. In current practice, this is approximated by the change in PROM score (for example the difference between preoperative and postoperative PROM score), but that practice assumes that the relationship between item responses and the latent construct is constant over time, which might not be the case [201]. Imagine asking someone with chronic pain to grade their pain severity from mild to terrible. After a number of years, that person's internal frame of reference might change, and they may provide a different response despite an unchanging level of pain. In 2010, Li Cai proposed

a multidimensional two-tier approach to measuring latent change [167]. Under this model, a single health construct measured at *n* time points is treated as *n* separate factors. This allows the items to load differently onto the construct at each time point. There is likely to be a substantial increase in the availability of longitudinal data for clinical psychometrics research in the near future, as smartphone-based data capture platforms become more prevalent. Low-burden CAT assessments might also facilitate high-frequency sampling and ecological momentary assessment (where a respondent completes assessments *in the moment*, as part of their usual daily routine). Methodological research, including MIRT modelling, could help us to make the most of these data.

8.3.3 Incorporating non-PROM data into latent variable models

In Chapter 7, I developed CAT assessments that used both PROM data and clinician-reported data. These did not offer efficiency gains over the modern test theory approach, but I still believe that this idea has great potential. Outside of medicine, accurate latent variable measurement through non-questionnaire data has been demonstrated [202]. For example, digital footprint data (such as facebook likes) can be used to measure personality traits, and these models can sometimes predict someone's scores on a personality questionnaire more accurately than that person's spouse [203]. Non-PROM data can almost certainly contain information about latent health constructs and future research should focus on how we might leverage this.

I have suggested that alternative forms of recursive partitioning which can handle non-binary data splits might provide more granular measurement than that demonstrated in Chapter 7. One much simpler approach, that fits with Bayesian and modern test theory approaches to measurement, might be to adjust a respondent's prior distribution based on clinical variables. In the CLEFT-Q CAT, we assume a standard normal prior for

each participant in each scale, even though we know that respondents with complete bilateral clefts have poorer lip scale scores than respondents with incomplete unilateral clefts. A second approach that is in keeping with the modern test theory framework might be to calibrate non-PROM observations as if they were items in the PROM scale. For example, we might collect 24-hour step count along with physical function item responses, categorise the step count data, and treat these categories as if it they were response options to an additional item.

Combining PROM and non-PROM data into CAT assessments may also provide solutions to DIF (where the relationship between item responses and a latent construct differs by population subgroups, see Chapter 1, section 1.4.6). For example, in the measurement of adolescent depression, boys might report tearfulness at higher levels of depression than girls [204]. Traditionally, this type of item might be omitted from a questionnaire as subgroup membership (in this case gender) can bias the measurement. However, if subgroup membership is known *a priori*, the item can simply be administered with subgroup-specific item parameters.

The range of observable behaviours that could contribute to latent variable measurement has grown rapidly as smartphones and wearable sensors have become more sophisticated and more abundant. Candidate indicators include heart rate, oxygen saturation, kinetic data relating to mobility and sleep patterns, spoken language, geolocation (for example time spent at a place of work or time spent in proximity to a person with certain latent characteristics), device usage, internet browsing habits, social media data, socioeconomic data, personal communications, and EHR data. These data might one day be used for cross-sectional latent variable measurements, to monitor disease progression, to predict treatment outcomes, or to target public health interventions. But collecting these data clearly risks violating a person's privacy and the law and ethics surrounding psychometric targeting are underdeveloped. When personality traits measured with social media data

are accounted for in consumer marketing campaigns, these can result in up to 50% more sales than unpersonalised advertising [5]. Although ethically debatable, similar tactics are also used in electoral campaigns [7]. I propose that it may be permissible under utilitarian ethics to use digital phenotyping and modern psychometrics to counter vaccine misinformation, support smoking cessation, improve access to research studies for underserved groups, or influence behaviours relating to climate change and social discrimination. As the potential applications of psychometrics grow, so should legal and ethical frameworks that safeguard the wellbeing of patients, research participants, and consumers.

8.4 Conclusion

Psychometrics has come a long way since Francis Galton, but the diffusion of modern psychometric theory and techniques into surgical research and practice has been slow. Psychometric theory is a highly specialist subject and not readily accessible to many of the most important stakeholders, such as funding bodies, policy makers, trialists, and clinicians. There are few academics with interests in both theoretical psychometrics and surgical research and practice, and a lack of outreach between these disciplines. As a result, even the most fundamental measurement concepts, such as dimensionality, interval scaling, and measurement error are siloed away from those using PROMs in surgical research and practice.

Besides the research aims and objects I have described, this work has the deeper, longer-term, and more strategic goal of bridging the fields of applied surgical measurement and modern psychometric theory. The most important messages that this thesis can convey to stakeholders are not about CAT. They are that:

1. There is a need to standardise and implement high quality outcome measurement in clinical practice and research.

2. While many PROMs are labelled as 'valid', their fitness for a given purpose can be highly variable.

3. Measurement science, including the creation, selection, and interpretation of questionnaires, is complex and should involve specialist psychometricians.

The outputs of this work have tried to convey these messages, and this has a greater potential long-term impact than the CLEFT-Q CAT platform. From my perspective, the CLEFT-Q CAT is a Trojan horse. Its potential to reduce assessment burden is easy to understand and appealing to a range of stakeholders. But the real agenda here is facilitating a move towards consistent outcome measurement, while driving-up standards in content and construct validity, and providing a springboard for the uptake of these techniques in other branches of plastic surgery and beyond.

A

Appendix A: Chapter 2 results

Contents

A.1 Demographics and clinical characteristics 153
A.2 Missing item response data 166
A.3 Item selection criteria 178

A.1 Demographics and clinical characteristics

Tables A.1 to A.12 present demographic and clinical characteristics for respondents to each CLEFT-Q scale, in the CLEFT-Q field test.

Table A.1: Lip scale respondents. IQR: inter-quartile range; CL: cleft lip; CP: cleft palate; CLA: cleft lip and alveolus; CLAP: cleft lip, alveolus and palate.

Lip scale	
Sample size	
Total participants born with cleft lip	1866
Included participants (complete response sets)	1704
How much do you like how your lips look overall?	
Not at all	128
A little bit	397
Quite a bit	653
Very much	520
Missing	6
Age	
Median (IQR)	14 (7)
Missing	0
Gender	
Male	1034
Female	669
Other	0
Country	
Australia	10
Canada	348
England	205
Ireland	81
USA	272
The Netherlands	158
India	193
Sweden	65
Turkey	49
Columbia	175
Chile	75
Spain	73
Missing	0
Cleft type	
CL	245
CLA	187
CLAP	1272
Missing	0
Unilateral lip type	
Complete	947
Incomplete	253
Lesser form	16
Missing	35
Bilateral lip type	
Complete	354
Incomplete	47
Lesser form	0
Asymmetric	26
Missing	16
Planned primary lip repair	
Yes	12
No	1692
Missing	0
Planned lip revision	
Yes	1345
No	359
Missing	0
Palate type	
Veau I	35
Veau II	42
Veau III	815
Veau IV	343
Missing	37

Table A.2: Cleft lip scar scale respondents. IQR: inter-quartile range; CL: cleft lip; CP: cleft palate; CLA: cleft lip and alveolus; CLAP: cleft lip, alveolus and palate.

Cleft lip scar scale	
Sample size	
Total participants born with cleft lip	1866
Included participants (complete response sets)	1645
How much do you like how your cleft lip scar looks overall?	
Not at all	225
A little bit	367
Quite a bit	505
Very much	543
Missing	5
Age	
Median (IQR)	14 (7)
Missing	0
Gender	
Male	1002
Female	642
Other	0
Country	
Australia	10
Canada	341
England	193
Ireland	84
USA	260
The Netherlands	144
India	193
Sweden	68
Turkey	49
Columbia	162
Chile	73
Spain	68
Missing	0
Cleft type	
CL	239
CLA	178
CLAP	1228
Missing	0
Unilateral lip type	
Complete	915
Incomplete	239
Lesser form	13
Missing	35
Bilateral lip type	
Complete	347
Incomplete	43
Lesser form	1
Asymmetric	27
Missing	15
Planned lip revision	
Yes	350
No	1295
Palate type	
Veau I	33
Veau II	41
Veau III	784
Veau IV	333
Missing	37

Table A.3: Teeth scale respondents. IQR: inter-quartile range; CL: cleft lip; CP: cleft palate; CLA: cleft lip and alveolus; CLAP: cleft lip, alveolus and palate; ABG: alveolar bone graft.

Teeth scale	
Sample size	
Total participants	2434
Included participants (complete response sets)	2227
How much do you like how your teeth look overall?	
Not at all	295
A little bit	500
Quite a bit	748
Very much	681
Missing	3
Age	
Median (IQR)	14 (7)
Missing	0
Gender	
Male	1231
Female	995
Other	0
Country	
Australia	25
Canada	526
England	309
Ireland	96
USA	348
The Netherlands	194
India	231
Sweden	92
Turkey	54
Columbia	184
Chile	85
Spain	83
Missing	0
Cleft type	
CL	248
CP	514
CLA	191
CLAP	1274
Missing	0
Lip type	
Unilateral	1262
Bilateral	441
Missing	10
Planned future ABG	
Yes	223
No	2004
Missing	0
Planned future orthodontic treatment	
Yes	777
No	1450
Missing	0
Palate type	
Veau I	236
Veau II	266
Veau III	863
Veau IV	354
Missing	71

Table A.4: Jaw scale respondents. IQR: inter-quartile range; CL: cleft lip; CP: cleft palate; CLA: cleft lip and alveolus; CLAP: cleft lip, alveolus and palate.

Jaw scale	
Sample size	
Total participants	2434
Included participants (complete response sets)	1443
How much do you like how your jaw looks overall?	
Not at all	118
A little bit	229
Quite a bit	552
Very much	539
Missing	5
Age	
Median (IQR)	16 (5)
Missing	0
Gender	
Male	775
Female	667
Other	0
Country	
Australia	12
Canada	345
England	205
Ireland	79
USA	178
The Netherlands	138
India	106
Sweden	62
Turkey	49
Columbia	137
Chile	71
Spain	61
Missing	0
Cleft type	
CL	146
CP	301
CLA	122
CLAP	874
Missing	0
Lip type	
Unilateral	821
Bilateral	314
Missing	7
Planned future orthognathic surgery	
Yes	256
No	1187
Missing	0
Palate type	
Veau I	147
Veau II	149
Veau III	574
Veau IV	250
Missing	55

Table A.5: Nose scale respondents. IQR: inter-quartile range; CL: cleft lip; CP: cleft palate; CLA: cleft lip and alveolus; CLAP: cleft lip, alveolus and palate.

Nose scale	
Sample size	
Total participants	2434
Included participants (complete response sets)	2092
How much do you like how your nose looks overall?	
Not at all	220
A little bit	491
Quite a bit	684
Very much	690
Missing	7
Age	
Median (IQR)	14 (7)
Missing	0
Gender	
Male	1156
Female	935
Other	0
Country	
Australia	24
Canada	468
England	282
Ireland	96
USA	331
The Netherlands	185
India	229
Sweden	91
Turkey	52
Columbia	179
Chile	79
Spain	76
Missing	0
Cleft type	
CL	226
CP	455
CLA	181
CLAP	1230
Missing	0
Unilateral lip type	
Complete	907
Incomplete	245
Lesser form	14
Missing	34
Bilateral lip type	
Complete	342
Incomplete	43
Lesser form	1
Asymmetric	28
Missing	13
Planned future rhinoplasty	
Yes	643
No	1449
Missing	0
Palate type	
Veau I	211
Veau II	234
Veau III	832
Veau IV	343
Missing	65

Table A.6: Nostril scale respondents. IQR: inter-quartile range; CL: cleft lip; CP: cleft palate; CLA: cleft lip and alveolus; CLAP: cleft lip, alveolus and palate.

Nostrils scale	
Sample size	
Total participants	2434
Included participants (complete response sets)	2290
How much do you like how your nostrils look overall?	
Not at all	289
A little bit	590
Quite a bit	648
Very much	673
Missing	10
Age	
Median (IQR)	13.5 (7)
Missing	0
Gender	
Male	1228
Female	981
Other	0
Country	
Australia	25
Canada	492
England	313
Ireland	100
USA	347
The Netherlands	199
India	229
Sweden	88
Turkey	52
Columbia	192
Chile	86
Spain	87
Missing	0
Cleft type	
CL	245
CP	471
CLA	188
CLAP	1306
Missing	0
Unilateral lip type	
Complete	966
Incomplete	256
Lesser form	15
Missing	34
Bilateral lip type	
Complete	369
Incomplete	46
Lesser form	0
Asymmetric	27
Missing	15
Planned future rhinoplasty	
Yes	694
No	1516
Missing	0
Palate type	
Veau I	211
Veau II	234
Veau III	832
Veau IV	343
Missing	157

Table A.7: Face scale respondents. IQR: inter-quartile range; CL: cleft lip; CP: cleft palate; CLA: cleft lip and alveolus; CLAP: cleft lip, alveolus and palate.

Face scale

Sample size	
Total participants	2434
Included participants (complete response sets)	2301
How much do you like how your face looks overall?	
Not at all	87
A little bit	385
Quite a bit	918
Very much	907
Missing	4
Age	
Median (IQR)	14 (7)
Missing	1
Gender	
Male	1277
Female	1022
Other	0
Country	
Australia	23
Canada	592
England	309
Ireland	198
USA	350
The Netherlands	198
India	232
Sweden	93
Turkey	54
Columbia	183
Chile	84
Spain	87
Missing	0
Cleft type	
CL	252
CP	526
CLA	195
CLAP	1328
Missing	0
Unilateral lip type	
Complete	976
Incomplete	258
Lesser form	17
Missing	43
Bilateral lip type	
Complete	373
Incomplete	47
Lesser form	1
Asymmetric	29
Missing	20
Palate type	
Veau I	237
Veau II	273
Veau III	889
Veau IV	376
Missing	79

Table A.8: Psychological function scale respondents. IQR: inter-quartile range; CL: cleft lip; CP: cleft palate; CLA: cleft lip and alveolus; CLAP: cleft lip, alveolus and palate.

Psychological function scale	
Sample size	
Total participants	2434
Included participants (complete response sets)	2187
Age	
Median (IQR)	14 (7)
Missing	1
Gender	
Male	1217
Female	968
Other	0
Country	
Australia	23
Canada	476
England	312
Ireland	95
USA	354
The Netherlands	197
India	231
Sweden	93
Turkey	54
Columbia	180
Chile	84
Spain	88
Missing	0
Cleft type	
CL	244
CP	494
CLA	179
CLAP	1270
Missing	0
Lip type	
Unilateral	1235
Bilateral	448
Missing	10
Planned future psychological therapy	
Yes	268
No	1919
Palate type	
Veau I	233
Veau II	260
Veau III	851
Veau IV	358
Missing	62

Table A.9: Social function scale respondents. IQR: inter-quartile range; CL: cleft lip; CP: cleft palate; CLA: cleft lip and alveolus; CLAP: cleft lip, alveolus and palate.

Social function scale	
Sample size	
Total participants	2434
Included participants (complete response sets)	2154
Age	
Median (IQR)	14 (7)
Missing	0
Gender	
Male	1199
Female	954
Other	0
Country	
Australia	24
Canada	468
England	304
Ireland	93
USA	351
The Netherlands	195
India	232
Sweden	91
Turkey	52
Columbia	174
Chile	81
Spain	89
Missing	0
Cleft type	
CL	233
CP	493
CLA	178
CLAP	1250
Missing	0
Lip type	
Unilateral	1200
Bilateral	450
Missing	11
Palate type	
Veau I	231
Veau II	256
Veau III	833
Veau IV	359
Missing	64

Table A.10: School function scale respondents. IQR: inter-quartile range; CL: cleft lip; CP: cleft palate; CLA: cleft lip and alveolus; CLAP: cleft lip, alveolus and palate.

School function scale

Sample size	
Total participants	2434
Included participants (complete response sets)	1527
Age	
Median (IQR)	12 (5)
Missing	0
Gender	
Male	866
Female	661
Other	0
Country	
Australia	15
Canada	260
England	233
Ireland	57
USA	312
The Netherlands	129
India	176
Sweden	80
Turkey	36
Columbia	105
Chile	57
Spain	67
Missing	0
Cleft type	
CL	175
CP	374
CLA	138
CLAP	839
Missing	0
Lip type	
Unilateral	856
Bilateral	290
Missing	7
Palate type	
Veau I	180
Veau II	195
Veau III	584
Veau IV	227
Missing	27

Table A.11: Speech distress scale respondents. IQR: inter-quartile range; CL: cleft lip; CP: cleft palate; CLA: cleft lip and alveolus; CLAP: cleft lip, alveolus and palate.

Speech distress scale	
Sample size	
Total participants born with a cleft palate	1967
Included participants (complete response sets)	1691
Age	
Median (IQR)	14 (7)
Missing	0
Gender	
Male	940
Female	751
Other	0
Country	
Australia	20
Canada	370
England	244
Ireland	63
USA	294
The Netherlands	142
India	163
Sweden	65
Turkey	46
Columbia	145
Chile	72
Spain	667
Missing	0
Cleft type	
CP	482
CLAP	1209
Missing	0
Current clinician-reported speech problem	
No	720
Mild	673
Moderate	238
Severe	27
Missing	33
Future primary palate repair	
No	1686
Yes	5
Missing	0
Future speech surgery	
No	1574
Yes	117
Missing	0
Future fistula repair	
No	1623
Yes	68
Missing	0
Future speech therapy	
No	1356
Yes	335
Missing	0
Palate type	
Veau I	229
Veau II	246
Veau III	816
Veau IV	340
Missing	60

Table A.12: Speech function scale respondents. IQR: inter-quartile range; CL: cleft lip; CP: cleft palate; CLA: cleft lip and alveolus; CLAP: cleft lip, alveolus and palate.

Speech function scale	
Sample size	
Total participants born with a cleft palate	1967
Included participants (complete response sets)	1637
Age	
Median (IQR)	14 (7)
Missing	0
Gender	
Male	905
Female	732
Other	0
Country	
Australia	20
Canada	359
England	234
Ireland	65
USA	293
The Netherlands	137
India	161
Sweden	59
Turkey	49
Columbia	118
Chile	73
Spain	69
Missing	0
Cleft type	
CP	464
CLAP	1173
Missing	0
Current clinician-reported speech problem	
No	707
Mild	646
Moderate	224
Severe	29
Missing	31
Future primary palate repair	
No	1632
Yes	5
Missing	0
Future speech surgery	
No	1521
Yes	116
Missing	0
Future fistula repair	
No	1566
Yes	71
Missing	0
Future speech therapy	
No	1328
Yes	309
Missing	0
Palate type	
Veau I	221
Veau II	239
Veau III	781
Veau IV	338
Missing	58

A.2 Missing item response data

Tables A.13 to A.24 present missing item-level response data for CLEFT-Q field test respondents, in each CLEFT-Q scale.

Table A.13: Missing lip scale responses.

Lip scale	
Missing responses	2114
Complete responses	19792
Number of missing items	**Number of respondents**
0	2130
1	66
2	9
3	3
4	0
5	2
6	0
7	2
8	1
9	221
Missing item	**Number of missing responses**
1	234
2	232
3	237
4	238
5	225
6	234
7	238
8	239
9	237

Table A.14: **Missing cleft lip scar scale responses.**

Cleft lip scar scale	
Missing responses	5110
Complete responses	11928
Number of missing items	**Number of respondents**
0	1645
1	59
2	8
3	3
4	1
5	1
6	2
7	715
Missing item	**Number of missing responses**
1	731
2	725
3	731
4	730
5	732
6	733
7	728

Table A.15: **Missing cleft lip scar scale responses.**

Teeth scale	
Missing responses	1118
Complete responses	18354
Number of missing items	**Number of respondents**
0	2227
1	67
2	7
3	2
4	1
5	1
6	3
7	4
8	122
Missing item	**Number of missing responses**
1	133
2	142
3	137
4	149
5	140
6	137
7	139
8	141

Table A.16: Missing jaw scale responses.

Jaw scale	
Missing responses	6749
Complete responses	10289
Number of missing items	**Number of respondents**
0	1443
1	29
2	1
3	2
4	0
5	0
6	1
7	958
Missing item	**Number of missing responses**
1	964
2	969
3	959
4	967
5	963
6	960
7	967

Table A.17: Missing nose scale responses.

Nose scale	
Missing responses	1964
Complete responses	27244
Number of missing items	**Number of respondents**
0	2092
1	165
2	23
3	5
4	2
5	1
6	2
7	0
8	1
9	0
10	4
11	3
12	136
Missing item	**Number of missing responses**
1	156
2	163
3	162
4	170
5	168
6	173
7	169
8	158
9	153
10	162
11	160
12	170

Table A.18: Missing nostrils scale responses.

Nostrils scale	
Missing responses	1045
Complete responses	13559
Number of missing items	**Number of respondents**
0	2210
1	51
2	6
3	3
4	3
5	5
6	156
Missing item	**Number of missing responses**
1	168
2	173
3	180
4	178
5	166
6	180

Table A.19: Missing face scale responses.

Face scale	
Missing responses	445
Complete responses	21461
Number of missing items	**Number of respondents**
0	2301
1	78
2	12
3	3
4	2
5	2
6	2
7	0
8	2
9	32
Missing item	**Number of missing responses**
1	43
2	43
3	59
4	46
5	53
6	45
7	48
8	59
9	49

Table A.20: Missing psychological function scale responses.

Psychological function scale	
Missing responses	1912
Complete responses	22428
Number of missing items	**Number of respondents**
0	2187
1	52
2	5
3	3
4	3
5	1
6	0
7	0
8	2
9	2
10	179
Missing item	**Number of missing responses**
1	182
2	191
3	188
4	191
5	193
6	188
7	196
8	197
9	192
10	194

Table A.21: Missing social function scale responses.

Social function scale	
Missing responses	2013
Complete responses	22327
Number of missing items	**Number of respondents**
0	2154
1	62
2	17
3	6
4	0
5	1
6	1
7	0
8	20
9	2
10	171
Missing item	**Number of missing responses**
1	201
2	200
3	206
4	203
5	200
6	208
7	198
8	205
9	193
10	199

Table A.22: **Missing school function scale responses.**

School function scale	
Missing responses	7879
Complete responses	16461
Number of missing items	**Number of respondents**
0	1527
1	124
2	6
3	2
4	1
5	1
6	0
7	0
8	0
9	2
10	771
Missing item	**Number of missing responses**
1	862
2	775
3	778
4	782
5	779
6	780
7	776
8	780
9	787
10	780

Table A.23: Missing speech distress scale responses.

Speech distress scale	
Missing responses	5553
Complete responses	18787
Number of missing items	**Number of respondents**
0	1819
1	56
2	7
3	4
4	0
5	0
6	1
7	1
8	0
9	2
10	544
Missing item	**Number of missing responses**
1	554
2	555
3	556
4	554
5	554
6	555
7	554
8	554
9	559
10	558

Table A.24: Missing speech function scale responses.

Speech function scale	
Missing responses	6969
Complete responses	22239
Number of missing items	**Number of respondents**
0	1764
1	77
2	16
3	3
4	1
5	0
6	1
7	2
8	1
9	2
10	1
11	1
12	565
Missing item	**Number of missing responses**
1	579
2	580
3	576
4	582
5	579
6	589
7	584
8	581
9	583
10	581
11	580
12	575

A.3 Item selection criteria

Tables A.25 to A.36 display the results to the simulations comparing item selection criteria, for each CLEFT-Q scale.

Table A.25: Comparison of item selection criteria for the lip scale. SE: standard error of measurement; RMSE: root mean square error; MEPV: minimum expected posterior variance; MPWI: maximum posterior weighted information; MI: maximum information; MEI: maximum expected information; MLWI: maximum likelihood weighted information

Lip scale (9 items)

	SE < 0.65			SE < 0.55			SE < 0.45		
Criterion	**Mean items administered**	**Correlation**	**RMSE**	**Mean items administered**	**Correlation**	**RMSE**	**Mean items administered**	**Correlation**	**RMSE**
MEPV	7.57	0.96	0.70	9.00	0.97	0.65	9.00	0.97	0.65
MPWI	7.58	0.97	0.69	9.00	0.97	0.65	9.00	0.97	0.65
MI	7.57	0.97	0.69	9.00	0.97	0.65	9.00	0.97	0.65
MEI	7.64	0.97	0.69	9.00	0.97	0.65	9.00	0.97	0.65
MLWI	7.57	0.96	0.70	9.00	0.97	0.65	9.00	0.97	0.65

Table A.26: Comparison of item selection criteria for the cleft lip scar scale. SE: standard error of measurement; RMSE: root mean square error; MEPV: minimum expected posterior variance; MPWI: maximum posterior weighted information; MI: maximum information; MEI: maximum expected information; MLWI: maximum likelihood weighted information

Cleft lip scar scale (7 items)

	SE < 0.65			SE < 0.55			SE < 0.45		
Criterion	**Mean items administered**	**Correlation**	**RMSE**	**Mean items administered**	**Correlation**	**RMSE**	**Mean items administered**	**Correlation**	**RMSE**
MEPV	6.65	0.97	0.69	7.00	0.97	0.67	7.00	0.97	0.67
MPWI	6.65	0.97	0.68	7.00	0.97	0.67	7.00	0.97	0.67
MI	6.65	0.97	0.68	7.00	0.97	0.67	7.00	0.97	0.67
MEI	6.66	0.97	0.68	7.00	0.97	0.67	7.00	0.97	0.67
MLWI	6.65	0.97	0.69	7.00	0.97	0.67	7.00	0.97	0.67

Table A.27: Comparison of item selection criteria for the teeth scale. SE: standard error of measurement; RMSE: root mean square error; MEPV: minimum expected posterior variance; MPWI: maximum posterior weighted information; MI: maximum information; MEI: maximum expected information; MLWI: maximum likelihood weighted information

Teeth scale (8 items)

	SE < 0.65			SE < 0.55			SE < 0.45		
Criterion	**Mean items administered**	**Correlation**	**RMSE**	**Mean items administered**	**Correlation**	**RMSE**	**Mean items administered**	**Correlation**	**RMSE**
MEPV	5.90	0.95	0.64	7.18	0.96	0.59	8.00	0.96	0.57
MPWI	5.86	0.95	0.64	7.18	0.96	0.59	8.00	0.96	0.57
MI	5.87	0.95	0.64	7.19	0.96	0.59	8.00	0.96	0.57
MEI	6.10	0.95	0.63	7.23	0.96	0.60	8.00	0.96	0.57
MLWI	5.90	0.95	0.64	7.18	0.96	0.59	8.00	0.96	0.57

Table A.28: Comparison of item selection criteria for the jaw scale. SE: standard error of measurement; RMSE: root mean square error; MEPV: minimum expected posterior variance; MPWI: maximum posterior weighted information; MI: maximum information; MEI: maximum expected information; MLWI: maximum likelihood weighted information

Jaw scale (7 items)

	SE < 0.65			SE < 0.55			SE < 0.45		
Criterion	**Mean items administered**	**Correlation**	**RMSE**	**Mean items administered**	**Correlation**	**RMSE**	**Mean items administered**	**Correlation**	**RMSE**
MEPV	6.88	0.97	0.72	7.00	0.97	0.72	7.00	0.97	0.72
MPWI	6.87	0.97	0.72	7.00	0.97	0.72	7.00	0.97	0.72
MI	6.87	0.97	0.72	7.00	0.97	0.72	7.00	0.97	0.72
MEI	6.87	0.97	0.72	7.00	0.97	0.72	7.00	0.97	0.72
MLWI	6.88	0.97	0.72	7.00	0.97	0.72	7.00	0.97	0.72

Table A.29: Comparison of item selection criteria for the nose scale. SE: standard error of measurement; RMSE: root mean square error; MEPV: minimum expected posterior variance; MPWI: maximum posterior weighted information; MI: maximum information; MEI: maximum expected information; MLWI: maximum likelihood weighted information

Nose scale (12 items)

	SE < 0.65			SE < 0.55			SE < 0.45		
Criterion	**Mean items administered**	**Correlation**	**RMSE**	**Mean items administered**	**Correlation**	**RMSE**	**Mean items administered**	**Correlation**	**RMSE**
MEPV	7.29	0.96	0.63	8.96	0.97	0.57	11.50	0.97	0.53
MPWI	7.27	0.96	0.64	8.96	0.97	0.57	11.50	0.97	0.53
MI	7.26	0.96	0.64	8.96	0.97	0.57	11.50	0.97	0.53
MEI	7.44	0.96	0.64	9.12	0.97	0.57	11.57	0.97	0.53
MLWI	7.30	0.96	0.63	8.96	0.97	0.57	11.50	0.97	0.53

Table A.30: Comparison of item selection criteria for the nostrils scale. SE: standard error of measurement; RMSE: root mean square error; MEPV: minimum expected posterior variance; MPWI: maximum posterior weighted information; MI: maximum information; MEI: maximum expected information; MLWI: maximum likelihood weighted information

Nostrils scale (6 items)

	SE < 0.65			SE < 0.55			SE < 0.45		
Criterion	**Mean items administered**	**Correlation**	**RMSE**	**Mean items administered**	**Correlation**	**RMSE**	**Mean items administered**	**Correlation**	**RMSE**
MEPV	6.00	0.96	0.74	6.00	0.96	0.74	6.00	0.96	0.74
MPWI	6.00	0.96	0.74	6.00	0.96	0.74	6.00	0.96	0.74
MI	6.00	0.96	0.74	6.00	0.96	0.74	6.00	0.96	0.74
MEI	6.00	0.96	0.74	6.00	0.96	0.74	6.00	0.96	0.74
MLWI	6.00	0.96	0.74	6.00	0.96	0.74	6.00	0.96	0.74

Table A.31: Comparison of item selection criteria for the face scale. SE: standard error of measurement; RMSE: root mean square error; MEPV: minimum expected posterior variance; MPWI: maximum posterior weighted information; MI: maximum information; MEI: maximum expected information; MLWI: maximum likelihood weighted information

Face scale (9 items)

	SE < 0.65			SE < 0.55			SE < 0.45		
Criterion	**Mean items administered**	**Correlation**	**RMSE**	**Mean items administered**	**Correlation**	**RMSE**	**Mean items administered**	**Correlation**	**RMSE**
MEPV	6.44	0.93	0.66	8.02	0.94	0.61	9.00	0.94	0.60
MPWI	6.43	0.93	0.66	8.01	0.94	0.61	9.00	0.94	0.60
MI	6.44	0.93	0.66	8.02	0.94	0.61	9.00	0.94	0.60
MEI	6.68	0.93	0.66	8.13	0.94	0.61	9.00	0.94	0.60
MLWI	6.44	0.93	0.66	8.02	0.94	0.61	9.00	0.94	0.60

Table A.32: Comparison of item selection criteria for the psychological function scale. SE: standard error of measurement; RMSE: root mean square error; MEPV: minimum expected posterior variance; MPWI: maximum posterior weighted information; MI: maximum information; MEI: maximum expected information; MLWI: maximum likelihood weighted information

Psychological function scale (10 items)

	SE < 0.65			SE < 0.55			SE < 0.45		
Criterion	**Mean items administered**	**Correlation**	**RMSE**	**Mean items administered**	**Correlation**	**RMSE**	**Mean items administered**	**Correlation**	**RMSE**
MEPV	8.02	0.94	0.69	9.61	0.95	0.65	10.00	0.95	0.64
MPWI	8.03	0.95	0.69	9.61	0.95	0.65	10.00	0.95	0.64
MI	8.02	0.94	0.69	9.60	0.95	0.65	10.00	0.95	0.64
MEI	8.18	0.95	0.68	9.62	0.95	0.65	10.00	0.95	0.64
MLWI	8.02	0.94	0.69	9.61	0.95	0.65	10.00	0.95	0.64

Table A.33: Comparison of item selection criteria for the social function scale. SE: standard error of measurement; RMSE: root mean square error; MEPV: minimum expected posterior variance; MPWI: maximum posterior weighted information; MI: maximum information; MEI: maximum expected information; MLWI: maximum likelihood weighted information

Social function scale (10 items)

	SE < 0.65			SE < 0.55			SE < 0.45		
Criterion	**Mean items administered**	**Correlation**	**RMSE**	**Mean items administered**	**Correlation**	**RMSE**	**Mean items administered**	**Correlation**	**RMSE**
MEPV	6.89	0.91	0.65	8.26	0.92	0.60	9.87	0.93	0.58
MPWI	6.89	0.91	0.65	8.26	0.92	0.60	9.87	0.93	0.57
MI	6.91	0.91	0.64	8.26	0.92	0.59	9.86	0.93	0.58
MEI	7.51	0.91	0.65	8.76	0.92	0.60	9.96	0.93	0.57
MLWI	6.91	0.91	0.64	8.26	0.92	0.60	9.87	0.93	0.58

Table A.34: Comparison of item selection criteria for the school function scale. SE: standard error of measurement; RMSE: root mean square error; MEPV: minimum expected posterior variance; MPWI: maximum posterior weighted information; MI: maximum information; MEI: maximum expected information; MLWI: maximum likelihood weighted information

School function scale (10 items)

	SE < 0.65			SE < 0.55			SE < 0.45		
Criterion	**Mean items administered**	**Correlation**	**RMSE**	**Mean items administered**	**Correlation**	**RMSE**	**Mean items administered**	**Correlation**	**RMSE**
MEPV	7.44	0.91	0.68	8.59	0.92	0.65	9.97	0.93	0.62
MPWI	7.43	0.91	0.67	8.60	0.92	0.64	9.97	0.93	0.62
MI	7.45	0.91	0.68	8.59	0.92	0.65	9.96	0.93	0.62
MEI	7.95	0.91	0.69	8.83	0.92	0.65	9.98	0.93	0.62
MLWI	7.41	0.91	0.68	8.59	0.92	0.65	9.97	0.93	0.62

Table A.35: Comparison of item selection criteria for the speech distress scale. SE: standard error of measurement; RMSE: root mean square error; MEPV: minimum expected posterior variance; MPWI: maximum posterior weighted information; MI: maximum information; MEI: maximum expected information; MLWI: maximum likelihood weighted information

Speech distress scale (10 items)

	SE < 0.65			SE < 0.55			SE < 0.45		
Criterion	**Mean items administered**	**Correlation**	**RMSE**	**Mean items administered**	**Correlation**	**RMSE**	**Mean items administered**	**Correlation**	**RMSE**
MEPV	6.89	0.93	0.64	7.85	0.94	0.60	9.12	0.94	0.57
MPWI	6.86	0.92	0.64	7.87	0.94	0.60	9.12	0.94	0.57
MI	6.87	0.93	0.63	7.86	0.94	0.60	9.12	0.94	0.57
MEI	7.16	0.93	0.64	8.04	0.93	0.60	9.14	0.94	0.57
MLWI	6.86	0.93	0.64	7.86	0.94	0.60	9.12	0.94	0.57

Table A.36: Comparison of item selection criteria for the speech function scale. SE: standard error of measurement; RMSE: root mean square error; MEPV: minimum expected posterior variance; MPWI: maximum posterior weighted information; MI: maximum information; MEI: maximum expected information; MLWI: maximum likelihood weighted information

Speech function scale (12 items)

	SE < 0.65			SE < 0.55			SE < 0.45		
Criterion	**Mean items administered**	**Correlation**	**RMSE**	**Mean items administered**	**Correlation**	**RMSE**	**Mean items administered**	**Correlation**	**RMSE**
MEPV	7.97	0.95	0.66	9.10	0.95	0.62	10.69	0.96	0.59
MPWI	7.79	0.95	0.65	9.08	0.95	0.61	10.69	0.96	0.58
MI	7.82	0.95	0.66	9.10	0.95	0.61	10.69	0.96	0.58
MEI	7.97	0.95	0.65	9.26	0.95	0.61	10.77	0.96	0.59
MLWI	7.80	0.95	0.65	9.09	0.95	0.61	10.71	0.96	0.58

B

Appendix B: Chapter 3 results

Contents

B.1 Simulation results including outliers 186
B.2 Simulation results excluding outliers 191
B.3 Score frequency distribution and standard error of measurement 196
B.4 Bland-Altman plots 201
B.5 Stopping rule workshop vote results 205

B.1 Simulation results including outliers

Tables B.1 to B.9 present the results of the Monte Carlo simulation described in Chapter 3, with outliers included.

Table B.1: Face scale CAT algorithm performance at different fixed-length stopping rules Scores are scaled from 0 to 100. Outliers have been included and missing data removed through listwise exclusion. RMSE: root mean square error; lower LA: lower 95% limit of agreement; upper LA: upper 95% limit of agreement.

Face scale (9 items)

Items administered	Correlation	RMSE	Lower LA	Upper LA
9	1.000	0.00	0.00	0.00
8	0.997	1.67	-3.52	2.92
7	0.989	3.19	-6.80	5.46
6	0.983	4.01	-8.48	7.00
5	0.972	5.07	-10.41	9.35
4	0.964	5.71	-11.56	10.79
3	0.949	6.76	-13.49	13.04
2	0.912	8.80	-17.42	17.09
1	0.812	12.64	-23.70	25.74

Table B.2: Jaw scale CAT algorithm performance at different fixed-length stopping rules Scores are scaled from 0 to 100. Outliers have been included and missing data removed through listwise exclusion. RMSE: root mean square error; lower LA: lower 95% limit of agreement; upper LA: upper 95% limit of agreement.

Jaw scale (7 items)

Items administered	Correlation	RMSE	Lower LA	Upper LA
7	1.000	0.00	0.00	0.00
6	0.997	2.24	-4.09	4.64
5	0.992	3.68	-6.82	7.57
4	0.985	5.23	-9.70	10.72
3	0.980	6.28	-11.68	12.86
2	0.967	8.24	-15.04	17.07
1	0.915	12.73	-23.20	26.39

Table B.3: Teeth scale CAT algorithm performance at different fixed-length stopping rules Scores are scaled from 0 to 100. Outliers have been included and missing data removed through listwise exclusion. RMSE: root mean square error; lower LA: lower 95% limit of agreement; upper LA: upper 95% limit of agreement.

Teeth scale (8 items)

Items administered	Correlation	RMSE	Lower LA	Upper LA
8	1.000	0.00	0.00	0.00
7	0.995	2.14	-3.87	4.46
6	0.989	3.17	-6.33	6.12
5	0.982	4.13	-8.20	7.98
4	0.968	5.47	-10.74	10.74
3	0.942	7.37	-14.98	13.84
2	0.894	9.94	-20.92	17.52
1	0.823	12.57	-26.46	22.12

Table B.4: School function scale CAT algorithm performance at different fixed-length stopping rules Scores are scaled from 0 to 100. Outliers have been included and missing data removed through listwise exclusion. RMSE: root mean square error; lower LA: lower 95% limit of agreement; upper LA: upper 95% limit of agreement.

School function scale (10 items)

Items administered	Correlation	RMSE	Lower LA	Upper LA
10	1.000	0.00	0.00	0.00
9	0.996	2.00	-4.20	3.54
8	0.991	2.92	-6.07	5.28
7	0.987	3.53	-7.28	6.52
6	0.975	4.97	-10.26	9.12
5	0.959	6.32	-13.09	11.56
4	0.937	7.84	-16.53	13.77
3	0.910	9.23	-19.04	16.98
2	0.857	11.59	-24.37	20.54
1	0.766	14.52	-30.93	24.84

Table B.5: Psychological function scale CAT algorithm performance at different fixed-length stopping rules Scores are scaled from 0 to 100. Outliers have been included and missing data removed through listwise exclusion. RMSE: root mean square error; lower LA: lower 95% limit of agreement; upper LA: upper 95% limit of agreement.

Psychological function scale (10 items)

Items administered	Correlation	RMSE	Lower LA	Upper LA
10	1.000	0.00	0.00	0.00
9	0.997	1.98	-3.70	4.03
8	0.994	2.72	-5.27	5.41
7	0.989	3.75	-7.20	7.53
6	0.985	4.45	-8.23	9.15
5	0.977	5.43	-9.92	11.26
4	0.969	6.32	-11.64	13.04
3	0.948	8.18	-15.80	16.32
2	0.909	10.65	-21.24	20.58
1	0.809	14.89	-29.54	28.94

Table B.6: Speech distress scale CAT algorithm performance at different fixed-length stopping rules Scores are scaled from 0 to 100. Outliers have been included and missing data removed through listwise exclusion. RMSE: root mean square error; lower LA: lower 95% limit of agreement; upper LA: upper 95% limit of agreement.

Speech distress scale (10 items)

Items administered	Correlation	RMSE	Lower LA	Upper LA
10	1.000	0.00	0.00	0.00
9	0.995	2.15	-4.48	3.85
8	0.973	5.44	-11.82	8.58
7	0.947	7.61	-16.58	11.79
6	0.904	10.44	-22.82	15.75
5	0.882	11.49	-25.09	17.46
4	0.864	11.91	-25.63	19.61
3	0.811	13.78	-29.38	23.52
2	0.748	15.45	-32.40	27.55
1	0.621	18.05	-33.75	36.86

Table B.7: Speech function scale CAT algorithm performance at different fixed-length stopping rules Scores are scaled from 0 to 100. Outliers have been included and missing data removed through listwise exclusion. RMSE: root mean square error; lower LA: lower 95% limit of agreement; upper LA: upper 95% limit of agreement.

Speech function scale (12 items)

Items administered	Correlation	RMSE	Lower LA	Upper LA
12	1.000	0.00	0.00	0.00
11	0.998	1.79	-3.86	2.90
10	0.992	3.10	-6.59	5.31
9	0.987	4.14	-8.91	6.83
8	0.981	4.98	-10.75	8.12
7	0.974	5.86	-12.74	9.19
6	0.964	6.84	-14.79	11.00
5	0.951	8.02	-17.46	12.42
4	0.933	9.33	-20.34	14.37
3	0.910	10.66	-23.10	17.06
2	0.835	14.12	-30.59	22.59
1	0.746	16.46	-33.98	30.24

Table B.8: Social function scale CAT algorithm performance at different fixed-length stopping rules
Scores are scaled from 0 to 100. Outliers have been included and missing data removed through listwise exclusion. RMSE: root mean square error; lower LA: lower 95% limit of agreement; upper LA: upper 95% limit of agreement.

Social function scale (10 items)

Items administered	Correlation	RMSE	Lower LA	Upper LA
10	1.000	0.00	0.00	0.00
9	0.998	1.40	-2.88	2.57
8	0.995	2.16	-4.14	4.32
7	0.988	3.45	-7.19	6.22
6	0.984	4.08	-8.61	7.19
5	0.961	6.28	-13.55	10.29
4	0.952	6.81	-14.38	11.95
3	0.935	7.89	-15.92	15.03
2	0.887	10.23	-20.89	19.16
1	0.770	14.01	-27.28	27.75

Table B.9: Social function scale CAT algorithm performance at different fixed-length stopping rules
Scores are scaled from 0 to 100. Outliers have been included and missing data imputed through a proportional odds model. RMSE: root mean square error; lower LA: lower 95% limit of agreement; upper LA: upper 95% limit of agreement.

Social function scale with missing data imputed (10 items)

Items administered	Correlation	RMSE	Lower LA	Upper LA
10	1.000	0.00	0.00	0.00
9	0.998	1.37	-2.82	2.54
8	0.994	2.35	-4.84	4.31
7	0.988	3.34	-7.08	5.77
6	0.983	4.06	-8.56	7.11
5	0.962	6.04	-13.05	9.82
4	0.952	6.66	-14.19	11.34
3	0.936	7.60	-15.24	14.57
2	0.887	9.99	-20.35	18.73
1	0.765	13.76	-26.62	27.36

B.2 Simulation results excluding outliers

Tables B.10 to B.18 present the results of the Monte Carlo simulation described in Chapter 3, with outliers excluded.

Table B.10: Face scale CAT algorithm performance at different fixed-length stopping rules Scores are scaled from 0 to 100. Outliers have been excluded and missing data removed through listwise exclusion. RMSE: root mean square error; lower LA: lower 95% limit of agreement; upper LA: upper 95% limit of agreement.

Face scale (9 items)

Items administered	Correlation	RMSE	Lower LA	Upper LA
9	1.000	0.00	0.00	0.00
8	0.997	1.66	-3.50	2.92
7	0.989	3.18	-6.80	5.34
6	0.984	3.90	-8.28	6.73
5	0.974	4.89	-9.98	9.13
4	0.966	5.54	-11.17	10.53
3	0.954	6.59	-13.17	12.69
2	0.919	8.58	-17.03	16.64
1	0.828	12.36	-22.84	25.40

Table B.11: Jaw scale CAT algorithm performance at different fixed-length stopping rules Scores are scaled from 0 to 100. Outliers have been excluded and missing data removed through listwise exclusion. RMSE: root mean square error; lower LA: lower 95% limit of agreement; upper LA: upper 95% limit of agreement.

Jaw scale (7 items)

Items administered	Correlation	RMSE	Lower LA	Upper LA
7	1.000	0.00	0.00	0.00
6	0.998	2.06	-3.85	4.19
5	0.995	3.08	-5.64	6.38
4	0.988	4.76	-8.84	9.76
3	0.983	6.05	-11.02	12.55
2	0.972	7.98	-14.32	16.67
1	0.931	12.26	-22.04	25.60

Table B.12: Teeth scale CAT algorithm performance at different fixed-length stopping rules Scores are scaled from 0 to 100. Outliers have been excluded and missing data removed through listwise exclusion. RMSE: root mean square error; lower LA: lower 95% limit of agreement; upper LA: upper 95% limit of agreement.

Teeth scale (8 items)

Items administered	Correlation	RMSE	Lower LA	Upper LA
8	1.000	0.00	0.00	0.00
7	0.995	2.10	-3.83	4.36
6	0.990	3.10	-6.19	5.96
5	0.983	4.08	-8.11	7.88
4	0.971	5.31	-10.44	10.39
3	0.949	7.00	-14.30	13.05
2	0.907	9.38	-19.79	16.43
1	0.831	12.40	-26.13	21.74

Table B.13: School function scale CAT algorithm performance at different fixed-length stopping rules Scores are scaled from 0 to 100. Outliers have been excluded and missing data removed through listwise exclusion. RMSE: root mean square error; lower LA: lower 95% limit of agreement; upper LA: upper 95% limit of agreement.

School function scale (10 items)

Items administered	Correlation	RMSE	Lower LA	Upper LA
10	1.000	0.00	0.00	0.00
9	0.996	1.97	-4.12	3.50
8	0.991	2.89	-5.97	5.29
7	0.988	3.35	-6.82	6.31
6	0.977	4.67	-9.55	8.71
5	0.963	5.92	-12.14	11.01
4	0.944	7.30	-15.23	13.13
3	0.918	8.74	-17.86	16.36
2	0.863	11.23	-23.46	20.17
1	0.779	13.99	-29.50	24.62

Table B.14: Psychological function scale CAT algorithm performance at different fixed-length stopping rules Scores are scaled from 0 to 100. Outliers have been excluded and missing data removed through listwise exclusion. RMSE: root mean square error; lower LA: lower 95% limit of agreement; upper LA: upper 95% limit of agreement.

Psychological function scale (10 items)

Items administered	Correlation	RMSE	Lower LA	Upper LA
10	1.000	0.00	0.00	0.00
9	0.997	1.94	-3.67	3.94
8	0.994	2.68	-5.12	5.38
7	0.989	3.64	-6.92	7.35
6	0.986	4.29	-7.86	8.86
5	0.979	5.21	-9.44	10.85
4	0.973	6.00	-10.82	12.51
3	0.952	7.83	-14.73	15.93
2	0.913	10.35	-20.42	20.22
1	0.812	14.68	-28.87	28.80

Table B.15: Speech distress scale CAT algorithm performance at different fixed-length stopping rules Scores are scaled from 0 to 100. Outliers have been excluded and missing data removed through listwise exclusion. RMSE: root mean square error; lower LA: lower 95% limit of agreement; upper LA: upper 95% limit of agreement.

Speech distress scale (10 items)

Items administered	Correlation	RMSE	Lower LA	Upper LA
10	1.000	0.00	0.00	0.00
9	0.990	2.57	-5.42	4.50
8	0.978	3.93	-7.25	8.09
7	0.960	5.72	-9.74	12.21
6	0.942	7.55	-10.87	16.59
5	0.921	9.89	-10.83	21.66
4	0.887	13.22	-12.14	28.57
3	0.816	18.23	-13.42	38.52
2	0.722	26.10	-10.52	51.92
1	0.610	34.17	0.37	60.68

Table B.16: Speech function scale CAT algorithm performance at different fixed-length stopping rules Scores are scaled from 0 to 100. Outliers have been excluded and missing data removed through listwise exclusion. RMSE: root mean square error; lower LA: lower 95% limit of agreement; upper LA: upper 95% limit of agreement.

Speech function scale (12 items)

Items administered	Correlation	RMSE	Lower LA	Upper LA
12	1.000	0.00	0.00	0.00
11	0.997	1.74	-2.96	3.71
10	0.990	3.06	-5.43	6.44
9	0.986	3.82	-5.77	8.36
8	0.978	5.08	-7.00	11.19
7	0.966	6.66	-7.32	14.60
6	0.952	8.32	-7.48	17.94
5	0.941	9.69	-8.00	20.72
4	0.922	11.20	-8.11	23.61
3	0.897	13.70	-8.28	28.36
2	0.842	18.07	-8.38	36.40
1	0.770	24.24	-4.71	45.80

Table B.17: Social function scale CAT algorithm performance at different fixed-length stopping rules Scores are scaled from 0 to 100. Outliers have been excluded and missing data removed through listwise exclusion. RMSE: root mean square error; lower LA: lower 95% limit of agreement; upper LA: upper 95% limit of agreement.

Social function scale (10 items)

Items administered	Correlation	RMSE	Lower LA	Upper LA
10	1.000	0.00	0.00	0.00
9	0.998	1.38	-2.84	2.56
8	0.995	2.13	-4.04	4.29
7	0.988	3.40	-7.05	6.19
6	0.984	4.01	-8.41	7.13
5	0.961	6.29	-13.57	10.33
4	0.952	6.84	-14.41	12.09
3	0.936	7.84	-15.78	15.00
2	0.891	10.06	-20.48	18.94
1	0.793	13.44	-25.77	27.00

Table B.18: Social function scale CAT algorithm performance at different fixed-length stopping rules Scores are scaled from 0 to 100. Outliers have been excluded and missing data imputed through a proportional odds model. RMSE: root mean square error; lower LA: lower 95% limit of agreement; upper LA: upper 95% limit of agreement.

Social function scale with missing data imputed (10 items)

Items administered	Correlation	RMSE	Lower LA	Upper LA
10	1.000	0.00	0.00	0.00
9	0.998	1.37	-2.82	2.53
8	0.994	2.38	-4.91	4.35
7	0.988	3.32	-7.02	5.78
6	0.984	3.96	-8.29	7.06
5	0.961	6.16	-13.27	10.10
4	0.948	6.97	-14.80	11.98
3	0.930	7.94	-16.02	15.09
2	0.881	10.20	-20.82	19.05
1	0.788	13.27	-25.13	26.86

B.3 Score frequency distribution and standard error of measurement

Figures B.1 to B.8 demonstrate the relative frequency distribution of respondents in the validation dataset, based on their respective scale scores. Standard error of measurement is superimposed to illustrate how measurement precision varies across latent construct levels.

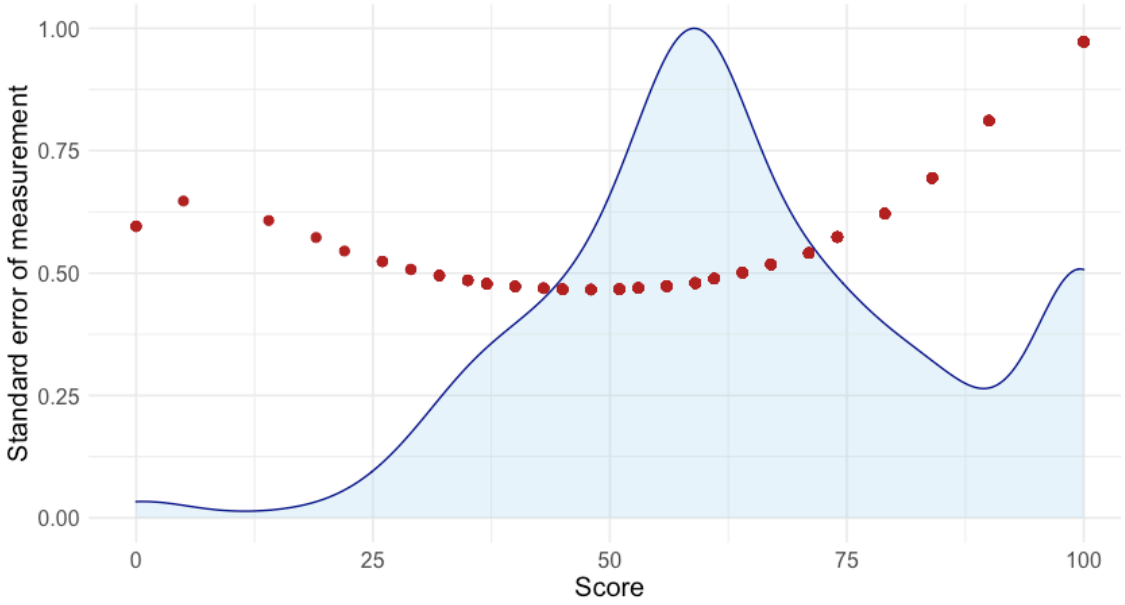


Figure B.1: Score frequency distribution and standard error of measurement for the face scale validation data. This kernel density plot demonstrates the relative frequency distribution of respondents based on their full-length assessment score, which has been scaled to range from 0 to 100. The red points demonstrate the standard error of measurement (SE) associated with scores across the range of latent construct levels.

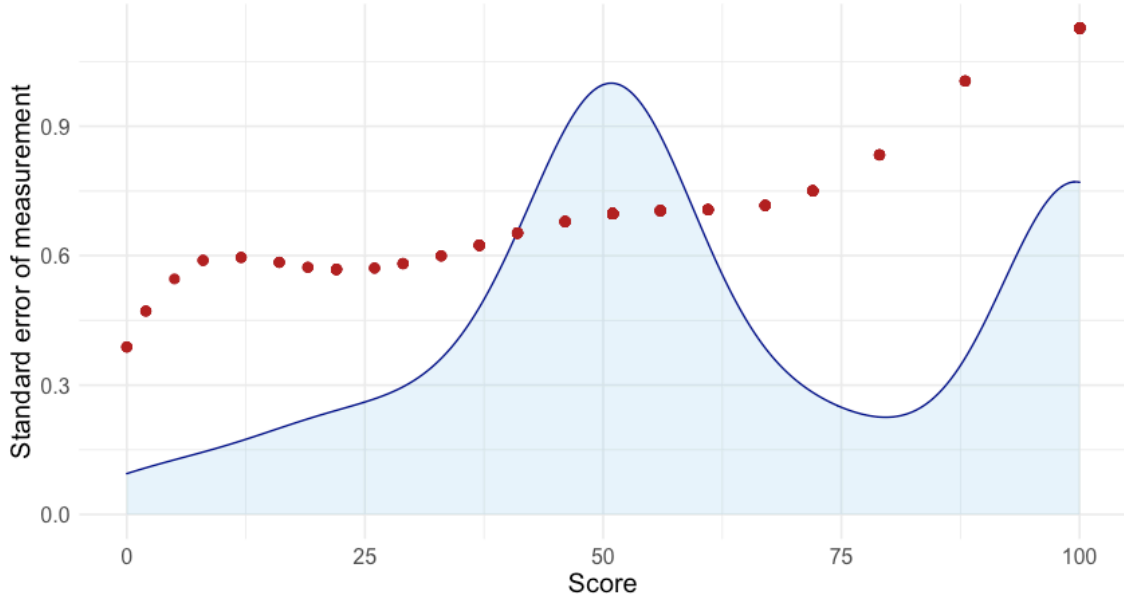


Figure B.2: Score frequency distribution and standard error of measurement for the jaw scale validation data. This kernel density plot demonstrates the relative frequency distribution of respondents based on their full-length assessment score, which has been scaled to range from 0 to 100. The red points demonstrate the standard error of measurement (SE) associated with scores across the range of latent construct levels.

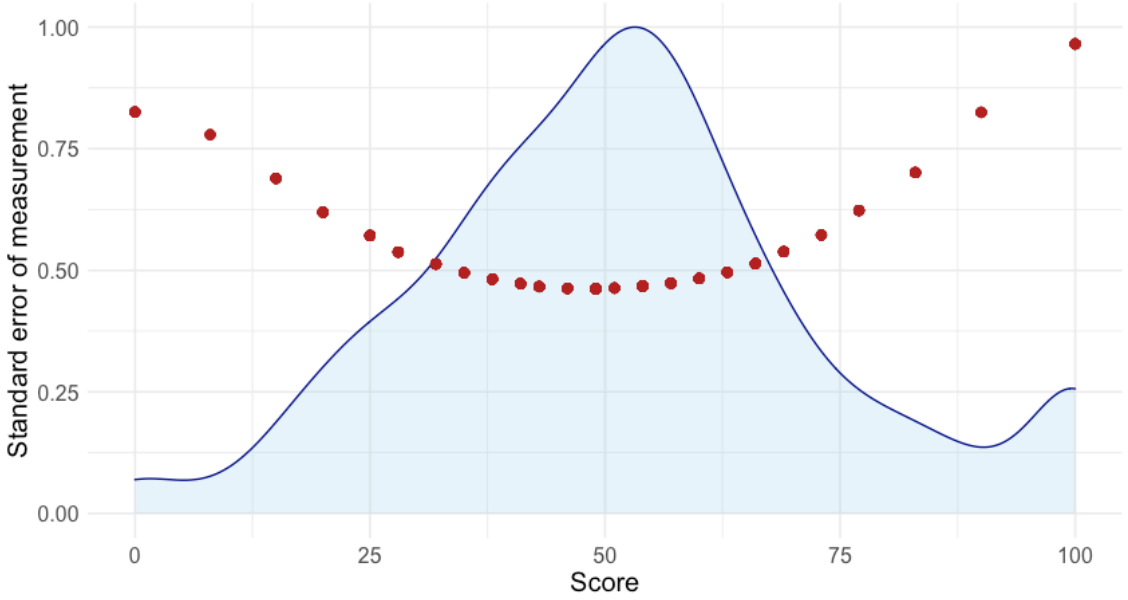


Figure B.3: Score frequency distribution and standard error of measurement for the teeth scale validation data. This kernel density plot demonstrates the relative frequency distribution of respondents based on their full-length assessment score, which has been scaled to range from 0 to 100. The red points demonstrate the standard error of measurement (SE) associated with scores across the range of latent construct levels.

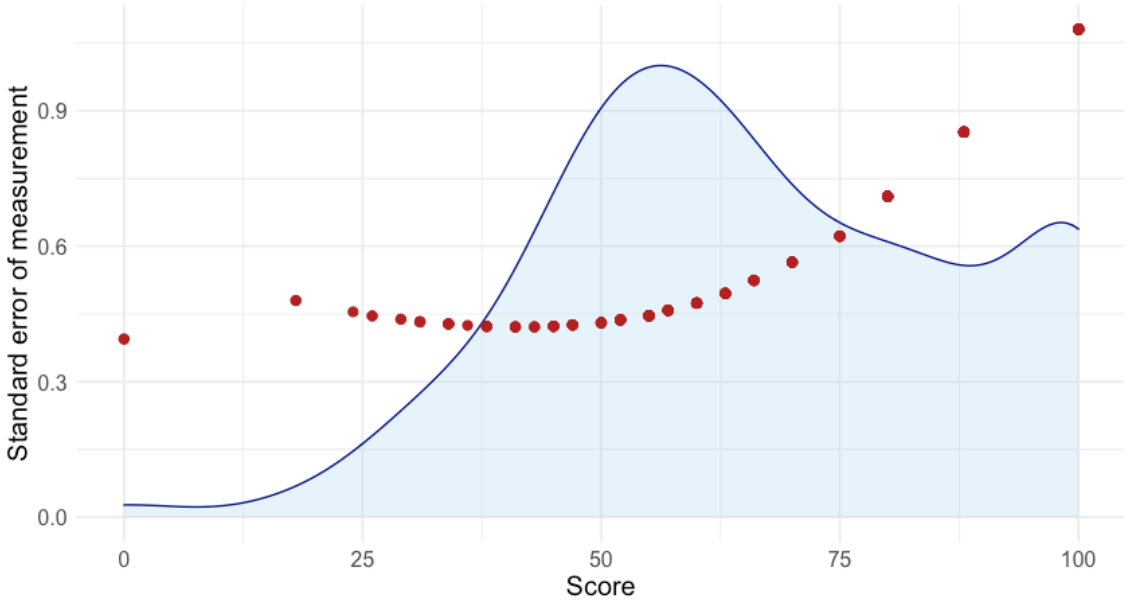


Figure B.4: Score frequency distribution and standard error of measurement for the school function scale validation data. This kernel density plot demonstrates the relative frequency distribution of respondents based on their full-length assessment score, which has been scaled to range from 0 to 100. The red points demonstrate the standard error of measurement (SE) associated with scores across the range of latent construct levels.

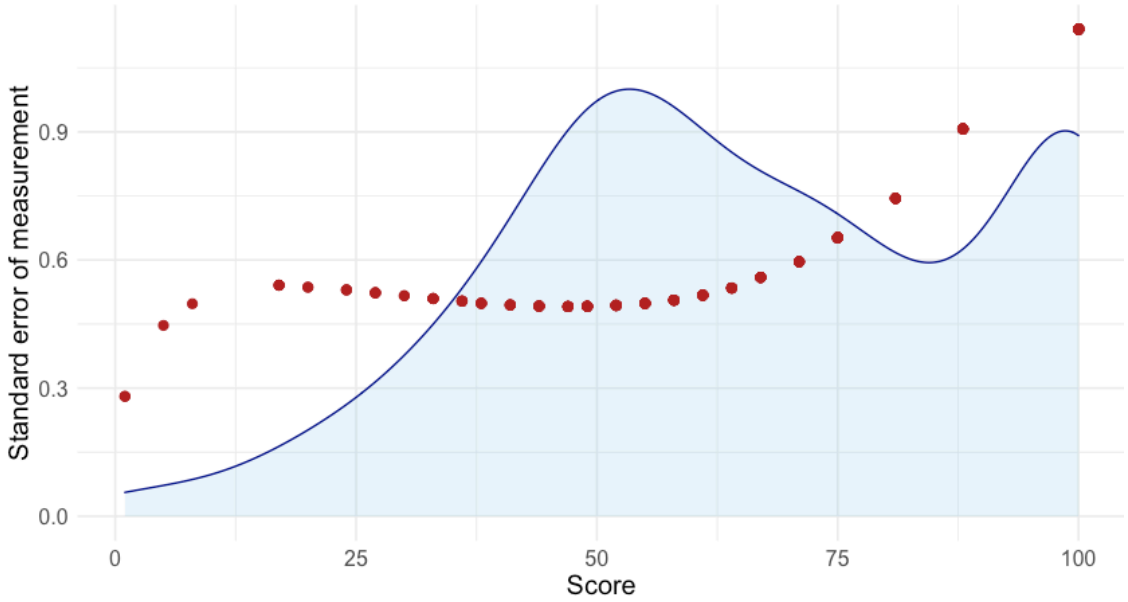


Figure B.5: Score frequency distribution and standard error of measurement for the psychological function scale validation data. This kernel density plot demonstrates the relative frequency distribution of respondents based on their full-length assessment score, which has been scaled to range from 0 to 100. The red points demonstrate the standard error of measurement (SE) associated with scores across the range of latent construct levels.

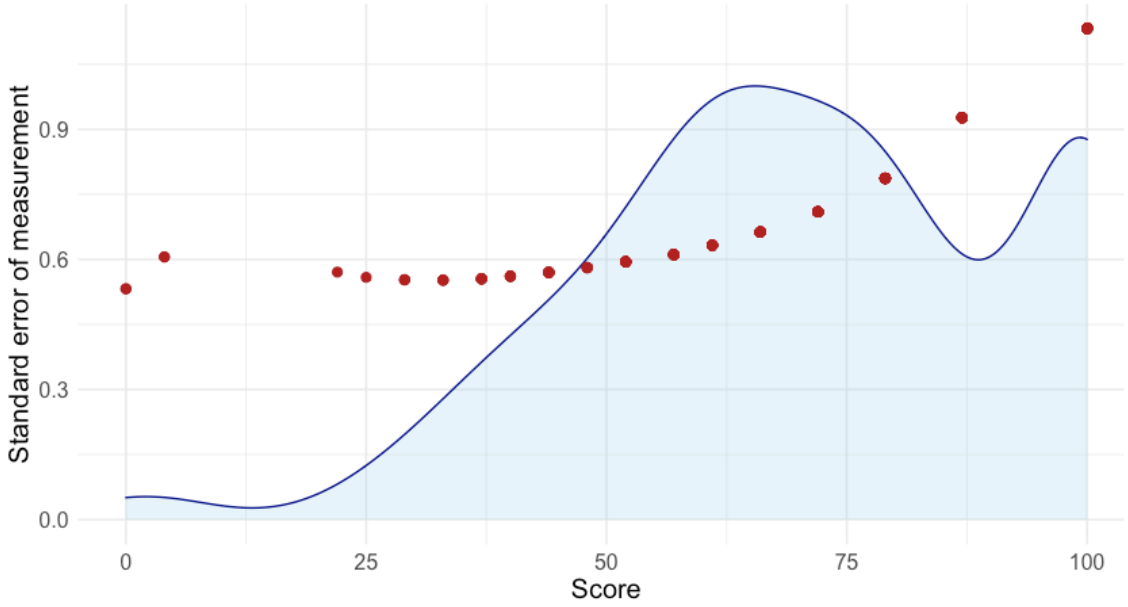


Figure B.6: Score frequency distribution and standard error of measurement for the speech distress scale validation data. This kernel density plot demonstrates the relative frequency distribution of respondents based on their full-length assessment score, which has been scaled to range from 0 to 100. The red points demonstrate the standard error of measurement (SE) associated with scores across the range of latent construct levels.

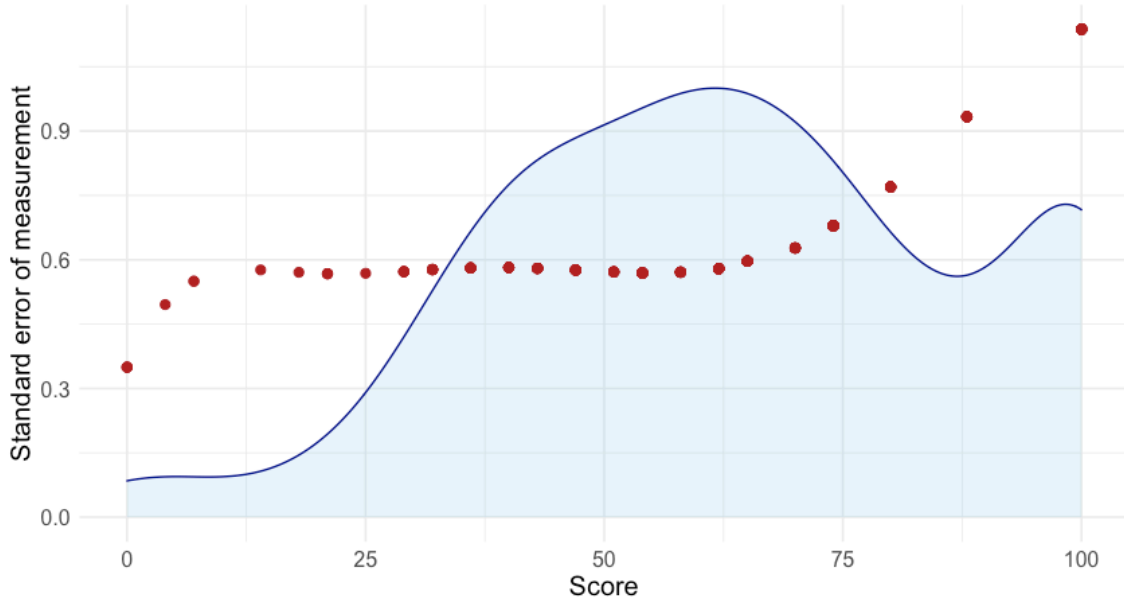


Figure B.7: Score frequency distribution and standard error of measurement for the speech function scale validation data. This kernel density plot demonstrates the relative frequency distribution of respondents based on their full-length assessment score, which has been scaled to range from 0 to 100. The red points demonstrate the standard error of measurement (SE) associated with scores across the range of latent construct levels.

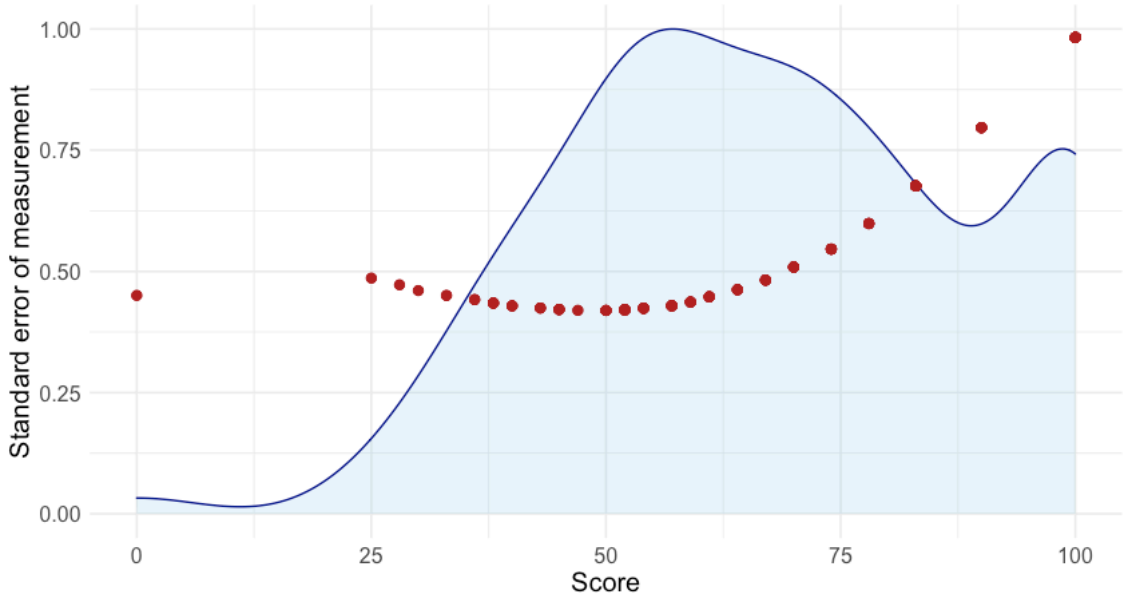


Figure B.8: Score frequency distribution and standard error of measurement for the social function scale validation data. This kernel density plot demonstrates the relative frequency distribution of respondents based on their full-length assessment score, which has been scaled to range from 0 to 100. The red points demonstrate the standard error of measurement (SE) associated with scores across the range of latent construct levels.

B.4 Bland-Altman plots

Figures B.9 to B.16 demonstrate the agreement between CAT score and linear assessment score, for each CLEFT-Q scale's CAT, at the stopping rules chosen in the consensus workshop.

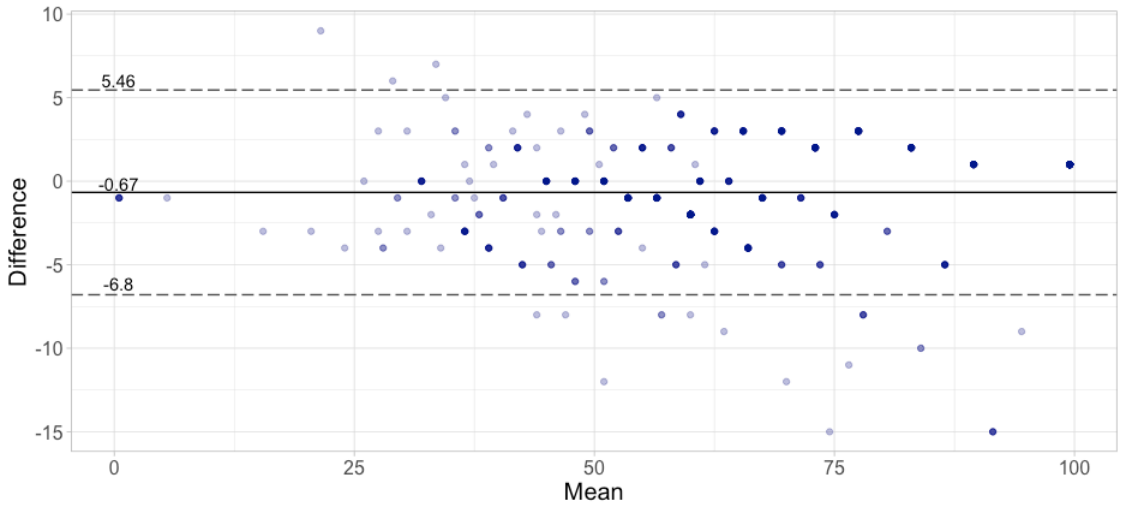


Figure B.9: Bland-Altman plot for the face scale CAT at 7 items. The x axis represents the mean of each respondent's CAT and linear assessment scores. The y axis shows the difference between CAT and linear assessment scores for each respondent (the CAT score has been subtracted from the linear assessment score). Points that overlie each other appear darker. The solid horizontal line is the mean absolute difference between CAT and linear assessment scores. The dashed horizontal lines represent the 95% limits of agreement.

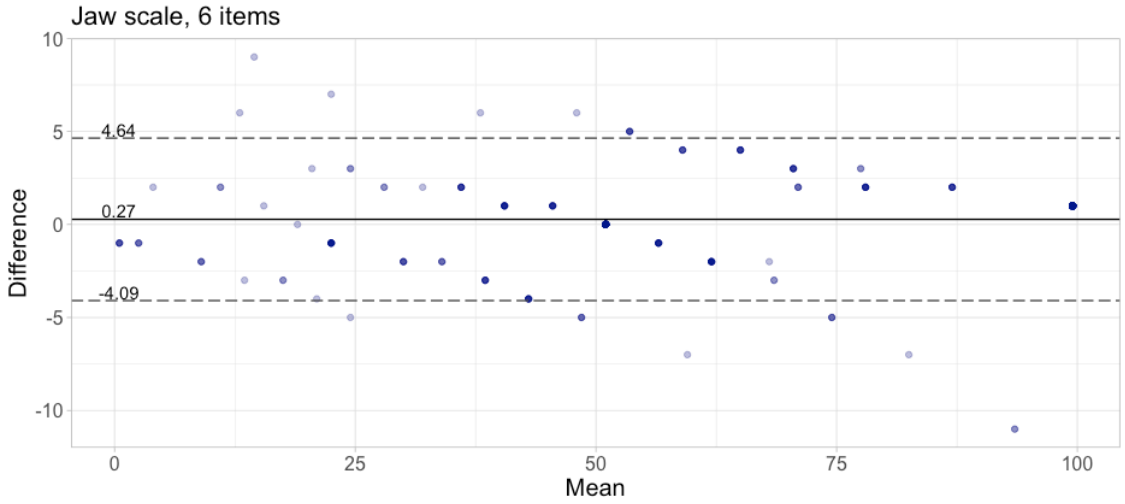


Figure B.10: Bland-Altman plot for the jaw scale CAT at 6 items. The x axis represents the mean of each respondent's CAT and linear assessment scores. The y axis shows the difference between CAT and linear assessment scores for each respondent (the CAT score has been subtracted from the linear assessment score). Points that overlie each other appear darker. The solid horizontal line is the mean absolute difference between CAT and linear assessment scores. The dashed horizontal lines represent the 95% limits of agreement.

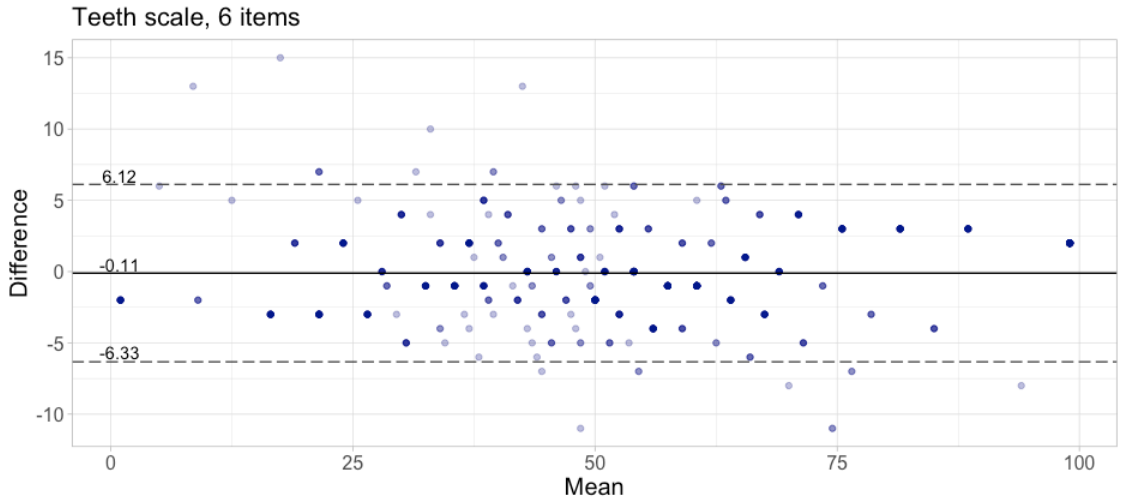


Figure B.11: Bland-Altman plot for the teeth scale CAT at 6 items. The x axis represents the mean of each respondent's CAT and linear assessment scores. The y axis shows the difference between CAT and linear assessment scores for each respondent (the CAT score has been subtracted from the linear assessment score). Points that overlie each other appear darker. The solid horizontal line is the mean absolute difference between CAT and linear assessment scores. The dashed horizontal lines represent the 95% limits of agreement.

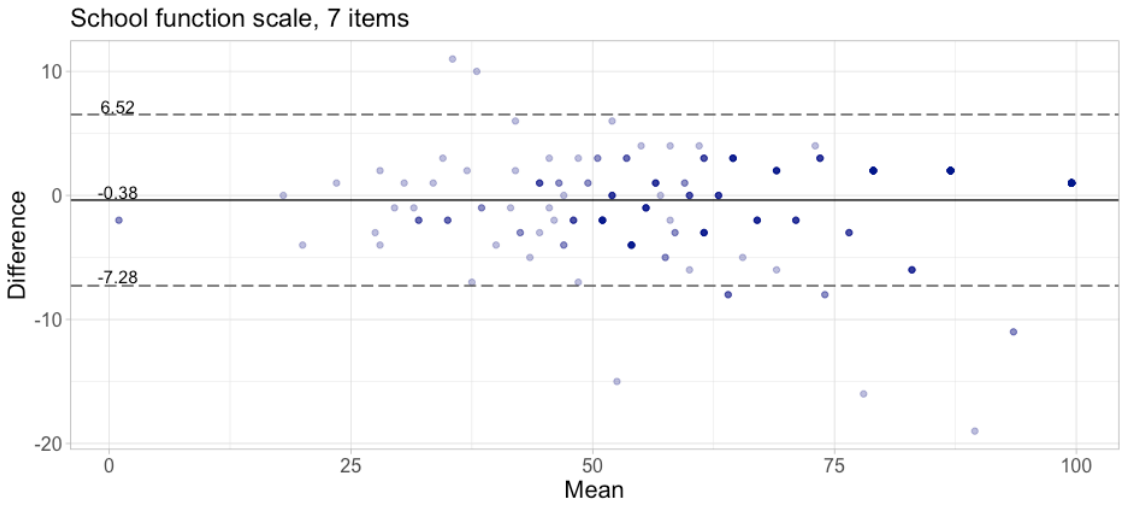


Figure B.12: Bland-Altman plot for the school function CAT at 7 items. The x axis represents the mean of each respondent's CAT and linear assessment scores. The y axis shows the difference between CAT and linear assessment scores for each respondent (the CAT score has been subtracted from the linear assessment score). Points that overlie each other appear darker. The solid horizontal line is the mean absolute difference between CAT and linear assessment scores. The dashed horizontal lines represent the 95% limits of agreement.

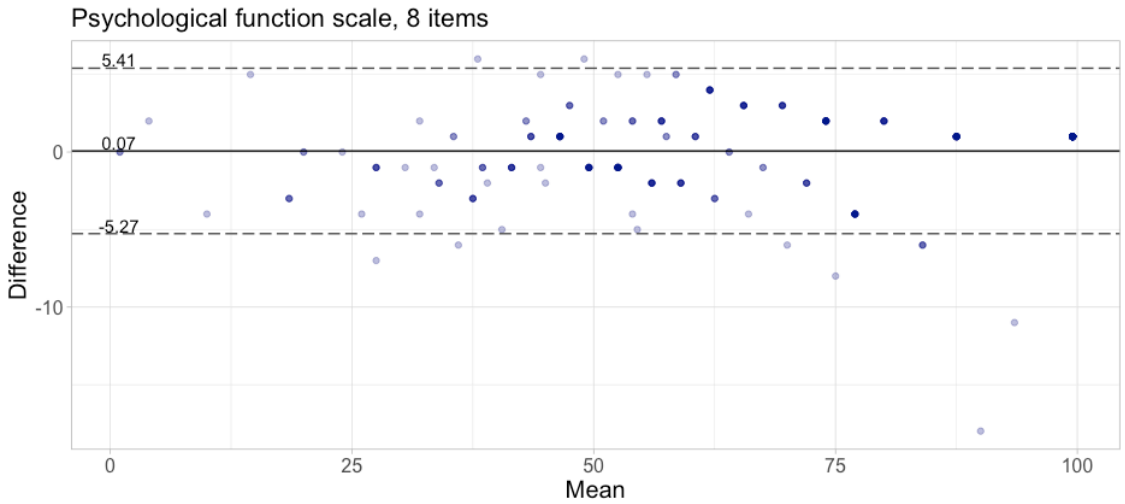


Figure B.13: Bland-Altman plot for the psychological function CAT at 8 items. The x axis represents the mean of each respondent's CAT and linear assessment scores. The y axis shows the difference between CAT and linear assessment scores for each respondent (the CAT score has been subtracted from the linear assessment score). Points that overlie each other appear darker. The solid horizontal line is the mean absolute difference between CAT and linear assessment scores. The dashed horizontal lines represent the 95% limits of agreement.

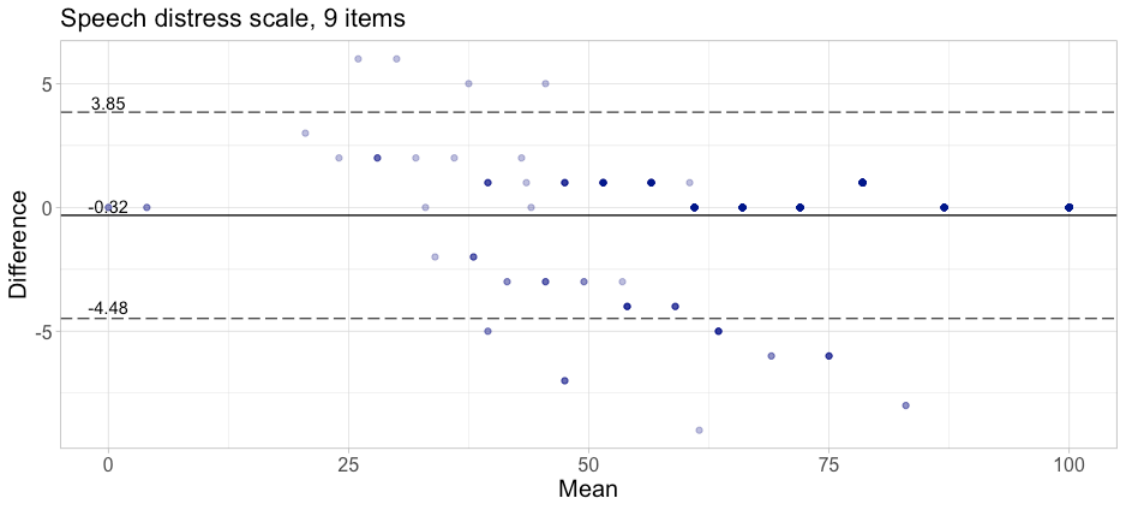


Figure B.14: Bland-Altman plot for the speech distress CAT at 9 items. The x axis represents the mean of each respondent's CAT and linear assessment scores. The y axis shows the difference between CAT and linear assessment scores for each respondent (the CAT score has been subtracted from the linear assessment score). Points that overlie each other appear darker. The solid horizontal line is the mean absolute difference between CAT and linear assessment scores. The dashed horizontal lines represent the 95% limits of agreement.

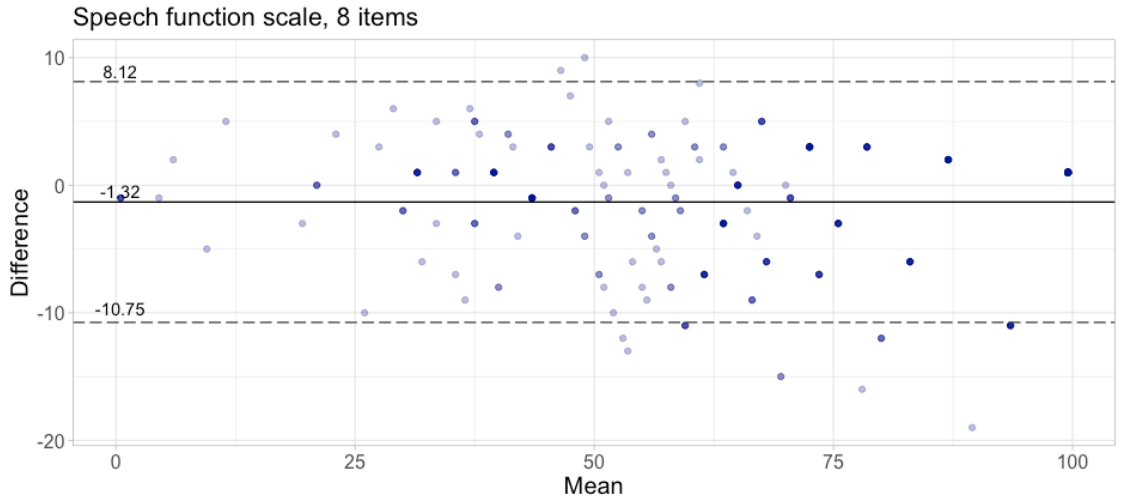


Figure B.15: Bland-Altman plot for the speech function CAT at 8 items. The x axis represents the mean of each respondent's CAT and linear assessment scores. The y axis shows the difference between CAT and linear assessment scores for each respondent (the CAT score has been subtracted from the linear assessment score). Points that overlie each other appear darker. The solid horizontal line is the mean absolute difference between CAT and linear assessment scores. The dashed horizontal lines represent the 95% limits of agreement.

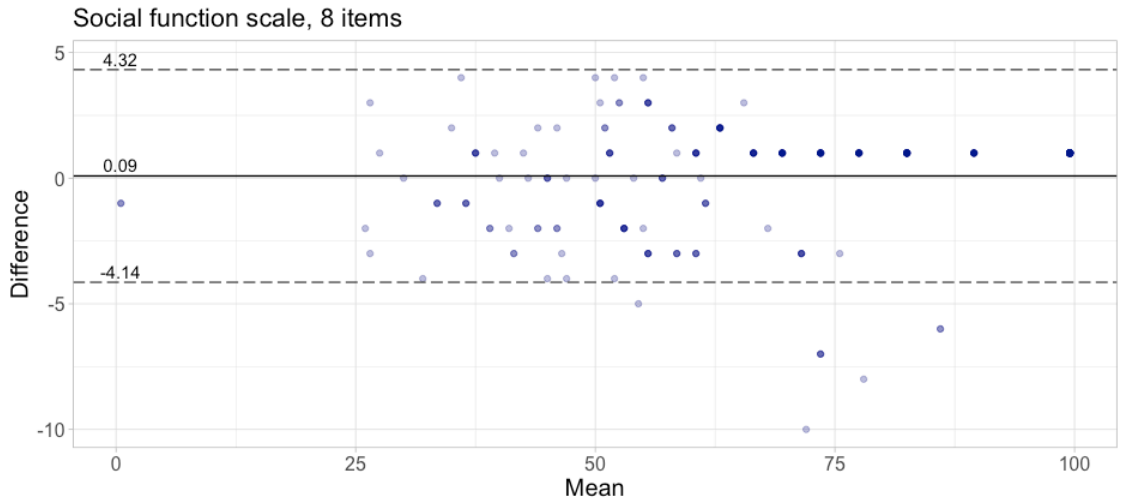


Figure B.16: Bland-Altman plot for the social function CAT at 8 items. The x axis represents the mean of each respondent's CAT and linear assessment scores. The y axis shows the difference between CAT and linear assessment scores for each respondent (the CAT score has been subtracted from the linear assessment score). Points that overlie each other appear darker. The solid horizontal line is the mean absolute difference between CAT and linear assessment scores. The dashed horizontal lines represent the 95% limits of agreement.

B.5 Stopping rule workshop vote results

Table B.19 presents the number of votes received for each stopping rule at the consensus workshop.

Table B.19: Number of votes cast per stopping rule.

Scale	Number of items	Number of votes
Face (9 items total)	3	1
	5	1
	6	1
	7	10
	9	1
Jaw (7 items total)	4	2
	5	6
	6	9
	7	3
Teeth (8 items total)	5	1
	6	15
	7	4
School function (10 items total)	3	1
	5	1
	7	12
	8	5
	9	1
Psychological function (10 items total)	4	1
	5	3
	6	4
	8	11
	9	1
Speech distress (10 items total)	6	1
	7	1
	8	6
	9	11
	10	1
Speech function (12 items total)	5	1
	6	1
	8	12
	9	1
	10	4
	11	1
Social function (10 items total)	6	2
	7	4
	8	13
	9	1

C

Appendix C: Chapter 4 interview schedules

Contents

C.1 Interview schedules 208

C.1 Interview schedules

Tables C.1 to C.4 present the interview schedules for patient and clinician interviews.

Table C.1: Interview schedule for patient interviews (part 1).

Patient interviews (part 1)

Focus area	Opening question and examples of additional probes
Introduction	Tell me about your last visit to see the cleft team.
Is the CLEFT-Q CAT a worthwhile adjunct to clinical practice?	What did you think of the CLEFT-Q CAT questionnaire? Do you think it changed anything about your conversation with the cleft team? What did it change? Did it change anything else? Did you like completing it? Do you think it's a good idea to ask other people to complete the CLEFT-Q CAT at their appointments, just like you did? Why do you think that?
How burdensome is the CLEFT-Q CAT?	How difficult was it to complete the CLEFT-Q CAT? What was difficult about it? Did it take a long time? Was it boring? What was boring about it? Did it make you tired? If you had the choice, would you rather do the CLEFT-Q CAT on an iPad (just like you did) or would you rather have a pen-and-paper version of the questionnaire, with slightly more questions? Why?

Table C.2: Interview schedule for patient interviews (part 2).

Patient interviews (part 2)

Focus area	Opening question and examples of additional probes
Facilitators and barriers to CLEFT-Q CAT implementation	Can you think of anything that might make you less likely to use the CLEFT-Q CAT? Can you think of anything that might make you more likely to use the CLEFT-Q CAT?
Areas for CLEFT-Q CAT improvement	If you could change anything about the CLEFT-Q CAT, what would you change? Why? Is there anything you really liked about the CLEFT-Q CAT?

Table C.3: Interview schedule for clinician interviews (part 1).

Clinician interviews (part 1)

Focus area	Opening question and examples of additional probes
Introduction	Tell me about your role in the cleft team.
	Have you used the CLEFT-Q CAT a lot?
Is the CLEFT-Q CAT a worthwhile adjunct to clinical practice?	What do you think of the CLEFT-Q CAT questionnaire?
	Do you think it changed any aspect of your clinical care, or that of your colleagues?
	What has it change?
	Has it change anything else?
	Do you think patients like completing it?
	How useful is it as an adjunct to clinical care?
	Do you think other cleft teams should be using it?
	Why do you think that?
How burdensome is the CLEFT-Q CAT?	How burdensome is the CLEFT-Q CAT, from your perspective?
	Has it changed your workload, or that of your colleagues?
	In what way?
	Does it make clinics faster or slower?
	Have patients given you feedback about the burden of completing it?
	Why?

Table C.4: Interview schedule for patient interviews (part 2).

Clinician interviews (part 2)

Focus area	Opening question and examples of additional probes
Facilitators and barriers to CLEFT-Q CAT implementation	Can you think of any barriers to cleft teams implementing the CLEFT-Q CAT? Can you think of anything that made it easier or more difficult to implement? What advice would you give other cleft teams that are thinking about using the CLEFT-Q CAT?
Areas for CLEFT-Q CAT improvement	If you could change anything about the CLEFT-Q CAT, what would you change? Why? Is there anything you really liked about the CLEFT-Q CAT?

D

Appendix D: Chapter 5 results

Contents

D.1 A comparison of Rasch and GRM scores 214
D.2 A comparison of CAT assessments built with Rasch models and GRMs 219

D.1 A comparison of Rasch and GRM scores

Figures D.1 to D.8 demonstrate the relationship between full-length CLEFT-Q scale scores, calculated via partial credit Rasch models and GRMs.

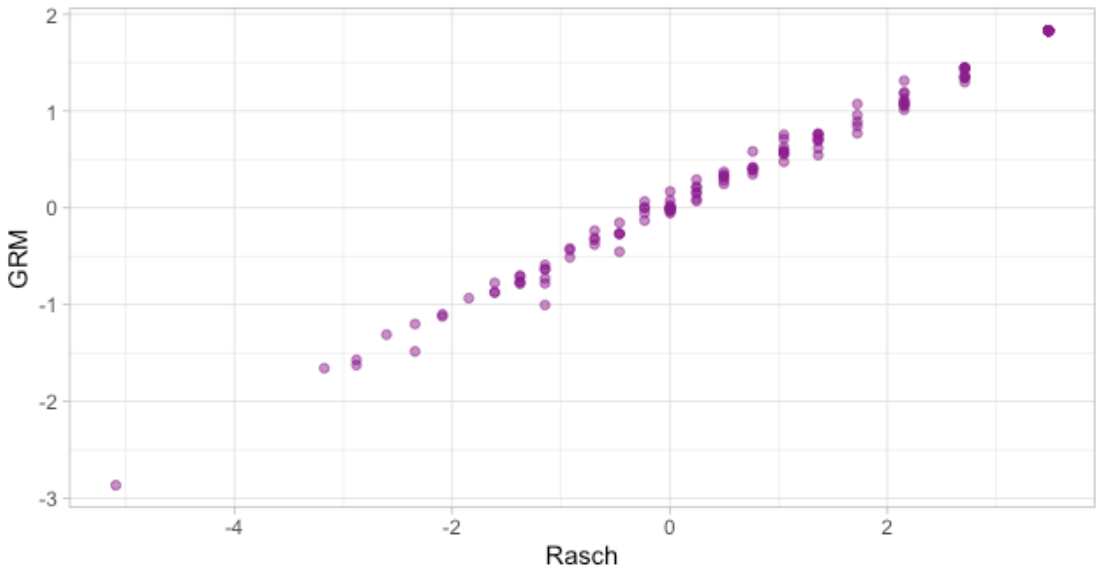


Figure D.1: The relationship between Rasch and graded response model (GRM) scores for the full-length face scale. For each respondent in the validation dataset, Rasch scores are presented as logits on the x axis and GRM scores are presented as logits on the y axis. There is a strong linear relationship between Rasch- and GRM-based scores for the full length face scale, but relative to GRM scores, Rasch scores are more widely spread around the sample mean (0 logits). This is most apparent at extreme ends of the scale.

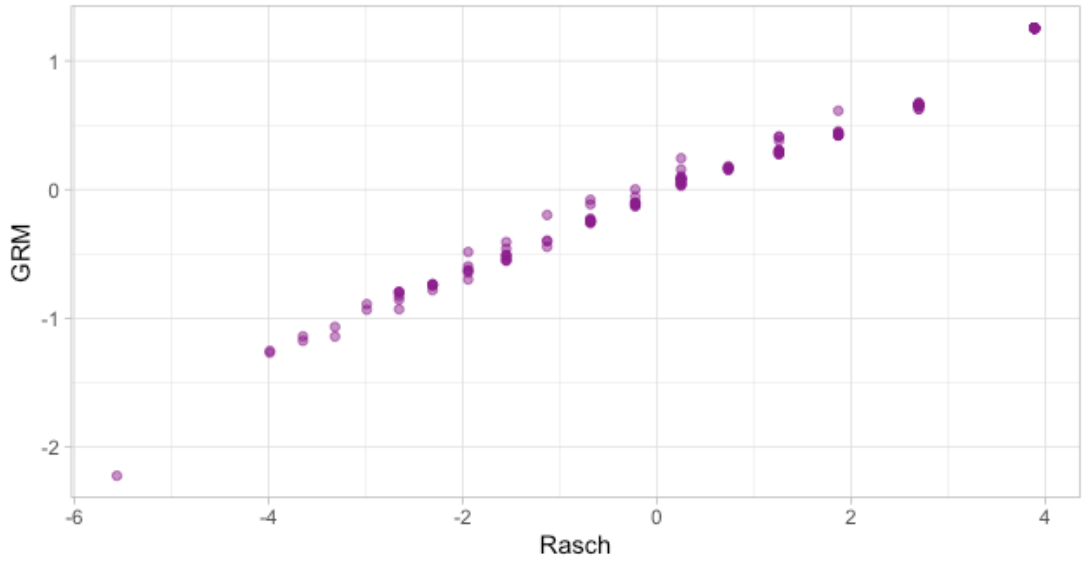


Figure D.2: The relationship between Rasch and graded response model (GRM) scores for the full-length jaw scale. For each respondent in the validation dataset, Rasch scores are presented as logits on the x axis and GRM scores are presented as logits on the y axis. There is a strong linear relationship between Rasch- and GRM-based scores for the full length face scale, but relative to GRM scores, Rasch scores are more widely spread around the sample mean (0 logits). This is most apparent at extreme ends of the scale.

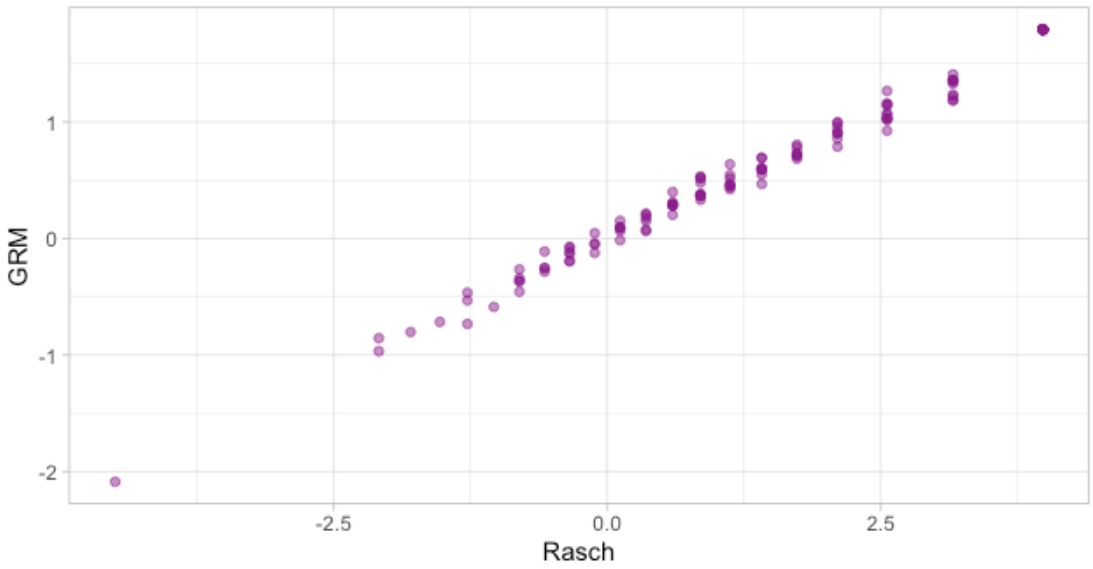


Figure D.3: The relationship between Rasch and graded response model (GRM) scores for the full-length teeth scale. For each respondent in the validation dataset, Rasch scores are presented as logits on the x axis and GRM scores are presented as logits on the y axis. There is a strong linear relationship between Rasch- and GRM-based scores for the full length face scale, but relative to GRM scores, Rasch scores are more widely spread around the sample mean (0 logits). This is most apparent at extreme ends of the scale.

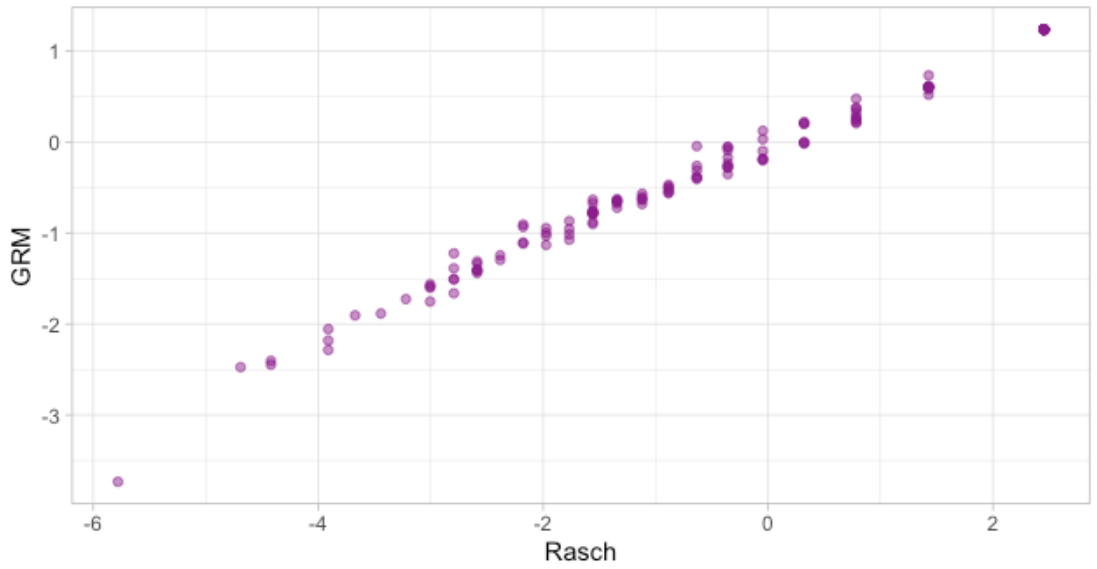


Figure D.4: The relationship between Rasch and graded response model (GRM) scores for the full-length school function scale. For each respondent in the validation dataset, Rasch scores are presented as logits on the x axis and GRM scores are presented as logits on the y axis. There is a strong linear relationship between Rasch- and GRM-based scores for the full length face scale, but relative to GRM scores, Rasch scores are more widely spread around the sample mean (0 logits). This is most apparent at extreme ends of the scale.

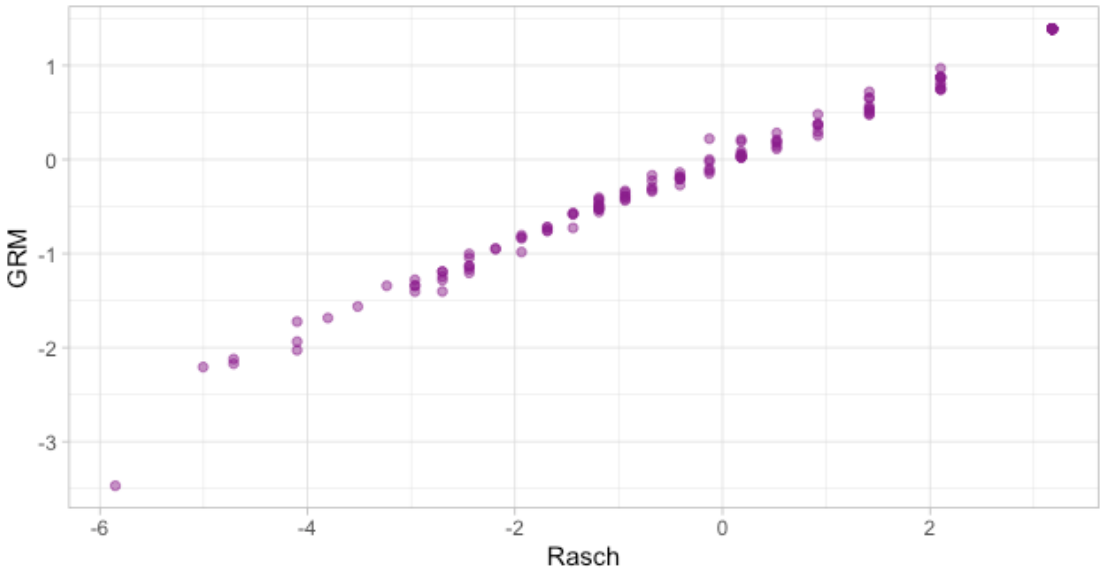


Figure D.5: The relationship between Rasch and graded response model (GRM) scores for the full-length psychological function scale. For each respondent in the validation dataset, Rasch scores are presented as logits on the x axis and GRM scores are presented as logits on the y axis. There is a strong linear relationship between Rasch- and GRM-based scores for the full length face scale, but relative to GRM scores, Rasch scores are more widely spread around the sample mean (0 logits). This is most apparent at extreme ends of the scale.

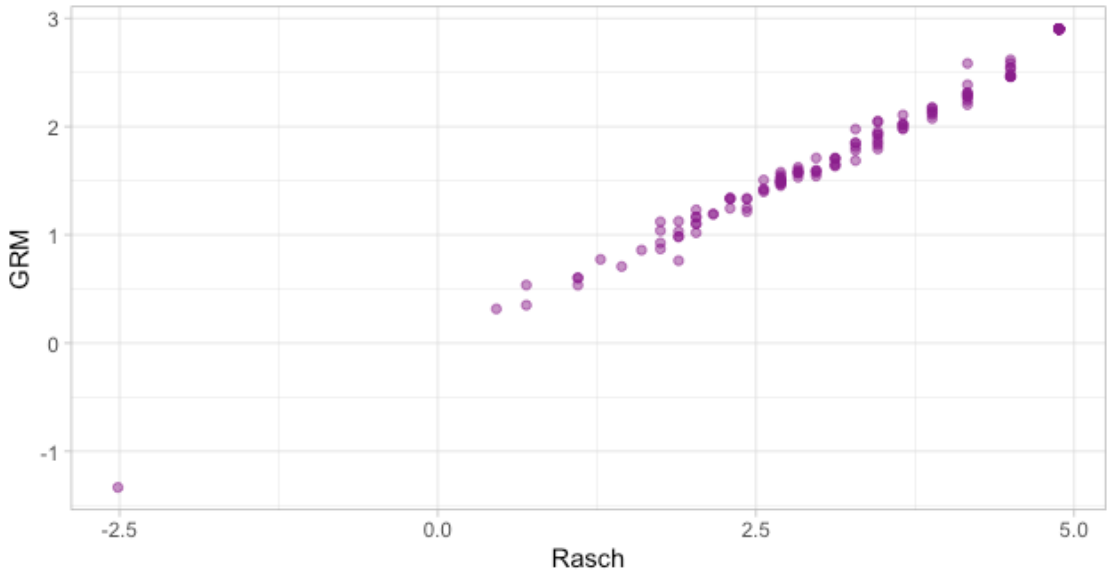


Figure D.6: The relationship between Rasch and graded response model (GRM) scores for the full-length speech distress scale. For each respondent in the validation dataset, Rasch scores are presented as logits on the x axis and GRM scores are presented as logits on the y axis. There is a strong linear relationship between Rasch- and GRM-based scores for the full length face scale, but relative to GRM scores, Rasch scores are more widely spread around the sample mean (0 logits). This is most apparent at extreme ends of the scale.

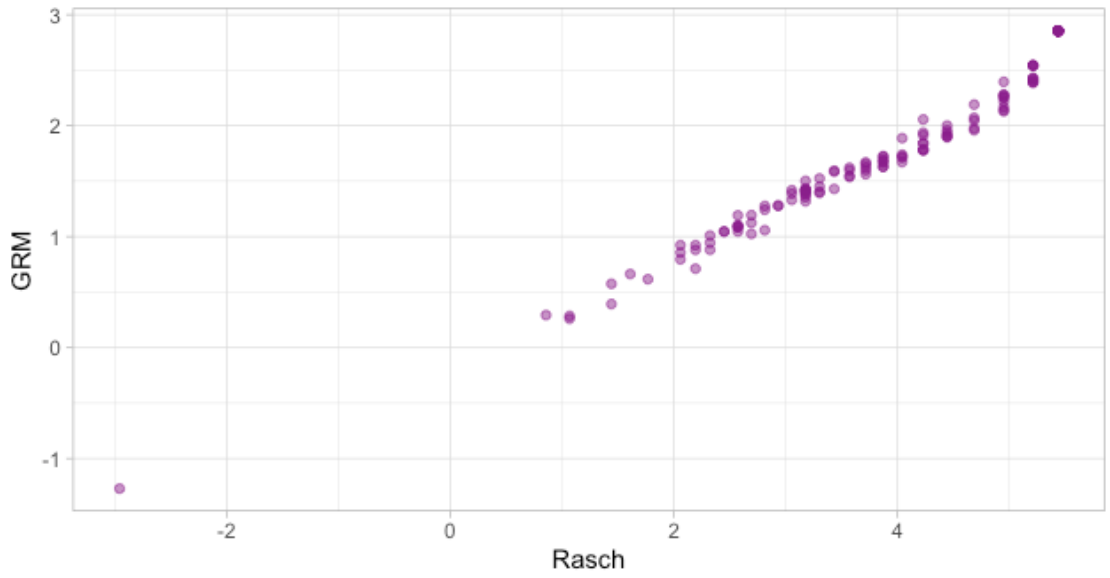


Figure D.7: The relationship between Rasch and graded response model (GRM) scores for the full-length speech function scale. For each respondent in the validation dataset, Rasch scores are presented as logits on the x axis and GRM scores are presented as logits on the y axis. There is a strong linear relationship between Rasch- and GRM-based scores for the full length face scale, but relative to GRM scores, Rasch scores are more widely spread around the sample mean (0 logits). This is most apparent at extreme ends of the scale.

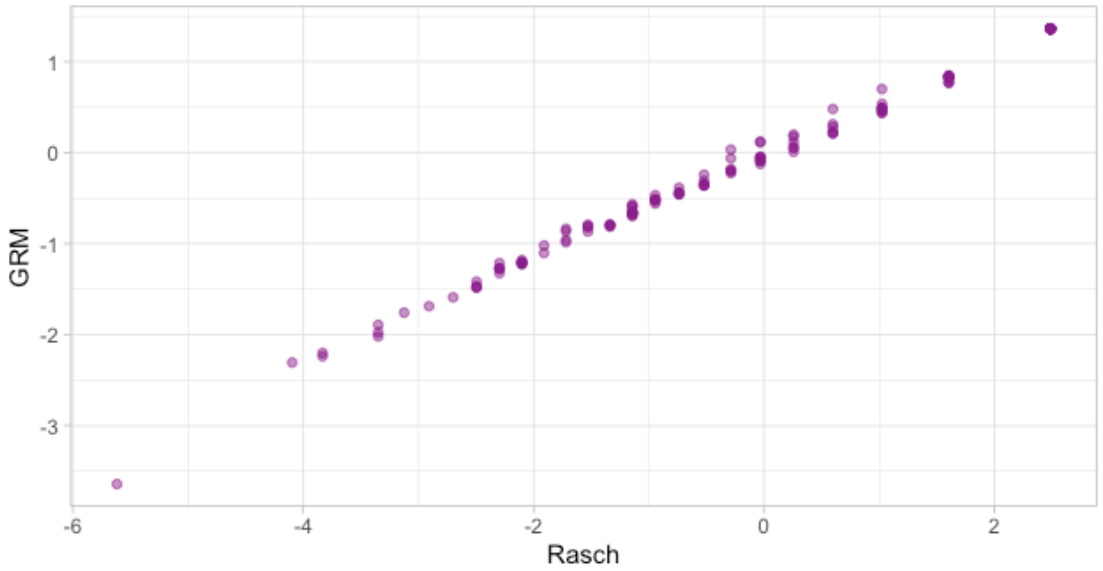


Figure D.8: The relationship between Rasch and graded response model (GRM) scores for the full-length social function scale. For each respondent in the validation dataset, Rasch scores are presented as logits on the x axis and GRM scores are presented as logits on the y axis. There is a strong linear relationship between Rasch- and GRM-based scores for the full length face scale, but relative to GRM scores, Rasch scores are more widely spread around the sample mean (0 logits). This is most apparent at extreme ends of the scale.

D.2 A comparison of CAT assessments built with Rasch models and GRMs

Tables D.1 to D.8 demonstrate the relationship between full-length CLEFT-Q scale scores derived from Rasch models, and the scores of CAT assessments, at different fixed length stopping rules, which have used either Rasch models of GRMs.

Table D.1: Comparison of CAT algorithms that use partial credit Rasch models and graded response models for scoring the CLEFT-Q face scale. GRM: graded response model; SE: standard error of measurement; RMSE: root mean square error; Lower LA: lower 95% limit of agreement; Upper LA: upper 95% limit of agreement. Comparisons are made to full-length scale scores derived from partial credit Rasch models.

Face scale (9 items)

Number of items	Model	Median SE	RMSE	Correlation	Upper LA	Lower LA
9	Rasch	0.501	0.000	1.000	0	0
	GRM	0.270	0.910	0.996	2.00	-1.34
8	Rasch	0.516	0.146	0.997	0.24	-0.28
	GRM	0.277	0.917	0.992	2.01	-1.35
7	Rasch	0.546	0.244	0.989	0.46	-0.57
	GRM	0.289	0.931	0.986	2.04	-1.38
6	Rasch	0.586	0.335	0.983	0.60	-0.70
	GRM	0.308	0.951	0.978	2.09	-1.39
5	Rasch	0.634	0.417	0.972	0.78	-0.87
	GRM	0.331	0.980	0.963	2.15	-1.45
4	Rasch	0.695	0.506	0.963	0.91	-0.98
	GRM	0.359	1.008	0.950	2.20	-1.51
3	Rasch	0.778	0.614	0.950	1.08	-1.13
	GRM	0.400	1.066	0.921	2.33	-1.62
2	Rasch	0.922	0.790	0.912	1.43	-1.46
	GRM	0.473	1.127	0.898	2.45	-1.76
1	Rasch	1.164	1.098	0.812	2.14	-2.00
	GRM	0.564	1.262	0.835	2.76	-1.91

Table D.2: Comparison of CAT algorithms that use partial credit Rasch models and graded response models for scoring the CLEFT-Q jaw scale. GRM: graded response model; SE: standard error of measurement; RMSE: root mean square error; Lower LA: lower 95% limit of agreement; Upper LA: upper 95% limit of agreement. Comparisons are made to full-length scale scores derived from partial credit Rasch models.

Jaw scale (7 items)

Number of items	Model	Median SE	MAE	RMSE	Correlation	Upper LA	Lower LA
7	Rasch	0.697	0.000	1.000	0.00	0.00	
	GRM	0.186	1.606	0.992	3.41	-2.76	
6	Rasch	0.743	0.212	0.997	0.43	-0.40	
	GRM	0.202	1.628	0.982	3.44	-2.84	
5	Rasch	0.801	0.347	0.992	0.71	-0.65	
	GRM	0.222	1.629	0.959	3.42	-2.89	
4	Rasch	0.881	0.486	0.985	1.00	-0.90	
	GRM	0.254	1.613	0.944	3.38	-2.88	
3	Rasch	0.998	0.593	0.980	1.20	-1.13	
	GRM	0.271	1.637	0.918	3.39	-2.98	
2	Rasch	1.180	0.786	0.966	1.62	-1.44	
	GRM	0.331	1.718	0.879	3.56	-3.13	
1	Rasch	1.555	1.214	0.915	2.52	-2.21	
	GRM	0.414	1.856	0.693	3.85	-3.39	

Table D.3: Comparison of CAT algorithms that use partial credit Rasch models and graded response models for scoring the CLEFT-Q teeth scale. GRM: graded response model; SE: standard error of measurement; RMSE: root mean square error; Lower LA: lower 95% limit of agreement; Upper LA: upper 95% limit of agreement. Comparisons are made to full-length scale scores derived from partial credit Rasch models.

Teeth scale (8 items)

Number of items	Model	Median SE	MAE	RMSE	Correlation	Upper LA
Lower LA						
8	Rasch	0.483	0.000	1.000	0.00	0.00
	GRM	0.223	1.215	0.995	2.62	-1.11
7	Rasch	0.512	0.168	0.995	0.34	-0.31
	GRM	0.228	1.228	0.989	2.65	-1.13
6	Rasch	0.554	0.254	0.990	0.48	-0.51
	GRM	0.239	1.245	0.984	2.69	-1.13
5	Rasch	0.591	0.333	0.982	0.64	-0.67
	GRM	0.253	1.273	0.973	2.74	-1.15
4	Rasch	0.655	0.443	0.968	0.86	-0.88
	GRM	0.269	1.282	0.946	2.77	-1.19
3	Rasch	0.747	0.591	0.942	1.11	-1.21
	GRM	0.302	1.282	0.926	2.77	-1.20
2	Rasch	0.879	0.801	0.894	1.39	-1.69
	GRM	0.383	1.350	0.888	2.92	-1.23
1	Rasch	1.149	1.019	0.821	1.79	-2.15
	GRM	0.414	1.444	0.812	3.14	-1.41

Table D.4: Comparison of CAT algorithms that use partial credit Rasch models and graded response models for scoring the CLEFT-Q school function scale. GRM: graded response model; SE: standard error of measurement; RMSE: root mean square error; Lower LA: lower 95% limit of agreement; Upper LA: upper 95% limit of agreement. Comparisons are made to full-length scale scores derived from partial credit Rasch models.

School function (10 items)

Number of items	Model	Median SE	MAE	RMSE	Correlation	Upper LA Lower LA
10	Rasch	0.495	0.000	1.000	0.00	0.00
	GRM	0.263	0.962	0.993	1.52	-2.09
9	Rasch	0.518	0.150	0.996	0.28	-0.31
	GRM	0.267	0.981	0.990	1.55	-2.13
8	Rasch	0.537	0.226	0.992	0.40	-0.48
	GRM	0.274	0.993	0.984	1.56	-2.16
7	Rasch	0.552	0.273	0.988	0.50	-0.56
	GRM	0.286	0.989	0.981	1.58	-2.14
6	Rasch	0.605	0.386	0.975	0.71	-0.80
	GRM	0.306	1.002	0.969	1.61	-2.17
5	Rasch	0.633	0.490	0.959	0.90	-1.01
	GRM	0.319	1.021	0.952	1.62	-2.22
4	Rasch	0.705	0.608	0.938	1.07	-1.28
	GRM	0.363	1.070	0.937	1.63	-2.34
3	Rasch	0.816	0.718	0.911	1.31	-1.50
	GRM	0.424	1.153	0.903	1.74	-2.52
2	Rasch	0.999	0.904	0.858	1.588	-1.91
	GRM	0.492	1.208	0.840	1.99	-2.60
1	Rasch	1.147	1.130	0.768	1.92	-2.41
	GRM	0.630	1.440	0.728	2.32	-3.11

Table D.5: Comparison of CAT algorithms that use partial credit Rasch models and graded response models for scoring the CLEFT-Q psychological function scale. GRM: graded response model; SE: standard error of measurement; RMSE: root mean square error; Lower LA: lower 95% limit of agreement; Upper LA: upper 95% limit of agreement. Comparisons are made to full-length scale scores derived from partial credit Rasch models.

Psychological function scale (10 items)

Number of items	Model	Median SE	MAE	RMSE	Correlation	Upper LA Lower LA
10	Rasch	0.541	0.000	1.000	0.00	0.00
	GRM	0.240	1.199	0.994	2.22	-2.47
9	Rasch	0.564	0.172	0.998	0.35	-0.33
	GRM	0.247	1.213	0.991	2.24	-2.50
8	Rasch	0.587	0.239	0.990	0.48	-0.47
	GRM	0.257	1.220	0.984	2.25	-2.51
7	Rasch	0.635	0.333	0.986	0.66	-0.65
	GRM	0.272	1.220	0.976	2.25	-2.52
6	Rasch	0.667	0.393	0.979	0.79	-0.75
	GRM	0.290	1.194	0.973	2.20	-2.46
5	Rasch	0.737	0.480	0.968	0.98	-0.90
	GRM	0.310	1.184	0.961	2.19	-2.44
4	Rasch	0.782	0.566	0.946	1.16	-1.05
	GRM	0.337	1.179	0.950	2.14	-2.45
3	Rasch	0.901	0.726	0.939	1.45	-1.40
	GRM	0.389	1.215	0.931	2.24	-2.50
2	Rasch	1.121	0.948	0.892	1.82	-1.90
	GRM	0.449	1.313	0.891	2.39	-2.73
1	Rasch	1.311	1.334	0.761	2.61	-2.63
	GRM	0.574	1.473	0.828	2.78	-2.99

Table D.6: Comparison of CAT algorithms that use partial credit Rasch models and graded response models for scoring the CLEFT-Q speech distress scale. GRM: graded response model; SE: standard error of measurement; RMSE: root mean square error; Lower LA: lower 95% limit of agreement; Upper LA: upper 95% limit of agreement. Comparisons are made to full-length scale scores derived from partial credit Rasch models.

Speech distress scale (10 items)

Number of items	Model	Median SE	MAE	RMSE	Correlation	Upper LA Lower LA
10	Rasch	0.710	0.000	1.000	0.00	0.00
	GRM	0.244	1.477	0.992	2.39	0.38
9	Rasch	0.716	0.179	0.995	0.33	-0.37
	GRM	0.251	1.481	0.987	2.39	0.39
8	Rasch	0.727	0.452	0.973	0.72	-0.98
	GRM	0.258	1.480	0.982	2.39	0.38
7	Rasch	0.809	0.643	0.947	0.97	-1.41
	GRM	0.270	1.468	0.971	2.38	0.38
6	Rasch	0.830	0.876	0.904	1.31	-1.92
	GRM	0.287	1.488	0.963	2.37	0.44
5	Rasch	0.858	0.957	0.883	1.45	-2.09
	GRM	0.307	1.531	0.942	2.43	0.46
4	Rasch	0.914	0.994	0.864	1.63	-2.14
	GRM	0.325	1.646	0.926	2.63	0.48
3	Rasch	1.078	1.146	0.811	1.96	-2.44
	GRM	0.348	1.801	0.900	2.92	0.45
2	Rasch	1.195	1.294	0.747	2.30	-2.72
	GRM	0.410	2.035	0.836	3.34	0.46
1	Rasch	1.531	1.499	0.622	3.05	-2.82
	GRM	0.531	2.370	0.791	3.85	0.60

Table D.7: Comparison of CAT algorithms that use partial credit Rasch models and graded response models for scoring the CLEFT-Q speech function scale. GRM: graded response model; SE: standard error of measurement; RMSE: root mean square error; Lower LA: lower 95% limit of agreement; Upper LA: upper 95% limit of agreement. Comparisons are made to full-length scale scores derived from partial credit Rasch models.

Speech function scale (12 items)

Number of items	Model	Median SE	MAE	RMSE	Correlation	Upper LA
Lower LA						
12	Rasch	0.582	0.000	1.000	0.00	0.00
	GRM	0.194	2.149	0.981	3.31	0.80
11	Rasch	0.607	0.155	0.998	0.28	-0.33
	GRM	0.197	2.167	0.979	3.32	0.83
10	Rasch	0.624	0.280	0.992	0.48	-0.60
	GRM	0.201	2.185	0.974	3.32	0.87
9	Rasch	0.669	0.374	0.987	0.62	-0.80
	GRM	0.207	2.186	0.967	3.32	0.87
8	Rasch	0.697	0.449	0.981	0.75	-0.96
	GRM	0.214	2.224	0.961	3.35	0.93
7	Rasch	0.760	0.525	0.974	0.84	-1.14
	GRM	0.223	2.252	0.951	3.38	0.95
6	Rasch	0.800	0.630	0.963	1.00	-1.34
	GRM	0.238	2.289	0.932	3.45	0.94
5	Rasch	0.900	0.722	0.951	1.12	-1.57
	GRM	0.251	2.328	0.907	3.54	0.92
4	Rasch	0.969	0.845	0.932	1.31	-1.84
	GRM	0.278	2.347	0.892	3.59	0.90
3	Rasch	1.140	0.965	0.910	1.54	-2.09
	GRM	0.307	2.418	0.883	3.73	0.88
2	Rasch	1.291	1.278	0.835	2.04	-2.77
	GRM	0.375	2.605	0.862	4.03	0.94
1	Rasch	1.689	1.487	0.746	2.73	-3.07
	GRM	0.532	2.879	0.783	4.44	1.05

Table D.8: Comparison of CAT algorithms that use partial credit Rasch models and graded response models for scoring the CLEFT-Q social function scale. GRM: graded response model; SE: standard error of measurement; RMSE: root mean square error; Lower LA: lower 95% limit of agreement; Upper LA: upper 95% limit of agreement. Comparisons are made to full-length scale scores derived from partial credit Rasch models.

Social function scale (10 items)

Number of items	Model	Median SE	MAE	RMSE	Correlation	Upper LA Lower LA
10	Rasch	0.482	0.000	1.000	0.00	0.00
	GRM	0.273	0.784	0.997	1.39	-1.65
9	Rasch	0.500	0.105	0.998	0.20	-0.21
	GRM	0.282	0.780	0.994	1.38	-1.64
8	Rasch	0.521	0.165	0.995	0.33	-0.32
	GRM	0.294	0.800	0.989	1.41	-1.69
7	Rasch	0.534	0.270	0.988	0.48	-0.57
	GRM	0.305	0.814	0.980	1.44	-1.71
6	Rasch	0.580	0.317	0.984	0.56	-0.67
	GRM	0.327	0.837	0.973	1.48	-1.76
5	Rasch	0.601	0.486	0.961	0.81	-1.05
	GRM	0.350	0.857	0.962	1.50	-1.81
4	Rasch	0.672	0.530	0.952	0.93	-1.2
	GRM	0.385	0.918	0.952	1.63	-1.93
3	Rasch	0.756	0.610	0.935	1.16	-1.24
	GRM	0.430	0.991	0.918	1.77	-2.08
2	Rasch	0.866	0.795	0.887	1.48	-1.63
	GRM	0.501	1.060	0.872	1.94	-2.19
1	Rasch	1.100	1.088	0.772	2.16	-2.12
	GRM	0.625	1.241	0.795	2.25	-2.58

E

Appendix E: Chapter 7 results

Contents

E.1 Comparative simulation study results 229

E.1 Comparative simulation study results

Tables E.1 to E.24 display the results to the comparative simulation study described in Chapter 7.

Table E.1: Comparison of Rasch-based CAT algorithms and tree-based CAT algorithms for the lip scale assessment (7 item limit). DT: decision tree; RMSE: root mean square error; MAE: mean absolute error.

Lip scale assessment (7 item limit)

	Variables used	Mean items	Correlation	RMSE	MAE
DT	Cleft type, age, and all lip scale items	6.40	0.97	0.61	0.42
Rasch	All lip scale items	7.00	0.99	0.27	0.21

Table E.2: Comparison of Rasch-based CAT algorithms and tree-based CAT algorithms for the lip scale assessment (5 item limit). DT: decision tree; RMSE: root mean square error; MAE: mean absolute error.

Lip scale assessment (5 item limit)

	Variables used	Mean items	Correlation	RMSE	MAE
DT	Lip scale items 1, 2, 3, 5, 6, 7, 8, and 9	4.95	0.96	0.73	0.52
Rasch	All lip scale items	5.00	0.99	0.44	0.34

Table E.3: Comparison of Rasch-based CAT algorithms and tree-based CAT algorithms for the cleft lip scar scale assessment (5 item limit). DT: decision tree; RMSE: root mean square error; MAE: mean absolute error.

Cleft lip scar scale assessment (5 item limit)

	Variables used	Mean items	Correlation	RMSE	MAE
DT	All scar scale items	4.96	0.97	0.64	0.43
Rasch	All scar scale items	5.00	0.99	0.33	0.26

Table E.4: Comparison of Rasch-based CAT algorithms and tree-based CAT algorithms for the cleft lip scar scale assessment (3 item limit). DT: decision tree; RMSE: root mean square error; MAE: mean absolute error.

Cleft lip scar scale assessment (3 item limit)

	Variables used	Mean items	Correlation	RMSE	MAE
DT	Scar scale items 2, 3, 5, 6, and 7	3.00	0.95	0.89	0.66
Rasch	All scar scale items	3.00	0.98	0.62	0.49

Table E.5: Comparison of Rasch-based CAT algorithms and tree-based CAT algorithms for the teeth scale assessment (6 item limit). DT: decision tree; RMSE: root mean square error; MAE: mean absolute error.

Teeth scale assessment (6 item limit)

	Variables used	Mean items	Correlation	RMSE	MAE
DT	Gender, teeth anchor question, and all teeth scale items	5.83	0.96	0.52	0.39
Rasch	All teeth scale items	6.00	0.99	0.27	0.20

Table E.6: Comparison of Rasch-based CAT algorithms and tree-based CAT algorithms for the teeth scale assessment (4 item limit). DT: decision tree; RMSE: root mean square error; MAE: mean absolute error.

Teeth scale assessment (4 item limit)

	Variables used	Mean items	Correlation	RMSE	MAE
DT	Teeth scale items 1, 3, 4, 5, 6, 7, 8, and teeth anchor question	4.00	0.94	0.66	0.50
Rasch	All teeth scale items	4.00	0.97	0.47	0.38

Table E.7: Comparison of Rasch-based CAT algorithms and tree-based CAT algorithms for the jaw scale assessment (5 item limit). DT: decision tree; RMSE: root mean square error; MAE: mean absolute error.

Jaw scale assessment (5 item limit)

	Variables used	Mean items	Correlation	RMSE	MAE
DT	All jaw scale items	4.96	0.98	0.59	0.37
Rasch	All jaw scale items	5.00	0.99	0.37	0.28

Table E.8: Comparison of Rasch-based CAT algorithms and tree-based CAT algorithms for the jaw scale assessment (3 item limit). DT: decision tree; RMSE: root mean square error; MAE: mean absolute error.

Jaw scale assessment (3 item limit)

	Variables used	Mean items	Correlation	RMSE	MAE
DT	Jaw scale items 1, 3, 4, 6, and 7	3.00	0.96	0.86	0.61
Rasch	All jaw scale items	3.00	0.98	0.64	0.51

Table E.9: Comparison of Rasch-based CAT algorithms and tree-based CAT algorithms for the nose scale assessment (9 item limit). DT: decision tree; RMSE: root mean square error; MAE: mean absolute error.

Nose scale assessment (9 item limit)

	Variables used	Mean items	Correlation	RMSE	MAE
DT	Gender, nose anchor question, and all nose scale items	7.32	0.97	0.56	0.40
Rasch	All nose scale items	9.00	0.99	0.25	0.20

Table E.10: Comparison of Rasch-based CAT algorithms and tree-based CAT algorithms for the nose scale assessment (6 item limit). DT: decision tree; RMSE: root mean square error; MAE: mean absolute error.

Nose scale assessment (6 item limit)

	Variables used	Mean items	Correlation	RMSE	MAE
DT	Nose anchor question and all nose scale items	5.87	0.96	0.63	0.47
Rasch	All nose scale items	6.00	0.98	0.44	0.36

Table E.11: Comparison of Rasch-based CAT algorithms and tree-based CAT algorithms for the nostrils scale assessment (5 item limit). DT: decision tree; RMSE: root mean square error; MAE: mean absolute error.

Nostrils scale assessment (5 item limit)

	Variables used	Mean items	Correlation	RMSE	MAE
DT	Nostril anchor question and all nostrils scale items	4.99	0.98	0.51	0.34
Rasch	All nostrils scale items	5.00	1.00	0.25	0.20

Table E.12: Comparison of Rasch-based CAT algorithms and tree-based CAT algorithms for the nostrils scale assessment (3 item limit). DT: decision tree; RMSE: root mean square error; MAE: mean absolute error.

Nostrils scale assessment (3 item limit)

	Variables used	Mean items	Correlation	RMSE	MAE
DT	All nostrils scale items	3.00	0.96	0.76	0.57
Rasch	All nostrils scale items	3.00	0.98	0.58	0.46

Table E.13: Comparison of Rasch-based CAT algorithms and tree-based CAT algorithms for the face scale assessment (7 item limit). DT: decision tree; RMSE: root mean square error; MAE: mean absolute error.

Face scale assessment (7 item limit)

	Variables used	Mean items	Correlation	RMSE	MAE
DT	Face anchor question and all face scale items	6.56	0.96	0.52	0.40
Rasch	All face scale items	7.00	0.99	0.24	0.19

Table E.14: Comparison of Rasch-based CAT algorithms and tree-based CAT algorithms for the face scale assessment (5 item limit). DT: decision tree; RMSE: root mean square error; MAE: mean absolute error.

Face scale assessment (5 item limit)

	Variables used	Mean items	Correlation	RMSE	MAE
DT	Face anchor question and all face scale items	6.56	0.96	0.52	0.40
Rasch	All face scale items	7.00	0.99	0.24	0.19

Table E.15: Comparison of Rasch-based CAT algorithms and tree-based CAT algorithms for the psychological function scale assessment (7 item limit). DT: decision tree; RMSE: root mean square error; MAE: mean absolute error.

Psychological function scale assessment (7 item limit)

	Variables used	Mean items	Correlation	RMSE	MAE
DT	Age, gender, and all psychological function scale items	6.63	0.97	0.51	0.36
Rasch	All psychological function scale items	7.00	0.99	0.34	0.25

Table E.16: Comparison of Rasch-based CAT algorithms and tree-based CAT algorithms for the psychological function scale assessment (5 item limit). DT: decision tree; RMSE: root mean square error; MAE: mean absolute error.

Psychological function scale assessment (5 item limit)

	Variables used	Mean items	Correlation	RMSE	MAE
DT	Age and all psychological function scale items	4.95	0.96	0.60	0.45
Rasch	All psychological function scale items	5.00	0.97	0.47	0.39

Table E.17: Comparison of Rasch-based CAT algorithms and tree-based CAT algorithms for the social function scale assessment (7 item limit). DT: decision tree; RMSE: root mean square error; MAE: mean absolute error.

Social function scale assessment (7 item limit)

	Variables used	Mean items	Correlation	RMSE	MAE
DT	Age, cleft type, laterality, and all social function scale items	6.64	0.97	0.42	0.31
Rasch	All social function scale items	7.00	0.99	0.25	0.19

Table E.18: Comparison of Rasch-based CAT algorithms and tree-based CAT algorithms for the social function scale assessment (5 item limit). DT: decision tree; RMSE: root mean square error; MAE: mean absolute error.

Social function scale assessment (5 item limit)

	Variables used	Mean items	Correlation	RMSE	MAE
DT	Social function scale items 1, 3, 4, 5, 6, 7, 8, 9, and 10	4.99	0.94	0.54	0.43
Rasch	All social function scale items	5.00	0.96	0.43	0.35

Table E.19: Comparison of Rasch-based CAT algorithms and tree-based CAT algorithms for the school function scale assessment (7 item limit). DT: decision tree; RMSE: root mean square error; MAE: mean absolute error.

School function scale assessment (7 item limit)

	Variables used	Mean items	Correlation	RMSE	MAE
DT	Age, cleft type, and all school function scale items	6.28	0.96	0.47	0.33
Rasch	All school function scale items	7.00	0.99	0.28	0.22

Table E.20: Comparison of Rasch-based CAT algorithms and tree-based CAT algorithms for the school function scale assessment (5 item limit). DT: decision tree; RMSE: root mean square error; MAE: mean absolute error.

School function scale assessment (5 item limit)

	Variables used	Mean items	Correlation	RMSE	MAE
DT	All school function scale items	4.93	0.95	0.54	0.41
Rasch	All school function scale items	5.00	0.96	0.46	0.37

Table E.21: Comparison of Rasch-based CAT algorithms and tree-based CAT algorithms for the speech distress scale assessment (7 item limit). DT: decision tree; RMSE: root mean square error; MAE: mean absolute error.

Speech distress scale assessment (7 item limit)

	Variables used	Mean items	Correlation	RMSE	MAE
DT	Age and all speech distress items	6.37	0.97	0.50	0.36
Rasch	All speech distress items	7.00	0.99	0.34	0.24

Table E.22: Comparison of Rasch-based CAT algorithms and tree-based CAT algorithms for the speech distress scale assessment (5 item limit). DT: decision tree; RMSE: root mean square error; MAE: mean absolute error.

Speech distress scale assessment (5 item limit)

	Variables used	Mean items	Correlation	RMSE	MAE
DT	Speech distress items 2, 3, 4, 5, 6, 7, 8, and 9	4.96	0.95	0.61	0.47
Rasch	All speech distress items	5.00	0.97	0.50	0.38

Table E.23: Comparison of Rasch-based CAT algorithms and tree-based CAT algorithms for the speech function scale assessment (9 item limit). DT: decision tree; RMSE: root mean square error; MAE: mean absolute error.

Speech function scale assessment (9 item limit)

	Variables used	Mean items	Correlation	RMSE	MAE
DT	Age and all speech function scale items	7.01	0.97	0.53	0.38
Rasch	All speech function scale items	9.00	0.99	0.28	0.19

Table E.24: Comparison of Rasch-based CAT algorithms and tree-based CAT algorithms for the speech function scale assessment (6 item limit). DT: decision tree; RMSE: root mean square error; MAE: mean absolute error.

Speech function scale assessment (6 item limit)

	Variables used	Mean items	Correlation	RMSE	MAE
DT	Age, cleft type, and all speech function scale items	5.80	0.97	0.56	0.43
Rasch	All speech function scale items	6.00	0.97	0.51	0.40

References

1. Romanes, G. J. Thought-reading. *Nature* **24,** 171–172 (1881).
2. Wiseman, R. *Paranormality: Why we see what isn't there* (MacMillan, London, 2011).
3. Rust, J., Koskinski, M. & Stillwell, D. *Modern psychometrics: The science of psychological assessment* Fourth (Routledge, Abingdon, 2021).
4. Wu, Y., Kosinski, M. & Stillwell, D. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences* **112,** 1036–1040 (2015).
5. Matz, S. C., Kosinski, M., Nave, G. & Stillwell, D. J. Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences* **114,** 12714–12719 (2017).
6. Arráiz, I., Bruhn, M. & Stucchi, R. Psychometrics as a tool to improve credit information. *The World Bank Economic Review* **30,** S67–S76 (2017).
7. Peters, M. A. Platform ontologies, the AI crisis and the ability to hack humans 'An algorithm knows me better than I know myself'. *Educational Philosophy and Theory* **52,** 593–601 (2020).
8. Weldring, T. & Smith, S. M. Article commentary: Patient-reported outcomes (pros) and patient-reported outcome measures (PROMs). *Health Services Insights* **6,** HSI–S11093 (2013).
9. Wormald, J. C. & Rodrigues, J. N. Outcome measurement in plastic surgery. *Journal of Plastic, Reconstructive & Aesthetic Surgery* **71,** 283–289 (2018).
10. Beard, D. J. *et al.* Arthroscopic subacromial decompression for subacromial shoulder pain (CSAW): a multicentre, pragmatic, parallel group, placebo-controlled, three-group, randomised surgical trial. *The Lancet* **391,** 329–338 (2018).
11. Beard, D. J. *et al.* The clinical and cost-effectiveness of total versus partial knee replacement in patients with medial compartment osteoarthritis (TOPKAT): 5-year outcomes of a randomised controlled trial. *The Lancet* **394,** 746–756 (2019).
12. Sabah, S. A., Alvand, A., Beard, D. J. & Price, A. J. Minimal important changes and differences were estimated for Oxford hip and knee scores following primary and revision arthroplasty. *Journal of Clinical Epidemiology* **143,** 159–168 (2022).
13. DeMuro, C. *et al.* Assessment of PRO label claims granted by the FDA as compared to the EMA (2006–2010). *Value in Health* **16,** 1150–1155 (2013).
14. *Health-related quality of life* 2020. https://www.nice.org.uk/Media/Default/About/what-we-do/our-programmes/nice-guidance/chte-methods-consultation/Health-related-quality-of-life-task-and-finish-group-report.docx.

15. Gibbons, C. *et al.* Routine provision of feedback from patient-reported outcome measurements to healthcare providers and patients in clinical practice. *Cochrane Database of Systematic Reviews* (2021).

16. Denis, F. *et al.* Two-year survival comparing web-based symptom monitoring vs routine surveillance following treatment for lung cancer. *Journal of the American Medical Association* **321,** 306–307 (2019).

17. Denis, F. *et al.* Randomized trial comparing a web-mediated follow-up with routine surveillance in lung cancer patients. *Journal of the National Cancer Institute* **109** (2017).

18. Strasser, F. *et al.* The effect of real-time electronic monitoring of patient-reported symptoms and clinical syndromes in outpatient workflow of medical oncologists: E-MOSAIC, a multicenter cluster-randomized phase III study (SAKK 95/06). *Annals of Oncology* **27,** 324–332 (2016).

19. Basch, E. *et al.* Symptom monitoring with patient-reported outcomes during routine cancer treatment: a randomized controlled trial. *Journal of Clinical Oncology* **34,** 557 (2016).

20. Engelen, V. *et al.* Reporting health-related quality of life scores to physicians during routine follow-up visits of pediatric oncology patients: Is it effective? *Pediatric Blood & Cancer* **58,** 766–774 (2012).

21. Wolfe, J. *et al.* Improving the care of children with advanced cancer by using an electronic patient-reported feedback intervention: results from the PediQUEST randomized controlled trial. *Journal of Clinical Oncology* **32,** 1119 (2014).

22. Murillo, M. *et al.* Impact of monitoring health-related quality of life in clinical practice in children with type 1 diabetes mellitus. *Quality of Life Research* **26,** 3267–3277 (2017).

23. Haverman, L., Engelen, V., van Rossum, M. A., Heymans, H. S. & Grootenhuis, M. A. Monitoring health-related quality of life in paediatric practice: development of an innovative web-based application. *BMC Pediatrics* **11,** 1–7 (2011).

24. Bele, S. *et al.* Patient-reported outcome measures in routine pediatric clinical care: a systematic review. *Frontiers in Pediatrics,* 364 (2020).

25. Greenhalgh, J. *et al.* Functionality and feedback: a realist synthesis of the collation, interpretation and utilisation of patient-reported outcome measures data to improve patient care. *Health Services and Delivery Research* **5,** 1–280 (2017).

26. Mokkink, L. B. *et al.* The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Quality of Life Research* **19,** 539–549 (2010).

27. Beelen, L. M. *et al.* Patient-reported outcome measures in lymphedema: a systematic review and COSMIN analysis. *Annals of Surgical Oncology* **28,** 1656–1668 (2021).

28. Roller, J. M., Courtemanche, R. J. & Courtemanche, D. J. A review of validated quality-of-life patient-reported outcome measures in pediatric plastic surgery. *Plastic and Reconstructive Surgery* **142,** 694e–707e (2018).

29. Wormald, J. C. *et al.* Site-specific patient-reported outcome measures for hand conditions: systematic review of development and psychometric properties. *Plastic and Reconstructive Surgery Global Open* **7** (2019).

30. Kosowski, T. R. *et al.* A systematic review of patient-reported outcome measures after facial cosmetic surgery and/or nonsurgical facial rejuvenation. *Plastic and Reconstructive Surgery* **123,** 1819–1827 (2009).

31. Pusic, A. *et al.* A systematic review of patient-reported outcome measures in head and neck cancer surgery. *Otolaryngology—Head and Neck Surgery* **136,** 525–535 (2007).

32. Cappelleri, J. C., Lundy, J. J. & Hays, R. D. Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics* **36,** 648–662 (2014).

33. Rusch, T., Lowry, P. B., Mair, P. & Treiblmaier, H. Breaking free from the limitations of classical test theory: Developing and measuring information systems scales using item response theory. *Information & Management* **54,** 189–203 (2017).

34. Fan, X. Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement* **58,** 357–381 (1998).

35. Jones, L. V. & Thissen, D. A history and overview of psychometrics. *Handbook of Statistics* **26,** 1–27 (2006).

36. Ostini, R. & Nering, M. *Polytomous Rasch models* (SAGE Publishing, 2005).

37. Partchev, I. *A visual guide to item response theory* 2004. https://www.stat.cmu.edu/~brian/PIER-methods/For%202013-02-28/Readings/Interactive%20pdf%20on%20Item%20Response%20Theory.pdf.

38. Wu, M. & Adams, R. *Applying the Rasch model to psycho-social measurement: A practical approach* (Educational Measurement Solutions Melbourne, 2007).

39. Mesbah, M. & Kreiner, S. Rasch models for ordered polytomous items. *Rasch Models in Health,* 27–42 (2012).

40. Henninger, M. & Meiser, T. Different approaches to modeling response styles in divide-by-total item response theory models (part 2): Applications and novel extensions. *Psychological Methods* **25,** 577 (2020).

41. Garcıa-Pérez, M. A. An analysis of (dis) ordered categories, thresholds, and crossings in difference and divide-by-total IRT models for ordered responses. *The Spanish Journal of Psychology* **20** (2017).

42. Andrich, D. Rating scales and Rasch measurement. *Expert Review of Pharmacoeconomics & Outcomes Research* **11,** 571–585 (2011).

43. Andrich, D. Sufficiency and conditional estimation of person parameters in the polytomous Rasch model. *Psychometrika* **75,** 292–308 (2010).

44. Masters, G. A Rasch Model for partial credit scoring. *Psychometrika* **47,** 149–174 (1982).

45. Samejima, F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement* (1969).

46. Auné, S. E., Abal, F. J. & Attorresi, H. F. Application of the graded response model to a scale of empathie behavior. *International Journal of Psychological Research* **12,** 49–56 (2019).

47. Battauz, M. Regularized estimation of the four-parameter logistic model. *Psych* **2,** 269–278 (2020).

48. Harvey, R. J. Improving measurement via item response theory: great idea, but hold the rasch. *The Counseling Psychologist* **44,** 195–204 (2016).

49. Chalmers, R. P. mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software* **48,** 1–29 (2012).

50. Chen, S.-K., Hou, L. & Dodd, B. G. A comparison of maximum likelihood estimation and expected a posteriori estimation in CAT using the partial credit model. *Educational and Psychological Measurement* **58,** 569–595 (1998).

51. Chen, J. & Choi, J. A comparison of maximum likelihood and expected a posteriori estimation for polychoric correlation using Monte Carlo simulation. *Journal of Modern Applied Statistical Methods* **8,** 32 (2009).

52. Frey, B. *Test information function* (SAGE Publishing, Thousand Oaks, 2018).

53. Martin, M. *et al.* Item response theory methods can improve the measurement of physical function by combining the modified health assessment questionnaire and the SF-36 physical function scale. *Quality of Life Research* **16,** 647–660 (2007).

54. Pek, J. & Van Zandt, T. Frequentist and Bayesian approaches to data analysis: Evaluation and estimation. *Psychology Learning & Teaching* **19,** 21–35 (2020).

55. Box, G. E. & Tiao, G. C. *Bayesian inference in statistical analysis* (John Wiley & Sons, 2011).

56. Kim, J. K. & Nicewander, W. A. Ability estimation for conventional tests. *Psychometrika* **58,** 587–599 (1993).

57. Natesan, P., Nandakumar, R., Minka, T. & Rubright, J. D. Bayesian prior choice in IRT estimation using MCMC and variational Bayes. *Frontiers in Psychology* **7,** 1422 (2016).

58. Chen, Y. & Gupta, M. R. *EM demystified: An expectation-maximization tutorial* in *Electrical Engineering* (2010).

59. Wu, C. J. On the convergence properties of the EM algorithm. *The Annals of Statistics,* 95–103 (1983).

60. Ranger, J. & Much, S. Analyzing the fit of IRT models with the Hausman test. *Frontiers in Psychology* **11,** 149 (2020).

61. Nguyen, T. H., Han, H.-R., Kim, M. T. & Chan, K. S. An introduction to item response theory for patient-reported outcome measurement. *The Patient-Patient-Centered Outcomes Research* **7,** 23–35 (2014).

62. Cassarly, C., Matthews, L. J., Simpson, A. N. & Dubno, J. R. Development and refinement of patient-reported outcome measures for hearing: A brief introduction to nonparametric item response theory. *American Journal of Audiology* **28,** 806–809 (2019).

63. Van Schuur, W. H. Mokken scale analysis: Between the Guttman scale and parametric item response theory. *Political Analysis* **11,** 139–163 (2003).

64. Reckase, M. D. The past and future of multidimensional item response theory. *Applied Psychological Measurement* **21,** 25–36 (1997).

65. Yong, A. G., Pearce, S., *et al.* A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology* **9,** 79–94 (2013).

66. Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A. & King, J. Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research* **99,** 323–338 (2006).

67. Sijtsma, K. & van der Ark, L. A. A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology* **70,** 137–158 (2017).

68. Teresi, J. A., Ramirez, M., Lai, J.-S. & Silver, S. Occurrences and sources of differential item functioning (DIF) in patient-reported outcome measures: Description of DIF methods, and review of measures of depression, quality of life and general health. *Psychology Science* (2008).

69. Choi, S. W., Gibbons, L. E. & Crane, P. K. Lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of Statistical Software* **39,** 1 (2011).

70. Christensen, K. B., Makransky, G. & Horton, M. Critical values for Yen's Q 3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement* **41,** 178–194 (2017).

71. Weiss, D. J. & Vale, C. D. Adaptive testing. *Applied Psychology* **36,** 249–262 (1987).

72. Harrison, C., Loe, B. S., Lis, P., Sidey-Gibbons, C., *et al.* Maximizing the potential of patient-reported assessments by using the open-source concerto platform with computerized adaptive testing and machine learning. *Journal of Medical Internet Research* **22,** e20950 (2020).

73. Magis, D. & Raıche, G. Random generation of response patterns under computerized adaptive testing with the R package catR. *Journal of Statistical Software* **48,** 1–31 (2012).

74. Loe, B. S., Stillwell, D. & Gibbons, C. Computerized adaptive testing provides reliable and efficient depression measurement using the CES-D scale. *Journal of Medical Internet Research* **19,** e7453 (2017).

75. Gibbons, C., Bower, P., Lovell, K., Valderas, J., Skevington, S., *et al.* Electronic quality of life assessment using computer-adaptive testing. *Journal of Medical Internet Research* **18,** e6053 (2016).

76. *Stopping rules* 2022. https://www.healthmeasures.net/resource-center/measurement-science/computer-adaptive-tests-cats/stopping-rules.

77. Chalmers, R. P. Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software* **71,** 1–38 (2016).

78. Seo, D. G. Overview and current management of computerized adaptive testing in licensing/certification examinations. *Journal of Educational Evaluation for Health Professions* **14** (2017).

79. Cella, D. *et al.* The patient-reported outcomes measurement information system (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology* **63,** 1179–1194 (2010).

80. Working group, P. *PROMIS reference populations* https://www.healthmeasures.net/score-and-interpret/interpret-scores/promis/reference-populations.

81. *EORTC CAT* 2022. https://qol.eortc.org/projectqol/eortc-cat/.

82. Kaasa, S. *et al.* The EORTC core quality of life questionnaire (QLQ-C30): validity and reliability when analysed with patients treated with palliative radiotherapy. *European Journal of Cancer* **31,** 2260–2263 (1995).

83. Harrison, C. J. & Sidey-Gibbons, C. J. in *Handbook of Quality of Life in Cancer* 133–140 (Springer, 2022).

84. Keulen, M. H., Teunis, T., Vagner, G. A., Ring, D. & Reichel, L. M. The effect of the content of patient-reported outcome measures on patient perceived empathy and satisfaction: a randomized controlled trial. *The Journal of Hand Surgery* **43,** 1141–e1 (2018).

85. Mellema, J. J., O'Connor, C. M., Overbeek, C. L., Hageman, M. G. & Ring, D. The effect of feedback regarding coping strategies and illness behavior on hand surgery patient satisfaction and communication: a randomized controlled trial. *Hand* **10,** 503–511 (2015).

86. Geerards, D., Pusic, A., Hoogbergen, M., Van Der Hulst, R., Sidey-Gibbons, C., *et al.* Computerized quality of life assessment: a randomized experiment to determine the impact of individualized feedback on assessment experience. *Journal of Medical Internet Research* **21,** e12212 (2019).

87. Smith, A. B., Hanbury, A. & Retzler, J. Item banking and computer-adaptive testing in clinical trials: Standing in sight of the PROMISed land. *Contemporary Clinical Trials Communications* **13,** 100306 (2019).

88. *HL7 FHIR implementation guide* 2019. http://hl7.org/fhir/us/patient-reported-outcomes/2019May/pro-overview.html.

89. Mossey, P. A., Little, J., Munger, R. G., Dixon, M. J. & Shaw, W. C. Cleft lip and palate. *The Lancet* **374,** 1773–1785 (2009).

90. Macho, P., Bohac, M., Fedeles Jr, J., Fekiacova, D. & Fedeles Sr, J. Impact of cleft lip and/or palate in children on family quality of life before and after reconstructive surgery. *Clinical Study* **370,** 373 (2017).

91. Dravet, C. *et al.* Epilepsy, antiepileptic drugs, and malformations in children of women with epilepsy: a French prospective cohort study. *Neurology* **42,** 75–82 (1992).

92. Little, J. *et al.* Folate and clefts of the lip and palate—a UK-based case-control study: Part II: Biochemical and genetic analysis. *The Cleft Palate-Craniofacial Journal* **45,** 428–438 (2008).

93. Honein, M. A. *et al.* Maternal smoking and environmental tobacco smoke exposure and the risk of orofacial clefts. *Epidemiology,* 226–233 (2007).

94. Loomfield, V. & Liao, C. Global Trends in the Rate of Cleft Lip and Palate: Bridging the Gap. *Paediatr Child Health* **20,** 75 (2015).

95. Pool, S. M., der Lek, L. M. v., de Jong, K., Vermeij-Keers, C. & Mouës-Vink, C. M. Embryologically based classification specifies gender differences in the prevalence of orofacial cleft subphenotypes. *The Cleft Palate-Craniofacial Journal* **58,** 54–60 (2021).

96. Allori, A. C., Mulliken, J. B., Meara, J. G., Shusterman, S. & Marcus, J. R. Classification of cleft lip/palate: then and now. *The Cleft Palate-Craniofacial Journal* **54,** 175–188 (2017).

97. Shkoukani, M. A., Chen, M. & Vong, A. Cleft lip - a comprehensive review. *Frontiers in Pediatrics* **1,** 53 (2013).

98. Yuzuriha, S., Oh, A. K. & Mulliken, J. B. Asymmetrical bilateral cleft lip: Complete or incomplete and contralateral lesser defect (minor-form, microform, or mini-microform). *Plastic and Reconstructive Surgery* **122,** 1494–1504 (2008).

99. Riff, K. W. W. *et al.* CLEFT-Q: detecting differences in outcomes among 2434 patients with varying cleft types. *Plastic and Reconstructive Surgery* **144,** 78e–88e (2019).

100. Veau, V. & Borel, S. *Division palatine: anatomie, chirurgie, phonetique; avec la collaboration de S. Borel* (Masson, 1931).

101. Georgievska-Jancheska, T. The relationship between the type of cleft and nasal air emission in speech of children with cleft palate or cleft lip and palate. *Open Access Macedonian Journal of Medical Sciences* **7,** 352 (2019).

102. Swan, M. C. & Fisher, D. M. in *Plastic Surgery-Principles and Practice* 196–211 (Elsevier, 2022).

103. Shaw, W. C. *et al.* The Eurocleft project 1996–2000: overview. *Journal of Cranio-Maxillofacial Surgery* **29,** 131–140 (2001).

104. Sandy, J. *et al.* The Clinical Standards Advisory Group (CSAG) cleft lip and palate study. *British Journal of Orthodontics* **25,** 21–30 (1998).

105. Ness, A. R. *et al.* Centralization of cleft care in the UK. Part 6: a tale of two studies. *Orthodontics & Craniofacial research* **18,** 56–62 (2015).

106. Rohrich, R. J., Love, E. J., Byrd, H. S. & Johns, D. F. Optimal timing of cleft palate closure. *Plastic and Reconstructive Surgery* **106,** 413–421 (2000).

107. Asher-Mcdade, C., Roberts, C., Shaw, W. C. & Gallager, C. Development of a method for rating nasolabial appearance in patients with clefts of the lip and palate. *The Cleft Palate-Craniofacial Journal* **28,** 385–391 (1991).

108. Debbie, S., Anne, H. & Pamela, G. GOS. SP. ASS.'98: an assessment for speech disorders associated with cleft palate and/or velopharyngeal dysfunction (revised). *International Journal of Language & Communication Disorders* **34,** 17–33 (1999).

109. Mars, M., Plint, D. A., Houston, W., Bergland, O. & Semb, G. The Goslon Yardstick: a new system of assessing dental arch relationships in children with unilateral clefts of the lip and palate. *The Cleft Palate Journal* **24,** 314–322 (1987).

110. Conroy, E. J. *et al.* A randomised controlled trial comparing palate surgery at 6 months versus 12 months of age (the TOPS trial): a statistical analysis plan. *Trials* **22,** 1–15 (2021).

111. Shaw, W. Global strategies to reduce the health care burden of craniofacial anomalies: report of WHO meetings on international collaborative research on craniofacial anomalies. *The Cleft Palate-Craniofacial Journal* **41,** 238–243 (2004).

112. Allori, A. C. *et al.* A standard set of outcome measures for the comprehensive appraisal of cleft care. *The Cleft Palate-Craniofacial Journal* **54,** 540–554 (2017).

113. Tsangaris, E. *et al.* Establishing content validity of the CLEFT-Q: a new patient-reported outcome instrument for cleft lip/palate. *Plastic and Reconstructive Surgery Global Open* **5** (2017).

114. Klassen, A. F., Stotland, M. A., Skarsgard, E. D. & Pusic, A. L. Clinical research in pediatric plastic surgery and systematic review of quality-of-life questionnaires. *Clinics in Plastic Surgery* **35,** 251–267 (2008).

115. Eckstein, D. A., Wu, R. L., Akinbiyi, T., Silver, L. & Taub, P. J. Measuring quality of life in cleft lip and palate patients: currently available patient-reported outcomes measures. *Plastic and Reconstructive Surgery* **128,** 518e–526e (2011).

116. Klassen, A. F. *et al.* Quality of life of children treated for cleft lip and/or palate: a systematic review. *Journal of Plastic, Reconstructive & Aesthetic Surgery* **65,** 547–557 (2012).

117. Riff, K. W. W. *et al.* International multiphase mixed methods study protocol to develop a cross-cultural patient-reported outcome instrument for children and young adults with cleft lip and/or palate (CLEFT-Q). *British Medical Journal Open* **7,** e015467 (2017).

118. Klassen, A. F. *et al.* Psychometric findings and normative values for the CLEFT-Q based on 2434 children and young adult patients with cleft lip and/or palate from 12 countries. *Canadian Medican Association Journal* **190,** E455–E462 (2018).

119. Q-Portfolio. *CLEFT-Q scales* https://qportfolio.org/wp-content/uploads/2022/03/CLEFT-Q-Translations.pdf.

120. Klassen, A. F. *et al.* Impact of completing CLEFT-Q scales that ask about appearance on children and young adults: an international study. *The Cleft Palate-Craniofacial Journal* **57,** 840–848 (2020).

121. Harrison, C. J. *et al.* Further construct validation of the CLEFT-Q: ability to detect differences in outcome for four cleft-specific surgeries. *Journal of Plastic, Reconstructive & Aesthetic Surgery* **72,** 2049–2055 (2019).

122. Stiernman, M., Klintö, K., Persson, M. & Becker, M. Comparison of corresponding scores from the Cleft Hearing Appearance and Speech Questionnaire (CHASQ) and CLEFT-Q in Swedish patients with cleft lip and/or palate. *The Cleft Palate-Craniofacial Journal* **58,** 736–745 (2021).

123. Miroshnychenko, A. *et al.* Psychometric validation of the CLEFT-Q patient reported outcome measure: a prospective study to examine cross-sectional construct validity. *The Cleft Palate-Craniofacial Journal,* 10556656211062837 (2021).

124. Miroshnychenko, A. *et al.* A prospective study to examine responsiveness and minimally important differences (MIDs) for the CLEFT-Q scales following three cleft-specific operations. *The Cleft Palate-Craniofacial Journal,* 10556656211064479 (2021).

125. Stock, N. M. *et al.* Achieving consensus in the measurement of psychological adjustment to cleft lip and/or palate at age 8+ years. *The Cleft Palate-Craniofacial Journal* **57,** 746–752 (2020).

126. Apon, I. *et al.* Barriers and facilitators to the international implementation of standardized outcome measures in clinical cleft practice. *The Cleft Palate-Craniofacial Journal* **59,** 5–13 (2022).

127. Weidler, E. M., Britto, M. T. & Sitzman, T. J. Facilitators and barriers to implementing standardized outcome measurement for children With cleft Lip and palate. *The Cleft Palate-Craniofacial Journal* **58,** 7–18 (2021).

128. Harrison, C. J., Rodrigues, J. N., Furniss, D. & Swan, M. C. Response to barriers and facilitators to the international implementation of standardized outcome measures in clinical cleft practice. *The Cleft Palate-Craniofacial Journal* **59,** 669–670 (2022).

129. Lord, F. M. Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika* **48,** 233–245 (1983).

130. Warm, T. A. Weighted likelihood estimation of ability in item response theory. *Psychometrika* **54,** 427–450 (1989).

131. Wang, T. Precision of Warm's weighted likelihood estimation of ability. *Trends in Cognitive Psychology,* 77 (2002).

132. Choi, S. W. & Swartz, R. J. Comparison of CAT item selection criteria for polytomous items. *Applied Psychological Measurement* **33,** 419–440 (2009).

133. Veerkamp, W. J. & Berger, M. P. Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics* **22,** 203–226 (1997).

134. Van der Linden, W. J. Bayesian item selection criteria for adaptive testing. *Psychometrika* **63,** 201–216 (1998).

135. Linden, W. J. & Pashley, P. J. in *Elements of Adaptive Testing* 3–30 (Springer, 2009).

136. Penfield, R. D. Applying Bayesian item selection approaches to adaptive tests using polytomous items. *Applied Measurement in Education* **19,** 1–20 (2006).

137. Linacre, J. Sample size and item calibration stability. *Rasch Measurement Transactions* **7,** 328 (1994).

138. Lameijer, C. *et al.* Graded response model fit, measurement invariance and (comparative) precision of the Dutch-Flemish PROMIS® Upper Extremity V2. 0 item bank in patients with upper extremity disorders. *BMC Musculoskeletal Disorders* **21,** 1–17 (2020).

139. Sedgwick, P. Limits of agreement (Bland-Altman method). *British Medical Journal* **346** (2013).

140. Harrison, C. *cleftqCATsim* 2021. https://github.com/MrConradHarrison/cleftqCATsim.

141. *Zoom* 2022. https://zoom.us/.

142. Pedersen, A. B. *et al.* Missing data and multiple imputation in clinical epidemiological research. *Clinical Epidemiology* **9,** 157 (2017).

143. Şahin, M. D. Effect of Item Order on Certain Psychometric Properties: A Demonstration on a Cyberloafing Scale. *Frontiers in Psychology,* 154 (2021).

144. *Concerto* 2022. https://www.psychometrics.cam.ac.uk/newconcerto.

145. Gibbons, C., Richards, S., Valderas, J. M., Campbell, J., *et al.* Supervised machine learning algorithms can classify open-text feedback of doctor performance with human-level accuracy. *Journal of Medical Internet Research* **19,** e6533 (2017).

146. Harrison, C. *The CLEFT-Q CAT Score Checker* 2022. https://conrad-harrison.shinyapps.io/CLEFT-Q-CAT-Score-Checker/.

147. Harrison, C. *CLEFT-Q CAT Score Checker* 2022. https://github.com/MrConradHarrison/CLEFT-Q-CAT-Score-Checker.

148. Fereday, J. & Muir-Cochrane, E. Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods* **5,** 80–92 (2006).

149. Walker, D. & Myrick, F. Grounded theory: An exploration of process and procedure. *Qualitative Health Research* **16,** 547–559 (2006).

150. Tong, A., Sainsbury, P. & Craig, J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care* **19,** 349–357 (2007).

151. Reeve, B. B. *et al.* Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care,* S22–S31 (2007).

152. Andersson, B. & Xin, T. Large sample confidence intervals for item response theory reliability coefficients. *Educational and Psychological Measurement* **78,** 32–45 (2018).

153. Sijtsma, K. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika* **74,** 107–120 (2009).

154. Cleanthous, S., Barbic, S. P., Smith, S. & Regnault, A. Psychometric performance of the PROMIS® depression item bank: a comparison of the 28-and 51-item versions using Rasch measurement theory. *Journal of Patient-Reported Outcomes* **3,** 1–12 (2019).

155. Stover, A. M., McLeod, L. D., Langer, M. M., Chen, W.-H. & Reeve, B. B. State of the psychometric methods: patient-reported outcome measure development and refinement using item response theory. *Journal of Patient-Reported Outcomes* **3,** 1–16 (2019).

156. Nolte, S., Coon, C., Hudgens, S. & Verdam, M. G. Psychometric evaluation of the PROMIS® Depression Item Bank: an illustration of classical test theory methods. *Journal of Patient-Reported Outcomes* **3,** 1–10 (2019).

157. Bjorner, J. B. State of the psychometric methods: comments on the ISOQOL SIG psychometric papers. *Journal of Patient-Reported Outcomes* **3,** 1–11 (2019).

158. Fischer, H. F. & Rose, M. Scoring depression on a common metric: a comparison of EAP estimation, plausible value imputation, and full Bayesian IRT modeling. *Multivariate Behavioral Research* **54,** 85–99 (2019).

159. Gorter, R., Fox, J.-P. & Twisk, J. W. Why item response theory should be used for longitudinal questionnaire data analysis in medical research. *BMC Medical Research Methodology* **15,** 1–12 (2015).

160. Gorter, R., Fox, J.-P., Apeldoorn, A. & Twisk, J. Measurement model choice influenced randomized controlled trial results. *Journal of Clinical Epidemiology* **79,** 140–149 (2016).

161. Chalmers, P. 2015. https://philchalmers.github.io/mirt/extra/mirt-Workshop-2015_Day-1.pdf.

162. Sheng, Y. & Wikle, C. K. Bayesian multidimensional IRT models with a hierarchical structure. *Educational and Psychological Measurement* **68,** 413–430 (2008).

163. Ip, E. H. Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *British Journal of Mathematical and Statistical Psychology* **63,** 395–416 (2010).

164. Hartig, J. & Höhler, J. Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation* **35,** 57–63 (2009).

165. Wu, M. & Adams, R. Modelling mathematics problem solving item responses using a multidimensional IRT model. *Mathematics Education Research Journal* **18,** 93–113 (2006).

166. Cheville, A. L. *et al.* Improving the delivery of function-directed care during acute hospitalizations: methods to develop and validate the Functional Assessment in Acute Care Multidimensional Computerized Adaptive Test (FAMCAT). *Archives of Rehabilitation Research and Clinical Translation* **3,** 100112 (2021).

167. Cai, L. A two-tier full-information item factor analysis model with applications. *Psychometrika* **75,** 581–612 (2010).

168. Harrison, C. *et al.* A comparative analysis of multidimensional computerized adaptive testing for the DASH and QuickDASH scores in Dupuytren's disease. *Journal of Hand Surgery (European Volume),* 17531934221081803 (2022).

169. Kyriazos, T. A. *et al.* Applied psychometrics: sample size and sample power considerations in factor analysis (EFA, CFA) and SEM in general. *Psychology* **9,** 2207 (2018).

170. Buchanan, E. *Advanced statistics* 2017. https://osf.io/dnuyv/..

171. Field, A. P. & Wilcox, R. R. Robust statistical methods: A primer for clinical psychology and experimental psychopathology researchers. *Behaviour Research and Therapy* **98,** 19–38 (2017).

172. Mindrila, D. Maximum likelihood (ML) and diagonally weighted least squares (DWLS) estimation procedures: A comparison of estimation bias with ordinal and multivariate non-normal data. *International Journal of Digital Society* **1,** 60–66 (2010).

173. Chalmers, P. *mirt - second order model* https://groups.google.com/g/mirt-package/c/Q6izYN0ab3Y.

174. DeMars, C. E. A tutorial on interpreting bifactor model scores. *International Journal of Testing* **13,** 354–378 (2013).

175. Chalmers, P. 2022. https://cran.r-project.org/web/packages/mirt/mirt.pdf.

176. Kenny, D. *Measuring Model Fit* 2020. http://www.davidakenny.net/cm/fit.htm.

177. Harrison, C. J. *et al.* Deeper understanding of appearance in orofacial clefts: A structural equation model of the CLEFT-Q appearance scales. *Plastic and Reconstructive Surgery Global Open* **9** (2021).

178. Becker, J.-M., Klein, K. & Wetzels, M. Hierarchical latent variable models in PLS-SEM: guidelines for using reflective-formative type models. *Long Range Planning* **45,** 359–394 (2012).

179. Ray, S. *et al.* 2022. https://cran.r-project.org/web/packages/seminr/seminr.pdf.

180. Therneau, T. M., Atkinson, E. J., *et al. An introduction to recursive partitioning using the RPART routines* tech. rep. (Technical Report Mayo Foundation, 1997).

181. Rodrıguez-Cuadrado, J., Delgado-Gómez, D., Laria, J. C. & Rodrıguez-Cuadrado, S. Merged Tree-CAT: A fast method for building precise computerized adaptive tests based on decision trees. *Expert Systems with Applications* **143,** 113066 (2020).

182. Michel, P. *et al.* Computerized adaptive testing with decision regression trees: an alternative to item response theory for quality of life measurement in multiple sclerosis. *Patient Preference and Adherence* **12,** 1043 (2018).

183. Delgado-Gomez, D., Baca-Garcia, E., Aguado, D., Courtet, P. & Lopez-Castroman, J. Computerized adaptive test vs. decision trees: development of a support decision system to identify suicidal behavior. *Journal of Affective Disorders* **206,** 204–209 (2016).

184. Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. *Classification and regression trees* (Routledge, 2017).

185. Delgado-Gómez, D., Laria, J. C. & Ruiz-Hernández, D. Computerized adaptive test and decision trees: a unifying approach. *Expert Systems with Applications* **117,** 358–366 (2019).

186. Kass, G. V. An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **29,** 119–127 (1980).

187. Jansen, M. C., van der Oest, M. J., Slijper, H. P., Porsius, J. T. & Selles, R. W. Item reduction of the Boston Carpal Tunnel Questionnaire using decision tree modeling. *Archives of Physical Medicine and Rehabilitation* **100,** 2308–2313 (2019).

188. Van der Oest, M. J., Porsius, J. T., MacDermid, J. C., Slijper, H. P. & Selles, R. W. Item reduction of the patient-rated wrist evaluation using decision tree modelling. *Disability and Rehabilitation* **42,** 2758–2765 (2020).

189. Harrison, C. J. *et al.* Developing machine learning algorithms to support patient-centered, value-based carpal tunnel decompression surgery. *Plastic and Reconstructive Surgery Global Open* **10** (2022).

190. Chapman, R. Expected a posteriori scoring in PROMIS®. *Journal of Patient-Reported Outcomes* **6,** 1–20 (2022).

191. Suzuki, S. J. *What is a Print?: Selections from the Museum of Modern Art* (The Museum of Modern Art, 2011).

192. Mejdahl, C. T., Schougaard, L. M. V., Hjollund, N. H., Riiskjær, E. & Lomborg, K. Patient-reported outcome measures in the interaction between patient and clinician–a multi-perspective qualitative study. *Journal of Patient-Reported Outcomes* **4,** 1–10 (2020).

193. Han, K. C. T. Components of the item selection algorithm in computerized adaptive testing. *Journal of Educational Evaluation for Health Professions* **15** (2018).

194. Harrison, C. J. & Rodrigues, J. N. *How to interpret clinical outcome data in plastic surgery research and clinical practice* 2020.

195. Cook, J. A. *et al.* DELTA2 guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial. *British Medical Journal* **363** (2018).

196. Rose, M. *et al.* The PROMIS Physical Function item bank was calibrated to a standardized metric and shown to improve measurement efficiency. *Journal of Clinical Epidemiology* **67,** 516–526 (2014).

197. Choi, S. W., Schalet, B., Cook, K. F. & Cella, D. Establishing a common metric for depressive symptoms: linking the BDI-II, CES-D, and PHQ-9 to PROMIS depression. *Psychological Assessment* **26,** 513 (2014).

198. Schalet, B. D. *et al.* Establishing a common metric for physical function: linking the HAQ-DI and SF-36 PF subscale to PROMIS® physical function. *Journal of General Internal Medicine* **30,** 1517–1523 (2015).

199. Schalet, B. D., Cook, K. F., Choi, S. W. & Cella, D. Establishing a common metric for self-reported anxiety: linking the MASQ, PANAS, and GAD-7 to PROMIS Anxiety. *Journal of Anxiety Disorders* **28,** 88–96 (2014).

200. Harrison, C. *Ecological momentary computerised adaptive testing* 2022. https://www.ndorms.ox.ac.uk/research/clinical-trials/current-trials-and-studies/emcat.

201. Drennan, J. & Hyde, A. Controlling response shift bias: the use of the retrospective pre-test design in the evaluation of a master's programme. *Assessment & Evaluation in Higher Education* **33,** 699–709 (2008).

202. Kosinski, M., Stillwell, D. & Graepel, T. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* **110,** 5802–5805 (2013).

203. Youyou, W., Kosinski, M. & Stillwell, D. Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences* **112,** 1036–1040 (2015).

204. Jelınek, M., Kveton, P., Buresová, I. & Klimusová, H. Measuring depression in adolescence: Evaluation of a hierarchical factor model of the Children's Depression Inventory and measurement invariance across boys and girls. *PLOS One* **16,** e0249943 (2021).